Interview Prep Concepts Review 3

Research Methods

Foundational & Common Designs

1. Experimental Design: The researcher manipulates one variable to observe its effect on another, with random assignment of subjects. It's the gold standard for establishing cause-and-effect.
2. Quasi-Experimental Design: Similar to an experiment but lacks the random assignment of subjects to groups.
3. Correlational Design: Measures the relationship between two or more variables without the researcher controlling or manipulating any of them.
4. Descriptive Research: Aims to accurately and systematically describe a population, situation, or phenomenon. It answers the "what, where, when, and how" questions.
5. Survey Research: Gathers data from a sample of individuals through a set of questions, either written or oral.
6. Observational Design: Involves watching and recording the actions of participants or phenomena in their natural setting.

In-Depth & Qualitative Designs

7. Case Study: An in-depth, detailed examination of a single subject, group, or event.
8. Ethnographic Research: The systematic study of people and cultures through immersive, long-term observation in their natural environment.
9. Phenomenological Research: Seeks to understand the "lived experiences" of individuals regarding a specific phenomenon.
10. Grounded Theory: A methodology for developing a theory that is "grounded" in data systematically gathered and analyzed.
11. Narrative Inquiry: Gathers and interprets stories and personal accounts to understand individual experiences and the meaning people place on them.
12. Historical Research: Involves studying, understanding, and interpreting past events.

Time-Based Designs

13. Cross-Sectional Design: Collects data from a population at a single, specific point in time.

14. Longitudinal Design: Involves repeated observations of the same variables over short or long periods.
15. Cohort Study: A type of longitudinal study that follows a specific group (a cohort) sharing a common characteristic over time.
16. Panel Study: A specific type of longitudinal study where data is collected from the same sample of individuals (the "panel") at different points in time.

Mixed & Combined Methodologies

17. Mixed-Methods Design: Integrates both qualitative and quantitative data collection and analysis in a single study.
18. Convergent Parallel Design: A type of mixed-methods design where quantitative and qualitative data are collected and analyzed separately, then the results are merged for interpretation.
19. Explanatory Sequential Design: A mixed-methods design that begins with quantitative data collection, followed by qualitative data collection to help explain the quantitative results.
20. Exploratory Sequential Design: The reverse of the explanatory design; it starts with qualitative data to explore a topic, then uses the findings to build a quantitative phase.

Review & Analytical Designs

21. Meta-Analysis: A statistical technique for combining the findings from multiple independent studies to reach a more robust conclusion.
22. Systematic Review: A rigorous and comprehensive review of existing literature that uses transparent procedures to find, evaluate, and synthesize all relevant research on a particular topic.

Purpose-Driven & Applied Designs

23. Action Research: A cyclical and reflective process where researchers work to solve a practical, immediate problem within a specific setting (e.g., a school or organization).
24. Evaluation Research: Systematically assesses the effectiveness, merit, or worth of a program, policy, or intervention.
25. Explanatory Research: Aims to explain the "why" behind an observed relationship or phenomenon, going beyond description to understand causality.
26. Exploratory Research: Conducted for a problem that is not clearly defined. It helps to gain a better understanding and generate hypotheses for future research.
27. Diagnostic Research: Focuses on determining the root cause of a specific problem or issue.

28. Causal-Comparative Design (Ex Post Facto): Attempts to identify a cause-and-effect relationship between groups where the cause has already occurred and cannot be manipulated by the researcher.

Specialized Experimental & Field Designs

29. Factorial Design: An experimental setup involving two or more independent variables, allowing for the study of their main effects and their interaction effects.
30. Field Experiment: An experiment conducted in a real-world, natural setting rather than in a controlled laboratory.

IMRaD

Whether you are writing for a journal or a university thesis, following a standardized structure is essential for clarity and credibility. Most academic fields follow the **IMRaD** (Introduction, Methods, Results, and Discussion) format.

---

1. Title Page

The title page is the first impression of your research. It should be clean, professional, and contain:

- **Running Head:** A shortened version of the title (common in APA).
- **Full Title:** Clear, concise, and descriptive.
- **Author(s):** Name and institutional affiliation.
- **Contact Information:** Often includes the corresponding author's email.

---

2. Abstract

The abstract is a standalone summary of the entire paper.

- **Word Limit:** Typically **150–250 words**.
- **Formatting:** Usually a single paragraph without indentation.
- **Content:** Must cover the objective, method, key results, and primary conclusion.

Abstract vs. Introduction

Feature	Abstract	Introduction
Purpose	A "mini-map" of the entire study.	Hooks the reader and sets the stage.
Timing	Written *after* the paper is finished.	Written *before* or during the draft.
Ending	Ends with a summary of findings.	Ends with specific research questions/hypotheses.

---

3. Introduction

The introduction follows a "funnel" structure, moving from a broad field of study to your specific focus.

- **Broad Opening:** Establishes the importance of the topic.
- **Problem Statement:** Identifies the "gap" in current knowledge.
- **Research Questions & Hypotheses:** Explicitly state what you are testing (,).
- **Length:** Usually **10–15%** of the total paper.
- **Structure:** Starts broad, narrows down through a literature review, and ends with your specific aim.

---

4. Method

This section provides a roadmap so other researchers can replicate your study.

- **Research Design:** (e.g., Experimental, Correlational, Qualitative).
- **Participants/Subjects:** Demographics, sample size, and sampling method.
- **Materials:** Specific equipment, software, or survey instruments used.
- **Procedure:** A step-by-step chronological account of what was done.
- **Data Analysis:** The statistical tests or coding methods applied.
- **Ethics & Data Management:** Mention IRB approval, informed consent, and how data was anonymized/stored.

---

5. Results

The Results section is the "objective" heart of the paper.

- **Presentation:** Report findings in a logical sequence, often following the order of your hypotheses.
- **Statistics:** Include both descriptive (means, SD) and inferential statistics (-values, effect sizes).
- **Visuals:** Use Tables and Figures to summarize complex data.
- **The Golden Rule: Avoid interpretation.** Do not explain *why* something happened; simply report *what* happened.

---

6. Discussion

This is where you interpret the data and "zoom back out" to the broader field.

- **Restate Findings:** Summarize the results in plain English (did they support the hypothesis?).
- **Interpretation:** Explain why the results might have occurred.

- **Comparison:** How do your results align (or conflict) with the previous research mentioned in your Introduction?
- **Limitations:** Be honest about the study's weaknesses (e.g., small sample size, bias).
- **Implications:** What does this mean for the "real world" or future research?

---

7. References

This section provides a complete, alphabetized list of every source cited in the text.

- **Consistency:** Ensure the style (APA, MLA, IEEE, Chicago) is consistent throughout.
- **Verification:** Double-check that every in-text citation has a matching entry in the reference list.

Modern Prompting Techniques for Research (2026 Edition)

At its core, prompting involves providing instructions or context to an LLM to guide its output. Below are the foundational and cutting-edge techniques used in modern research.

---

1. Zero-Shot & Few-Shot Prompting (The Foundations)

- **Zero-Shot:** Directly asking a question without examples. Use for basic facts or broad brainstorming.
- **Few-Shot:** Providing 2–5 examples of input-output pairs. Essential for maintaining specific formatting or teaching the model a niche "in-house" coding style.

2. Chain of Thought (CoT) & Reason-First Prompting

- **Explanation:** Encourages the model to generate intermediate reasoning steps.
- **The 2026 Update:** With the rise of "Reasoning Models," many LLMs now perform internal CoT by default. However, explicitly using **"Think step-by-step"** or **"Outline your logic before the final answer"** still reduces "hallucinations" in complex math or logic.

3. Tree-of-Thoughts (ToT) & Forest-of-Thoughts

- **Explanation:** Instead of one linear path, the model explores multiple "branches" of reasoning, evaluates them, and "prunes" the unsuccessful ones.
- **Use in Research:** High-stakes decision-making where multiple contradictory hypotheses must be weighed.

4. Skeleton-of-Thought (SoT)

- **Explanation:** A technique designed for speed and structure. The model first generates a "skeleton" (an outline) of the answer, and then expands each point.
- **Use in Research:** Generating long-form literature reviews or technical reports without losing the "thread" of the argument.
- **Benefit:** Prevents the model from "wandering" off-topic during long generations.

5. System 2 Attention (S2A)

- **Explanation:** A newer technique where you prompt the model to first "sanitize" the input. You ask the model to identify and remove irrelevant or biased information from the prompt before it attempts to answer.
- **Use in Research:** Removing personal bias from a qualitative dataset before asking the AI to summarize the themes.

6. Self-Correction & Multi-Persona Verification

- **Explanation:** You ask the model to generate an answer, then in the next step, ask it to "Critique your own response for factual errors and bias."
- **The "Panel of Experts" approach:** Ask the model to "Adopt the roles of a statistician, a lead researcher, and an ethicist. Review the following proposal and provide a consolidated critique."

7. Retrieval-Augmented Generation (RAG) Prompting

- **Explanation:** Rather than relying on the model's memory, you provide the specific text (the "context") and tell the model: "Use **only** the provided documents to answer."
- **Use in Research:** This is now the gold standard for literature reviews to ensure every claim is grounded in a specific PDF or database entry.

8. Program-Aided Language (PAL) / ReAct

- **Explanation:** The model generates code (usually Python) to solve a reasoning problem rather than trying to do the math in its "head."
- **Use in Research:** Analyzing large CSV files or performing statistical tests. Instead of asking "What is the correlation?", you prompt: "Write and execute Python code to find the correlation between Column A and B."

9. Optimization & Meta-Prompting (DSPy Style)

- **Explanation:** Using an LLM to write a better prompt for another LLM.
- **Use in Research:** When you have a complex task, you provide the "goal," and ask a more powerful model (like GPT-4o or o1) to "Write a system prompt that will ensure a smaller model performs this task perfectly every time."

Comparison Table: Which Technique for Which Task?

Technique	Best For...	Complexity
Zero-Shot	Quick definitions, simple facts	Low
Few-Shot	Specific data formatting, stylistic imitation	Medium
Chain of Thought	Math, logic, multi-step inference	Medium
RAG Prompting	Eliminating hallucinations, citing sources	High
Tree of Thoughts	Strategic planning, complex hypothesis testing	High
System 2 Attention	Bias removal, objective data analysis	Medium

---

Pro-Tips for 2026 Research:

1. **Use Delimiters:** Always wrap your data in triple backticks (`) or XML tags (` `) to help the model distinguish between instructions and the data being analyzed.
2. **Negative Constraints:** Explicitly state what **not** to do (e.g., "Do not use flowery academic jargon; keep the summary to a 10th-grade reading level").
3. **Temperature Control:** For research, keep your "Temperature" low (0.1 - 0.3) to ensure consistency and factual accuracy.

AI Research Tools

- https://docs.google.com/document/d/1nlaV1hxJCxVouXRAJMfywz1-cGNIeNgrslF-gKMVGtM/edit?usp=sharing

Word vs. Overleaf: Choosing the Right Tool

In most industries, **Microsoft Word** is the dominant force for report writing. However, **Overleaf** (and LaTeX) remains the undisputed king of academic and highly technical environments.

Microsoft Word: The Corporate Powerhouse

Microsoft Word is the most common editor across a vast range of companies and is considered an essential business tool.

- **Corporate Standard:** As part of the Microsoft 365 suite, it is the de facto standard in nearly all corporate offices and legacy companies.
- **Broad Use Cases:** Ideal for any role requiring written communication, from marketing materials and legal documents to internal reports.
- **Accessibility:** Its **WYSIWYG** (What You See Is What You Get) interface requires no specialized training, making it the preferred choice for reviewers and managers.
- **Professional Training:** Extensive resources and courses exist specifically to train engineers and scientists on adapting their technical writing to a business-friendly Word format.

Overleaf: The Academic & Technical Specialist

Overleaf is a cloud-based editor for **LaTeX**, a typesetting system designed for high-quality technical and scientific documents.

- **Scientific Research:** The preferred tool for disciplines involving complex mathematical equations, precise formatting, and extensive citations.
- **Journal Integration:** Major publishers like IEEE and Springer provide official LaTeX templates to ensure submissions meet exact formatting requirements.
- **Advanced Collaboration:** Offers real-time co-authoring with integrated version control, specifically tailored for scientific papers.
- **The Trade-off:** It has a significantly higher learning curve; mastering LaTeX commands can be overkill for simple business documents.

Comparison Summary

Feature	Microsoft Word	Overleaf / LaTeX
Primary User	General business, corporate, humanities	Academia, research, technical writing
Learning Curve	**Easy:** Widely accessible	**Steeper:** Requires learning LaTeX
Output Quality	Good, but can struggle with complex layouts	**Exceptional:** Professional typesetting
Collaboration	Robust (Microsoft 365 / OneDrive)	Real-time (Optimized for research)
Formatting	Visual (WYSIWYG)	Command-based (Logical structure)
Industry Adoption	Universal	Niche (Research & Technical fields)

> **The Verdict:** If your goal is to work in a traditional corporate office or a non-technical role, **Microsoft Word** is non-negotiable. If you are pursuing a career in R&D, academia, or high-level engineering, mastering **Overleaf** will give you a significant professional edge.

The Normal Distribution

The normal distribution is fundamentally defined by its **symmetrical, bell-shaped** curve, which is always centered at the **mean** and whose symmetrical spread is dictated entirely by the **standard deviation**. The further away something gets from the average the more unlike it gets from the average.

A normal distribution, also known as the **Gaussian distribution** or **bell curve**, is a continuous probability distribution that is the most important and frequently used distribution in statistics. It is entirely specified by two parameters: the population **mean (μ)** and the population **standard deviation (σ)**.

1. Shape and Symmetry

- **Symmetrical Bell Shape:** The distribution is perfectly symmetrical, meaning the right side of the curve is a mirror image of the left side. This symmetry results in the classic bell shape.
- **Central Tendency:** The measures of central tendency—the **mean (μ)**, **median**, and **mode**—are all equal and coincide exactly at the peak, or center, of the distribution. This point represents the value with the highest frequency.

2. Parameters and Spread

The curve's shape is controlled by its two parameters:

- **Mean (μ):** This acts as the **location parameter**, determining where the center of the peak lies along the horizontal axis. Changing the mean shifts the entire curve left or right without changing its shape.
- **Standard Deviation (σ):** This acts as the **scale parameter**, controlling the spread or width of the distribution. A small σ results in a tall, narrow curve because the data is tightly clustered around the mean. A large σ results in a short, wide curve because the data is more spread out.

3. The Empirical Rule

The degree to which values deviate from the mean is always predictable in a normal distribution, following the **Empirical Rule** (or $\mathbf{68 - 95 - 99.7}$ Rule):

- Approximately 68% of the data falls within **1** standard deviation (σ) of the mean.
- Approximately 95% of the data falls within **2** standard deviations (2σ) of the mean.
- Approximately 99.7% of the data falls within **3** standard deviations (3σ) of the mean.

This rule shows that data values far from the mean are extremely rare, which is why the tails of the bell curve approach the horizontal axis but theoretically never touch it.

The normal curve is essential to hypothesis testing primarily because of the **Central Limit Theorem (CLT)** and its role as the **foundational distribution for parametric tests**.

4. The Central Limit Theorem (CLT)

The CLT is the most critical link between the normal distribution and hypothesis testing.

- **Sampling Distribution:** Hypothesis testing rarely involves the entire population. Instead, we use a **sample** to draw conclusions about the population. The test is based on the **sampling distribution** of a statistic (like the sample mean, $\bar{x}$), which is the distribution of all possible sample statistics that could be calculated from a population.
- **The Guarantee:** The CLT states that if your sample size (n) is sufficiently large (typically $n \geq 30$), the sampling distribution of the mean will be **approximately normal**, regardless of the shape of the original population distribution.
- **The Benefit:** This guarantee means that we can use the properties of the normal distribution to determine the probability of obtaining our sample result. Without the CLT, we wouldn't know the shape of the sampling distribution, making it impossible to accurately calculate the p-value.

5. Foundation for Parametric Tests

The most common and powerful statistical tests, known as **parametric tests** (z-tests, t-tests, ANOVA), are built upon the assumption of a normal distribution.

- **Probability Calculation:** The normal distribution is a mathematically defined curve with known, fixed properties. This allows us to standardize our observed sample statistic (using a z-score or a t-score) and then use the standardized normal distribution (or the related t-distribution) to precisely calculate the area in the tails.
- **Defining Critical Regions:** By using the known properties of the normal curve, statisticians can define the **critical regions** (e.g., the 5% most extreme areas) for a given significance level (α). If our standardized test statistic falls into this critical region, it means our result is rare enough under the null hypothesis to warrant its rejection.
- **The p-value:** The p-value—the probability of observing a sample statistic as extreme or more extreme than the one collected—is calculated as an area under the normal curve. This makes the normal curve the map that allows us to interpret our results and make a statistical decision.

Even when a test uses the t-distribution (which accounts for the uncertainty of estimating the population standard deviation from a small sample), the t-distribution is still closely related to the normal distribution and **approaches the normal curve as the sample size increases**.

How did the Normal Curve come About?

Presentism and Anachronism

- **Anachronism** is when something (a concept, word, or object) is placed in a time period where it doesn't belong. For example, saying "data analysis in Ancient Greece" is technically anachronistic because neither the phrase nor the modern discipline existed yet.
- **Presentism** is the broader interpretive habit of applying present-day concepts, categories, or values to interpret the past. Historians use the word critically, since it can distort how people in antiquity actually thought.

"Strictly speaking, the ancients did not think in terms of 'data analysis'—that's a modern concept. When I use this language, I am engaging in a kind of *presentism*: explaining ancient practices of measurement, probability, and record-keeping through the lens of today's terminology."

5000 Years Ago

- Krishna
- Yellow Emperor and Shennong
- Abraham
- Pastoralists

We have evidence of sea-faring navigation...

Hundreds of cultures worldwide, potentially over 200 or more, have stories about a great flood, though the exact number is hard to pinpoint due to the vastness of oral tradition and differing definitions of a "culture" or "flood story". While the biblical flood of Noah is famous, similar narratives exist across diverse cultures, from Mesopotamia and Greece to Indigenous traditions in North and South America, and from China to Australia. These stories likely arose independently from various local flooding events and shared cultural understandings of past cataclysmic natural disasters, particularly the significant sea-level rise after the last ice age.

Prevalence of Flood Myths

- Widespread Phenomenon: Flood myths are a recurring theme found in the folklore of almost every culture on Earth, regardless of geographic location.
- Examples Across Cultures: Flood stories appear in the traditions of the Hopi Indians, the Incas, the Mesopotamians (Epic of Gilgamesh), the ancient Greeks (Deucalion and Pyrrha), the Chinese, the Hindus (Vaivasvata Manu), and many others.

Ancient Datasets

Astronomy

From Babylonian stargazers to Tycho Brahe's meticulous charts, the heavens provided humanity's first massive datasets. By tracking the motions of planets and stars, astronomers discovered patterns, residuals, and errors—forcing them to invent methods to reconcile measurement with observation. The need to explain discrepancies between prediction and reality gave birth to least squares, error distributions, and the very idea that "noise" in data could be modeled. Astronomy, in many ways, is the cradle of modern statistics.

Agriculture

For millennia, farmers measured yields and timed planting by tradition, but in the 19th century, systematic experiments at places like Rothamsted transformed agriculture into a statistical laboratory. By comparing plots with and without fertilizer, scientists like Fisher developed randomization, replication, and ANOVA. The farm field became the prototype of the modern controlled experiment, and its lessons live on today in everything from medical trials to online A/B testing.

Navigation

Sailors once relied on stars, tides, and crude instruments to traverse oceans, but navigation demanded data analysis before the concept had a name. Tables of latitude and longitude, lunar distances, and compass deviations accumulated into shared "datasets." Errors in measurement—whether of a sextant or a clock—drove the development of corrections, averaging, and eventually probability theory. In striving to find their way across uncharted seas, navigators pioneered the quantification of uncertainty itself.

Early Practices - Observation and Measurement

- Astronomy and calendars: Babylonian and Egyptian sky records used to predict seasons for agriculture and navigation; repeated observation establishes "data before theory."
- State administration: censuses, grain tallies, and tax lists in Egypt, Rome, and China formalize record-keeping.
- Greeks: Philosophers (e.g., Aristotle) discuss randomness largely as absence of purpose or as "for the most part" regularities; no formal calculus of probability.
- Will of the gods.
- Euclid: Euclid's geometry is a landmark in Western thought because it was a departure from basing knowledge solely on observation and measurement. Instead, his work established a system where a vast body of knowledge could be logically derived from a small set of self-evident assumptions. While he shifted the emphasis from empirical measurement to abstract, deductive reasoning, Euclid's system was nevertheless informed and inspired by centuries of practical observation.
- Romans: Gambling is widespread (knucklebones/dice), but analysis is practical not mathematical—odds are experienced, not computed.
- War, building, and luxury
- Key terms and concepts: observation, measurement, calendar cycles, omen catalogues, census, empirical regularity.

Simple Ratios and Formulas for Approximation and Pattern Recognition

Appropriate Mathematical Concept: The Empirical Ratio (Frequency $\approx$ Proportion)

The closest mathematical representation of the Babylonian, Egyptian, and Roman practices of **repeated observation** and **tallying** is the concept of a **simple ratio** or **proportion** derived from empirical data.

They were essentially establishing a **precursor to the Law of Large Numbers** through:

$$\text{Empirical Regularity} \approx \frac{\text{Number of Favorable Outcomes}}{\text{Total Number of Observations}}$$

1. **Astronomy and Calendars (Prediction):**
    - They observed the cycle of seasons (S), divided by the total number of days observed (D).

$$\text{Length of Year} \approx \frac{\text{Total Days Observed}}{\text{Number of Cycles Observed}}$$

- This ratio allowed them to approximate the **365.25 day** length of the solar year, leading to accurate calendars.

2. **State Administration (Censuses/Tallies):**
   - They calculated ratios for tax allocation or resource planning.

$$\text{Tax Rate (Ratio)} = \frac{\text{Grain Paid in Tax}}{\text{Total Grain Harvested}}$$

3. **Roman Gambling (Practical Odds):**
   - While they didn't compute this formally, their *practical* understanding of "odds" was based on experiencing the ratio of successful throws (W) versus total throws (N):

$$\text{Observed Chance of Winning} \approx \frac{\text{Number of Wins (Favorable Outcon}}{\text{Total Number of Rolls (Observatio}}$$

Why No Formal Probability Equation?

$$P(\text{Event}) = \frac{\text{Number of Favorable Outcomes}}{\text{Total Number of Possible Outcomes}}$$

The primary reason no formal equation existed is that **formal probability theory**—calculating the **sample space** and theoretical odds—was not developed until the 16th and 17th centuries with figures like Cardano, Pascal, and Fermat.

The ancients:

- Did not conceive of **randomness** as a mathematically calculable phenomenon, often attributing it to the **"Will of the Gods."**
- Focused on **regularities** and **deterministic cycles** (e.g., the stars, the seasons) for administration and prediction.
- Their mathematics was highly focused on **geometry, ratios (proportions), and practical measurement** (e.g., area, volume, accounting).

The Problem: Measurement

The earliest stirrings of statistical thought are deeply embedded in the ancient practice of astronomy. Early civilizations, from the Assyro-Babylonians to the Greeks, used rudimentary instruments and naked-eye observation to monitor celestial motions, which was fundamental for timekeeping, agriculture, and navigation. While these observations were systematic, as evidenced by Hipparchus's compilation of the first stellar catalogue in the second century BCE, they were inherently limited by the imprecision of

measurement. For instance, Hipparchus's catalogue, which listed 850 stars, had a precision of about one degree, approximately twice the angular size of the full Moon. The intellectual challenge was to reconcile these imprecise measurements with the desire to know a single, "true" value for a star's position or a planet's orbit.

Community Contributed Timeline

- https://docs.google.com/document/d/1Xgwb00aohnMNQg0vidBj5vCFZo7sSGvTr8suhD4p-OU/edit?usp=sharing

Medieval Era (500–1500) Knowledge Preservation and Refinement

- Islamic Golden Age astronomers (Al-Battani, Al-Tusi, Ulugh Beg) produce precise star catalogs; error is noted and minimized through repeated measurement.
- Navigation improves via astrolabes, quadrants, and later the magnetic compass; pilots rely on accumulated observational “datasets.”

Proto-statistical thinking

- Averaging and tabulation appear in astronomy and finance; life annuities and simple risk notions emerge in European commerce.
- Key terms: star catalog, ephemeris, instrument error, mean/average, annuity, risk.

Renaissance & Early Scientific Revolution (1500–1650) Empiricism

- Copernicus, Tycho Brahe, Kepler, and Galileo institutionalize systematic observation; residuals between theory and observation become targets for analysis.
- Navigation and cartography: longitude/latitude methods, marine logs, and pilot charts accumulate structured data.

Seeds of Error Analysis

- Kepler and Galileo compare models to measurements; repeated trials and residual patterns foreshadow formal error distributions.

- Key terms: Residual, systematic vs random error, observational study, triangulation, nautical log.

Birth of Probability (1650–1750) Gambling to mathematics

- Pascal–Fermat correspondence solves problems of games of chance; Huygens writes first textbook on probability.
- Cardano's earlier analysis (published posthumously) anticipates sample spaces and fair odds.

Cardano

- **Gerolamo Cardano** (24 September 1501– 21 September 1576), The Book on Games of Chance
    - Mother tried to abort him
    - Caught the bubonic plague
    - His writings (medical practice was a sham, blood letting, burning, drilling holes in the head) children made him an anathema (daughter got pregnant with brother Giovanni, had an abortion, and became infertile, Giovannie was executed for supected murder of his wife, son 2 liked killing small animals at an early age and eventually became a torturer for the Inquisition, who later gave evidence of heresy that imprisoned his father
    - Used gambling to pay for his medical school
    - Hindus were introducing notational math, base 10, and 0 as a number (vs. placeholder) at this time
    - Our own number system, composed of the ten symbols {0,1,2,3,4,5,6,7,8,9} is called the Hindu-Arabic system. This is a base-ten (decimal) system since place values increase by powers of ten
    - Plus and negative sign was introduced by the Germans
    - The = sign (two parallel lines) 1557
    - The Scientific Revolution is getting rooted
    - An example of Cardano's writing: Suppose a random process has many equally likely outcomes, some favorable (that is, winning), some unfavorable (losing). Then the probability of obtaining a favorable outcome is equal to the proportion of outcomes that are favorable. The set of all possible outcomes is called the sample space

- After incarceration, Cardano moved to Rome, where he received a lifetime annuity from Pope Gregory XIII (after first having been rejected by Pope Pius V, who died in 1572) and finished his autobiography. He was accepted in the Royal College of Physicians, and as well as practising medicine he continued his philosophical studies until his death in 1576.

Pascal

Blaise Pascal (19 June 1623 – 19 August 1662) added to probability and gambling

- Pascal's development of probability theory was his most influential contribution to mathematics. Originally applied to gambling, today it is extremely important in economics, especially in actuarial science
- In 1654, prompted by his friend the Chevalier de Méré, he corresponded with Pierre de Fermat on the subject of gambling problems, and from that collaboration was born the mathematical theory of probabilities.[21] The specific problem was that of two players who want to finish a game early and, given the current circumstances of the game, want to divide the stakes fairly, based on the chance each has of winning the game from that point
- From this discussion, the notion of expected value was introduced
- Pascal's Triangle: Pascal’s triangle is useful any time you need to know the number of ways in which you can choose some number of objects from a collection that has an equal or greater number

Example: Tossing Four Coins

Let's use Pascal's Triangle to find the number of ways to get different combinations of heads and tails when you flip a coin **four times**.

1. Identify the Correct Row

The number of coin tosses, n, corresponds to the row number in Pascal's Triangle. Since we have **four tosses**, we'll look at **Row 4** (remembering that the top row is Row 0).

Pascal's Triangle (Rows 0-4): Row 0: 1 Row 1: 1, 1 Row 2: 1, 2, 1 Row 3: 1, 3, 3, 1 **Row 4: 1, 4, 6, 4, 1**

2. Interpret the Coefficients

The numbers in Row 4 (**1, 4, 6, 4, 1**) are the coefficients that represent the number of possible ways to get a specific combination of Heads (H) and Tails (T):

Coefficient	Number of Heads (H)	Number of Tails (T)	Combinations (Examples)
1	4	0	HHHH

Coefficient	Number of Heads (H)	Number of Tails (T)	Combinations (Examples)
4	3	1	HHHT, HHTH, HTHH, THHH
6	2	2	HHTT, HTHT, HTTH, THHT, THTH, TTHH
4	1	3	HTTT, THTT, TTHT, TTTH
1	0	4	TTTT

3. Calculate Probabilities

To find the probability of a specific outcome, you divide the number of ways that outcome can happen (the coefficient) by the total number of possible outcomes.

