Probability Theory Review

Roi Yehoshua

Northeastern University

Agenda

- Axioms of probability
- Random variables
- Common probability distributions
- Conditional probabilities and Bayes’ rule
- Joint and marginal probability distributions
- Covariance and correlation
- Random vectors and multivariate distributions
- Limit theorems
- Maximum likelihood estimation

Probability Basics

- Consider an experiment that can result in several possible outcomes
- The **sample space** is the set containing all these possible outcomes
  - e.g., if the experiment is tossing a die, the sample space is $\Omega = \{1, 2, 3, 4, 5, 6\}$
- An **event** *A* is any subset of the sample space
  - e.g., the event that the die shows an even number is

$$A = \{x \in \Omega \mid x \text{ is even}\} = \{2, 4, 6\}$$

- The **probability** of an event, denoted $P(A)$, quantifies the uncertainty of its occurrence on a scale from 0 to 1
- Two approaches for defining probabilities: frequentist and Bayesian

The Frequentist Approach

- Defines probabilities in terms of **long-run relative frequencies**
- Suppose an experiment is repeated *n* times under the same conditions
- Let $n(A)$ denote the number of times the event A occurs
- Then the probability of A is defined as the frequency of the event in the limit:

$$P(A) = \lim_{n \to \infty} \frac{n(A)}{n}$$

- For example, the probability of rolling any number i with a fair die is:

$$P(i) = \frac{1}{6}, \quad \text{for all } i \in \{1, 2, 3, 4, 5, 6\}$$

The Bayesian Approach

- Defines probability as a **degree of belief** or **subjective certainty** about the occurrence of an event
- Does not rely on repeated trials, thus can be applied to singular events
- Probabilities are updated as new information becomes available
- Based on Bayes' theorem, which relates prior belief, likelihood, and posterior belief

Likelihood   Prior probability

Posterior probability $$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$ Normalizing constant

- For example, in a medical setting, we might assign a prior probability to a disease based on population data, then update it after observing test results

Axioms of Probability

1. The probability of an event is nonnegative:

$$P(A) \geq 0$$

2. The probability of the entire sample space is 1:

$$P(\Omega) = 1$$

3. For any sequence of mutually exclusive events $A_1, A_2, \ldots, A_n$ the probability of at least one of these events occurring is the sum of their probabilities:

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

For example, from Axiom 3, it follows that the probability of rolling an even number is

$$P(\{2,4,6\}) = P(\{2\}) + P(\{4\}) + P(\{6\}) = \frac{1}{2}$$

Propositions Derived from the Axioms

- Probability of an empty set: $P(\emptyset) = 0$
- Probability bounds: For any event A

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

- Monotonicity: For any two events $A \subseteq B$

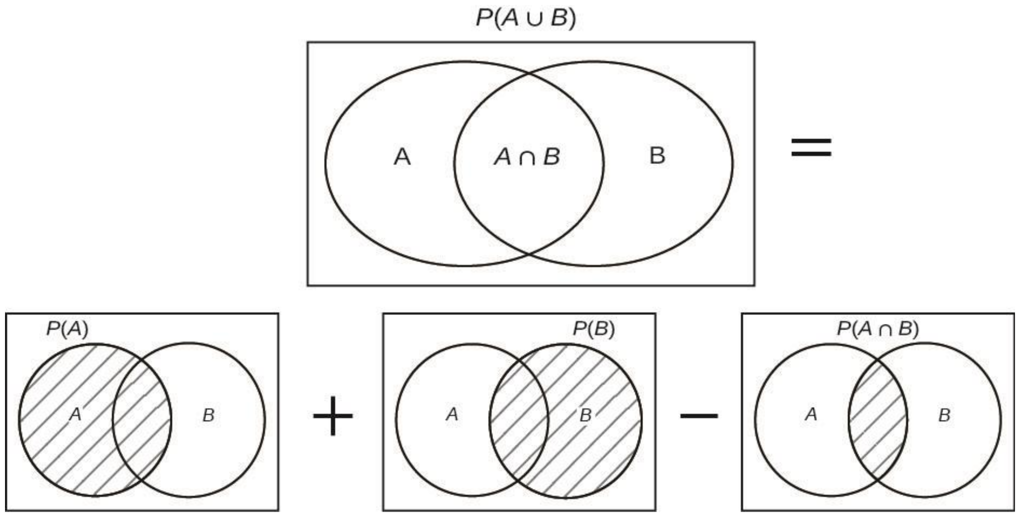
$$P(A) \leq P(B)$$

- Complement rule:
  - The **complement** of an event A, denoted by A^c or A', is the subset of outcomes in the sample space that are not in the event A
  - The probability of A^c is: $P(A^C) = 1 - P(A)$

The Addition Rule

- For any two events *A* and *B*

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



- Example: what is the probability of drawing a card that is either a heart or a face card from a deck of 52 cards?

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) = \frac{13}{52} + \frac{12}{52} - \frac{3}{52} = \frac{22}{52} = \frac{11}{26}$$

Inclusion-Exclusion Principle

- Generalizes the addition rule to any number of events

$$P\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} P(A_i) - \sum_{1 \le i < j \le n} P(A_i \cap A_j) + \sum_{1 \le i < j < k \le n} P(A_i \cap A_j \cap A_k)$$
$$- \cdots + (-1)^{n+1} P(A_1 \cap \cdots \cap A_n)$$

- For example, for three events we get:

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

Random Variables

- We are often interested in a numerical quantity based on the experiment outcome
- A **random variable** is a function that assigns a real number to each possible outcome

$$X : \Omega \to \mathbb{R}$$

- The range of *X* is the set of all real values it can take:

$$\mathcal{R}_X = \{x \in \mathbb{R} \mid \exists \omega \in \Omega \text{ such that } X(\omega) = x\}$$

- For example, if we flip a fair coin 3 times and define *X* as the number of observed heads, its range is:

$$\mathcal{R}_X = \{0, 1, 2, 3\}$$

Probability Distributions

- A **probability distribution** of random variable *X*, denoted *