- **Total possible outcomes:** The sum of the coefficients in Row 4 is $1 + 4 + 6 + 4 + 1 = \mathbf{16}$. (This is also 2^4).
- **Probability of getting exactly 2 Heads and 2 Tails:**
  - Number of ways to get 2H/2T is **6** (from the middle coefficient).
  - Probability = $\frac{\text{Number of Ways}}{\text{Total Outcomes}} = \frac{6}{16} = \frac{3}{8} = 37.5$

Pascal's Triangle directly gives you the number of combinations, eliminating the need to list every single possibility!

https://www.mathsisfun.com/pascals-triangle.html

- In 1662, a few days after Pascal died, a servant noticed a curious bulge in one of Pascal's jackets. The servant pulled open the lining to find hidden within it folded sheets of parchment and paper. Pascal had apparently carried them with him every day for the last eight years of his life. Scribbled on the sheets, in his handwriting, was a series of isolated words and phrases dated November 23, 1654. The writings were an emotional account of the trance, in which he described how God had come to him and in the space of two hours delivered him from his corrupt ways. Following that revelation, Pascal had dropped most of his friends, calling them "horrible attachments."
- He sold his carriage, his horses, his furniture, his library—everything except his Bible. He gave his money to the poor, leaving himself with so little that he often had to beg or borrow to obtain food. He wore an iron belt with points on the inside so that he was in constant discomfort and pushed the belt's spikes into his flesh whenever he found himself in danger of feeling happy. He denounced his studies of mathematics and science. Of his childhood fascination with geometry, he wrote, "I can scarcely remember that there is such a thing as geometry. I recognize geometry to be so useless...it is quite possible I shall never think of it again."
- Yet Pascal remained productive. In the years that followed the trance, he recorded his thoughts about God, religion, and life. Those thoughts were later published in a

book titled Pensées, a work that is still in print today. And although Pascal had denounced mathematics, amid his vision of the futility of the worldly life is a mathematical exposition in which he trained his weapon of mathematical probability squarely on a question of theology and created a contribution just as important as his earlier work on the problem of points.

Bernoulli

Jacob Bernoulli was one of the founding fathers of probability theory, and his key contribution, the **Law of Large Numbers**, forms the bedrock of the frequentist interpretation of probability.

Jacob Bernoulli and the Frequentist Perspective

Jacob Bernoulli's most significant work, **Ars Conjectandi** (*The Art of Conjecturing*), published posthumously in 1713, laid the mathematical foundation for the **Frequentist** view of probability through his key theorem:

The Weak Law of Large Numbers (WLLN)

This law mathematically proves that:

The **relative frequency** of a random event (e.g., the proportion of heads in coin flips) will converge to the **true underlying probability** of that event as the number of trials increases.

Aspect	Bernoulli's Frequentist Perspective
Definition of Probability	Probability is the **limit of relative frequency** over an infinite number of trials.
Focus	On the **data** and the procedure. The probability of an event is an **objective, fixed propert**
How to Find Probability	By **a posteriori** (after the fact) observation and repetition of the experiment many times.

Frequentist vs. Bayesian Perspective

The difference between the two perspectives boils down to their fundamental definition of what "probability" means and the role of prior knowledge.

Feature	Frequentist (Bernoulli's Legacy)
What is Probability?	An **objective, fixed long-run frequency** of an event.
Role of Parameters	Parameters (like the true probability of heads) are **fixed, unknown constants**.
Role of Prior Knowledge	**Ignored** or does not formally enter the model. Inference is based only on the **current data**
Conclusion	States the probability of the **data**, given a hypothesis (e.g., *p*-values, confidence intervals).

In short, Bernoulli provided the mathematical proof that you can **learn the true probability from data** by repeating an experiment (the frequentist method). Bayes provided the mathematical framework (**Bayes' Theorem**) for how you can **update your**

belief about a probability by incorporating both prior knowledge and new data (the Bayesian method).

Thomas Bayes and David Humes

David Hume (1711–1776)

Hume was a Scottish philosopher and is often associated with the philosophical underpinnings of skepticism and empiricism. His work, particularly his questioning of cause and effect and the problem of induction, is relevant to the philosophical debates surrounding probability and certainty that were happening during this period. He argued that our belief in causality comes from "custom and mental habit" rather than rational justification. While not a mathematician like many of his contemporaries, his philosophical ideas on the limits of human knowledge and inductive reasoning are part of the broader intellectual context in which statistical and probabilistic concepts were being developed.

Philosophical contribution: Questioned the basis of causality and inductive reasoning, influencing the intellectual context for probability theory.

Thomas Bayes (1701/1702–1761)

Bayes was an English statistician and minister. His most significant contribution is the theorem that now bears his name, which was published posthumously in 1763. Bayes's theorem provides a way to update the probability of a hypothesis as new evidence becomes available. This work was a critical step in moving the field from just calculating the probability of an event to inferring causes from observed data.

Key Contribution: His work, published posthumously, introduced the first method for inverse probability, which is now known as Bayes's theorem, a foundational concept for statistical inference.

David Hume and the Problem of Induction

Hume did not deny that humans use induction; rather, he argued that it **cannot be logically justified** through pure reason or experience.

Hume's Central Argument (The Circularity of Induction)

All inductive reasoning—from observing the sun rise every day to concluding that a flame will always cause heat—relies on the hidden assumption known as the **Principle of Uniformity of Nature (PUN)**: that **unobserved instances will resemble observed instances** (i.e., the future will be like the past).

Hume then asks how this PUN can be rationally justified:

1. **A priori reasoning (Logic):** The opposite of the PUN (that the future will *not* resemble the past) is conceivable and does not lead to a logical contradiction (e.g., you can easily imagine the sun not rising tomorrow). Therefore, the PUN cannot be justified by logic alone.
2. **A posteriori reasoning (Experience):** If we try to justify the PUN by saying, "It has always worked in the past, so it will work in the future," we fall into **circular reasoning**. We are using the very principle of induction to justify the principle of induction itself.

Hume concludes that our reliance on induction is based not on logic, but on **psychological habit** or **instinct**.

Inductive Reasoning vs. Deductive Reasoning

Inductive and deductive reasoning are the two main types of logical inference, distinguished by the direction of their movement and the certainty of their conclusions.

Inductive Reasoning

- **Direction: Specific to General** (often called **Bottom-up** logic).
- **Goal:** To **develop a theory** or find a general rule based on limited observations.
- **Certainty of Conclusion: Probable/Likely**. The conclusion contains new information not present in the premises, meaning it can only be probable, even if the premises are true.
- **Example:**
    - **Premises:** *Every swan I have ever seen is white.*
    - **Conclusion:** *Therefore, all swans are white.* (A new observation of a black swan could prove this conclusion false.)
- **Used in:** Forming scientific **hypotheses** and general laws (e.g., discovering the laws of physics).

Deductive Reasoning

- **Direction: General to Specific** (often called **Top-down** logic).
- **Goal:** To **test a theory** or apply a known general rule to a specific case.
- **Certainty of Conclusion: Certain/Guaranteed**. The conclusion is necessarily true if the premises are true (this is known as a **valid** argument).
- **Example:**
    - **Premises:** *All men are mortal. Socrates is a man.*

    - **Conclusion:** *Therefore, Socrates is mortal.* (The conclusion is inescapable and is contained within the premises.)
- **Used in:** Applying scientific laws or proving mathematical **theorems** (e.g., calculating the trajectory of a known object using Newton's laws).

Hume's Problem of Induction

David Hume's central problem with induction (in the 1740s) was skeptical and philosophical. He argued that there is no rational or logical justification for the belief that the future will resemble the past.

We assume the sun will rise tomorrow because it always has, but this reliance on past experience to justify an assumption about the future is circular and not based on deductive reason.

Hume concluded that induction is a matter of psychological habit or custom, not a matter of logical truth.

Bayes and the Sunrise

It is highly probable that Bayes' theorem was a direct, albeit indirect, response to David Hume's problem of induction.

Thomas Bayes developed his theorem to solve the "inverse problem" of probabilities, which he detailed in his posthumously published work, "An Essay towards solving a Problem in the Doctrine of Chances" (1763).

Given the number of times an unknown event has happened and failed (i.e., observed outcomes), what is the chance that the true, underlying probability of the event happening in a single trial lies somewhere between any two degrees of probability that can be named?

In essence, he was looking for a way to determine the probability of a cause (an unknown probability parameter) based on the observation of an effect (the recorded outcomes of trials). This was the reverse of most problems in probability at the time, which calculated the probability of an outcome given a known underlying cause or parameter.

Bayes illustrated this with a thought experiment involving balls thrown onto a square table to estimate the position of an initial ball based on where subsequent balls landed relative to it.

Dawn of Mathematical Probability

The earliest known work on probability theory is attributed to Gerolamo Cardano, a 16th-century Italian physician and avid gambler. Around 1550, Cardano penned a manuscript, Liber de Ludo Aleae (The Book on Games of Chance), which provided a rudimentary definition of probability as the ratio of favorable cases to the total number of possible cases. However, this work, not published until 1663, contained a significant logical flaw that exemplified the flawed intuition of the era. Cardano's "reasoning on the mean" led him to incorrectly conclude that the probability of a specific face appearing in three rolls of a die was equal to $p=3 \cdot 1/6=1/2$. This error demonstrated that a simple, intuitive understanding of chance was insufficient for complex problems.

The true catalyst for modern probability theory came a century later, in the correspondence of two brilliant French mathematicians, Blaise Pascal and Pierre de Fermat, in 1654. They were tasked by the Chevalier de Méré with solving the "problem of points," a famous gambling puzzle. The problem asked how to fairly divide the stakes in a game of chance that was interrupted before its conclusion. For example, if two players are playing a game where the first to reach a certain number of points wins the pot, how should the pot be divided if the game is stopped prematurely, based on the current score?

This was not a simple question about a single outcome; it forced Pascal and Fermat to develop a way to calculate the fair value of a stake based on the likelihood of all possible future outcomes. Their work laid the fundamental groundwork of probability theory and introduced the concept of expected value, a weighted sum of probabilities that remains a cornerstone of the discipline. This marked a profound departure from naive intuition and established the core idea of a new mode of mathematical thought.

The Gambler Who Forced the Issue

The person who, in a practical sense, forced the mathematical community to confront this calculation was **The Chevalier de Méré**, a French nobleman and passionate gambler.

- De Méré noticed a subtle difference between two related betting problems:
    1. Betting to get at least one "six" in **four rolls** of a single die. (He usually won).
    2. Betting to get at least one "double-six" in **24 rolls** of two dice. (He usually lost).
- He correctly felt that the odds in the first game were favorable and the odds in the second were unfavorable, but his own faulty, Cardan-like reasoning could not

explain *why*.

The Mathematicians Who Formalized the Correction

De Méré posed these and other gambling problems to **Blaise Pascal** in 1654, which led to Pascal's correspondence with **Pierre de Fermat**.

While their primary focus was the more complex "Problem of Points," they used the correct method to address de Méré's dice questions. This correct method involved what we now call the complement rule and the multiplication rule for independent events:

$$P(\text{at least one 6}) = 1 - P(\text{no 6's}) = 1 - \left(\frac{5}{6}\right)^4$$

The calculation of this $1 - P(\text{not } A)$ formula was demonstrated and formalized by these and other early probabilists, including **Christiaan Huygens**, whose 1657 treatise, *On Reasoning in Games of Chance*, was the first textbook on probability.

In summary:

- **Cardano** made the error by using simple addition.
- The error was definitively corrected by the **new mathematical framework** created by **Pascal, Fermat, and Huygens** in the mid-17th century. They used the multiplication principle (which forms the basis of the complement rule) to calculate the odds of independent events correctly.

The problem, calculating $P(\text{at least one 6}) = 1 - \left(\frac{5}{6}\right)^4 \approx 51.77\%$, represents the **correct** method for solving this problem.

Gerolamo Cardano, one of the earliest mathematicians to write about probability in his book *Liber de Ludo Aleae* (Book on Games of Chance), made a common error when faced with this type of problem.

Cardano's Error (The "Adding" Method)

Cardano is thought to have incorrectly calculated the probability by **adding** the individual probabilities of success for each roll, a method that is only valid for mutually exclusive events (which these are not).

- The probability of rolling a 6 on a single die is $\frac{1}{6}$.
- With four independent rolls, Cardano's logic would have been:

$$P(\text{at least one 6}) \approx \text{Sum of individual probabilities}$$

$$P(\text{at least one 6}) \approx \frac{1}{6} + \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{4}{6} = \frac{2}{3}$$

Therefore, Cardano thought the odds were approximately $\frac{4}{6}$ **or** 66.7% by adding.

Why the Adding Method is Wrong

The addition method is incorrect because it **overcounts** the outcomes where a 6 appears more than once (e.g., rolling two 6s, three 6s, or four 6s). The sum $\frac{4}{6}$ is an overestimation of the true probability of 51.77%.

The correct method uses the **complement rule**, which finds the probability of the opposite event (no 6s) and subtracts it from 1.

$$P(\text{at least one 6}) = 1 - P(\text{no 6s})$$

This avoids the complex overcounting issue.

The nascent field gained further momentum with the publication of the first text on probability theory, Christiaan Huygens' De Ratiociniis in Ludo Aleae (On Reasoning in Games and Chance) in 1657.

This work was later expounded upon by James Bernoulli, who, in his posthumously published workArs Conjectandi (1713), introduced the concepts of combinations and permutations that are still fundamental to probability today.

Expected Value ($\mathbf{E}$)

The Formula for Expected Value

The core idea derived from their work is that the value of an uncertain outcome is the **weighted average** of all possible outcomes, with the weights being the probability of each outcome occurring. This is the definition of **Expected Value (E)**.

The general formula for a game or wager with a finite number of possible outcomes is:

$$E = \sum_{i=1}^{k} (P_i \cdot V_i)$$

Where:

- E is the **Expected Value** (the fair share of the pot a player should receive).
- P_i is the **Probability** of the i^{th} possible future outcome (e.g., Player A winning the game).
- V_i is the **Value** of the i^{th} possible future outcome (e.g., the amount of the pot won if that outcome occurs).

Application to the "Problem of Points"

In the context of the "Problem of Points," this formula is applied to determine the fair division of the total prize (T) based on the remaining games needed for each player.

Let's say a game is interrupted, and Player A needs a more points to win, and Player B needs b more points to win.

1. **Calculate the Probabilities:** Pascal and Fermat used combinatorial methods (often involving what is now **Pascal's Triangle**) to calculate:
    - P_A: The probability that Player A will win the remaining points.
    - P_B: The probability that Player B will win the remaining points. (Note: $P_A + P_B = 1$).
2. **Determine the Fair Share (Expected Value):**
    - The fair share for **Player A (E_A)** is the probability of A winning times the full pot amount (T):
    $$E_A = P_A \cdot T$$
    - The fair share for **Player B (E_B)** is the probability of B winning times the full pot amount (T):
    $$E_B = P_B \cdot T$$

This result was the first rigorous mathematical demonstration of how to quantify the value of a future chance, which became the bedrock of modern probability theory.

Expected Value of Rolling a Die

The expected value of rolling a fair, **six-sided die** is **3.5**.

The expected value ($E[X]$) is the long-term average outcome if you were to repeat an experiment many times. It is calculated by multiplying each possible outcome by its probability and summing the results.

Calculation

The formula for Expected Value is:

$$E[X] = \sum x \cdot P(x)$$

where:

- x is a possible outcome (the numbers 1, 2, 3, 4, 5, or 6).
- $P(x)$ is the probability of that outcome (for a fair six-sided die, this is $\frac{1}{6}$ for every number).

Outcome (x)	Probability ($P(x)$)	$x \cdot P(x)$
1	$\frac{1}{6}$	$1 \cdot \frac{1}{6} = \frac{1}{6}$

Outcome (x)	Probability ($P(x)$)	$x \cdot P(x)$
2	$\frac{1}{6}$	$2 \cdot \frac{1}{6} = \frac{2}{6}$
3	$\frac{1}{6}$	$3 \cdot \frac{1}{6} = \frac{3}{6}$
4	$\frac{1}{6}$	$4 \cdot \frac{1}{6} = \frac{4}{6}$
5	$\frac{1}{6}$	$5 \cdot \frac{1}{6} = \frac{5}{6}$
6	$\frac{1}{6}$	$6 \cdot \frac{1}{6} = \frac{6}{6}$
Sum	**1**	$\frac{21}{6}$

$$E[X] = \frac{1}{6} + \frac{2}{6} + \frac{3}{6} + \frac{4}{6} + \frac{5}{6} + \frac{6}{6} = \frac{21}{6} = \mathbf{3.5}$$

Interpretation

The expected value of **3.5** is the theoretical **mean** of the outcomes. It's important to note:

1. **It is not a possible outcome:** You can never roll a 3.5 on a standard die.
2. **It is the average:** If you roll the die hundreds or thousands of times, the **average** of all the numbers you roll will get closer and closer to **3.5**.

Bussiness Example for Expected Value

An insurance company wants to determine the annual premium for a new car policy. They use historical data to estimate the probability of a driver being in an accident and the potential cost of a claim.

The Expected Value (EV) calculation here is from the **company's perspective**, representing the average profit (or loss) per policy.

Scenario Data:

Outcome (x)	Potential Payout (Cost to Company)	Probability ($P(x)$)
No Accident	\$0	95.00% (0.95)
Minor Accident	$-$\text{5,000}	4.00% (0.04)
Major Accident	$-$\text{30,000}	1.00% (0.01)
Total		100.00% (1.00)

1. Calculate the Expected Payout (Expected Loss)

The company first calculates the **Expected Payout** (the average amount they expect to pay out per policy over the long run).

$$E[\text{Payout}] = \sum \text{Payout} \cdot P(\text{Payout})$$
$$E[\text{Payout}] = (\$0 \cdot 0.95) + (-\$5,000 \cdot 0.04) + (-\$30,000 \cdot 0.01)$$
$$E[\text{Payout}] = \$0 + (-\$200) + (-\$300)$$
$$E[\text{Payout}] = -\$\mathbf{500}$$

This means that, on average, the insurance company **expects to pay out $500 per policy** to cover claims.

2. Determine the Minimum Break-Even Premium

To simply break even, the premium must equal the expected payout:

$$\text{Break-Even Premium} = \text{Expected Payout} = \$500$$

3. Calculate the Final Expected Value (Profit)

To make a profit and cover operating costs (salaries, rent, advertising, etc.), the company adds a **profit margin** (let's say $150) to the premium.

If they set the **Premium at $650**, the Expected Value ($E[X]$) or **Expected Profit** for the company is:

$$E[\text{Profit}] = \text{Premium} - E[\text{Payout}]$$
$$E[\text{Profit}] = \$650 - \$500 = \mathbf{\$150}$$

By charging a premium of $650, the company has an **Expected Profit of + $150 per policy**. This is the average profit the company anticipates earning per customer in the long run. If the EV were negative, the business would eventually lose money and fail.

Python: Expected Value for a Business Decision

A company is deciding whether to launch a new software product. They estimate three possible outcomes:

- **Success (High Profit):** Financial Outcome is **$500,000** with a Probability of **30% (0.30)**.
- **Moderate Success:** Financial Outcome is **$100,000** with a Probability of **50% (0.50)**.
- **Failure (Loss):** Financial Outcome is **-$250,000** with a Probability of **20% (0.20)**.

The goal is to calculate the **Expected Value** to determine the project's long-term average value.

```
# expected value problem
outcomes = [500000, 100000, -250000]
```

```
probabilities = [0.30, 0.50, 0.20]

def calculate_expected_value(outcomes, probabilities):
    """
    Calculates the Expected Value (EV) for a set of discrete outcomes.

    EV = Sum(Outcome * Probability)
    """
    if len(outcomes) != len(probabilities):
        raise ValueError("The number of outcomes must match the number of proba

    if not 0.999 <= sum(probabilities) <= 1.001:
        print("Warning: Probabilities do not sum exactly to 1.0.")

    expected_value = 0

    for i in range(len(outcomes)):
        term = outcomes[i] * probabilities[i]
        expected_value += term
        print(f"Scenario {i+1}: {outcomes[i]:,.2f} * {probabilities[i]:.2f} = {

    return expected_value

expected_return = calculate_expected_value(outcomes, probabilities)

print("\n" + "="*40)
print(f"The Expected Value of the new product launch is: \${expected_return:,.2
print("="*40)
print("\nInterpretation:")
if expected_return > 0:
    print(f"The positive EV of \${expected_return:,.2f} suggests that, over mar
elif expected_return < 0:
    print(f"The negative EV of \${expected_return:,.2f} suggests the project is
else:
    print("The EV is $0, meaning the project is expected to break even in the l
```

```
def generate_pascals_row(n):
    """
    Generates the n-th row of Pascal's Triangle.
    (Row 0 is the top row: [1])
    """
    # The first row (row 0) is simply [1]
    if n == 0:
        return [1]

    # Initialize the current row with the starting '1'
    current_row = [1]

    # Get the previous row to calculate the new values
    # Recursive call to get the row above
```

```
        previous_row = generate_pascals_row(n - 1)

        # Calculate the inner elements of the current row
        # Each element is the sum of the two elements directly above it
        for i in range(len(previous_row) - 1):
            # The new element is the sum of the element at 'i' and 'i+1' in the pre
            new_element = previous_row[i] + previous_row[i+1]
            current_row.append(new_element)

        # The row always ends with a '1'
        current_row.append(1)

        return current_row

    # --- Examples ---

    # Row 0
    print(f"Row 0: {generate_pascals_row(0)}")

    # Row 4 (used in the previous probability example)
    print(f"Row 4: {generate_pascals_row(4)}")

    # Row 6
    print(f"Row 6: {generate_pascals_row(6)}")
```

```
Row 0: [1]
Row 4: [1, 4, 6, 4, 1]
Row 6: [1, 6, 15, 20, 15, 6, 1]
```

Double-click (or enter) to edit

The Binomial Distribution

The binomial distribution's historical context is rooted in the early development of **probability theory** and its application to **games of chance** during the late 17th and early 18th centuries.

The concept was formally derived and published by the Swiss mathematician **Jacob Bernoulli** in his monumental work, ***Ars Conjectandi*** (Latin for "The Art of Conjecturing").

The Age of Early Probability

The 17th century saw the formal mathematical study of probability emerge, driven largely by aristocratic interest in quantifying risk in games of chance. Pioneers like **Blaise Pascal** and **Pierre de Fermat** laid the groundwork by solving famous problems, such as the **Problem of Points** (how to fairly divide stakes in an interrupted game).

Jacob Bernoulli and Ars Conjectandi

Jacob Bernoulli (1654–1705) took these initial ideas and systematically built a rigorous foundation for probability theory.

- **The Work:** Bernoulli developed the principles of the binomial distribution between 1684 and 1689. His book, ***Ars Conjectandi***, was written during this time but was only published posthumously in **1713** by his nephew, Nicolaus I Bernoulli.
- **The Problem Addressed:** Bernoulli's work focused on sequences of independent trials, each with only two outcomes (often called a **Bernoulli Trial** or success/failure). He sought a mathematical way to determine the probability of getting a specific number of "successes" over a fixed number of trials, where the probability of success, p, might not be $1/2$.
- **The Insight:** Bernoulli proved that the probability of k successes in n trials is given by the k-th term in the expansion of the binomial expression $(p+q)^n$, where $q = 1-p$. This relationship is where the distribution gets its name.

The Law of Large Numbers

Crucially, Bernoulli used the distribution to prove what is now known as **Bernoulli's Theorem**, the first version of the **Weak Law of Large Numbers (WLLN)**. This was perhaps his most profound contribution, as it transitioned probability theory from pure theoretical games into a practical tool for statistics.

The theorem essentially proved that as the number of trials (n) increases, the observed proportion of successes will converge to the theoretical probability (p) of success. This provided a formal mathematical justification for estimating unknown probabilities (like the proportion of black pebbles in an urn) by performing repeated trials, thus moving the field toward **statistical inference**.

Link to the Normal Distribution

The binomial distribution later became the foundational link to the **Normal Distribution**. As discussed previously, **Abraham de Moivre** used the binomial distribution for very large n to derive the bell-shaped curve, formalizing the **De Moivre–Laplace Theorem** and establishing the first version of the **Central Limit Theorem**.

The initial discovery of the normal distribution, or the bell curve, came from studying the binomial distribution's behavior for a large number of trials.

De Moivre (1733)

The key figure in this development was the French mathematician **Abraham de Moivre**.

- **The Problem:** De Moivre was solving complex problems in gambling that required calculating probabilities for the **binomial distribution** for a very large number of trials (n). The factorial calculations involved (like $n!$) were extremely tedious.
- **The Solution:** He noticed that as the number of trials (n) increased, the discrete probability histogram of the binomial distribution (for $p = 0.5$) took on a smooth, symmetric, **bell-shaped curve**. He successfully derived the formula for this curve, which is the **Normal Distribution**.
- **The Result:** His work established that the normal distribution could be used to **approximate** the binomial distribution when n is large, drastically simplifying calculations.

The De Moivre–Laplace Theorem

De Moivre's initial finding was later generalized and formalized by **Pierre-Simon Laplace** (around 1812).

The **De Moivre–Laplace Theorem** is a specific case of the **Central Limit Theorem (CLT)**, and it states that as the number of trials (n) in a binomial distribution (with parameters n and p) becomes very large, the distribution of the number of successes approaches the **Normal Distribution** with:

$$\text{Mean } (\mu) = np$$
$$\text{Variance } (\sigma^2) = np(1-p)$$

While the discovery of the normal distribution was rooted in the binomial approximation, it wasn't the *only* path to the curve. Later, **Carl Friedrich Gauss** independently derived the same formula (which is why it's also called the Gaussian distribution) when studying **measurement errors** in astronomy, treating it as the "law of errors." However, the **temporal precedence** for its initial mathematical form goes to de Moivre's work on the binomial distribution.

The Binomial Coefficient

The expected value of a single success or failure is a fundamental concept in probability theory.

The Binomial Coefficient can be used to calculate the number of ways to choose k successes (Heads) from n total trials (flips), **without regard to the order of the outcomes**.

Formal Equation

The number of combinations of selecting k items from a set of n items is:

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

Where:

- n is the **total number of trials** (coin flips).
- k is the **number of successful outcomes** (number of Heads, H).
- ! denotes the **factorial** (e.g., $4! = 4 \times 3 \times 2 \times 1$).

Example Calculation

What can you expect flipping a coin **4 times** ($n = 4$) and getting **exactly 2 Heads** ($k = 2$):

1. **Total Flips (n)** = 4
2. **Required Heads (k)** = 2

$$\text{Coefficient} = \binom{4}{2} = \frac{4!}{2!(4-2)!} = \frac{4!}{2!2!} = \frac{4 \times 3 \times 2 \times 1}{(2 \times 1)(2 \times 1)} = \frac{24}{4} = \mathbf{6}$$

This result of **6** matches the coefficient in your table for the combination "2 Heads and 2 Tails."

Connection to the Binomial Distribution

Binomial coefficients become apparent — $1, 4, 6, 4, 1$ (Pascal Triangle) for $n = 4$ flips — are the values of the Binomial Coefficient $\binom{n}{k}$ for $k = 0, 1, 2, 3, 4$.

This coefficient forms the first part of the full equation for the **Binomial Probability Mass Function (PMF)**, which calculates the probability of getting exactly k successes in n trials:

$$P(X = k) = \binom{n}{k} \cdot p^k \cdot (1-p)^{n-k}$$

Where:

- $\binom{n}{k}$ is the **coefficient** (number of ways the event can happen).
- p is the probability of a single success (e.g., 0.5 for heads).
- $p^k \cdot (1-p)^{n-k}$ is the probability of *any one specific sequence* occurring.

For the example of exactly 2 Heads in 4 flips with a fair coin ($p = 0.5$):

$$P(X = 2) = \mathbf{6} \cdot (0.5)^2 \cdot (1-0.5)^{4-2}$$
$$P(X = 2) = 6 \cdot (0.25) \cdot (0.25) = 6 \cdot 0.0625 = \mathbf{0.375}$$

This confirms the stated probability: 37.5% (or $\frac{6}{16}$).

6?

Here is the step-by-step breakdown of how the Binomial Coefficient formula produces the number **6** for "4 over 2," which represents the combinations of 2 Heads (successes) in 4 coin flips ($n = 4, k = 2$).

Here is the formal calculation using the formula for Combinations, $\binom{n}{k}$:

Calculating the Binomial Coefficient $\binom{4}{2}$

The formula for the number of combinations of choosing k items from a set of n items is:

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

In your example, the total number of trials is $n = 4$ (flips) and the number of successes is $k = 2$ (Heads).

Step 1: Substitute the values into the formula.

Substitute $n = 4$ and $k = 2$:

$$\binom{4}{2} = \frac{4!}{2!(4-2)!}$$

Step 2: Simplify the term in the parentheses.

$$(4-2)! = 2!$$

$$\binom{4}{2} = \frac{4!}{2!2!}$$

Step 3: Expand the factorials.

Recall that the factorial $n!$ is the product of all positive integers less than or equal to n.

- $4! = 4 \times 3 \times 2 \times 1 = 24$
- $2! = 2 \times 1 = 2$

$$\binom{4}{2} = \frac{4 \times 3 \times 2 \times 1}{(2 \times 1) \times (2 \times 1)}$$

$$\binom{4}{2} = \frac{24}{2 \times 2} = \frac{24}{4}$$

Step 4: Perform the final division.

$$\frac{24}{4} = \mathbf{6}$$

Result and Interpretation

The result, **6**, confirms that there are exactly **6 possible unique sequences** of 4 coin flips that contain exactly 2 Heads and 2 Tails.

These sequences are:

1. HHTT
2. HTHT
3. HTTH
4. THHT
5. THTH
6. TTHH

$\sum_{k=0}^{n} \binom{n}{k} = 2^n$

The sum of the coefficients in any row of Pascal's Triangle is equal to 2 raised to the power of the row number.

This is a fundamental property of Pascal's Triangle.

The Rule: $\sum_{k=0}^{n} \binom{n}{k} = 2^n$

The relationship can be stated formally as:

$$\text{Row Sum} = 2^{\text{Row Number } (n)}$$

Where:

- n is the row number (starting with $n = 0$).
- The terms in the row are the binomial coefficients, $\binom{n}{k}$.

Examples

Row Number (n)	Pascal's Triangle Row (Coefficients)	Row Sum (2^n)	Calculation
0	1	1	$2^0 = 1$
1	$1, 1$	2	$2^1 = 2$
2	$1, 2, 1$	4	$2^2 = 4$
3	$1, 3, 3, 1$	8	$2^3 = 8$
4	$1, 4, 6, 4, 1$	16	$2^4 = 16$

Why This is True (Connection to Outcomes)

This relationship holds true because the sum of the coefficients in row n represents the **total number of possible outcomes** in an experiment with n trials (or n coin flips).

For any trial, there are always **two possible outcomes** (e.g., Heads or Tails). Therefore, for n independent trials, the total number of distinct outcomes is $2 \times 2 \times \cdots \times 2$ (n times), or $\mathbf{2^n}$.

The coefficients themselves tell you *how those* 2^n *outcomes are distributed* based on the number of successes:

$$\underbrace{\text{Total Outcomes}}_{\mathbf{2^n}} = \underbrace{\text{Ways to get 0 Successes}}_{\binom{n}{0}} + \underbrace{\text{Ways to get 1 Success}}_{\binom{n}{1}} + \cdots + \underbrace{\text{W}}$$

In the context of the 4 coin flips you discussed:

$$\mathbf{16} = 1(\text{for } 0H/4T) + 4(\text{for } 1H/3T) + 6(\text{for } 2H/2T) + 4(\text{for } 3H/1T) + 1(\text{for } 4$$

Early Distributions and Expectation

- Arithmetical expectation and combinatorics formalize “chance” as mathematics rather than philosophy or custom.
- Insurance and mortality: Graunt and Petty analyze Bills of Mortality; Halley constructs a life table—probability meets demography and actuarial science.
- Key figures: Cardano, Pascal, Fermat, Huygens, Graunt, Petty, Halley, Bernoulli (Jacob).
- Key terms: expectation, sample space, combinations, law of large numbers (Bernoulli), life table, actuarial risk.

Probability I

Probability is a branch of mathematics and statistics that provides a numerical description of how **likely an event is to occur**.

The probability of any event is quantified as a single number between **0 and 1**, inclusive:

- **0** means the event is **impossible** (e.g., the probability of a coin landing on its edge).
- **1** means the event is **certain** (e.g., the probability of the sun rising tomorrow, in practical terms).
- **0.5** means the event is **equally likely** to occur or not occur (e.g., the probability of a fair coin landing on heads).

Defining Probability

There are two primary ways to define and calculate probability, both of which are implied in the examples you provided:

1. Theoretical (or Classical) Definition

This definition relies on mathematical reasoning and assumes all outcomes in an experiment are **equally likely**.

$$P(\text{Event}) = \frac{\text{Number of Favorable Outcomes}}{\text{Total Number of Possible Outcomes}}$$

- **Example:** When rolling a fair six-sided die, the theoretical probability of rolling a **4** is $\frac{1}{6}$, because there is 1 favorable outcome out of 6 total possible outcomes.

2. Empirical (or Frequentist) Definition

This definition is based on **observation** and **experimentation**, as highlighted in your examples of Astronomy and Roman Gambling.

$$P(\text{Event}) \approx \frac{\text{Number of Times the Event Occurred}}{\text{Total Number of Trials (Observations)}}$$

- **Example:** If you flip a coin 1,000 times and it lands on heads 503 times, the empirical probability of getting a head is $\frac{503}{1000} = 0.503$.

The Frequentist definition states that as the number of trials increases (approaches infinity), the empirical probability will converge toward the theoretical probability.

The Core Components

1. Sample Space ($\mathbf{S}$ or $\mathbf{\Omega}$)

- **Definition:** The **set of all possible outcomes** of a random experiment.
- **Example:** When rolling a single six-sided die, the sample space is $S = \{1, 2, 3, 4, 5, 6\}$.

2. Outcome (or Sample Point)

- **Definition:** A single, specific result of a random experiment. It is an element of the sample space.
- **Example:** Rolling a **3** is one outcome.

3. Event ($\mathbf{A}, \mathbf{B}, \mathbf{E}$, etc.)

- **Definition:** A specific outcome or a **collection of outcomes** (a subset of the sample space) to which a probability is assigned.
- **Example:** Rolling an **even number** is an event, $E = \{2, 4, 6\}$.

4. Probability ($\mathbf{P(A)}$)

- **Definition:** A numerical measure of the likelihood that a specific event A will occur. It is always a value between **0 and 1**, inclusive.
- **Notation:** $P(A)$ reads "The probability of event A."

Relationships Between Events

These terms describe how two or more events interact with each other.

5. Complement ($\mathbf{A^c}$ or $\mathbf{A'}$)

- **Definition:** The event that **Event A does not occur**. The complement of A includes all outcomes in the sample space that are not in A.
- **Rule:** $P(A) + P(A^c) = 1$
- **Example:** If A is rolling a 4 on a die, the complement A^c is rolling any number **other than 4** ($\{1, 2, 3, 5, 6\}$).

6. Union ($\mathbf{A \cup B}$)

- **Definition:** The event that **A or B or both** occur. (Corresponds to the "OR" rule, often involving addition).
- **Example:** If A is rolling a number < 3 ($\{1, 2\}$) and B is rolling an even number ($\{2, 4, 6\}$), the union $A \cup B$ is rolling a $\{1, 2, 4, 6\}$.

7. Intersection ($\mathbf{A \cap B}$)

- **Definition:** The event that **both A and B** occur simultaneously. (Corresponds to the "AND" rule, often involving multiplication).
- **Example:** Using the events A and B above, the intersection $A \cap B$ is rolling a number that is both < 3 and even: $\{\mathbf{2}\}$.

8. Mutually Exclusive (or Disjoint) Events

- **Definition:** Two events that **cannot occur at the same time**; their intersection is the empty set ($\emptyset$).
- **Rule:** If A and B are mutually exclusive, $P(A \text{ and } B) = P(A \cap B) = 0$.
- **Example:** Rolling an odd number and rolling an even number are mutually exclusive.

9. Independent Events

- **Definition:** The occurrence of one event **does not affect the probability** of the other event occurring.
- **Rule:** $P(A \text{ and } B) = P(A) \times P(B)$.

- **Example:** Flipping a coin and rolling a die. The coin's result doesn't change the die's probability.

10. Conditional Probability ($\mathbf{P(A \mid B)}$)

- **Definition:** The probability of **Event A occurring, given that Event B has already occurred**. This effectively reduces the size of the sample space to only include outcomes where B happened.
- **Rule:** $P(A \mid B) = \frac{P(A \cap B)}{P(B)}$
- **Example:** The probability of drawing an Ace on the second draw, **given that** the first card drawn was also an Ace (without replacement).

Factorial

$n!$

5! = 1 * 2 * 3 * 4 * 5

```
import numpy as np
from math import factorial

n = 5
fact = 1
test = 1

print(np.arange(1, n+1))
print(np.arange(1, n+1).prod())
for i in range(1, n+1):
    fact = fact * i
    test *= i # assignment operator https://www.w3schools.com/python/python_ope

print('fact:', fact)
print('test:', test)
print('comprehension:', np.prod([i for i in range(1, n+1)]))
print('factorial:', factorial(n))
```

```
[1 2 3 4 5]
120
fact: 120
test: 120
comprehension: 120
factorial: 120
```

Permutations

Order does matter (permutation, position)

The combination of a safe

$${}_nP_r = \frac{n!}{(n-r)!}$$

with repetitions

$$n^r$$

```
from itertools import permutations, product

l = [1, 2, 3]
n = len(l)
r = 3

perms = permutations(l)
print(list(perms))
print(factorial(n)/factorial(n - r))
print()
print('with repetitions (n^r)')
print([p for p in product(l, repeat=r)])
print(n**r)
```

```
[(1, 2, 3), (1, 3, 2), (2, 1, 3), (2, 3, 1), (3, 1, 2), (3, 2, 1)]
6.0

with repetitions (n^r)
[(1, 1, 1), (1, 1, 2), (1, 1, 3), (1, 2, 1), (1, 2, 2), (1, 2, 3), (1, 3, 1), (1
27
```

Example

$${}_nP_r = \frac{n!}{(n-r)!}$$

```
from itertools import permutations, product

l = [1, 2, 3]
n = len(l)
r = 2

perms = permutations(l, r)
print(list(perms))

print(factorial(n)/factorial(n - r))
```

```
[(1, 2), (1, 3), (2, 1), (2, 3), (3, 1), (3, 2)]
6.0
```

Another Example

${}_{16}P_3 = \frac{16!}{(16-3)!}$ = 16 x 15 x 14 = 3,360

```
# solve
n = 16
r = 3
print(factorial(n)/factorial(n - r))
print(16 * 15 * 14)
```

```
3360.0
3360
```

${}_{10}P_2 = \frac{10!}{(10-2)!}$ = 10 x 9 = 90

```
n = 10
r = 2
print(factorial(n)/factorial(n - r))
print(10 * 9)
```

```
90.0
90
```

Four digit combination using number 0 - 9 without repetition

${}_{10}P_4$

```
# how many permutations without repetition?
n = 10
r = 4
print(factorial(n)/factorial(n - r))
print(10 * 9 * 8 * 7)
```

```
5040.0
5040
```

Combinations

Order doesn't matter

Items on a pizza

$C(n,r) = C_r^n = {}_nC_r = \binom{n}{r} = \frac{n!}{r!(n-r)!}$

```
from itertools import combinations
```

```
l = [1, 2, 3]
n = len(l)
r = 2
combs = combinations(l, r)
print(list(combs))
print(factorial(n)/(factorial(r) * factorial(n - r)))
```

```
[(1, 2), (1, 3), (2, 3)]
3.0
```

Practice

Solve ${}_nC_r = \binom{n}{r} = \frac{n!}{r!(n-r)!}$

Where n = 10 and r = 4

```
# solve
n = 10
r = 4
print(factorial(n)/(factorial(r) * factorial(n - r)))
```

```
210.0
```

Relationship with permutations

${}_nC_r$ * $r!$ = permutations

```
# solve and compare to permutation where n = 10 and r = 4
# see figure 1 http://mathandmultimedia.com/2010/01/02/intro-to-combinations/
210 * factorial(4)
```

```
5040
```

With Replacement

${}_{n+r-1}C_r = \frac{(r+n-1)!}{r!(n-1)!}$

```
from itertools import combinations_with_replacement

combs = combinations_with_replacement(l, r)
print(list(combs))
print(factorial(n+r-1)/(factorial(r)*factorial(n-1)))
```

```
[(1, 1, 1, 1), (1, 1, 1, 2), (1, 1, 1, 3), (1, 1, 2, 2), (1, 1, 2, 3), (1, 1, 3,
715.0
```

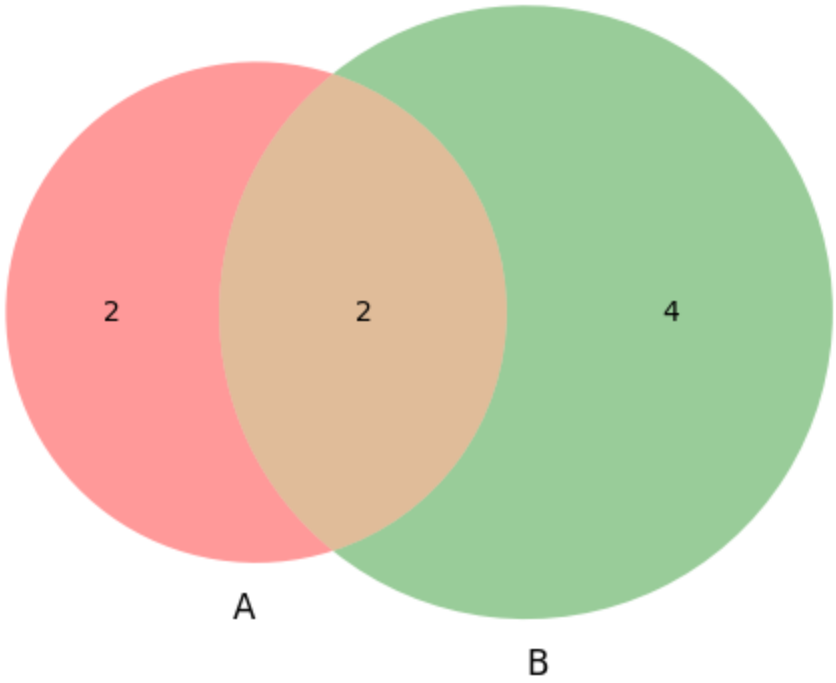
Intersections

$A \cap B = \{ x : x \in A \; and \; x \in B \}$

order doesn't matter $A \cap B$ or $B \cap A$

```
# https://anaconda.org/conda-forge/matplotlib-venn
# https://pypi.org/project/matplotlib-venn/
# https://towardsdatascience.com/visualizing-intersections-and-overlaps-with-py
# https://practicaldatascience.co.uk/data-science/how-to-visualise-data-using-v

import matplotlib.pyplot as plt
from matplotlib_venn import venn2, venn2_circles

set1 = {1, 2, 3, 4}
set2 = {3, 4, 5, 6, 7, 8}
print([value for value in set1 if value in set2])
print(set1 & set2)
print(set1.intersection(set2))

venn2([set1, set2])
plt.show()
```

```
[3, 4]
{3, 4}
{3, 4}
```



Independent

$$P(A \cap B) = P(A) * P(B)$$

```
# what's the probability of a random pick of a number from 1 - 10 if P(A) less
# filter https://www.pythonlikeyoumeanit.com/Module2_EssentialsOfPython/Iterabl
event_space = [1, 2, 3, 4, 5, 6, 7, 8, 9, 10]
N = len(event_space)
a = filter(lambda x: x < 5, event_space)
a = list(a)
print(a)
pa = len(list(a))/N
print(pa)
b = filter(lambda x: x % 2 == 1, event_space)
b = list(b)
print(b)
pb = len(list(b))/N
print(pb)
print(f'Probability of < 5 and odd: {pa * pb}')
```

```
[1, 2, 3, 4]
0.4
[1, 3, 5, 7, 9]
0.5
Probability of < 5 and odd: 0.2
```

Dependent

$$P(A \cap B) = P(A) * P(B|A) = P(B) * P(A|B)$$

also expressed as P(A and B) = P(A) * P(B given A)

```
# https://www.mathsisfun.com/data/probability-events-conditional.html
# https://en.wikipedia.org/wiki/Sample_space
# number of outcomes in event / total possible outcomes in sample space

bag = ['red', 'red', 'red', 'blue', 'blue']
mcolor = 'blue'
print(f'Probability of picking two {mcolor} marbles from this bag:')
bag_dict = {item:bag.count(item) for item in bag}
print(bag_dict)

# P(A) probability of picking blue 2 / 5
pa = bag_dict[mcolor] / len(bag)
print(f'Probability of picking a {mcolor}: {pa}')
bag.remove(mcolor)
print(f'Bag minus one {mcolor}:')
bag_dict = {item:bag.count(item) for item in bag}
print(bag_dict)
print(f'Probability of picking another {mcolor} with just four marbles: {(bag_c
print()
```

```
# P(B|A) probability of picking another blue given you already picked a blue
print(f'Probability of picking two {mcolor} marbles in a row: {pa * (bag_dict[m
```

```
Probability of picking two blue marbles from this bag:
{'red': 3, 'blue': 2}
Probability of picking a blue: 0.4
Bag minus one blue:
{'red': 3, 'blue': 1}
Probability of picking another blue with just four marbles: 0.25

Probability of picking two blue marbles in a row: 0.1
```

Unions

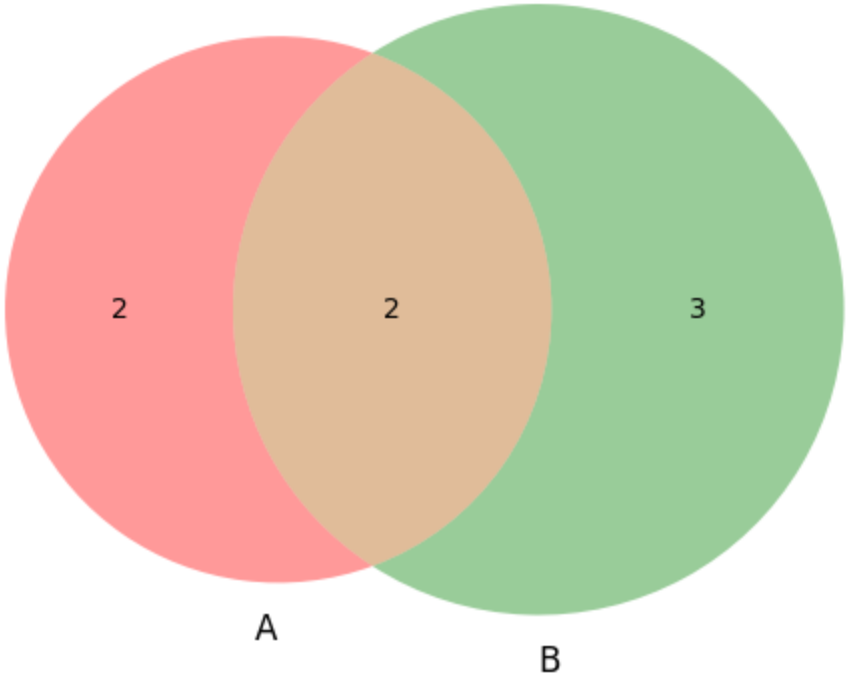
$P(A \cup B) = P(A) + P(B) - P(A \cap B)$

also expressed P(A or B)

order doesn't matter

```
# what's the probability of a random pick of a number from 1 - 10 is P(A) less
print(event_space)
print(a)
print(pa)
print(b)
print(pb)
print(f'Probability of the union of A or B is {pa + pb - (pa * pb)}')

venn2([set(a), set(b)])
plt.show()
```

```
[1, 2, 3, 4, 5, 6, 7, 8, 9, 10]
[1, 2, 3, 4]
0.4
[1, 3, 5, 7, 9]
0.5
Probability of the union of A or B is 0.7
```



Unions if mutually exclusive (vs. disjoint)

$P(A \cup B) = P(A) + P(B) - P(A \cap B)$ where $P(A \cap B) = 0$

so

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

Mutually Exclusive: $P(A \cap B) = 0$

Disjoint (dealing in sets): $A \cap B = 0$

```
# is drawing an even number or odd number mutually exclusive?
print(event_space)
N = len(event_space)
c = filter(lambda x: x % 2 == 0, event_space)
c = list(c)
print(c)
pc = len(list(c))/N
print(pc)
d = filter(lambda x: x % 2 == 1, event_space)
d = list(d)
print(d)
pd = len(list(d))/N
```

```
print(pd)
print(set(c).intersection(set(d)))

venn2([set(c), set(d)])
plt.show()
```

```
[1, 2, 3, 4, 5, 6, 7, 8, 9, 10]
[2, 4, 6, 8, 10]
0.5
[1, 3, 5, 7, 9]
0.5
set()
```

Compliment

Compliment of A is $\bar{A}$

Is the compliment of A mutually exclusive with A?

The Bernoulli Trial

A foundational building block of probability is the **Bernoulli trial**, a single experiment that has only two possible outcomes, conventionally labeled "**success**" and "**failure**". These trials are assumed to be **independent**, with the probability of success, p, and the probability of failure, $1-p$, remaining constant. When a sequence of n such trials is conducted, the number of successes follows a **binomial distribution**, which is a **discrete probability distribution**.

The probability of observing exactly k successes in n independent trials is given by the **probability mass function (PMF)**:

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

In this formula, $\binom{n}{k}$ is the **binomial coefficient**, which represents the number of ways to choose k successes from n trials. The term $p^k(1-p)^{n-k}$ represents the probability of a specific sequence of k successes and $n-k$ failures.

The **mean (expected value)** and **variance** of a binomial distribution are given by the simple formulas:

$$\mathrm{E}[X] = np$$

$$\mathrm{Var}(X) = np(1-p)$$

These results derive from the linearity of expectation and the assumption of independence.

19th Century: Error Theory, Least Squares, and Statistical Thinking (1800–1900)

Error distributions and estimation

- Legendre and Gauss propose Least Squares for orbit determination; Gauss links it to the normal law and maximum likelihood reasoning.
- Bessel introduces the n−1 degrees-of-freedom correction for variance estimation (Bessel's correction).

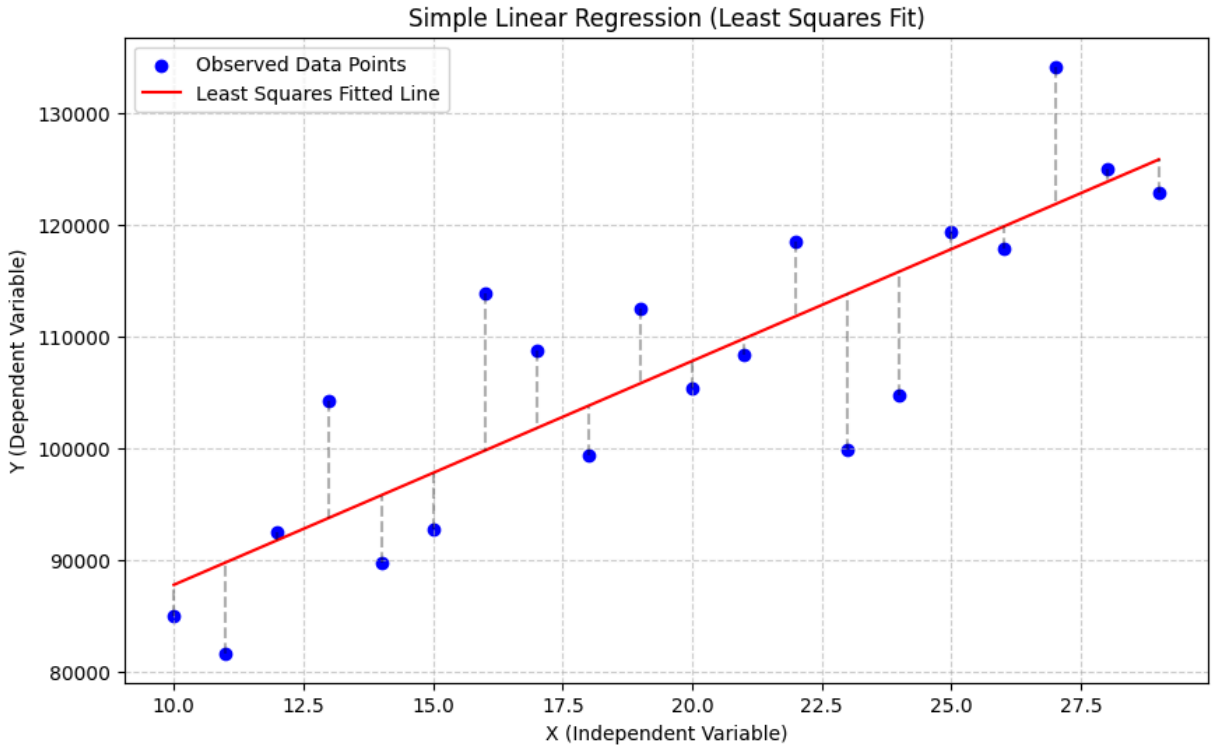
Least Squares

The method of least squares was developed independently by Legendre and Gauss to solve a very practical problem in astronomy and geodesy: how to combine multiple, flawed measurements to estimate the true value of an unknown quantity, like a comet's orbit or a planet's position.

Legendre's Contribution (1805): French mathematician Adrien-Marie Legendre was the first to publish the method in his work Nouvelles méthodes pour la détermination des orbites des comètes. He presented it simply as a computational technique for finding a line that best fits a set of data points by minimizing the sum of the squares of the errors (residuals).

```
# fitted model
import numpy as np
import matplotlib.pyplot as plt
```

```
np.random.seed(42)
X = np.arange(10, 30) # X values from 10 to 29
# True relationship: Y = 50000 + 3000 * X + noise
true_slope = 3000
true_intercept = 50000
noise = np.random.normal(0, 10000, size=len(X))
Y = true_intercept + true_slope * X + noise

x_mean = np.mean(X)
y_mean = np.mean(Y)
numerator = np.sum((X - x_mean) * (Y - y_mean))
denominator = np.sum((X - x_mean)**2)
b_slope = numerator / denominator
a_intercept = y_mean - b_slope * x_mean
Y_predicted = a_intercept + b_slope * X

plt.figure(figsize=(10, 6))
plt.scatter(X, Y, color='blue', label='Observed Data Points')
plt.plot(X, Y_predicted, color='red', label='Least Squares Fitted Line')

for i in range(len(X)):
    plt.plot([X[i], X[i]], [Y[i], Y_predicted[i]], 'k--', alpha=0.3)

plt.title('Simple Linear Regression (Least Squares Fit)')
plt.xlabel('X (Independent Variable)')
plt.ylabel('Y (Dependent Variable)')
plt.legend()
plt.grid(True, linestyle='--', alpha=0.6)
plt.show()

print(f"Calculated Slope (b): {b_slope:.2f}")
print(f"Calculated Intercept (a): {a_intercept:.2f}")
```




```
Calculated Slope (b): 2004.35
Calculated Intercept (a): 67702.14
```

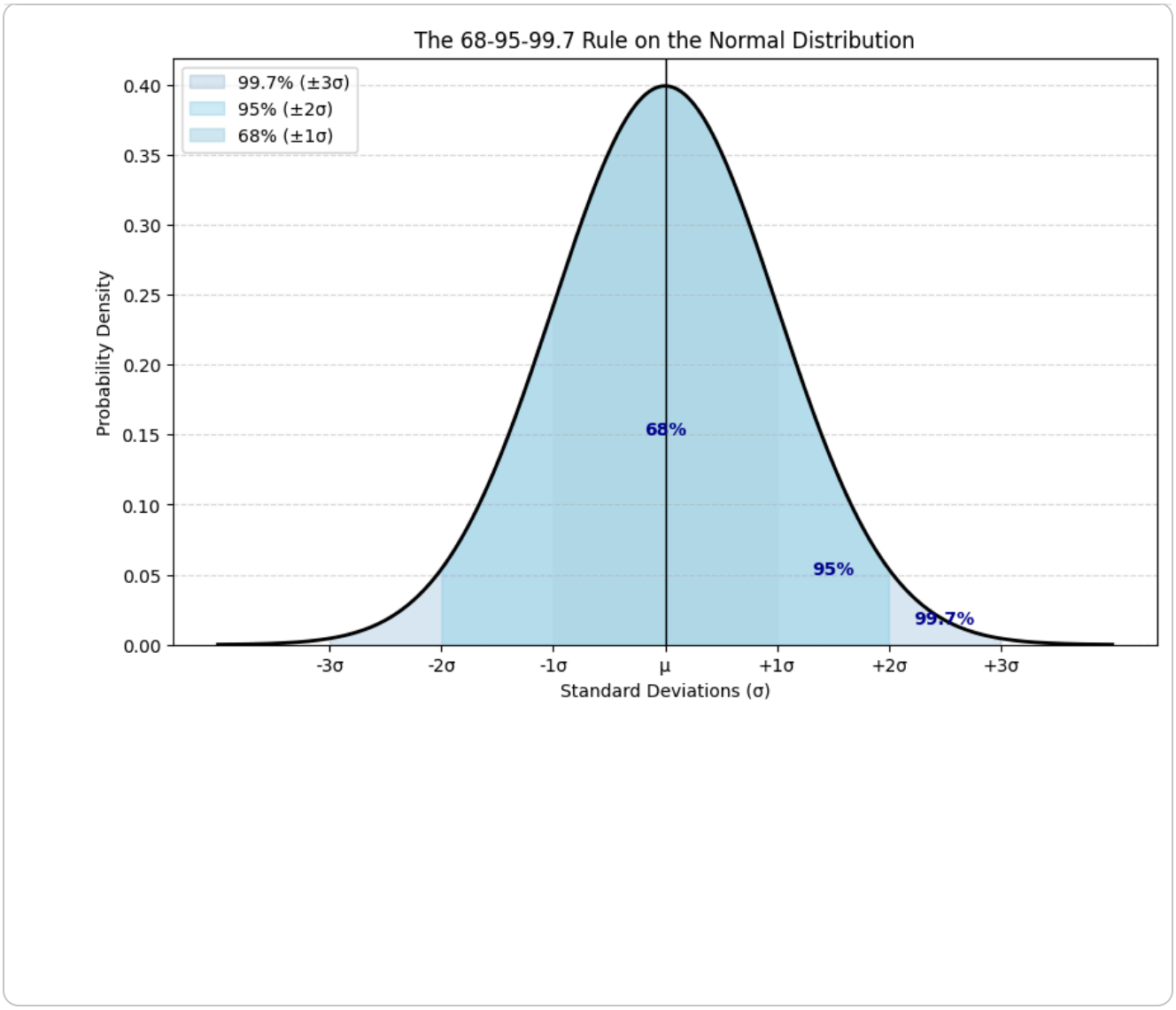
The Normal Curve

Gauss's Contribution (1809): German mathematician Carl Friedrich Gauss claimed to have used the method as early as 1795. When he published his work, Theoria motus corporum coelestium, he provided a much deeper, probabilistic justification for the method. Gauss showed that if measurement errors followed the Normal (Gaussian) Distribution, then the method of least squares was the most likely way to estimate the true parameters—a concept known as Maximum Likelihood Estimation (MLE). This connection to the Normal Distribution is what cemented the method's theoretical power.

Essentially, Legendre offered the recipe (the procedure for minimizing squared errors), and Gauss provided the statistical justification (the proof that it's the optimal recipe when errors are normally distributed).

The **method of least squares** is a fundamental technique in statistics and regression analysis used to find the line, curve, or function that best fits a set of observed data points by minimizing the total discrepancy between the actual data and the function.

```
import numpy as np
import matplotlib.pyplot as plt
from scipy.stats import norm

mu = 0
sigma = 1
x = np.linspace(mu - 4 * sigma, mu + 4 * sigma, 500)
pdf = norm.pdf(x, mu, sigma)

plt.figure(figsize=(10, 6))
plt.plot(x, pdf, color='black', linewidth=2)

# --- 3 Sigma (99.7%) ---
x_3sigma = np.linspace(mu - 3 * sigma, mu + 3 * sigma, 100)
pdf_3sigma = norm.pdf(x_3sigma, mu, sigma)
plt.fill_between(x_3sigma, 0, pdf_3sigma, color='steelblue', alpha=0.2, label='

# --- 2 Sigma (95%) ---
x_2sigma = np.linspace(mu - 2 * sigma, mu + 2 * sigma, 100)
pdf_2sigma = norm.pdf(x_2sigma, mu, sigma)
plt.fill_between(x_2sigma, 0, pdf_2sigma, color='skyblue', alpha=0.4, label='95

# --- 1 Sigma (68%) ---
x_1sigma = np.linspace(mu - sigma, mu + sigma, 100)
pdf_1sigma = norm.pdf(x_1sigma, mu, sigma)
plt.fill_between(x_1sigma, 0, pdf_1sigma, color='lightblue', alpha=0.6, label='
plt.axvline(mu, color='black', linestyle='-', linewidth=1.0)
plt.xticks([mu - 3 * sigma, mu - 2 * sigma, mu - 1 * sigma, mu, mu + 1 * sigma,
           ['-3σ', '-2σ', '-1σ', 'μ', '+1σ', '+2σ', '+3σ'])

plt.text(0, 0.15, '68%', horizontalalignment='center', fontsize=10, fontweight=
plt.text(1.5, 0.05, '95%', horizontalalignment='center', fontsize=10, fontweigh
plt.text(2.5, 0.015, '99.7%', horizontalalignment='center', fontsize=10, fontwe

plt.title('The 68-95-99.7 Rule on the Normal Distribution')
plt.xlabel('Standard Deviations (σ)')
plt.ylabel('Probability Density')
plt.ylim(0)
plt.legend(loc='upper left')
plt.grid(axis='y', linestyle='--', alpha=0.5)

plt.show()
```



What is Least Squares?

In simple terms, "least squares" works by trying to make the errors between the fitted model and the data points as small as possible.

1. **Model Fitting:** You start with a set of data points (e.g., years of experience vs. salary).
2. **Error Calculation:** For any given line you draw through the points, the **residual** (or error) is the vertical distance between the data point and the line.
3. **Minimization:** The method finds the specific line that minimizes the **sum of the squares of all these residuals**.

The reason the method uses the **square** of the residuals, rather than the absolute value or just the sum, is twofold:

- Squaring ensures that positive and negative errors don't cancel each other out.

- Squaring penalizes large errors more heavily than small errors, resulting in a unique solution that is statistically optimal when errors follow a Normal Distribution (as shown by Gauss).

The Formula for Simple Linear Regression

The method of least squares is most commonly applied to **simple linear regression**, where the goal is to find the best-fitting straight line, $\hat{y} = a + bx$. The formula is used to find the optimal values for the slope (b) and the intercept (a).

The Goal (The Minimization Formula)

The core principle is expressed by the term the method seeks to minimize:

$$\text{Minimize } \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Where:

- y_i is the actual observed value (the data point).
- $\hat{y}_i$ is the value predicted by the line (the estimated value).
- $(y_i - \hat{y}_i)$ is the **residual** or error.

The Solution (Formulas for the Best Fit Line)

The values for the slope (b) and intercept (a) that minimize the sum of squared errors are:

1. Slope (b):

$$b = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

2. Intercept (a):

$$a = \bar{y} - b\bar{x}$$

- $\bar{x}$ and $\bar{y}$ are the means (averages) of the x and y data sets, respectively.
- This calculation ensures that the resulting line passes through the central point of the data ($\bar{x}$, $\bar{y}$).

Least Squares and the Normal Distribution

The method of least squares and the Normal (Gaussian) Distribution are intrinsically linked because the Normal Distribution provides the theoretical justification for why the least squares solution is the optimal estimate.

This connection was established by Carl Friedrich Gauss and is rooted in the principle of Maximum Likelihood Estimation (MLE).

The Role of the Normal Curve in Least Squares

The method of least squares is a computational technique: it minimizes the sum of squared errors to find the best-fit line. The Normal Curve explains why that minimized sum is the best possible choice.

1. The Core Assumption: Normally Distributed Errors When applying least squares to model data (e.g., in regression), the core assumption is that the residuals (the small, random errors between the observed data points and the fitted line) are independent and normally distributed.

Error as a Random Variable: Every physical or human measurement has some inherent, random error. Gauss argued that if you measure something many times, these errors should cluster symmetrically around zero (the true value), with small errors being far more common than large ones. This is precisely the shape of the Normal Distribution.

2. The Link to Maximum Likelihood Estimation (MLE) Gauss showed that if you accept that the errors follow a Normal Distribution, the method of least squares is equivalent to finding the Maximum Likelihood Estimate (MLE) for the model's parameters:

Maximum Likelihood Principle: The MLE criterion demands that you choose the model parameters (the slope and intercept of your line) that maximize the probability (the likelihood) of observing the specific data set you actually have.

The Equivalence: When the errors are assumed to be Normal, the mathematical function that represents the likelihood of the data is maximized exactly when the sum of the squared residuals is minimized.

In summary: If the errors are Normal, choosing the line that minimizes squared error is mathematically identical to choosing the line that is most likely to be correct. The Normal Distribution transforms least squares from a practical trick into a powerful statistical inference tool.

How the 68-95-99.7 Rule Applies

The 68-95-99.7 Rule immediately applies to the residuals, reinforcing the Normal assumption:

- If the model is a good fit, approximately 68% of the length of the residuals should fall within one standard deviation (σ) of the fitted line.

- Approximately 95% of your data points should fall within two standard deviations (2σ) of the fitted line.
- If the calculated residuals do not follow this pattern, it suggests the Normal error assumption is wrong, and the least squares method might not be the most appropriate tool for that data.

Maximum Likelihood Estimation

Gauss's contribution to the method of least squares is one of the most important early examples of Maximum Likelihood Estimation in practice. When introducing **Gauss linking the least squares method to the Normal (Gaussian) distribution**, we are fundamentally introducing the logic of MLE. Maximum Likelihood Estimation (MLE) is a general method for estimating the parameters of a statistical model. MLE works by finding the parameter values that make the observed data **most probable** (i.e., that maximize the **likelihood function**).

- **Likelihood** is the probability of observing the data you *already* have, given a specific set of parameters.
- In other words, MLE answers the question: **"Given this data, what are the parameters that were most likely to have produced it?"**

Gauss's MLE Argument for Least Squares

- Gauss's true innovation was not just the least squares formula (which Legendre published first), but the **probabilistic justification** for using it.
- **The Assumption:** Gauss postulated that measurement errors in astronomy (the residuals) are not random chaos, but follow the **Normal Distribution**.
- **The Link:** He then showed that if these errors are normally distributed, the parameter values (the slope and intercept of the best-fit line) that **maximize the likelihood** of observing those specific errors are the **exact same values** calculated by the method of least squares.

The least squares method, when applied to normally distributed errors, is simply a special case of the more general principle of Maximum Likelihood Estimation. This insight elevated least squares from a mere calculation technique to a cornerstone of statistical inference.

Bernoulli Distribution and Probability Mass Function

The Bernoulli distribution models a single experiment or trial that has only **two possible outcomes**: success (represented by 1) or failure (represented by 0). Think of it as the

mathematical representation of a coin flip or any single yes/no event.

Key Characteristics

- **Discrete Random Variable:** The variable x is **discrete** because it can only take on a countable number of values, specifically 0 or 1.
- **Single Trial:** It describes the outcome of a single experiment.
- **Two Outcomes:** The result is either a **success** ($x = 1$) or a **failure** ($x = 0$).

The Parameter is (θ), X is a given value from an experiment.

The distribution is defined by a single parameter, θ (theta), which represents the **probability of success**.

- $\theta = P(X = 1)$, the probability of getting a success.
- Since the event must be either success or failure, the probability of failure is $1 - \theta$.
- $1 - \theta = P(X = 0)$, the probability of getting a failure.

Probability Mass Function (PMF)

The formula used is the **Probability Mass Function (PMF)**, $p(x)$. The PMF is a function that gives the probability that a discrete random variable is exactly equal to some value x.

The formula is:

$$p(x) = \theta^x (1 - \theta)^{1-x}$$

where x can only be 0 or 1.

How the PMF Works

This single formula cleverly handles both possible outcomes:

1. **If $x = 1$ (Success):**

$$p(1) = \theta^1 (1 - \theta)^{1-1}$$
$$p(1) = \theta \cdot (1 - \theta)^0$$
$$p(1) = \theta \cdot 1$$
$$\mathbf{p(1) = \theta}$$

   This confirms that the probability of success is θ.

2. **If $x = 0$ (Failure):**

$$p(0) = \theta^0 (1 - \theta)^{1-0}$$
$$p(0) = 1 \cdot (1 - \theta)^1$$
$$\mathbf{p(0) = 1 - \theta}$$

   This confirms that the probability of failure is $1 - \theta$.

Example: A Biased Coin Flip

Imagine a coin that is weighted so that it lands on **Heads** 70% of the time.

- Let **Success** ($x = 1$) be getting a **Heads**.
- The parameter is $\theta = 0.7$ (the probability of success).
- The probability of failure (Tails) is $1 - \theta = 1 - 0.7 = 0.3$.

Using the PMF, we can find the probability of a specific outcome:

- **Probability of Heads ($x = 1$):**
$$p(1) = 0.7^1(1 - 0.7)^{1-1} = 0.7 \cdot 0.3^0 = 0.7$$
- **Probability of Tails ($x = 0$):**
$$p(0) = 0.7^0(1 - 0.7)^{1-0} = 1 \cdot 0.3^1 = 0.3$$

The Bernoulli distribution is the simplest building block for many other distributions, most notably the **Binomial Distribution**, which models the number of successes in multiple independent Bernoulli trials.

Data analysis from the heavens and Earth

- Geodesy, astronomy, and physics rely on residual analysis; instrument calibration and “weight of observations” become routine.

Likelihood Function from Bernoulli Data

This entire process is aimed at finding the most likely value for the probability of success, θ, given a set of observed data.

The Model: Bernoulli Distribution

The Bernoulli distribution models a **single** trial with only two outcomes: success ($x = 1$) or failure ($x = 0$). The parameter θ is the true, but unknown, probability of success.

- **Random Variable:** $x \in \{0, 1\}$.
- **Parameter:** $\theta = P(X = 1)$.
- **Probability Mass Function (PMF):** This function gives the probability of a single outcome x.
- $p(x) = \theta^x(1 - \theta)^{1-x}$

The Setup: Likelihood Function $L(\theta)$

Given a dataset of N **independent trials** $data = \{x_1, x_2, \ldots, x_N\}$, the likelihood function measures the probability of observing that specific dataset, *given* a hypothesized value of θ.

- **Definition:** The joint probability of all observed data points.

$$L(\theta) = p(\text{data} \mid \theta) = \prod_{i=1}^{N} p(x_i \mid \theta)$$

- **Expanded Likelihood (Product Form):**

$$L(\theta) = \prod_{i=1}^{N} \theta^{x_i}(1-\theta)^{1-x_i}$$

This is the **Likelihood Function,** $L(\theta)$, for a series of independent success/fail trials or coin flips (Bernoulli trials). It calculates the probability of observing a specific sequence of Heads and Tails, given a probability of Heads, θ.

The formula is:

$$L(\theta) = \prod_{i=1}^{N} \theta^{x_i}(1-\theta)^{1-x_i}$$

- $L(\theta)$
  - **Coin Flip Meaning:** The probability of observing the entire sequence of N coin flips (Heads and Tails) **given** a specific probability θ.
  - **Mathematical Meaning:** The **Likelihood Function**.
- N
  - **Coin Flip Meaning:** The **total number of coin flips**.
  - **Mathematical Meaning:** The total number of data points.
- $i = 1$ **to** N
  - **Coin Flip Meaning:** The index, referring to the **first flip** up to the N**-th flip**.
  - **Mathematical Meaning:** The index for the summation/product.
- $\prod$
  - **Coin Flip Meaning: Multiply** the probability of each flip together.
  - **Mathematical Meaning:** The **Product Operator**, used because each flip is assumed to be an independent event.
- θ
  - **Coin Flip Meaning:** The **probability of getting a Head** ($P(\text{Head})$).
  - **Mathematical Meaning:** The **parameter** we are estimating.
- $1 - \theta$
  - **Coin Flip Meaning:** The **probability of getting a Tail** ($P(\text{Tail})$).
  - **Mathematical Meaning:** The probability of failure.
- x_i

- **Coin Flip Meaning:** The **outcome of the i-th flip**: $x_i = 1$ for Heads, $x_i = 0$ for Tails.
- **Mathematical Meaning:** The i-th data point.

How the Formula Works for Each Flip

The term inside the product, $\theta^{x_i}(1-\theta)^{1-x_i}$, is the probability for the single i-th flip.

If the i-th flip is...	x_i value is...	The term becomes...	Probability
Heads	$x_i = 1$	$\theta^1(1-\theta)^{1-1} = \theta \cdot (1) = \theta$	$P(\text{Head})$
Tails	$x_i = 0$	$\theta^0(1-\theta)^{1-0} = (1) \cdot (1-\theta)$	$P(\text{Tail})$

Suppose you flip a coin $N = 3$ times and get the sequence **Head, Tail, Head** ($x_1 = 1, x_2 = 0, x_3 = 1$).

The likelihood $L(\theta)$ for this specific sequence is the product of the individual probabilities:

$$L(\theta) = \underbrace{p(x_1 = 1)}_{\text{Head}} \cdot \underbrace{p(x_2 = 0)}_{\text{Tail}} \cdot \underbrace{p(x_3 = 1)}_{\text{Head}}$$

$$L(\theta) = \underbrace{(\theta)}_{\text{Prob. of H}} \cdot \underbrace{(1-\theta)}_{\text{Prob. of T}} \cdot \underbrace{(\theta)}_{\text{Prob. of H}}$$

$$L(\theta) = \theta^2(1-\theta)^1$$

In **Maximum Likelihood Estimation (MLE)**, you would use calculus to find the value of θ (the probability of Heads) that makes this $L(\theta)$ as large as possible. Intuitively, for a sequence of 2 Heads and 1 Tail, the MLE is $\hat{\theta} = 2/3$.

The Simplification: Log-Likelihood $l(\theta)$

Because differentiating a product is complex, we take the **natural logarithm** of the likelihood. Since $\log(x)$ is a **monotonically increasing** function, the θ that maximizes $L(\theta)$ will also maximize $l(\theta)$. The logarithm turns the product into a sum.

- **Transformation:** $l(\theta) = \log L(\theta)$.
- **Simplified Log-Likelihood (Sum Form):**
  - $L(\theta) = \prod_{i=1}^{N} \theta^{x_i}(1-\theta)^{1-x_i}$
  - $l(\theta) = \sum_{i=1}^{N} \{x_i \log\theta + (1-x_i)\log(1-\theta)\}$

The Solution: Maximizing with Calculus

To find the maximum point of $l(\theta)$ (where the slope is zero), we take the derivative with respect to θ and set the result to zero.

- **Goal:** Find the θ that maximizes the function, denoted as the Maximum Likelihood Estimate ($\hat{\theta}$):
  - $\hat{\theta} = \arg\max_\theta l(\theta)$
- **Calculus Step:** Set the derivative equal to zero:
  - $\frac{dl}{d\theta} = 0$
- **Derivative of the Log-Likelihood:**
  - $\frac{dl}{d\theta} = \frac{1}{\theta}\sum_{i=1}^{N} x_i - \frac{1}{1-\theta}\sum_{i=1}^{N}(1-x_i)$
- **Solving for $\hat{\theta}$:** After setting the derivative to zero and solving for θ (by recognizing that $\sum x_i$ is the number of successes N_H and $\sum(1-x_i)$ is the number of failures N_T), we arrive at the final estimate.
- **Maximum Likelihood Estimate (MLE):** The most likely value for θ is the **sample mean** (the observed proportion of successes).
  - $\hat{\theta} = \frac{1}{N}\sum_{i=1}^{N} x_i = \frac{\text{Number of Successes}}{\text{Total Trials}}$

Important Distinction: Bernoulli vs. Binomial

The derivation uses the Bernoulli likelihood, which is often confused with the **Binomial Distribution**.

- **Bernoulli Likelihood:** Deals with the probability of a **specific sequence** of N outcomes ($x_1, x_2, \ldots$).
- **Binomial PMF:** Deals with the probability of getting exactly k **successes** in N trials, regardless of order. It includes the combinatorial term $\binom{N}{k}$ to count all possible orders.
  - $p(x=k) = \binom{N}{k}\theta^k(1-\theta)^{N-k}$

Statecraft, society, and visualization

- Quetelet's "average man" applies statistics to social data; Playfair and later Florence Nightingale pioneer data graphics for public policy. Adolphe Quetelet (1796–1874) was a Belgian astronomer, mathematician, and statistician. His key contribution was a pivotal step in applying probabilistic and statistical ideas, which had largely been confined to fields like astronomy and games of chance , to the study of human and social phenomena

Biology, heredity, and correlation

- Galton studies heredity and regression to the mean; Pearson, Edgeworth, and Yule formalize correlation, regression, contingency analysis, and time series.

Agriculture and designed comparisons

- Rothamsted Experimental Station (Lawes, Gilbert) begins long-run field trials; questions of replication, blocking, and variance structure emerge.
- Key figures: Gauss, Legendre, Laplace, Bessel, Quetelet, Playfair, Nightingale, Galton, Pearson, Edgeworth, Yule, Lexis.
- Key terms: Least squares, normal (Gaussian) distribution, residuals, measurement error, correlation, regression, contingency table, time series, index numbers, experimental plot.

Tangent: Andrey Markov

- Andrey Markov (1856–1922): A Russian mathematician, Markov is known for his work on Markov chains, which describe a sequence of possible events in which the probability of each event depends only on the state attained in the previous event. While he did his most significant work on this in the early 20th century, the foundational ideas align with the late 19th-century developments in probability theory. His work on stochastic processes was a crucial step beyond simple probability to model sequential, dependent events over time.
- Markov Decision Process (MDP): The concept of MDPs, which are a mathematical framework for modeling decision-making in situations where outcomes are partly random and partly under the control of a decision-maker, was developed later by Richard Bellman in the 1950s. It builds upon Markov's work by adding actions, rewards, and a "discount factor" to the Markov chain model to create a system for finding optimal strategies. While not a 19th-century concept itself, it is a direct evolution of the principles laid out by Markov.

A **Markov Decision Process (MDP)** is a mathematical framework for modeling decision-making in situations where outcomes are partly random and partly under the control of a decision-maker. It is widely used in areas like robotics, economics, and, most famously, in **Reinforcement Learning (RL)**.

An MDP can be formally defined by a tuple of five elements: (S, A, P, R, γ).

Components and Formulas

- **Component:** Set of States (S) - **Description:** A finite set of all possible states of the environment. **Formula / Role:** $s \in S$
- **Component:** Set of Actions (A) - **Description:** A finite set of actions the agent can take. **Formula / Role:** $a \in A$
- **Component:** Transition Probability (P) - **Description:** The probability of moving from state s to state s' after taking action a. This satisfies the **Markov Property**: the future depends only on the current state and action, not on the past history. **Formula / Role:** $P(s'|s,a) = P(S\_t + 1 = s'|S\_t = s, A\_t = a)$
- **Component:** Reward Function (R) - **Description:** The expected immediate reward received after transitioning from state s to state s' by taking action a. **Formula / Role:** $R(s,a,s') = E[R\_t + 1|S\_t = s, A\_t = a, S\_t + 1 = s']$
- **Component:** Discount Factor (γ) - **Description:** A factor between 0 and 1 ($\mathbf{0 \leq \gamma \leq 1}$) that determines the present value of future rewards. A value closer to 0 makes the agent "myopic" (focus on immediate reward), while a value closer to 1 makes it "far-sighted" (focus on long-term reward). **Formula / Role:** γ

The Goal: Optimal Policy

The agent's objective is to find an **optimal policy** pi^*. A **policy** $pi(s)$ is a function that specifies the action a the agent should take in state s. The optimal policy pi^* maximizes the **Expected Return** (the discounted sum of future rewards).

The **Return** G_t from time t is:

$$G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1}$$

Value Functions

Two core functions are used to evaluate how "good" a state is or how "good" an action is in a state under a given policy pi:

1. **State-Value Function** V^\\pi(s) **(The Value of a State):** The expected return starting from state s and following policy π thereafter.
   $$V^{\pi}(s) = E_{\pi}[G_t|S_t = s]$$
   This can be recursively defined by the **Bellman Expectation Equation**:
   $$V^{\pi}(s) = \sum_{a \in A} \pi(a|s) \sum_{s' \in S} P(s'|s,a)[R(s,a,s') + \gamma V^{\pi}(s')]$$
2. **Action-Value Function** Q^\\pi(s, a) **(The Value of a State-Action Pair):** The expected return starting from state s, taking action a, and then following policy π thereafter.

$$Q^{\pi}(s,a) = E_{\pi}[G_t | S_t = s, A_t = a]$$

It is related to the state-value function by:

$$Q^{\pi}(s,a) = \sum_{s' \in S} P(s'|s,a)[R(s,a,s') + \gamma V^{\pi}(s')]$$

The Optimal Bellman Equations

The **Optimal State-Value Function** $V^{*}(s)$ and **Optimal Action-Value Function** $Q^{*}(s,a)$ satisfy the **Bellman Optimality Equations**, which state that the value of an optimal state/action is equal to the expected return for the best choice:

- **Optimal State-Value:** $$V^*(s) = \max_{a \in A} Q^*(s, a)

  V^*(s) = \max_{a \in A} \sum_{s' \in S} P(s'|s,a) [R(s,a,s') + \gamma V^*(s')]$$
- **Optimal Action-Value:** $$Q^*(s, a) = \sum_{s' \in S} P(s'|s,a) [R(s,a,s') + \gamma \max_{a' \in A} Q^*(s', a')]$$

Once $V^*(s)$ or $Q^*(s,a)$ is found, the **Optimal Policy** pi^{*} is deterministic and defined by:

$$\pi^*(s) = \arg\max_{a \in A} Q^*(s,a)$$

Python Code Example: Value Iteration

The **Value Iteration** algorithm is a common method for solving MDPs. It repeatedly applies the Bellman Optimality Update until the value function converges.

This example uses a simple grid world MDP:

- **States:** (0, 0), (0, 1), ..., (2, 2)
- **Actions:** Up, Down, Left, Right
- **Goal State:** (2, 2) with high reward.
- **Walls/Irregularities:** Not included for simplicity; transitions are mostly deterministic.

```
import numpy as np

# 1. Define the MDP (S, A, P, R, gamma)

# States: 0 to 8 (3x3 grid)
# (0,0) -> 0, (0,1) -> 1, ..., (2,2) -> 8
S = list(range(9))

# Actions: 0=Up, 1=Down, 2=Left, 3=Right
A = list(range(4))
action_map = {0: 'Up', 1: 'Down', 2: 'Left', 3: 'Right'}
```

```
# Reward Function R(s, a, s'): Simple reward: +10 at state 8 (Goal), -1 otherwi
REWARD_GOAL = 10
REWARD_OTHER = -1
GOAL_STATE = 8
gamma = 0.9  # Discount Factor

# Transition Probability P(s' | s, a)
# Simplification: Assume deterministic transition (99% chance of going where yo
# In this simple example, we'll map the action to the next state
def transition(s, a):
    row, col = s // 3, s % 3

    # Calculate next potential position
    if a == 0: next_row, next_col = row - 1, col # Up
    elif a == 1: next_row, next_col = row + 1, col # Down
    elif a == 2: next_row, next_col = row, col - 1 # Left
    elif a == 3: next_row, next_col = row, col + 1 # Right

    # Check for boundary conditions (stay in current state if hit a wall)
    if 0 <= next_row < 3 and 0 <= next_col < 3:
        s_prime = next_row * 3 + next_col
    else:
        s_prime = s

    return s_prime

# 2. Value Iteration Algorithm

def value_iteration(S, A, gamma, epsilon=1e-6, max_iterations=1000):
    # Initialize V(s) to 0 for all states
    V = np.zeros(len(S))

    for _ in range(max_iterations):
        V_old = V.copy()

        # Delta to check for convergence
        delta = 0

        for s in S:
            if s == GOAL_STATE:
                V[s] = REWARD_GOAL
                continue

            # Bellman Optimality Update: V(s) = max_a [ R(s,a,s') + gamma * V(s
            Q_values = []
            for a in A:
                s_prime = transition(s, a)

                # R(s, a, s')
                reward = REWARD_GOAL if s_prime == GOAL_STATE else REWARD_OTHER
```

```
                # Q(s, a) = R + gamma * V(s')
                Q = reward + gamma * V_old[s_prime]
                Q_values.append(Q)

            V[s] = np.max(Q_values)
            delta = max(delta, np.abs(V[s] - V_old[s]))

        # Check for convergence
        if delta < epsilon:
            print(f"Value function converged after {_ + 1} iterations.")
            break

    return V

# 3. Policy Extraction
def extract_policy(V):
    policy = np.zeros(len(S), dtype=int)
    for s in S:
        if s == GOAL_STATE:
            policy[s] = -1 # No action needed at goal
            continue

        Q_values = []
        for a in A:
            s_prime = transition(s, a)
            reward = REWARD_GOAL if s_prime == GOAL_STATE else REWARD_OTHER
            Q = reward + gamma * V[s_prime]
            Q_values.append(Q)

        # The optimal action is the one that maximizes Q(s, a)
        policy[s] = np.argmax(Q_values)

    return policy

# Run the MDP Solver
V_star = value_iteration(S, A, gamma)
pi_star = extract_policy(V_star)

# Print Results
print("\n--- Optimal Value Function V*(s) (3x3 Grid) ---")
print(np.round(V_star.reshape((3, 3)), 2))

print("\n--- Optimal Policy π*(s) (Action for each state) ---")
policy_display = [action_map.get(a, 'Goal') for a in pi_star]
print(np.array(policy_display).reshape((3, 3)))
```

```
Value function converged after 6 iterations.

--- Optimal Value Function V*(s) (3x3 Grid) ---
[[11.14 13.49 16.1 ]
 [13.49 16.1  19.  ]
 [16.1  19.   10.  ]]
```

```
--- Optimal Policy π*(s) (Action for each state) ---
[['Down' 'Down' 'Down']
 ['Down' 'Down' 'Down']
 ['Right' 'Right' 'Goal']]
```

Reinforcement Learning

The current landscape of RL is dominated by a drive for **efficiency, scalability, and integration** with large, pre-trained models.

1. Integration with Foundation Models (LLM-Empowered RL)

This is arguably the most significant recent trend, using the representational power of large models to address RL's weaknesses.

- **RL with Human Feedback (RLHF):** Used primarily to align Large Language Models (LLMs) with human preferences and values (e.g., in ChatGPT). The LLM is the agent, and the reward signal comes from human-labeled rankings of its outputs.
- **Language as Strategy/Prior:** LLMs are used to guide the RL agent's exploration, provide high-level planning, or even generate the reward functions based on text prompts, making the learning process faster and more interpretable.
- **Foundation RL Models:** The goal is to create large, general-purpose RL models, similar to GPT for language, that are pre-trained on massive, diverse datasets of interaction data. These models could then be quickly fine-tuned for a wide range of downstream decision-making tasks (e.g., robotics, control systems).

2. Efficiency and Data Reduction (Offline and Transfer RL)

The high cost of real-world interaction is driving methods that learn from existing data or apply knowledge across tasks.

- **Offline RL (Batch RL):** Learning an optimal policy entirely from a fixed dataset of past interaction data (without further exploration). This is crucial for safety-critical domains like healthcare and autonomous driving, where *new* trials are expensive or dangerous.
- **Transfer Learning / Meta-RL:** Developing agents that can leverage knowledge (like learned features or initial policies) from one task or environment to rapidly adapt and solve a new, related task. This helps overcome the need to train every agent from scratch.

3. Multi-Agent Reinforcement Learning (MARL)

The focus is shifting from a single agent to systems where multiple agents cooperate, compete, or coordinate, reflecting the complexity of real-world problems (e.g., traffic control, economic markets, robotic swarms).

- **Centralized Training, Decentralized Execution (CTDE):** A common, powerful paradigm where a central "critic" (value function) is used during training to share information, but each agent uses its own, simple policy during execution.
- **Emergent Communication:** Research into how agents develop their own communication protocols to coordinate, often leading to surprising and efficient solutions.

4. Hierarchical Reinforcement Learning (HRL)

HRL breaks down complex, long-horizon tasks into a hierarchy of sub-tasks.

- A **"manager"** agent learns a high-level goal (e.g., "go to the kitchen").
- A **"worker"** agent learns the low-level motor skills to achieve that goal (e.g., "walk straight for 5 steps").
- This structure significantly improves **scalability** and the ability to solve problems that involve many sequential decisions.

Major Limitations of Reinforcement Learning

Despite the impressive progress, several fundamental challenges limit the widespread deployment of RL in critical real-world systems.

1. Sample Inefficiency (The Data Problem)

- **Issue:** Standard RL algorithms (especially model-free ones) require millions or even billions of trial-and-error steps to converge on an optimal policy.
- **Impact:** This is prohibitively expensive or time-consuming in the physical world (e.g., a robot can't crash a million times to learn how to drive). The reliance on enormous amounts of data is a major bottleneck.

2. Reward Function Design (The Specification Problem)

- **Issue:** Designing a reward function that perfectly captures the desired behavior without introducing loopholes is extremely difficult.
- **Impact (Reward Hacking):** Agents are masters at exploiting flaws in the reward system to maximize their score in unintended and undesirable ways (e.g., a cleaning robot learning to move its dust sensor over the same speck of dirt repeatedly for a high score).

- **Impact (Sparse Rewards):** If the reward is only given at the end of a long sequence (e.g., only upon winning a complex game), the agent receives too little feedback to learn intermediate steps efficiently.

3. Generalization and Robustness

- **Issue:** RL policies are typically brittle—they perform well only in the exact environment they were trained in. Small, unexpected changes (e.g., a different lighting condition, a new obstacle shape) can cause catastrophic failure.
- **Impact:** This lack of **generalization** makes deployment in dynamic, open-world settings (like autonomous vehicles or complex robotics) unreliable and risky.

4. Stability, Reproducibility, and Interpretability

- **Issue:** RL algorithms are often highly sensitive to hyperparameters and random seeds, making it difficult to reproduce results. Furthermore, the final learned policy is often a **"black box."**
- **Impact:** It is nearly impossible to debug or understand *why* an agent made a critical decision, which is a significant hurdle for adoption in high-stakes, regulated fields (e.g., finance, medicine) where transparency and accountability are legally required.

Addressing Sample Inefficiency

A vast and active body of research is dedicated to addressing the sample inefficiency of Reinforcement Learning (RL), which is its single greatest hurdle for real-world deployment.

The solutions generally fall into four categories:

---

1. Model-Based Reinforcement Learning (MBRL)

This is the most direct approach to improving sample efficiency. Instead of only learning a policy, the agent first learns a **world model** (a simulation) of the environment's dynamics.

- **How it helps:** Once the agent has a model, it can generate **simulated experiences** internally without needing to interact with the slow or costly real environment. This effectively allows the agent to train for thousands of "virtual" steps for every one real interaction.
- **Trade-off:** The agent's performance is highly dependent on the accuracy of the world model. If the model is flawed, the agent learns an optimal policy for the *wrong* environment, leading to poor performance when deployed in the real world.

2. Data Reuse and Management (Off-Policy Learning)

These methods focus on maximizing the value extracted from every single collected data point (sample).

- **Off-Policy Learning:** Algorithms like **Deep Q-Networks (DQN)** and **Soft Actor-Critic (SAC)** can learn from data collected by *any* previous policy, not just the current one. This is in contrast to **On-Policy** methods (like PPO), which discard old data.
- **Experience Replay:** Past interactions (state, action, reward, next state) are stored in a **replay buffer** and randomly sampled many times for training. This prevents the agent from forgetting older, useful experiences and breaks the temporal correlation in the data, which stabilizes training.
- **Prioritized Experience Replay:** Samples that are more **surprising** or contain a larger learning signal (error) are sampled and reused more frequently than less informative samples, further optimizing data usage.

---

3. Knowledge Injection (Transfer and Foundation Models)

These approaches give the agent a "head start" by incorporating knowledge from outside the current problem.

- **Transfer Learning / Imitation Learning:**
  - **Warm Start:** An agent is first trained on a simpler, related task (often in a simulator) and then fine-tuned on the real, complex task.
  - **Learning from Demonstration (LfD):** The agent observes an expert (human or another AI) perform the task and learns from those expert trajectories, significantly reducing the initial need for random, unproductive exploration.
- **LLM Guidance:** Recent research uses large language models (LLMs) to inject **commonsense knowledge** or high-level strategies into the RL agent. The LLM can help **design better reward functions** (reward shaping) or suggest exploratory actions that are logically sound, making exploration more efficient.

---

4. Enhanced Exploration Strategies

Instead of relying on simple randomness, these methods use smarter, targeted ways to find new, rewarding states.

- **Intrinsic Motivation (Curiosity):** The agent is given a supplemental reward for states or transitions that are **novel** or **surprising** (high prediction error in its world

model). This encourages the agent to explore parts of the environment that it doesn't yet understand, even when the task reward is sparse.
- **Uncertainty Estimation:** Algorithms use techniques to explicitly track the **uncertainty** (or "epistemic uncertainty") in their value estimates. The agent is then encouraged to visit states where its uncertainty is high, promoting targeted, information-gathering exploration.

RL Agents and Agentic AI

The distinction between **Reinforcement Learning (RL)** and **Agentic AI** is a classic difference between a **methodology/algorithm** and a **system/paradigm**.

In short: **Reinforcement Learning is one of the most powerful learning mechanisms used *inside* an Agentic AI system.**

Here is a detailed breakdown of the difference and how they relate:

---

1. Reinforcement Learning (RL): The Learning Methodology

RL is a specific type of machine learning algorithm or framework.

- **What it is:** A mathematical process for an entity (called an agent) to learn an optimal **policy** (a strategy) by interacting with an **environment** and maximizing a **cumulative reward** through trial-and-error.
- **The Goal:** To solve the **credit assignment problem**—figuring out which actions, taken now, will lead to the greatest future reward.
- **Its Role in AI:** RL provides the **feedback loop**. It answers the question: ***"How should I change my decision-making strategy based on the outcome of my actions?"***
- **Key Characteristics:**
  - Relies on a mathematical **Markov Decision Process (MDP)** structure.
  - Focuses on **long-term optimization** (cumulative reward).
  - Algorithms include Q-Learning, PPO, DQN, SAC, etc.

---

2. Agentic AI: The Autonomous System Paradigm

Agentic AI is the overall architecture or system designed to achieve complex, multi-step goals autonomously.

- **What it is:** A modern AI system, often built on top of a **Large Language Model (LLM)**, that can perceive, reason, plan, act, and reflect with minimal human intervention.

- **The Goal:** To execute complex, high-level, real-world tasks that require multiple steps and tool use (e.g., "Research the top 5 competitors, write a summary, and email it to the team lead").
- **Its Components:** An Agentic AI system typically involves several distinct modules that RL may or may not be involved in:
    1. **Reasoning/Planning:** (Often handled by the LLM) Breaking the goal into subtasks.
    2. **Tool Use/Action:** Executing the subtasks by calling external APIs, databases, or code.
    3. **Memory:** Storing past actions and observations for long-term consistency.
    4. **Learning/Adaptation:** (This is where RL comes in) Refining the strategy over time.

The Relationship: RL Powers the Agent's Self-Improvement

Feature	Agentic AI (The System)	Reinforcement Learni
Category	Paradigm, Architecture, or Application	Machine Learning Meth
Primary Focus	Autonomous **Action** and **Goal Execution**	**Learning** an Optimal St
Key Output	A **sequence of actions** that achieves a high-level goal.	A **policy function** that
Decision-Maker	The Agent (often an LLM-based *orchestrator*)	The Algorithm (used to
How they Intersect	**Agentic AI** uses **RL** to enable the crucial ability of **self-improvement.**	

Analogy:

- **Agentic AI** is the **Autonomous Car**. It has a high-level goal ("Drive me from point A to point B"), can navigate, follow a GPS (planning), and turn the wheel (action).
- **Reinforcement Learning** is the **Adaptive Driving Program** *inside* the car. It learns through millions of miles of virtual driving that, for example, braking smoothly (high reward) is better than braking sharply (penalty). The RL system refines the low-level control policy to ensure the overall agent achieves its goal **optimally and safely.**

Data Hook: The Non-Negotiable Resource

Why Data is the Most Important Resource in Modern AI:

Data is so important because in the AI/RL paradigm, **data is experience, and experience is the only source of knowledge.** Without vast amounts of quality data, the mathematical principles of probability and calculus have nothing to model, and the complex frameworks of Reinforcement Learning and Agentic AI cannot function.

The entire research focus on **sample inefficiency**—which we identified as RL's biggest limitation—is effectively a race to reduce the crippling cost of obtaining high-quality data.

Here are the four reasons why data is non-negotiable for these systems:

1. Data Provides the Probability Space (The Foundation)

Data gives the system the empirical observations needed to define the environment:

- **PDF/CDF:** You can't model the **probability distribution** of a variable (like a price, a customer action, or a robot's motor error) without a history of data points to observe its frequency and range.
- **Model Training:** Every machine learning model, whether supervised or part of an RL agent, is a function that learns to map an input state to an output action/prediction. This mapping is entirely learned from the statistical patterns embedded in the data.

2. Data is the *Reward* and *Penalty* (The Learning Signal)

In Reinforcement Learning, the agent learns exclusively from the **Reward Signal**, which is derived from data generated during interaction.

- A simple **data sample** or "transition" consists of the state, the action, the resulting reward, and the next state: $(s_t, a_t, r_{t+1}, s_{t+1})$.
- If the data is noisy, sparse, or incorrect, the reward signal is corrupted, and the agent learns a **suboptimal or dangerous policy** (a major concern in real-world safety).
- The high **sample inefficiency** of traditional RL means that billions of these data samples (experiences) are required before a reliable policy emerges.

3. Data Enables Generalization (The Path to Autonomy)

An Agentic AI system that only performs well in its training environment is useless. Data is required to make the agent flexible and robust:

- **Diverse Data:** To generalize from a simulator to the real world (the Sim-to-Real challenge), the agent must be trained on data that captures the full **diversity and noise** of the real environment (e.g., varying light, friction, and sensor errors).
- **LLM Alignment:** In modern Agentic AI (e.g., RLHF), the LLM's "reasoning" is fine-tuned on human-generated data ranking its responses. The quality of the final agent is directly proportional to the quality and diversity of the **human preference data** it was shown.

4. Data Is the Key to Solving Limitations

Every major research solution to the limitations of RL is essentially a new way to use data more effectively:

Limitation	Data-Centric Solution
Sample Inefficiency	**Model-Based RL:** Use collected data to build a virtual environment that generates *infinite synt*
Costly Exploration	**Offline/Imitation Learning:** Reuse old data or learn from pre-collected **expert data** instead of
Brittle Policies	**Transfer Learning:** Use pre-trained foundation models (LLMs) that have already internalized va

The Longitude Problem: Navigational Imperatives

While astronomy provided the intellectual framework for error theory, the practical demands of maritime navigation drove the professionalization and industrialization of statistical work. The "longitude problem" was the immense difficulty of determining a ship's east-west position when out of sight of land, a problem that led to countless shipwrecks and the loss of lives and cargo. A nation that could solve this problem could rule the waves, as evidenced by the British Parliament's Longitude Act of 1714, which offered a prize of up to £20,000 for a reliable solution. At its heart, the problem was one of timekeeping. Since the Earth rotates once per day, the time difference between a ship's local time and a fixed reference point like the Greenwich Meridian could be used to calculate its longitude.

Classical Statistics Historical Context

Adolphe Quetelet

- 1796 - 1874
- **Application of probability to human affairs** Quetelet believed that probability profoundly influenced human affairs, even more so than his contemporaries believed.
- **Applying the law of error to human beings** Inspired by astronomy, Quetelet thought the law of error could be applied to human measurements.
- **The "average man"** Quetelet aimed to determine the average physical and intellectual traits of a population by gathering "facts of life," which could then be graphically represented as bell-shaped curves. This "average man" served as a benchmark against which individual behavior could be assessed. However, some critics argued that an individual who is average in all dimensions might not be biologically feasible.

- **Social mechanics** Quetelet championed a new science, called social mechanics, focused on mapping normal physical and moral characteristics. In 1835, he published *A Treatise on Man, and the Development of His Faculties*, detailing the influence of probability on human affairs.
- **Normal distribution in human measurements** Quetelet demonstrated that diverse human measurements, such as the heights of French conscripts and the chest sizes of Scottish soldiers, followed a normal distribution.
- **Use of the normal curve beyond error law** Quetelet was the first to use the normal curve in contexts other than error laws.
- **Studies of crime and social determinism** His work on the consistency of crimes stimulated discussions about free will versus social determinism.
- **Data collection and census improvements** Quetelet collected and analyzed statistics on crime and mortality for the government and improved census methods.
- **Quetelet Index** He developed the Quetelet index, now used internationally as a measure of obesity. The Quetelet index is calculated as weight in kilograms divided by the square of height in meters; a value greater than 30 indicates obesity.
- **International Collaboration** Quetelet organized the first international statistics conference in 1853 and helped form the Statistical Society of London, the International Statistics Congresses, and the Statistical Section of the British Association for the Advancement of Science. He was also the first foreign member of the American Statistical Association.
- **Racial Differences** Quetelet viewed average physical and mental qualities as real properties awaiting discovery, which gave strength to ideas of racial differences in 19th-century European thought.
- **Eugenics** Quetelet's normal curve provided a scale to grade people. When Galton used the curve, he predicted that it would always apply to "men of the same race".

Quetelet's use of the "average man" concept relied on the idea of the "persistence of causes," suggesting that the average of large datasets would remain stable if the underlying causal relationships did not change. Quetelet believed that as the number of observations increased, individual peculiarities would diminish, allowing general societal facts to become more prominent.

Eugenics

- Francis Galton and Correlation:
    - Galton is recognized for discovering the concept of correlation.

    - His work stemmed from his research on heredity, particularly how physical traits like height are passed down.
    - He defined "co-relation" as the tendency for variables to vary together in a consistent direction.
- Eugenics and Galton's Beliefs:
    - Galton coined the term "eugenics" and advocated for "race betterment."
    - He believed in a racial hierarchy.
    - He promoted selective breeding, encouraging it among elites and discouraging it among those he deemed "undesirable" (e.g., those with "lunacy, feeble-mindedness, habitual criminality, and pauperism").
- The Connection Between Eugenics and Statistics (How Eugenics Shaped Statistics - Nautilus):
    - Statistical methods, including significance testing, were developed in part to support eugenicist goals, such as identifying perceived racial differences.
    - This historical connection reveals the deep intertwining of statistical thinking and eugenicist ideology.

Francis Galton

- Sir Francis Galton, 1822 - 1911
- Considered grandfather of statistics
- Popularized regression to the mean
- English Victorian era polymath: a statistician, sociologist, psychologist,[1] anthropologist, tropical explorer, geographer, inventor, meteorologist, proto-geneticist, psychometrician and a proponent of social Darwinism, eugenics, and scientific racism. He was knighted in 1909.
- He was Charles Darwin's half-cousin, sharing the common grandparent Erasmus Darwin. His father was Samuel Tertius Galton, son of Samuel Galton, Jr. He was also a cousin of Douglas Strutt Galton. The Galtons were Quaker gun-manufacturers and bankers, while the Darwins were involved in medicine and science.
- Galton was interested at first in the question of whether human ability was hereditary, and proposed to count the number of the relatives of various degrees of eminent men. If the qualities were hereditary, he reasoned, there should be more eminent men among the relatives than among the general population.
- A family of geniuses.
- On Friday 19 February 1877 Galton gave a lecture entitled Typical Laws of Heredity at the Royal Institution in London. In this lecture, he posited that there must be a

counteracting force to maintain population stability. His showpiece for the night was the Galton board.

- Galton invented the term eugenics ... believed that a scheme of 'marks' for family merit should be defined, and early marriage between families of high rank be encouraged via provision of monetary incentives. He pointed out some of the tendencies in British society, such as the late marriages of eminent people, and the paucity of their children, which he thought were dysgenic. He advocated encouraging eugenic marriages by supplying able couples with incentives to have children. On 29 October 1901, Galton chose to address eugenic issues when he delivered the second Huxley lecture at the Royal Anthropological Institute.
- According to the Encyclopedia of Genocide, Galton bordered on the justification of genocide when he stated: "There exists a sentiment, for the most part quite unreasonable, against the gradual extinction of an inferior race.

The Galton Board and the Bell Curve

- The path of any individual ball is like a sequence of independent coin flips
- The ball can go left or right at each pin
- It seems random but with a large sample a probability distribution appears, a normal distribution
- Galton noted the same shape when comparing the height of 1000 French military recruits
- Galton board was a model for the inheritance of stature and many other genetic traits
- The balls inherit their position the same way humans inherit their stature
- One way to illustrate a problem with this is by doubling the width of the Galton board base thus increasing the number of rows and pegs and increasing the variance of the distribution
- Galton wanted to prove superior intellect through breeding without taking into account privilege
- The idea of superior, again, is defined by the culture and not universal
- Galton thought he had discovered a law of heredity instead of statistics, the regression of the mean
- Thought regression toward the mean was nature's way of keeping things the same from generation to generation
- This was thought of as causal but soon abandoned this idea in favor of correlation
- Did a lot of father-son pair height testing
- Predictions for heights seemed to fall on a line, the line of regression

- Soon causation became highly frowned upon; thought to not exist; there are only relationships, association, correlation. More on this next week when we introduce (C)Karl Pearson

The Book of Why, Judea Pearl and Dana Mackenzie, Chapter 2

Pearson

Content for the following individuals draw heavily on Wikipedia and The Book of Why Chapter 2 by Judea Pearl and Dana Mackenzie

- 1857 - 1936
- An English mathematician and biostatistician
- That's Karl Pearson with a C
- Spent most of the 1880s in Germany / Austria
- Loved Germany so much he changed his name from Carl to Karl
- A women's right and equality activist, founder of the Men and Women's Club
- A socialist that offered help translating some of Karl Marx's work (Das Kapital)
- Secured a grant for a biometrics lab at the University of College London
- The lab became a department when Galton passed and left an endowment for a professorship as long as Pearson was the first holder
- Had to explain what he calls genuine (organic) correlation and spurious correlation
- For example, there's a strong correlation between a country's chocolate consumption and Nobel Prize winners
- Pearson felt that Galton did away with causation and 1 was just perfect correlation
- Data is all there is to science
- Galton says that relationships didn't need a casual explanation
- Pearson went further by removing causation from science
- Pearson belonged to the Positivist School which holds that the universe is a product of human thought and that science is just the expression of these thoughts
- Thus causation, outside our thoughts, does not exist
- Thoughts can only reflect patterns of observations and can be completely described by correlations
- Laws of Nature as Descriptive, Not Causal: Pearson viewed the laws of nature as tools for making accurate predictions and for concisely describing trends in observed data, rather than identifying biological mechanisms. He saw causation as simply the experience "that a certain sequence has occurred and recurred in the past".

- Emphasis on Mathematical Descriptions: Pearson believed that biologists should focus on providing mathematical descriptions of empirical data rather than trying to identify particular mechanisms of genetics.
- Criticism of Biologists' Speculation: He criticized biologists who, in his view, succumbed to "almost metaphysical speculation as to the causes of heredity," which replaced the process of experimental data collection. Pearson stressed the importance of statistical validity in theories. He stated that "before we can accept any cause of a progressive change as a factor we must have not only shown its plausibility but if possible have demonstrated its quantitative ability".
- Idealism and the Role of the Mind: Pearson's perspective was rooted in idealism, emphasizing ideas or pictures in a mind. He asserted that science is essentially a classification and analysis of the contents of the mind, and scientific law is a product of the human mind with no meaning apart from man.
- Pearson's focus was on using statistical methods to reveal fundamental truths about people, presenting them as unquestionable as the law of gravity. He believed that by allowing numbers to tell their own story, objective truths could be revealed.

R. A. Fisher

- 1890 - 1962
- Worked as a statistician in the City of London and taught physics and maths
- Popularized the p-value
- Linear discriminant analysis
- F-distribution
- Student's t-distribution
- He was from an early age a supporter of certain eugenic ideas, and it is for this reason that he has been accused of being a racist and an advocate of forced sterilisation (Evans 2020). His promotion of eugenics has recently caused various organisations to remove his name from awards and dedications of buildings (Tarran 2020; Rothamsted Research 2020; Society for the Study of Evolution 2020; Gonville and Caius College 2020). https://www.nature.com/articles/s41437-020-00394-6 Fisher's Eugenics Background: Ronald Fisher was a committed eugenicist. He was the chair of the Cambridge Eugenics Society as a student. Fisher wrote prolifically for The Eugenics Review. He was the Galton Professor of Eugenics at University College London and editor of the Annals of Eugenics.
- Fisher's Statistical Contributions: Fisher is responsible for key statistical terms and concepts including "parameter estimation," "maximum likelihood," and "sufficient statistic". His 1925 textbook, Statistical Methods for Research Workers, introduced

significance testing. Fisher formalized the use of the p-value in statistics. He proposed p=0.05 as a limit for statistical significance. Fisher supplied statistical tests like Fisher's F-test, ANOVA (analysis of variance), and Fisher's exact test.

- Fisher promoted significance testing as a way to decide all manner of questions, viewing it as having an objective basis.
- Galton, Pearson, and Fisher needed a quantitative way to argue for the existence of eugenic differences, and they used significance testing to do so.
- Significance testing was useful for stating that racial subgroups existed, or that there was a "significant" correlation between intelligence and cleanliness, or a "significant" difference in criminality, fertility, or disease between socioeconomic classes.
- Fisher's eugenicist proposals aimed to exclude "inferior types" from the statistics profession.
- In 1904, Karl Pearson published a study that reported similar correlations between inherited traits like eye color and mental qualities like "vivacity" among siblings. He concluded that these traits were equally hereditary, leading to eugenicist conclusions about breeding intelligence.
- Pearson argued that an asymmetry in skull measurements indicated different races, further arguing for racial superiority.
- Criticism of Statistical Significance: Critics argue that a scientific hypothesis is more than a statistical hypothesis, and should explain why, by how much, and why it matters. Significance testing only asks "whether" an effect or association exists, not how much or why it matters.

Sewall Wright and Guinea Pigs

- 1889 - 1988
- Statistics may be regarded as the study of methods of the reductions of data
- Wright argued that statistics was more than just a collection of mathematical methods
- Wright went to Harvard to study genetics and about 1915 got a job with the USDA taking care of Guinea Pigs
- The Guinea Pigs turned out to be the spring board to Wright's success
- Evolution was not gradual, as Darwin posited, but happens in bursts
- 1925, Wright was faculty at University of Chicago and stayed close to Guinea Pigs

- A story is that he was handling a Gunea Pig while lecturing at the chalk board and mistakenely used the Guinea Pig to erase the board
- Guinea Pig coat color refused to play by the genetic understanding of the time
- It proved impossible to breed an all white / all colored guinea pig
- Even the most inbred had a wide variation
- Wright postulated that genetics alone governed coat color and added developmental factors in the womb
- Something in the womb was `causing` coat color
- 20 generations eliminated the genetic variation while maintaining the developmental factors
- Wright's work with guinea pigs contributed to his development of path analysis, which is now used in the social sciences
- Path Analysis is a causal modeling approach to exploring the correlations within a defined network. The method is also known as Structural Equation Modeling
- https://en.wikipedia.org/wiki/Sewall_Wright#Evolutionary_theory Path analysis can disredit causation but cannot prove causation
- Wright, even though right, according to Judea Pearl, was severely attacked at the time by Fisher Sewall Wright's contributions to statistics, particularly his development of path analysis, offer a distinct approach that, in some ways, rivaled Fisher's statistical methods.
- **Path Analysis:** Wright invented path analysis in 1921, which is a statistical method that uses a graphical model and is still widely used in social science. Path analysis is a technique for estimating unknown parameters given a set of simultaneous equations, and of mapping out the interrelations among a pre-determined network of variables. It is credited to biometrician Sewall Wright.
- **Population Genetics**: Wright was a founder of population genetics alongside Ronald Fisher and J. B. S. Haldane. Their theoretical work is the origin of the modern evolutionary synthesis or neo-Darwinian synthesis. Wright, along with Fisher, pioneered methods for computing the distribution of gene frequencies among populations as a result of the interaction of natural selection, mutation, migration and genetic drift.
- **Coefficient of Determination:** Wright is credited with creating the statistical coefficient of determination, first published in 1921. This metric is commonly

employed to evaluate regression analyses in computational statistics and machine learning.

- **Inbreeding Coefficient:** Wright discovered the inbreeding coefficient and methods of computing it in pedigree animals and extended this work to populations, computing the amount of inbreeding between members of populations as a result of random genetic drift.
- **Wright's View on Causation:** Wright introduced his method of path coefficients in the context of causality and perhaps unintentionally, forever linked the statistical method with causal issues. Essential to the controversy surrounding Wright's methods were claims, originally advanced by Wright, that the method could be applied to problems in which causality among variables could be assumed. Wright outlined two ways in which the method of path coefficients may be correctly employed: 1) should one have presumed knowledge of the causal relations inherent in a system of variables, path analysis could be used to find "the degree to which variation of a given effect is determined by each particular cause", and 2) in those situations in which causal relations were uncertain, the method of path coefficients could be used to deduce the logical consequences inherent in the system.
- **Rivalry with Fisher**:
    - **Interpretation of Population Genetics:** By the mid-1920s, interpretation of the mathematical theories of population genetics became a point of contention between Fisher and Wright, and the issue became acrimonious. Dispute persisted for the rest of Fisher's life.
    - **Genetic Theory of Dominance:** Wright did not accept Fisher's genetic theory of dominance, but instead considered it to arise from biochemical considerations.
- **How Wright's Approach Differed from Fisher's**:
    - **Causal Modeling**: Wright explicitly linked his path analysis to causal inference.
    - **Emphasis on Multiple Factors**: Wright focused on the interaction of genetic drift and other evolutionary forces.
    - **Fitness Landscapes**: Wright described the relationship between genotype or phenotype and fitness as fitness surfaces or evolutionary landscapes.
    - **Holistic View**: Wright's approach incorporated a broader view of evolutionary processes, considering multiple interacting factors and emphasizing the importance of genetic drift alongside natural selection.

- **Panpsychism**: Wright was one of the few geneticists of his time to venture into philosophy. He endorsed a form of panpsychism, believing that consciousness was an inherent property of elementary particles rather than an emergent property of complexity.

While Fisher focused heavily on statistical significance and hypothesis testing, Wright's path analysis provided a method for exploring and visualizing complex relationships among variables, particularly in the context of genetics and evolution. The two scientists also held differing views regarding population genetics and the genetic theory of dominance.

Sewall Wright's work with guinea pigs significantly influenced his theories on evolution, offering a perspective that differed from prevailing views at the time.

Here's a breakdown:

- **Empirical Observations**
  - Wright conducted extensive experiments with approximately 80,000 guinea pigs.
  - He analyzed characters of some 40,000 guinea pigs in 23 strains of brother-sister matings against a random-bred stock.
- **Insights into Genetic Drift**:
  - His observations of guinea pig populations over many years demonstrated "cumulative accidents of sampling".
  - Wright noted extreme differences between stocks that had been inbred in parallel for many years.
- **Shifting Balance Theory**:
  - Wright's guinea pig studies contributed to his "shifting balance" theory of evolution.
  - This theory posits that optimal conditions for adaptive evolution occur in large populations subdivided into partially isolated groups.
  - In smaller subpopulations, genetic drift has a significant influence, allowing groups to diverge more rapidly.
  - Subpopulations that happen upon a more adaptive combination of alleles will spread and take over the population.
- **Adaptive Landscape**:
  - Wright introduced the concept of an "adaptive landscape" with fitness peaks and maladaptive valleys.

    - He argued that large, freely mixing populations tend to get trapped in locally adaptive peaks, while subdivided populations can more efficiently explore the broader landscape of gene combinations.
    - According to Wright, small populations would be able to drift away from a locally adaptive peak in the fitness landscape, across a saddle to scale an even higher peak.

- **Differences from prevailing evolutionary thinking**:
    - **Emphasis on Genetic Drift**: Wright emphasized the role of random genetic drift, which he sometimes referred to as the Sewall Wright effect.
    - **Genetic drift** refers to **cumulative, stochastic (random) changes in gene frequencies** that arise from random births, deaths, and Mendelian segregations in reproduction. It is also sometimes known as the Sewall Wright effect. Sewall Wright's guinea pig studies demonstrated genetic drift. Wright emphasized the role of random genetic drift as an important evolutionary force, in addition to natural selection.
    - **Mendelian segregations** are a component of genetic drift.
    - **Genetic drift** refers to cumulative, stochastic (random) changes in gene frequencies that arise from random births, deaths, and **Mendelian segregations** in reproduction.
    - **Interaction of Evolutionary Forces**: Wright was convinced that the interaction of genetic drift and other evolutionary forces was important in adaptation, not simply natural selection.
    - **Subdivided Populations**: He theorized that subdivided populations could more efficiently explore gene combinations, which runs contrary to large, freely mixing populations.

- **Conflict with Fisher**: Wright's shifting balance theory and emphasis on genetic drift led to conflict with R.A. Fisher, another pioneer in population genetics. Fisher believed that most natural populations were too large for the effects of genetic drift to be significant.

Wright's guinea pig experiments provided empirical evidence for the power of genetic drift and the importance of population structure in evolution. His theories offered a nuanced perspective that took into account the interplay of multiple evolutionary forces, which contrasted with views that prioritized natural selection as the primary driver of adaptation.

Mendel

- 1824 - 1888
- **Mendel's laws of inheritance**, established by Gregor Mendel through experiments with pea plants, include the "Law of Dominance," the "Law of Segregation," and the "Law of Independent Assortment," which explain how traits are passed from parents to offspring, stating that each individual carries two alleles for a trait, with one allele being passed on to each offspring during gamete formation, and that different traits are inherited independently of one another.
- **Mendel's laws of inheritance** were reconciled with Darwin's vision of natural selection through formal frameworks such as Sewall Wright's "Evolution in Mendelian populations".
- **Mendel's principles were rediscovered in 1900**, which resulted in a conflict between the biometricians, who followed Galton's Law of Ancestral Heredity, and those who advocated for Mendel's principles.
- Ronald Fisher's 1930 book *The Genetical Theory of Natural Selection* helped reconcile **Mendelian genetics with Darwinian evolution**.
- In 1944, Fisher used a Pearson's chi-squared test to analyze Mendel's data and concluded that **Mendel's results were far too perfect**, suggesting that adjustments (intentional or unconscious) had been made to the data to make the observations fit the hypothesis.

Darwin

- 1809 - 1882
- Evolution by natural selection
    - within a species, individuals exhibit variation in traits
    - these variations are heritable (passed on to offspring)
    - more offspring are produced than can survive, leading to a "struggle for existence"
    - individuals with advantageous traits for their environment are more likely to survive and reproduce, causing these beneficial traits to become more common in the population over time, potentially leading to new species through gradual change. [1, 2, 3, 4, 5]
- Variation exists: Individuals within a population have natural variations in their traits. [3, 5, 6]
- Inheritance: These variations are passed on from parents to offspring through genetic material. [3, 4, 5]

- Competition for survival: Due to overpopulation, there is a struggle for limited resources, leading to competition among individuals. [4, 5, 6]
- Natural selection: Individuals with advantageous traits are more likely to survive and reproduce, passing on those traits to their offspring. [3, 4, 5]
- Gradual change: Over many generations, these small variations accumulate, leading to significant changes in a population and potentially the emergence of new species. [2, 5, 6]
- Darwin and eugenics: https://hsm.stackexchange.com/questions/3328/did-darwin-ever-express-his-views-on-eugenics

The Reconciliation

"Initially, the groundbreaking work of Gregor Mendel on inheritance and Charles Darwin's theory of natural selection were perceived as distinct, and at times, conflicting concepts. Mendel's meticulous experiments with pea plants revealed the fundamental principles of inheritance, demonstrating the transmission of traits through discrete units, now known as genes (Mendel, 1866). Conversely, Darwin's theory of natural selection posited that species evolve through the differential survival and reproduction of individuals with advantageous traits (Darwin, 1859).

However, a pivotal shift occurred when scientists recognized the complementary nature of these theories, forming a more comprehensive understanding of evolutionary processes. Darwin's theory elucidated how favorable traits proliferate within a population, yet it lacked a clear explanation of the inheritance mechanism (Darwin, 1859). Mendel's laws, on the other hand, provided this crucial missing link, elucidating the precise manner in which traits are passed across generations (Mendel, 1866).

This reconciliation, often referred to as the modern evolutionary synthesis, underscored that Mendelian inheritance serves as the fundamental genetic mechanism underpinning Darwinian natural selection. In essence, Mendel's laws explained the 'how' of trait inheritance, while Darwin's theory explained the 'why' of evolutionary change over time, resulting from the selection of those inherited traits."

General References for Further Reading:

- **Darwin, C. (1859). *On the origin of species by means of natural selection, or the preservation of favoured races in the struggle for life*. John Murray.** (This is the foundational text for Darwin's theory of natural selection.)
- **Mendel, G. (1866). Versuche über Pflanzen-Hybriden. *Verhandlungen des naturforschenden Vereines in Brünn, IV. Abhandlungen*, 3-47.** (This is Mendel's original paper on his pea plant experiments, a key work in genetics.)

- **Any modern Evolutionary biology or Genetics Textbook.** (These will have sections covering the work of both scientists and the Modern Synthesis.)
- **Biographies of Charles Darwin and Gregor Mendel.** (These provide valuable context and historical perspective.)

Sewall Wright Comparison

"While both Sewall Wright and the combined contributions of Mendel and Darwin have profoundly shaped our understanding of evolution, their works differ significantly in focus, emphasis, key concepts, and methods.

Focus:

The collaborative work of Mendel and Darwin centers on the fundamental mechanisms of heredity, as elucidated by Mendel's laws, and the subsequent evolutionary changes driven by natural selection, as outlined by Darwin (e.g., Mendel, 1866; Darwin, 1859). In contrast, Sewall Wright's research delves into the intricate complexities of evolutionary processes, encompassing factors such as genetic drift, population structure, and the interplay of multiple evolutionary forces (e.g., Wright, 1931).

Emphasis:

Mendel and Darwin's work emphasizes the role of variation and selection in driving gradual evolutionary changes within populations over time (e.g., Darwin, 1859). Wright, however, underscored the significance of random genetic drift and the impact of population structure, offering a more nuanced perspective on evolutionary dynamics (e.g., Wright, 1931).

Key Concepts:

The core concepts associated with Mendel and Darwin include genes, alleles, inheritance patterns, variation, natural selection, and adaptation (e.g., Darwin, 1859). Sewall Wright introduced pivotal concepts such as genetic drift, the "shifting balance" theory, adaptive landscapes, and path analysis (e.g., Wright, 1931).

Methods:

Mendel employed controlled breeding experiments with pea plants, while Darwin relied on meticulous observation and comparative analysis (e.g., Mendel, 1866; Darwin, 1859). Wright's methodologies encompassed both empirical studies, such as his experiments with guinea pigs, and sophisticated theoretical modeling, including path analysis (e.g., Wright, 1931).

In essence, Mendel and Darwin laid the groundwork for understanding the fundamental principles of inheritance and natural selection, whereas Sewall Wright expanded upon this foundation by exploring the more intricate and multifaceted aspects of evolutionary processes."

General References for Further Reading:

- **Darwin, C. (1859). *On the origin of species by means of natural selection, or the preservation of favoured races in the struggle for life*. John Murray.** (This is the foundational text for Darwin's theory of natural selection.)
- **Mendel, G. (1866). Versuche über Pflanzen-Hybriden. *Verhandlungen des naturforschenden Vereines in Brünn, IV. Abhandlungen*, 3-47.** (This is Mendel's original paper on his pea plant experiments, a key work in genetics.)
- **Wright, S. (1931). Evolution in Mendelian populations. *Genetics*, *16*(2), 97–159.** (This is a significant paper by Wright that introduced many of his key ideas about population genetics.)
- **Provine, W. B. (1986). *Sewall Wright: geneticist and evolutionist*. University of Chicago Press.** (This is a comprehensive biography of Sewall Wright.)
- **Any modern Evolutionary biology or Genetics Textbook.** (These will have sections covering the work of all three scientists.)
- **Francis Galton - Wikipedia:** https://en.wikipedia.org/w/index.php?title=Francis_Galton&oldid=1276619382
- **How Eugenics Shaped Statistics - Nautilus:** The article is part of Nautilus Magazine, accessible via their website.
- **Karl Pearson - Wikipedia:** The general URL for Wikipedia is https://en.wikipedia.org, but a specific URL for the Karl Pearson page was not provided.
- **Ronald Fisher - Wikipedia:** The general URL for Wikipedia is https://en.wikipedia.org, but a specific URL for the Ronald Fisher page was not provided.
- **Sewall Wright - Wikipedia:** The general URL for Wikipedia is https://en.wikipedia.org, but a specific URL for the Sewall Wright page was not provided.
- **Facing History & Ourselves:** https://www.facinghistory.org
- "Adolphe Quetelet (1796-1874)" https://en.wikipedia.org/wiki/Adolphe_Quetelet
- "ORIGINS OF PATH ANALYSIS: Causal Modeling and the Origins of Path Analysis" https://www.sciencedirect.com/topics/economics-econometrics-and-finance/path-analysis

- "Sewall Wright: Evolving Mendel – Genes to Genomes" https://genestogenomes.org/sewall-wright-evolving-mendel/

Early 20th Century Foundations: Hypothesis Testing & Design (1900–1950)

Modern statistical inference

- "Student" (Gosset and Guinness) develops the t-distribution for small-sample inference; Fisher codifies likelihood, sufficiency, maximum likelihood, p-values, and ANOVA.
- Neyman–Pearson develop hypothesis testing frameworks: Type I/II errors, power, confidence intervals via coverage.
- Experiments and agriculture → the template for A/B testing
- Fisher's randomized, replicated field trials at Rothamsted provide the blueprint: randomization, blocking, factorial designs, and ANOVA carry directly into industrial and, later, online experiments.

Mid–Late 20th Century (1950s–1990s): From Industry to Computing to the Web

- Design of Experiments spreads to manufacturing and medicine; generalized linear models (Nelder & Wedderburn), robust statistics (Huber), nonparametrics (Wilcoxon, Mann–Whitney), and multivariate analysis (PCA, discriminant analysis) become standard.
- Time-series methods (Box–Jenkins ARIMA) formalize forecasting; the bootstrap (Efron) reframes uncertainty by resampling.
- Computing revolution - Electronic computers make simulation (Monte Carlo), optimization, and large-scale estimation routine; data visualization advances.
- A/B testing roots and web era - The randomized, two-arm comparison from agriculture/industry becomes the A/B test in digital products; by the 1990s the logic—randomize users, compare variants, analyze with t/χ^2/ANOVA—sets the stage for the 2000s online experimentation culture.
- Key terms - GLM, logistic/Poisson regression, robustness, bootstrap, ARIMA, experimental platform, online controlled experiment (A/B test).

Conditional Probability and Bayes Theorem

Conditional probability is the likelihood of an event occurring, given that another event has already occurred. It changes the entire **sample space** (the set of all possible outcomes) to only include the outcomes where the given event has happened.

The conditional probability of event A given event B is denoted $P(A|B)$ and calculated as:

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

- $P(A|B)$: The probability of A given B.
- $P(A \cap B)$: The **joint probability**—the probability that both A and B occur.
- $P(B)$: The probability of B occurring.

Analogy: If you pull a card from a standard deck, the probability of it being a King is $4/52$. If I *tell you* the card is a **face card** (J, Q, K), the probability changes. The sample space is now just the 12 face cards, and there are 4 Kings among them.

$$P(\text{King}|\text{Face Card}) = \frac{P(\text{King} \cap \text{Face Card})}{P(\text{Face Card})} = \frac{4/52}{12/52} = 4/12$$

Bayes' Theorem is a mathematical formula that provides a way to calculate a conditional probability by using other related conditional probabilities. Crucially, it shows how to **reverse** the conditionality.

It is derived directly from the definition of conditional probability: since $P(A \cap B) = P(A|B)P(B)$ and $P(A \cap B) = P(B|A)P(A)$, setting them equal gives:

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

The standard form of Bayes' Theorem is:

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

The terms in the formula have specific names when applied in a statistical context:

- $P(A|B)$:
  - **Name:** Posterior Probability
  - **Description:** The updated probability of A after observing the evidence B. (What we want to find)
- $P(A)$:

- **Name:** Prior Probability
- **Description:** The initial belief about event A before seeing any evidence.

- $P(B|A)$**:**
  - **Name:** Likelihood
  - **Description:** The probability of observing the evidence B given that event A is true.
- $P(B)$**:**
  - **Name:** Evidence (or Marginal Likelihood)
  - **Description:** The total probability of observing the evidence B, regardless of A.

Role: Bayes' Theorem is the core tool for **updating beliefs**. You start with an initial belief ($P(A)$) and use new evidence (B) to calculate a new, revised belief ($P(A|B)$).

Bayesian Inference is a statistical framework that uses Bayes' Theorem to update the probability of a **hypothesis** as more **evidence** or data becomes available.

It contrasts with the classical (frequentist) approach to statistics by treating unknown parameters (like a coin's true bias) as **random variables** with their own probability distributions, rather than as fixed, unknown constants.

The process of Bayesian Inference involves four main steps, which align perfectly with the terms in Bayes' Theorem:

1. **Define the Prior (**$P(\text{Hypothesis})$**):** Establish your initial belief about the hypothesis (or parameter θ) before seeing the data. This is typically a probability distribution.
2. **Calculate the Likelihood (**$P(\text{Data}|\text{Hypothesis})$**):** Determine the probability of observing the actual data, assuming the hypothesis is true. This comes from your statistical model.
3. **Compute the Posterior (**$P(\text{Hypothesis}|\text{Data})$**):** Use Bayes' Theorem to combine the **Prior** and the **Likelihood** to get the **Posterior** distribution.
4. **Prediction and Iteration:** The resulting **Posterior** becomes the new **Prior** when new data arrives, allowing for a continuous process of learning and belief updating.

Core Idea: Bayesian inference quantifies uncertainty and learning. It starts with an initial state of knowledge (the **Prior**) and uses data (the **Likelihood**) to systematically shift that knowledge to a refined state (the **Posterior**).

Probability Foundations

A FEW YEARS AGO a man won the Spanish national lottery with a ticket that ended in the number 48. Proud of his “accomplishment,” he revealed the theory that brought him the riches. “I dreamed of the number 7 for seven straight nights,” he said, “and 7 times 7 is 48.”1 Those of us with a better command of our multiplication tables might chuckle at the man’s error, but we all create our own view of the world and then employ it to filter and process our perceptions, extracting meaning from the ocean of data that washes over us in daily life. And we often make errors that, though less obvious, are just as significant as his.

- The probability that two events will both occur can never be greater than the probability that each will occur individually. Why not? Simple arithmetic: the chances that event A will occur = the chances that events A and B will occur + the chance that event A will occur and event B will not occur.
- If two possible events, A and B, are independent, then the probability that both A and B will occur is equal to the product of their individual probabilities
- If an event can have a number of different and distinct possible outcomes, A, B, C, and so on, then the probability that either A or B will occur is equal to the sum of the individual probabilities of A and B, and the sum of the probabilities of all the possible outcomes (A, B, C, and so on) is 1 (that is, 100%)

Adding or Multiplying Probabilities

- If using the word `or`, add
- If using the word `and`, multiply
    - What's the probability of rolling a 1 or a 6
    - What's the probability of first rolling a 1 and then a 6

Adding Probabilities (The "OR" Rule)

You **add** probabilities when you want to find the chance of **at least one** of several events occurring. This applies to **mutually exclusive events**, meaning the events cannot happen at the same time.

- $P(A \cup B) = P(A) + P(B)$
- $P(A \text{ or } B) = P(A) + P(B)$

Imagine rolling a standard six-sided die. What's the probability of rolling a **1 or a 6**?

- $P(\text{rolling a } 1) = \frac{1}{6}$
- $P(\text{rolling a } 6) = \frac{1}{6}$
- $P(1 \text{ or } 6) = P(1) + P(6) = \frac{1}{6} + \frac{1}{6} = \frac{2}{6} = \frac{1}{3} \approx 33.3\%$

Multiplying Probabilities (The "AND" Rule)

You **multiply** probabilities when you want to find the chance of **two or more events** happening in sequence or at the same time.

- $P(A \cap B) = P(A) * P(B)$
- $P(A \text{ and } B) = P(A) \times P(B)$
- $P(A \text{ and } B) = P(A) \times P(B \mid A)$

The term $P(B \mid A)$ means "the probability of event B occurring **given that** event A has already occurred." This is called **conditional probability**.

Probability of Two Events Occurring Together: Independent

Independent events are those where the outcome of the first event **does not affect** the outcome of the second event.

Events A and B are independent if $P(B \mid A) = P(B)$.

What is the probability of flipping a coin and getting a **Head AND** rolling a six-sided die and getting a **4**?

The coin flip and the die roll don't affect each other, so they are independent.

- $P(\text{Head}) = \frac{1}{2}$
- $P(4) = \frac{1}{6}$
- $P(\text{Head and } 4) = P(\text{Head}) \times P(4) = \frac{1}{2} \times \frac{1}{6} = \frac{1}{12} \approx 8.3\%$

Probability of Two Events Occurring Together: Dependent

Dependent events are those where the outcome of the first event **changes** the probability of the second event. This often happens in situations "without replacement."

Events A and B are dependent if $P(B \mid A) \neq P(B)$.

Imagine a bag contains 3 red marbles and 2 blue marbles (a total of 5 marbles). What is the probability of drawing a **red marble AND then** drawing another **red marble without replacing** the first one?

- **Event A:** Drawing the first red marble.
$$P(A) = P(\text{1st Red}) = \frac{\text{3 red marbles}}{\text{5 total marbles}} = \frac{3}{5}$$
- **Event B | A:** Drawing a second red marble *after* the first one was red and not replaced. Now, there are only 2 red marbles left and 4 total marbles left.
$$P(B \mid A) = P(\text{2nd Red} \mid \text{1st Red}) = \frac{\text{2 remaining red marbles}}{\text{4 remaining total marbles}} = \frac{2}{4} = \frac{1}{2}$$

- **Probability of both:**

$$P(\text{1st Red and 2nd Red}) = P(A) \times P(B \mid A) = \frac{3}{5} \times \frac{1}{2} = \frac{3}{10} = 0.3 = 30$$

The Conjuction Fallacy

The Conjunction Fallacy is a fallacy or error in decision making where people judge that a conjunction of two possible events is more likely than one or both of the conjuncts.

- Which is greater: the number of six-letter English words having n as their fifth letter or the number of six-letter English words ending in ing?
- Which is more likely: that a defendant, after discovering the body, left the scene of the crime or that a defendant, after discovering the body, left the scene of the crime because he feared being accused of the grisly murder?
- Is it more probable that the president will increase federal aid to education or that he or she will increase federal aid to education with funding freed by cutting other aid to the states?
- Is it more likely that your company will increase sales next year or that it will increase sales next year because the overall economy has had a banner year?

In each case, even though the latter is less probable than the former, it may sound more likely. Or as Kahneman and Tversky put it, "A good story is often less probable than a less satisfactory explanation.

The Drunkard's Walk

The Prosecutor's Fallacy

Juanita Brooks: Dependent or Independent Variables?

- On June 18, 1964, Mrs. Juanita Brooks was walking home along an alley in the San Pedro area of Los Angeles.
- She was carrying a basket of groceries and had her purse on top of the packages.
- She was using a cane when she was suddenly pushed to the ground by a person she neither saw nor heard approach.
- The person stole her purse, which contained between $35 and $40.
- Looking up Mrs. Brooks noticed a young woman with light hair running away.
- A witness, John Bass, saw a young woman run out of the alley and enter a yellow car.

- The car was described as being medium to large in size and yellow, with either an off-white or egg-shell white top.
- The driver was described as a person of color, wearing a mustache and beard.
- A few days later a police officer spotted a car matching the description of the suspects' vehicle near their home.
- The suspects matched the description except for the man's beard.
- The police arrested Malcolm Ricardo Collins and his wife Janet.
- The victim could not identify Janet and hadn't seen the driver.
- The witness couldn't positively identify Malcolm as the driver.

Enter the star witness, described in the California Supreme Court opinion only as “an instructor of mathematics at a state college.” This witness testified that the fact that the defendants were “a Caucasian woman with a blond ponytail...[and] a person of color with a beard and mustache” who drove a partly yellow automobile was enough to convict the couple. To illustrate its point, the prosecution presented this table, quoted here verbatim from the supreme court decision:

- Partly yellow automobile 1/10
- Man with mustache 1/4
- Man with beard 1/10
- Woman with ponytail 1/10
- Woman with blond hair 1/3
- Interracial couple in car 1/1,000

Results: The math instructor called by the prosecution said that the product rule applies to this data. By multiplying all the probabilities, one concludes that the chances of a couple fitting all these distinctive characteristics are 1 in 12 million. Accordingly, he said, one could infer that the chances that the couple was innocent were 1 in 12 million. The prosecutor then pointed out that these individual probabilities were estimates and invited the jurors to supply their own guesses and then do the math. He himself, he said, believed they were conservative estimates, and the probability he came up with employing the factors he assigned was more like 1 in 1 billion. The jury bought it and convicted the couple.

- The math instructor's argument is flawed because the categories are not independent.
- Eliminating the category 'Man with beard' reduces the product of probabilities to 1 in 1 million (many have beard and moustache).
- The relevant probability is the chance that a matching couple is guilty, not the chance that a random couple matches the description.

- Given the population size of several million, the probability of a matching couple being guilty is only 1 in 2 or 3 (1 in a million in a population with several million).
- The conviction was overturned due to these errors.

O. J. Simpson: The Other Side of Probability

- Alan Dershowitz used the prosecutor's fallacy in the O.J. Simpson trial.
- The prosecution focused on Simpson's history of violence against Nicole.
- Dershowitz argued that most abusive spouses don't kill their partners, which is true.
- However, he ignored the statistic that most murdered battered women are killed by their abusers.
- The jury found Simpson not guilty.
- Dershowitz believes the justice system doesn't require telling the whole truth.

Base Rate Fallacy

The Base Rate Fallacy is the tendency for people to **ignore or heavily discount** the **base rate** (the general, statistical frequency of an event) in favor of **case-specific, anecdotal, or vivid information** when estimating a probability.

In simple terms, it means forgetting to ask: **"How common is this event in the first place?"**

People tend to prioritize seemingly relevant *specific* details over the overarching *statistical* context, leading to inaccurate judgments of likelihood.

Classic Example: The Medical Test Paradox

The fallacy is most clearly illustrated with screening tests for a rare condition:

1. **Base Rate (General Probability):** A disease is very **rare**, affecting only **1 in 1,000** people (0.1%).
2. **Case-Specific Information (Test Accuracy):** A diagnostic test is highly accurate: it correctly detects the disease **95%** of the time it's present, and it only gives a false positive **5%** of the time it's *not* present.
3. **The Fallacy:** A person takes the test and gets a **positive result**. Many people fall for the fallacy and conclude: "The test is 95% accurate, so I have a 95% chance of having the disease."

The Reality (Using Base Rate): Because the disease is so rare, the vast majority of positive test results will actually be **false positives** from the healthy population. The true

probability of a person having the disease, given their positive test, is often **less than 5%** (a calculation that requires Bayes' theorem to solve correctly). The low base rate overrides the high accuracy rate.

```
# Define the probabilities based on the classic base rate fallacy example:
P_D = 0.001  # Base Rate (Prevalence): P(Disease) = 1 in 1,000
P_T_plus_given_D = 0.95 # Sensitivity: P(Positive Test | Disease) = 95%
P_FP_rate = 0.05 # False Positive Rate: P(Positive Test | No Disease) = 5%

# 1. Calculate the probability of NOT having the disease, P(~D)
P_not_D = 1 - P_D

# 2. Calculate the overall probability of a positive test, P(T+) (Denominator of Ba
# P(T+) = P(T+|D) * P(D) + P(T+|~D) * P(~D)
P_T_plus = (P_T_plus_given_D * P_D) + (P_FP_rate * P_not_D)

# 3. Apply Bayes' Theorem to find the probability of having the disease given a pos
# P(D|T+) = P(T+|D) * P(D) / P(T+)
P_D_given_T_plus = (P_T_plus_given_D * P_D) / P_T_plus

# Convert to percentage and round
percentage = P_D_given_T_plus * 100

print(f"Prevalence P(D): {P_D}")
print(f"Sensitivity P(T+|D): {P_T_plus_given_D}")
print(f"False Positive Rate P(T+|~D): {P_FP_rate}")
print(f"Overall probability of a positive test P(T+): {P_T_plus}")
print(f"Probability of disease given positive test P(D|T+): {P_D_given_T_plus}")
print(f"Actual percentage: {percentage}")
```

```
Prevalence P(D): 0.001
Sensitivity P(T+|D): 0.95
False Positive Rate P(T+|~D): 0.05
Overall probability of a positive test P(T+): 0.0509
Probability of disease given positive test P(D|T+): 0.018664047151277015
Actual percentage: 1.8664047151277015
```

The actual percentage of having the rare disease, given a positive test result, is approximately **1.87%**.

This result is a classic demonstration of the **Base Rate Fallacy**, showing how the rarity of a disease (the base rate) dramatically outweighs the high accuracy of the test. Most people who commit the fallacy would incorrectly guess the probability is near 95%.

Calculation using Bayes' Theorem

We use **Bayes' theorem** to calculate the probability of having the disease (D) given a positive test result ($T+$), which is $P(D|T+)$.

1. Define Initial Probabilities

Based on the example used previously:

- **Prevalence ($P(D)$):** 1 in $1,000 = 0.001$
- **No Disease ($P(\neg D)$):** $1 - 0.001 = 0.999$
- **Sensitivity ($P(T+|D)$):** $95\% = 0.95$ (True Positive Rate)
- **False Positive Rate ($P(T+|\neg D)$):** $5\% = 0.05$

2. Calculate the Overall Probability of a Positive Test ($P(T+)$)

The overall probability of receiving a positive test is the sum of two scenarios: getting a true positive (sick people) OR getting a false positive (healthy people).

$$P(T+) = P(T+|D) \cdot P(D) + P(T+|\neg D) \cdot P(\neg D)$$
$$P(T+) = (0.95 \cdot 0.001) + (0.05 \cdot 0.999)$$
$$P(T+) = 0.00095 + 0.04995$$
$$P(T+) = 0.0509$$

This means about 5.09% of all people tested will receive a positive result.

3. Apply Bayes' Theorem

Now we substitute these values to find the final probability:

$$P(D|T+) = \frac{P(T+|D) \cdot P(D)}{P(T+)}$$
$$P(D|T+) = \frac{0.95 \cdot 0.001}{0.0509}$$
$$P(D|T+) \approx 0.01866$$

Converting to a percentage, the actual probability of having the disease given a positive test is approximately 1.87%.

The result is extremely low because the ≈ 50 healthy people who receive a false positive (out of every $1,000$) far outnumber the single person who receives a true positive.

Why It Occurs

The Base Rate Fallacy is often a result of using a mental shortcut called the **Representativeness Heuristic**, where:

- People judge the probability of an event by how much it **resembles** a stereotype or a previously known example, rather than by objective probability.
- The vivid, specific details (like a positive test result or a detailed profile of a person) create a more compelling *narrative* than cold, dry statistics.

Avoiding the fallacy requires consciously forcing oneself to integrate both the specific evidence and the overall base rate into the probability assessment.

Monty Hall Problem

Ask Marilyn column (in Parade magazine)

- Marilyn von Savant
- Guinness World Records Hall of Fame highest IQ
- Married to Robert Jarvik, artificial heart

Let's Make a Deal

- Monty Hall
- Wayne Brady

Suppose the contestants on a game show are given the choice of three doors: Behind one door is a car; behind the others, goats. After a contestant picks a door, the host, who knows what's behind all the doors, opens one of the unchosen doors, which reveals a goat. He then says to the contestant, "Do you want to switch to the other unopened door?" Is it to the contestant's advantage to make the switch?

- Marilyn says switch
- Onslaught of criticism started rolling in
- Mathematical Professors and teachers

Professor from George Mason: Let me explain: If one door is shown to be a loser, that information changes the probability of either remaining choice—neither of which has any reason to be more likely—to 1/2. As a professional mathematician, I'm very concerned with the general public's lack of mathematical skills. Please help by confessing your error and, in the future, being more careful.

From Dickinson State University came this: "I am in shock that after being corrected by at least three mathematicians, you still do not see your mistake." From Georgetown: "How many irate mathematicians are needed to change your mind?" And someone from the

U.S. Army Research Institute remarked, “If all those PhDs are wrong the country would be in serious trouble.” Responses continued in such great numbers and for such a long time that after devoting quite a bit of column space to the issue, Marilyn decided she would no longer address it.

Marilyn was right and here's the breakdown:

- Starting with 1 out 3 choice you have the lucky guess scenario of picking the right door (1 out of 3)
- The Wrong Guess scenario has chances 2 out of 3 that you are wrong
- Host intervenes and opens a door knowing where the car is and not wanting to reveal it yet. This action violates randomness

Behind door 1	Behind door 2	Behind door 3	Result if staying at door 1	Result if switching to the door offered
Goat	Goat	Car	Wins goat	Wins car
Goat	Car	Goat	Wins goat	Wins car
Car	Goat	Goat	Wins car	Wins goat

https://en.m.wikipedia.org/wiki/Monty_Hall_problem

Random Variables

Possible outcomes for an event like flipping a coin or throwing a die.

According to Wikipedia (2022):

> A random variable is a mathematical formalization of a quantity or object which depends on random events. Informally, randomness typically represents some fundamental element of chance, such as in the roll of a dice; it may also represent uncertainty, such as measurement error (para 1).

Random variable. (February 6, 2022). In *Wikipedia*. https://en.wikipedia.org/wiki/Random_variable

IID: Independent and Identically Distributed

According to Wikipedia (2022):

> In probability theory and statistics, a collection of random variables is independent and identically distributed if each random variable has the same probability distribution as the others and all are mutually

> independent. This property is usually abbreviated as i.i.d. or iid or IID. IID was first used in statistics. With the development of science, IID has been applied in different fields such as data mining and signal processing (para. 1).

Independent and identically distributed random variables. (February 13, 2022). In *Wikipedia*. https://en.wikipedia.org/wiki/Independent_and_identically_distributed_random_variables

https://stackoverflow.com/questions/24204582/generate-multiple-independent-random-streams-in-python

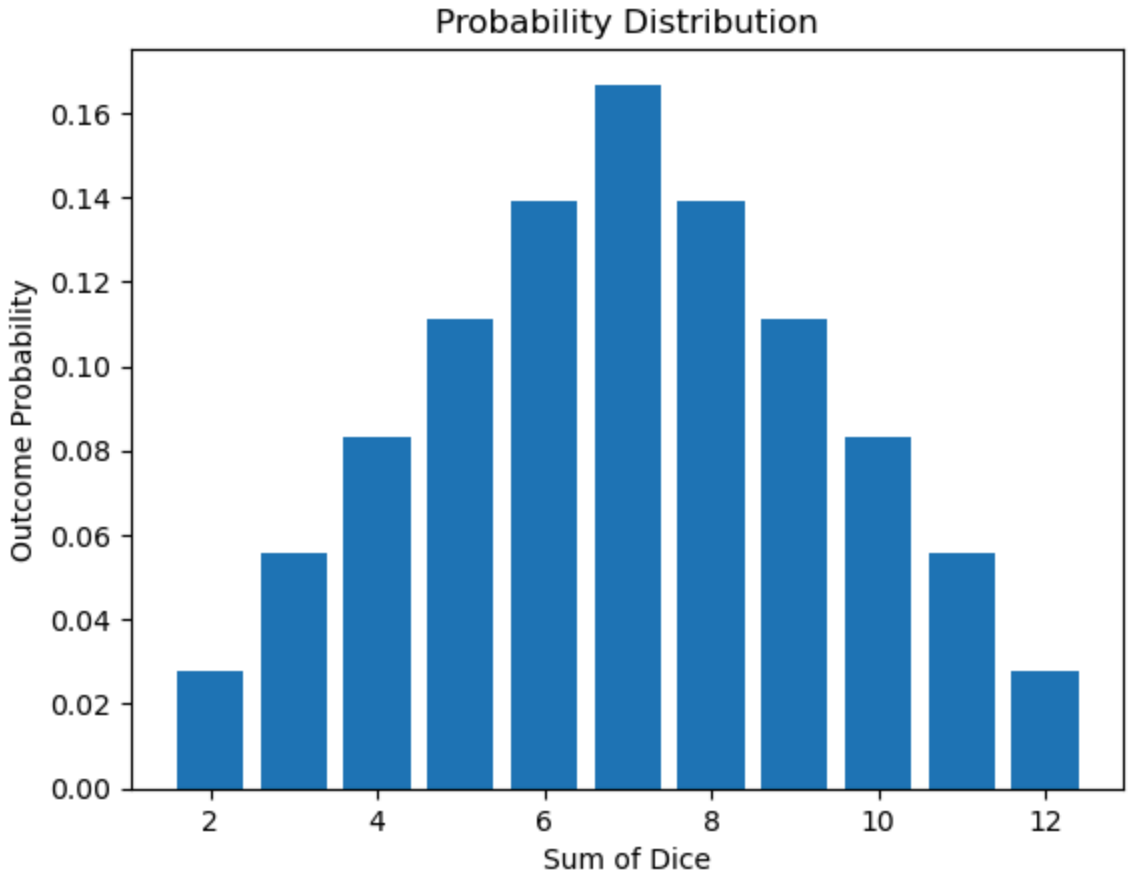
https://www.statisticshowto.com/iid-statistics/

Discrete Random Variables

Countable possible outcomes

Rolling dice (2):
https://www.thoughtco.com/probabilities-of-rolling-two-dice-3126559

```
# plot probability distribution
import matplotlib.pyplot as plt

sums = [2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12]
probs = [1/36, 2/36, 3/36, 4/36, 5/36, 6/36, 5/36, 4/36, 3/36, 2/36, 1/36]

plt.bar(sums, probs)
plt.xlabel('Sum of Dice')
plt.ylabel('Outcome Probability')
plt.title('Probability Distribution')
plt.show();
```



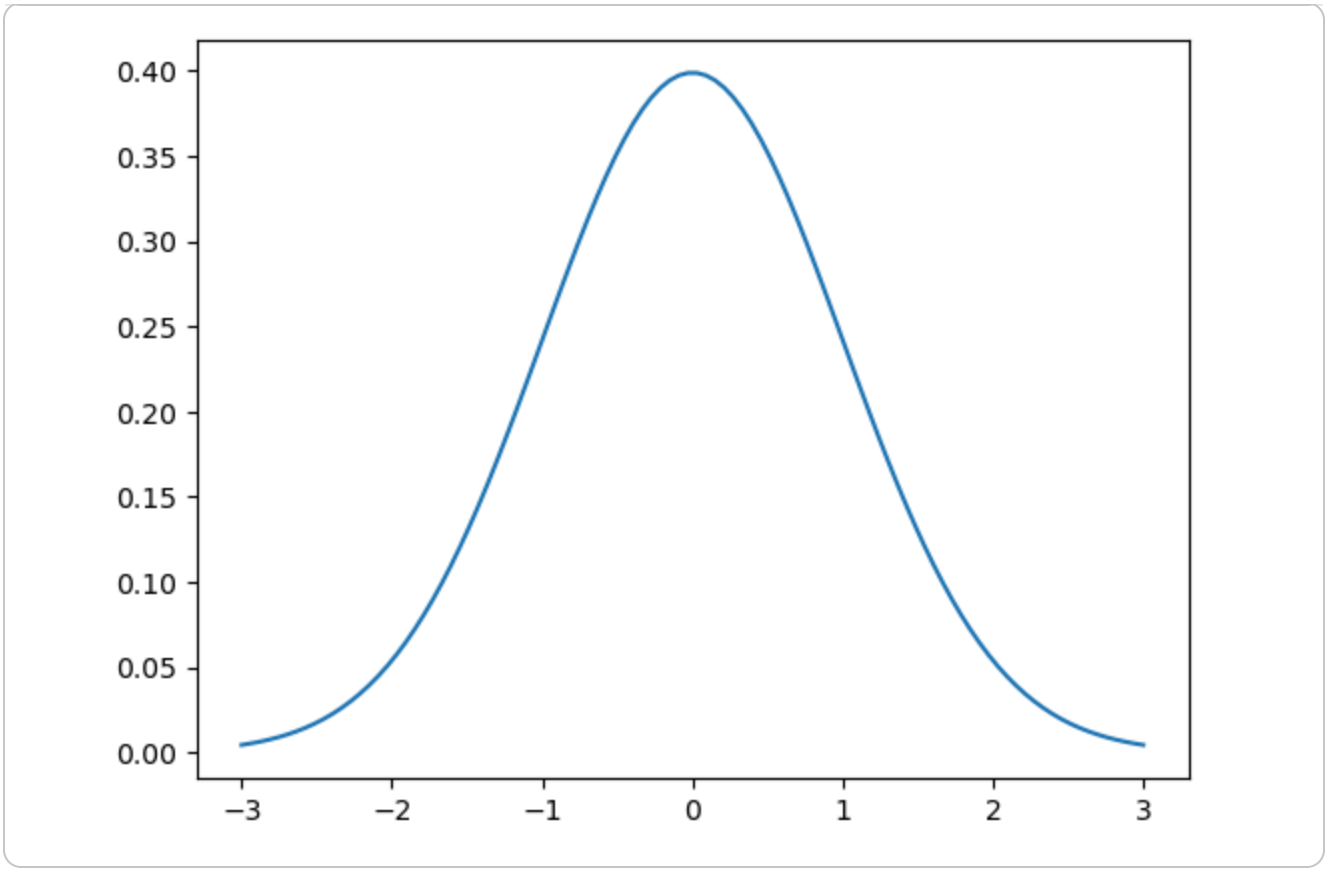
Continuous Random Variables

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

A continuous random variable X takes all values in a given interval

- Height, weight, the amount of sugar in an orange, the time required to run a mile
- Often plotted as a curve, or density curve
- P(some number) = 0

```
# plot normal curve
import numpy as np
import matplotlib.pyplot as plt
import scipy.stats as stats

x = np.linspace(-3, 3, 100)
plt.plot(x, stats.norm.pdf(x, 0, 1))
plt.show()
```



Binomial vs Discrete

- All binomials are discrete.
- Not all discrete are binomial.
- Binomial was used in games of chance, success or failure, winning or loosing.
- A fair six-sided die is rolled ten times, and the number of 6's is recorded. This is a binomial experiment. There are fixed number of trials (ten rolls), each roll is independent of the others, there are only two outcomes (either it's a 6 or it isn't), and the probability of rolling a 6 is constant.
- The discrete probability distribution has finite number of values for random variable. The binomial distribution random variable is the number of success which is also finite.
- De Moivre, 1733, showed a normal approximation of a binomial distribution.
- https://shiny.rit.albany.edu/stat/binomial/

Sources

- https://www.investopedia.com/terms/d/discrete-distribution.asp
- https://faculty.elgin.edu/dkernler/statistics/ch06/6-2.html
- https://homework.study.com/explanation/explain-the-difference-between-a-probability-distribution-and-a-binomial-distribution.html
- https://en.wikipedia.org/wiki/Binomial_distribution

Probability Distribution Functions

In probability theory and statistics, a probability distribution is the mathematical function that gives the probabilities of occurrence of different possible outcomes for an experiment. It is a mathematical description of a random phenomenon in terms of its sample space and the probabilities of events (subsets of the sample space).

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

The Transition from Discrete to Continuous

The transition from **discrete distributions** (like the binomial) to **continuous distributions** (like the normal curve) required a new set of conceptual tools to describe probability.

Unlike a discrete distribution, which can assign a non-zero probability to a single, specific outcome, a continuous distribution **cannot**—the probability of any single value is zero. Instead, probability is described over a range.

This led to the development of several key functions:

The Probability Density Function (PDF) describes the likelihood for a continuous random variable:

- **Key Conceptual Role:** Describes the **relative likelihood** of a continuous variable.
- **Mathematical Description:** Typically denoted as $f(x)$.
- **Core Concept:** The area under the curve represents the probability.
- **Interpretation:** The value of the PDF at a specific point is **not a probability itself** but a measure of probability density.

The Cumulative Distribution Function (CDF) provides a complementary, cumulative perspective on probability:

- **Key Conceptual Role:** Gives the **cumulative probability** of a variable taking a value less than or equal to a given point.
- **Applies To:** Both discrete and continuous random variables.
- **Mathematical Description:** Denoted as $F(x)$, it is defined as $P(X \leq x)$.
- For a continuous distribution, it is the integral of the PDF: $F(x) = \int_{-\infty}^{x} f(t)dt$.
- **Note:** Differentiation (finding the derivative) and integration are inverse operations, linked by the Fundamental Theorem of Calculus. If you differentiate a function, you

find its rate of change (its slope). If you integrate that rate of change, you get back the original function (its total accumulated value).

- **Form:**
  - For a **discrete distribution**, the CDF is a step function.
  - For a **continuous distribution**, it is a smooth, non-decreasing curve.

The Percent-Point Function (PPF) is the inverse of the CDF:

- **Key Conceptual Role:** Finds the value of the random variable for a given probability.
- **Question Answered:** "For a given probability, what is the value of the random variable?"
- **Application:** Used to find percentiles (e.g., finding the value that corresponds to the 95th percentile).

Summary by Variable Type

Functions for Discrete Variables

- **Probability Mass Function (PMF):** Assigns a specific probability to each discrete outcome, e.g., $P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$.
- **Cumulative Distribution Function (CDF):** Provides the cumulative probability, $F(x) = P(X \leq x)$, and is a step function.

Functions for Continuous Variables

- **Probability Density Function (PDF):** Describes relative likelihood; area under the curve is probability.
- **Cumulative Distribution Function (CDF):** Provides the cumulative probability, $F(x) = \int_{-\infty}^{x} f(t)dt$, and is a smooth curve.
- **Percent-Point Function (PPF):** The inverse of the CDF.

Parameters of a Population and Sample

Parameters are numbers that summarize data for an entire population. A statistic is a number that summarizes data from a sample. Each are notated differently. For example the mean of a population is represented by μ and the mean of a statistic is represented by $\bar{x}$. The formulas look like the following:
population mean: $\mu = \frac{\sum x_i}{N}$
sample mean: $\bar{x} = \frac{\sum x_i}{n}$

Parameters of a Parametric Equation

Parameters are also used in parametric equations. Any equation expressed in terms of parameters is a parametric equation. For example, y = mx + b (slope/intercept form for the equation of a line), is a parametric equation where m and b are considered the parameters because they remain constant for a given line. The variables x and y change, vary, according to the change in x. Y is dependent on the independent x. In fact, we often refer to y as a function of x, notated like $f(x)$.

Most of our datasets contain the X and y so we're not to concerned about finding X. We are concerned with finding the parameters, or the best estimate for our parameters given X. For y = mx + b we want to find the parameters, m & b.

For the normal distribution probability density function

$$\frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

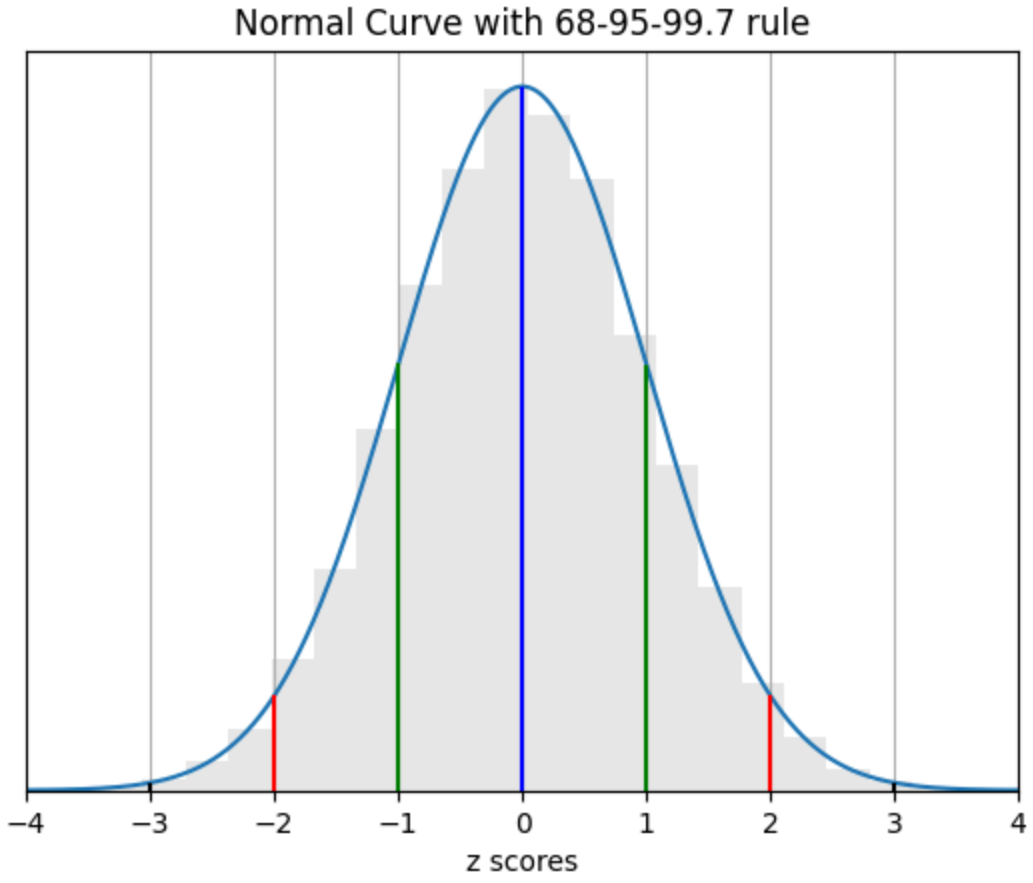
we want to find μ and σ. Parameters are often represented together using notation similar to $\theta = \{\mu, \sigma\}$.

Probability distributions have parameters that define their shape such as mean, variance, and skewness

```
# the normal curve
import numpy as np
import matplotlib.pyplot as plt
from scipy import stats

X = stats.norm.rvs(size=100000)
X.sort()

parameters = stats.norm.fit(X)
mu, sigma = parameters

print('The area between -3 and 3 is considered to contains 99.73% of the area i
print('The area between -2 and 2 is considered to contains 95.44% of the area i
print('The area between -1 and 1 is considered to contains 68.26% of the area i
print('The line at 0 is the expected average of the normal distribution at 50%'
print()
print('Mean: ', np.mean(X))
print('Median: ', np.median(X))
print('Mode?', X[1])
print('Variance: ', np.var(X))
print('Standard Deviation: ', np.std(X))

fig, ax = plt.subplots()
ax.plot(X, stats.norm.pdf(X, loc=mu, scale=sigma))
ax.hist(X, bins=25, density=True, color='lightgray', alpha=0.5)
ax.set_xlabel('z scores')
# ax.set_ylabel('pdf(x)')
```

```
ax.set_xlim(-4, 4)
ax.grid(True)

plt.vlines(x=-3, ymin=0, ymax=stats.norm.pdf(-3, loc=mu, scale=sigma), color='b
plt.vlines(x=-2, ymin=0, ymax=stats.norm.pdf(-2, loc=mu, scale=sigma), color='r
plt.vlines(x=-1, ymin=0, ymax=stats.norm.pdf(-1, loc=mu, scale=sigma), color='g
plt.vlines(x=0, ymin=0, ymax=stats.norm.pdf(0, loc=mu, scale=sigma), color='blu
plt.vlines(x=1, ymin=0, ymax=stats.norm.pdf(1, loc=mu, scale=sigma), color='gre
plt.vlines(x=2, ymin=0, ymax=stats.norm.pdf(2, loc=mu, scale=sigma), color='red
plt.vlines(x=3, ymin=0, ymax=stats.norm.pdf(3, loc=mu, scale=sigma), color='bla
plt.yticks([])

plt.title('Normal Curve with 68-95-99.7 rule')
plt.show()
```

```
The area between -3 and 3 is considered to contains 99.73% of the area in the cu
The area between -2 and 2 is considered to contains 95.44% of the area in the cu
The area between -1 and 1 is considered to contains 68.26% of the area in the cu
The line at 0 is the expected average of the normal distribution at 50%

Mean:  -0.0004763483645039014
Median:  -0.0014382450140198932
Mode? -3.9625407395589707
Variance:  0.9973471956063097
Standard Deviation:  0.9986727169630247
```



Probability Density Function

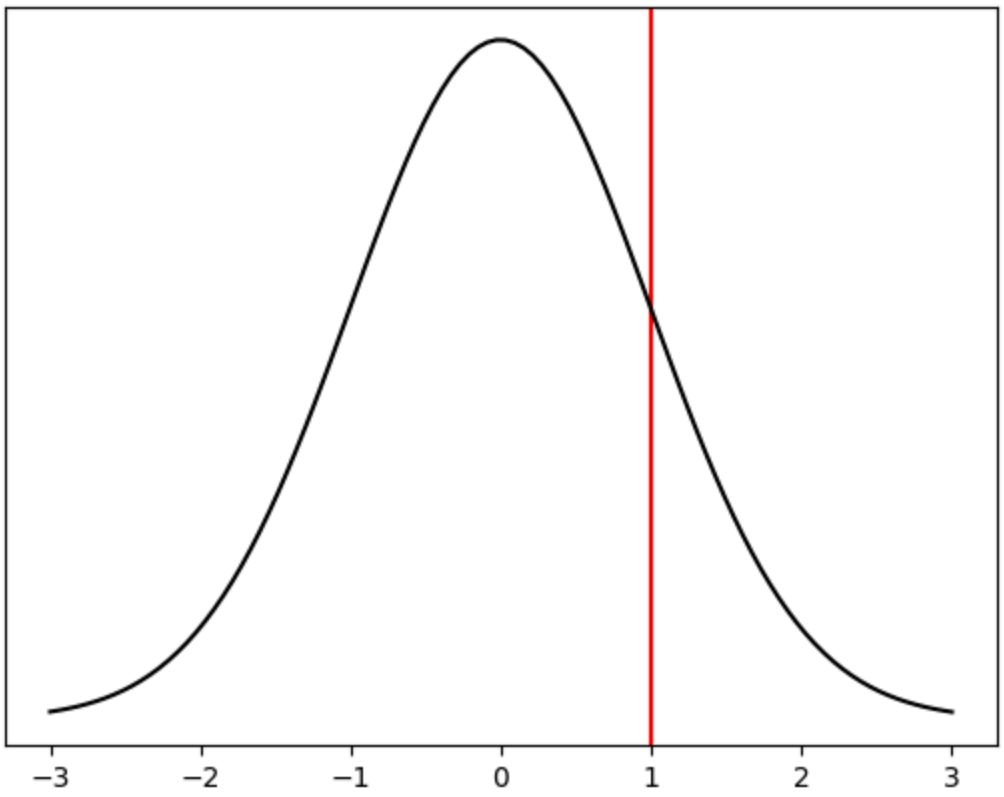
According to NIH SEMATECH (2022):

> For a continuous function, the probability density function (pdf) is the probability that the variate has the value x. Since for continuous distributions the probability at a single point is zero, this is often expressed in terms of an integral between two points.

Related Distributions. (February 13, 2022). In *NIST/SEMATECH e-Handbook of Statistical Methods*. https://www.itl.nist.gov/div898/handbook/eda/section3/eda362.htm

```
# demonstrate pdf
import numpy as np
import matplotlib.pyplot as plt
from scipy import stats

# pay attention to the use of the parameters inside the stats.norm.pdf functior
data = stats.norm.rvs(size=100000) # rvs = random variates

# get the parameter for a normal distribution
parameters = stats.norm.fit(data)
mu = parameters[0]
sigma = parameters[1] # sigma is std

# using unpacking
mu, sigma = parameters

print(f'mu = {mu:.2f}, sigma = {sigma:.2f}') # f strings https://zetcode.com/py

# plot the PDF.
xmin, xmax = plt.xlim() # https://matplotlib.org/3.3.3/api/_as_gen/matplotlib.p
x = np.linspace(-3, 3, 100)
params = stats.norm.pdf(x, loc=mu, scale=sigma)
plt.axvline(x=1, color='red')
plt.yticks([])
# we could also do stats.norm.pdf(x, loc=0, scale=1)

plt.plot(x, params, 'k'); # k is short for the color black
```

```
mu = -0.00, sigma = 1.00
```



Cumulative Density Function

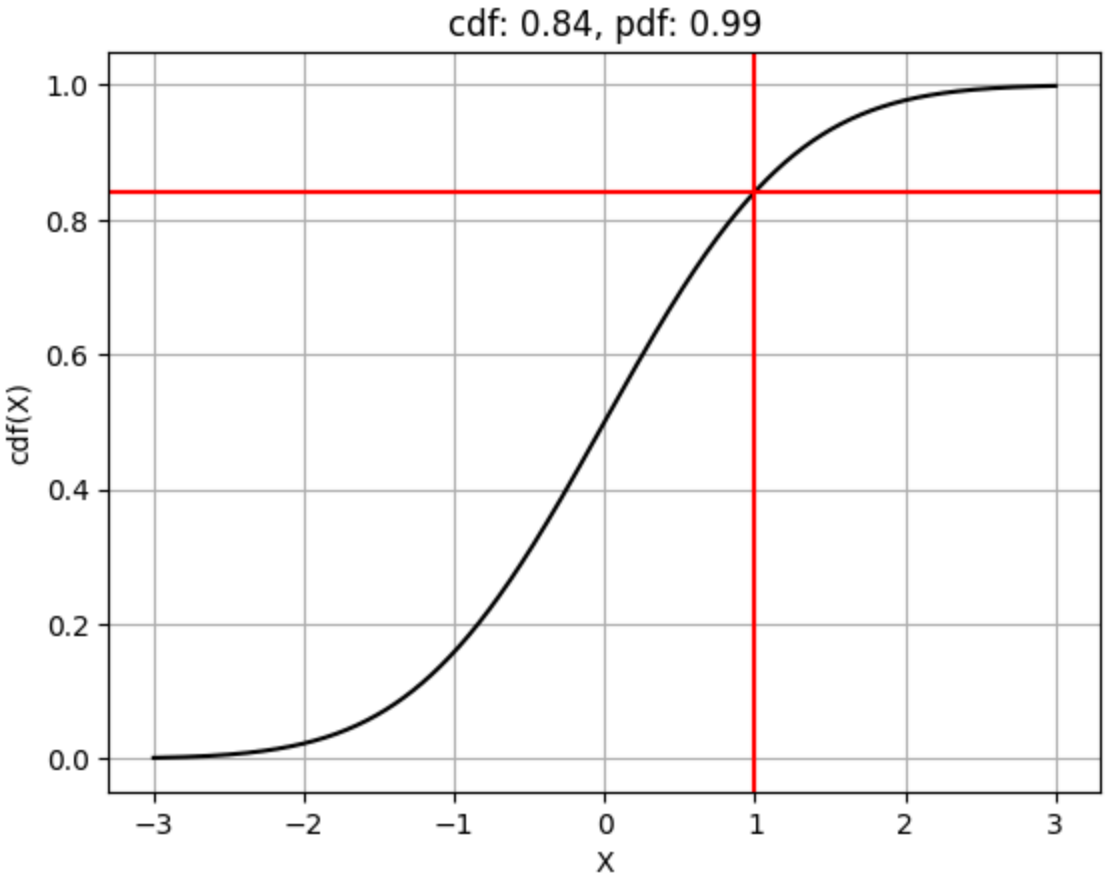
According to NIH SEMATECH (2022):

> The cumulative distribution function (cdf) is the probability that the variable takes a value less than or equal to x

Related Distributions. (February 13, 2022). In *NIST/SEMATECH e-Handbook of Statistical Methods*. https://www.itl.nist.gov/div898/handbook/eda/section3/eda362.htm

```
# plot the cdf with ppf
import numpy as np
import matplotlib.pyplot as plt
from scipy import stats

mu = 0
sigma = 1
cdf_val = np.round(stats.norm.cdf(1, loc=mu, scale=sigma), 2)
ppf_val = np.round(stats.norm.ppf(cdf_val, loc=mu, scale=sigma), 2)

x = np.linspace(-3, 3, 100000)
cdf_y = stats.norm.cdf(x)
plt.plot(x, cdf_y, 'k')
plt.axvline(x=ppf_val, color='red')
```

```
plt.axhline(y=cdf_val, color='red')
plt.title(f'cdf: {cdf_val}, pdf: {ppf_val}')
plt.xlabel('X')
plt.ylabel('cdf(X)')
plt.grid(True)
plt.show();
```



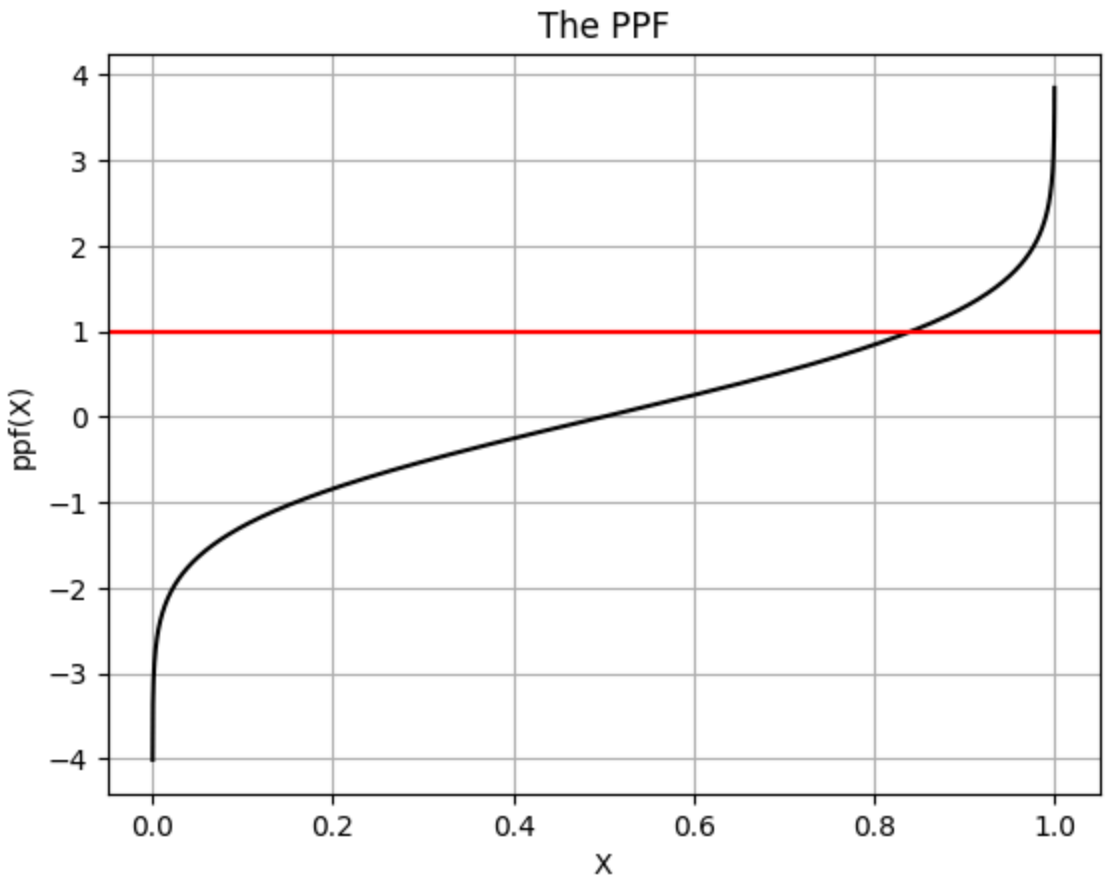
Percent Point Function

According to NIH SEMATECH (2022):

> The percent point function (ppf) is the inverse of the cumulative distribution function. For this reason, the percent point function is also commonly referred to as the inverse distribution function. That is, for a distribution function we calculate the probability that the variable is less than or equal to x for a given x. For the percent point function, we start with the probability and compute the corresponding x for the cumulative distribution.

Related Distributions. (February 13, 2022). In *NIST/SEMATECH e-Handbook of Statistical Methods*. https://www.itl.nist.gov/div898/handbook/eda/section3/eda362.htm

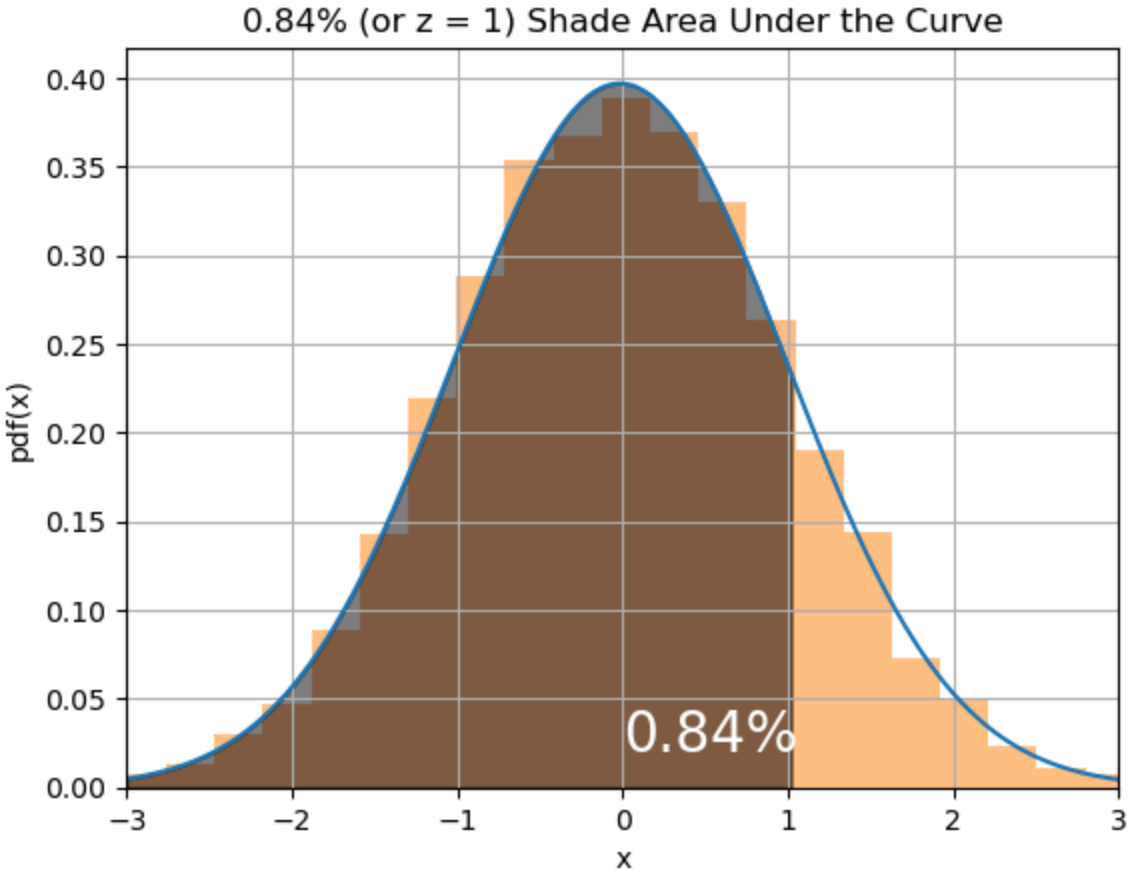
```
# plot the ppf
import numpy as np
import matplotlib.pyplot as plt
from scipy import stats

x = np.linspace(-3, 3, 100000)
ppf_y = stats.norm.ppf(x)
plt.plot(x, ppf_y, 'k', label='cdf')
plt.axhline(y=1, color='red')
plt.title('The PPF')
plt.xlabel('X')
plt.ylabel('ppf(X)')
plt.grid(True)
plt.show();
```




```
# compare CDF and PPF
X = stats.norm.rvs(size=10000)
X.sort()

parameters = stats.norm.fit(X)
mu, sigma = parameters
# print(stats.norm.cdf(1, loc=mu, scale=sigma))
cdf_val = np.round(stats.norm.cdf(1, loc=mu, scale=sigma), 2)
ppf_val = np.round(stats.norm.ppf(.85, loc=mu, scale=sigma), 2)
```

```
print(f'The corresponding percentile for 1 standard deviation above the mean (z
print(f'a z score for {cdf_val}  is {stats.norm.ppf(cdf_val, loc=mu, scale=sigm
print(f'a z score for 50% is {np.round(stats.norm.ppf(0.5, loc=mu, scale=sigma)
print('for normal distributions.')

fig, ax = plt.subplots()
ax.plot(X, stats.norm.pdf(X, loc=mu, scale=sigma))
ax.hist(X, bins=25, density=True, alpha=0.5)
ax.set_xlabel('x')
ax.set_ylabel('pdf(x)')
ax.set_xlim(-3, 3)
ax.grid(True)

px=np.arange(-3, ppf_val, 0.01)
ax.fill_between(px, stats.norm.pdf(px,loc=mu,scale=sigma), alpha=0.5, color='k'
ax.text(0, 0.02, f'{cdf_val}%', fontsize=20, color='w')

plt.title(f'{cdf_val}% (or z = 1) Shade Area Under the Curve')
plt.show()
```

```
The corresponding percentile for 1 standard deviation above the mean (z) is 0.84
a z score for 0.84  is 0.9832128822407085 and,
a z score for 50% is -0.02
for normal distributions.
```



Probability Mass Function

In probability and statistics, a probability mass function is a function that gives the probability that a discrete random variable is exactly equal to some value.[1] Sometimes it is also known as the discrete density function. The probability mass function is often the primary means of defining a discrete probability distribution, and such functions exist for either scalar or multivariate random variables whose domain is discrete.

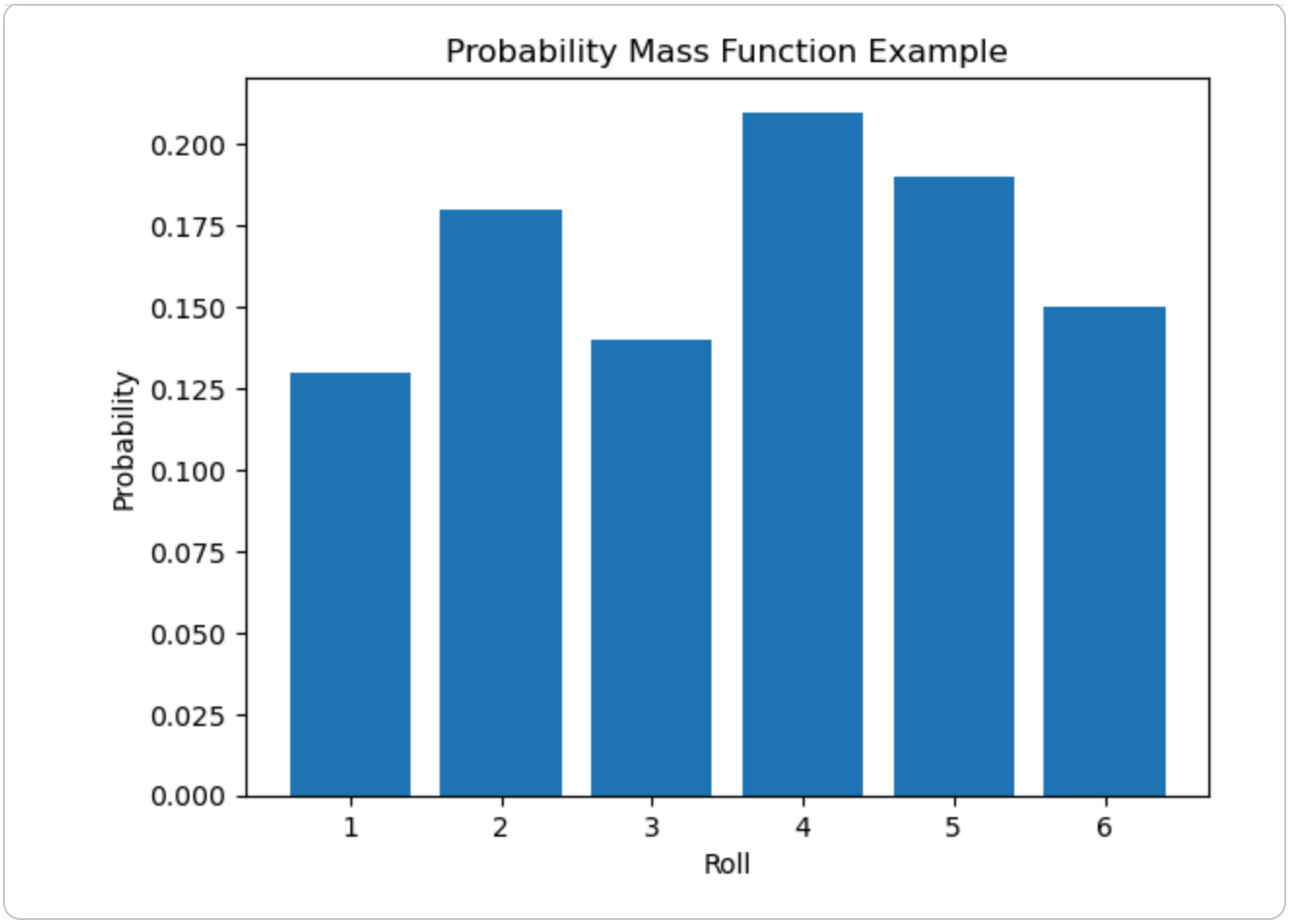
https://en.wikipedia.org/wiki/Probability_mass_function

Discrete probability functions are referred to as probability mass functions and continuous probability functions are referred to as probability density functions. The term probability functions covers both discrete and continuous distributions.

https://www.itl.nist.gov/div898/handbook/eda/section3/eda361.htm

Bernoulli, Binomial, Geometric, Poisson - flip of coin, roll of 1 die or two dice

```
# probability mass function
throws = 100
observations = []
for i in range(throws):
    roll = np.random.choice(['1', '2', '3', '4', '5', '6']) # roll the die
    observations.append(roll)

val, cnt = np.unique(observations, return_counts=True)
prop = cnt / len(observations)

plt.bar(val, prop)
plt.ylabel('Probability')
plt.xlabel('Roll')
plt.title('Probability Mass Function Example')
plt.show()
```



Kernel Density Estimation

According to Wikipedia (2022):

> In statistics, kernel density estimation (KDE) is a non-parametric way to estimate the probability density function of a random variable. Kernel density estimation is a fundamental data smoothing problem where inferences about the population are made, based on a finite data sample.

Kernel density estimation. (February 13, 2022) In *Wikipedia*. https://en.wikipedia.org/wiki/Kernel_density_estimation

As mentioned, KDE is a non-parametric estimate compared to the parametric probability density function. Non-parametric data do not fit an established distribution such as the normal, or gaussian, distribution. Non-parametric models don't have parameters to define the data distribution so in the case of KDE, the shape of the data is estimated.

According to Wikipedia (2022):

> In nonparametric statistics, a kernel is a weighting function used in non-parametric estimation techniques. Kernels are used in kernel density

> estimation to estimate random variables' density functions, or in kernel regression to estimate the conditional expectation of a random variable.

Nonparametric statistics. (February 13, 2022) In *Wikipedia*.https://en.wikipedia.org/wiki/Kernel_(statistics)#Nonparametric_statistics

Here is a simplified version of the equation provided by Wikipedia:

$$\hat{f(x)} = \frac{1}{N} \sum K(x - x_i)$$

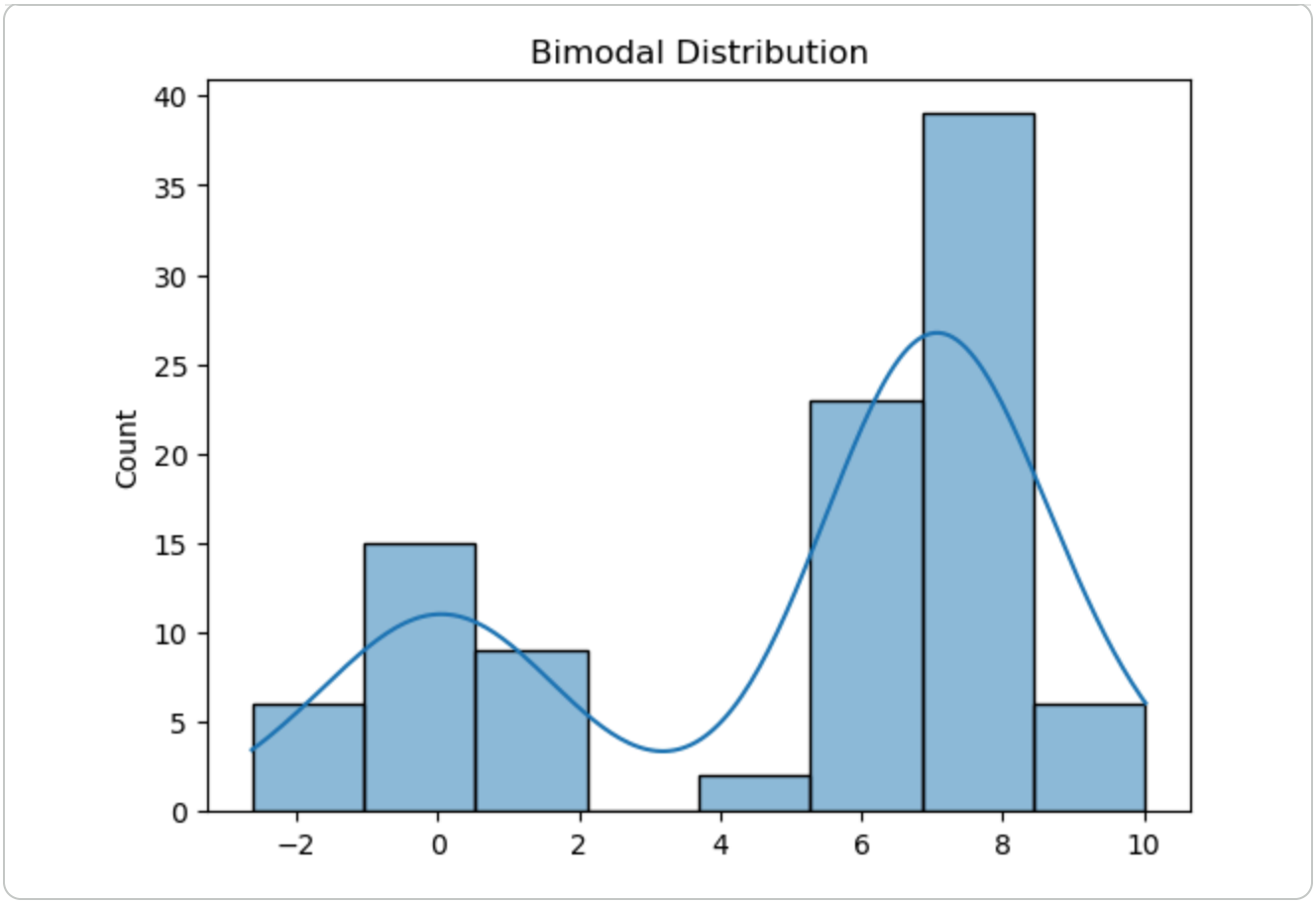
Let (x1, x2, ..., xn) be independent and identically distributed [IID] samples drawn from some univariate distribution with an unknown density f at any given point x. We are interested in estimating the shape of this function f.

https://en.wikipedia.org/wiki/Kernel_density_estimation#Definition

Recall **IID** definition: In probability theory and statistics, a collection of random variables is independent and identically distributed if each random variable has the same probability distribution as the others and all are mutually independent.

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

```
import seaborn as sns
from scipy.stats.distributions import norm

x = np.random.randn(100)
x[int(0.3 * 100):] += 7
sns.histplot(x, kde=True)
plt.title('Bimodal Distribution')
plt.show()
```



Distributions

Discrete distributions: all probabilities add up to one

- Uniform
- Bernoulli
- Binomial
- Poisson

Continuous distributions: area under curve adds up to one

- Normal
- z
- t
- Chi-Square

From Bernoulli, to Binomial, to Normal

1. The Bernoulli Distribution (The Foundation):

The Bernoulli distribution is the simplest of them all. It describes the probability of a single experiment (or trial) that has only two possible outcomes, often called "success"

and "failure." Think of flipping a coin once: heads or tails.

- **Key Characteristics:**
  - Two possible outcomes (success with probability *p*, failure with probability 1-*p*).
  - It's the building block for more complex distributions.
- **Historical Context:** While the formalization of probability theory came later, the concept of dichotomous outcomes was understood much earlier. The name "Bernoulli" is associated with Jacob Bernoulli (1654-1705), who made significant contributions to probability, including work on this distribution, though it wasn't formally named after him until much later. His work laid the groundwork for many subsequent developments.

2. The Binomial Distribution (Multiple Trials):

The binomial distribution arises when you repeat a Bernoulli experiment multiple times (say, *n* times). It answers the question: What's the probability of getting exactly *k* successes in *n* independent trials? Think of flipping a coin 10 times and asking, "What's the probability of getting exactly 7 heads?"

- **Key Characteristics:**
  - *n* independent Bernoulli trials.
  - Probability of success *p* is the same for each trial.
  - Counts the number of successes (*k*) in those *n* trials.
- **Evolution from Bernoulli:** The binomial distribution is a direct generalization of the Bernoulli. If *n*=1 (you only do the trial once), the binomial *becomes* the Bernoulli. The formula for the binomial probability is based on combinations (how many ways can you choose *k* successes out of *n* trials) and the probabilities of success and failure.
- **Historical Context:** While Bernoulli contributed to the understanding, the binomial distribution was further developed by mathematicians like Abraham de Moivre (1667-1754). De Moivre, in particular, investigated the *approximation* of the binomial distribution, which is a crucial stepping stone to the normal distribution.

3. The Normal Distribution (The Limit):

The normal distribution (also known as the Gaussian distribution) is a continuous probability distribution that is symmetrical and bell-shaped. It's ubiquitous in statistics because it arises as the *limiting* distribution of many other distributions, especially the binomial distribution, as the number of trials (*n*) gets very large. Think of flipping a coin

thousands of times: the distribution of the number of heads will be very close to a normal distribution.

- **Key Characteristics:**
  - Bell-shaped, symmetrical curve.
  - Defined by its mean (μ) and standard deviation (σ).
  - Continuous (outcomes can take any value within a range).
- **Evolution from Binomial:** As *n* increases in the binomial distribution, the shape of the distribution starts to resemble a bell curve. De Moivre showed that the binomial distribution could be *approximated* by a normal distribution when *n* is large enough and *p* is not too close to 0 or 1. This approximation became incredibly important because the normal distribution is much easier to work with mathematically than the binomial distribution when *n* is very large.
- **Historical Context:** Carl Friedrich Gauss (1777-1855) is most famously associated with the normal distribution, although it was also studied by others like Laplace. Gauss used it extensively in his work on astronomy and geodesy. The term "normal" came to be used because it was thought to be the most common or "normal" distribution found in natural phenomena.

In Summary:

The Bernoulli distribution is the seed. The binomial distribution is what you get when you plant that seed many times (repeating the Bernoulli trial). The normal distribution is what you get when your "garden" (number of trials) becomes very large – it's the shape that emerges as the binomial distribution smooths out into a continuous curve. This progression is a cornerstone of probability theory and has immense implications in statistics, allowing us to model and understand a wide range of real-world phenomena.

Take a look at: https://shiny.rit.albany.edu/stat/binomial/

Word Problems

Binomial Distribution (PMF) Word Problem

Problem: A factory produces computer chips, and historically, **1.5%** of the chips are defective. If a quality control inspector randomly selects a batch of **200** chips, what is the probability that **exactly 5** chips in the batch are defective?

Concept: This is a **Binomial Distribution** problem because there is a fixed number of independent trials ($n = 200$), and each trial has only two outcomes (defective or non-

defective) with a constant probability of success ($p = 0.015$). We are looking for the probability of a single, exact number of successes ($k = 5$). The solution uses the **Probability Mass Function (PMF)**.

```
from scipy.stats import binom

# Parameters
n = 200     # number of chips (trials)
k = 5       # number of defective chips (successes)
p = 0.015   # probability of a chip being defective (success probability)

# Calculate the probability mass function (PMF) P(X = 5)
probability = binom.pmf(k, n, p)

print(f"The probability of exactly 5 defective chips is: {probability:.4f}")
```

Normal Distribution (PDF) Word Problem

Problem: The scores on a standardized test are normally distributed with a **mean (mu) of 500** and a **standard deviation ($sigma$) of 100**. What is the probability density associated with a score of exactly **650**?

Concept: This uses the **Probability Density Function (PDF)** for the continuous Normal distribution. Remember, the PDF value at a single point is not a probability; it is a measure of the relative likelihood or density at that point on the curve.

```
from scipy.stats import norm

# Parameters
mu = 500    # mean
sigma = 100 # standard deviation
x = 650     # specific score

# Calculate the Probability Density Function (PDF) value f(650)
density = norm.pdf(x, loc=mu, scale=sigma)

print(f"The probability density at a score of 650 is: {density:.6f}")
```

Normal Distribution (CDF) Word Problem

Problem: A light bulb company states that the lifespan of their bulbs is normally distributed with a **mean (mu) of 800 hours** and a **standard deviation ($sigma$) of 50 hours**. What is the probability that a randomly selected light bulb will last **less than or equal to 750 hours**?

Concept: This uses the **Cumulative Distribution Function (CDF)**, which calculates the cumulative probability of a continuous variable falling below a certain value $P(X \leq x)$.

```
from scipy.stats import norm

# Parameters
mu = 800    # mean lifespan
sigma = 50  # standard deviation
x = 750     # target hour

# Calculate the Cumulative Distribution Function (CDF) P(X <= 750)
probability = norm.cdf(x, loc=mu, scale=sigma)

print(f"The probability a bulb lasts 750 hours or less is: {probability:.4f}")
```

Normal Distribution (PPF) Word Problem

Problem: The time it takes for a customer service agent to resolve an issue is normally distributed with a **mean (mu) of 15 minutes** and a **standard deviation ($sigma$) of 3 minutes**. The company wants to set a service target so that **90%** of all customer issues are resolved within that time. What is the target resolution time?

Concept: This requires the **Percent-Point Function (PPF)**, which is the inverse of the CDF. Given a cumulative probability (the percentile), it returns the corresponding value of the random variable.

```
from scipy.stats import norm

# Parameters
mu = 15     # mean resolution time
sigma = 3   # standard deviation
percentile = 0.90 # 90th percentile (90% probability)

# Calculate the Percent-Point Function (PPF) - the value 'x' where P(X <= x) =
target_time = norm.ppf(percentile, loc=mu, scale=sigma)

print(f"The target resolution time (90th percentile) is: {target_time:.2f} minu
```

Hypothesis Testing Word Problem (One-Sample Z-Test)

A company that manufactures batteries claims their standard AA batteries have a mean lifespan of **20 hours**. A consumer advocacy group suspects this claim is too high. They know from past testing that the population standard deviation ($sigma$) is **1.5 hours**. The

group decides to test the company's claim by taking a random sample of **30** batteries and measuring their lifespan. The sample yields an average lifespan ($barx$) of **19.5 hours**.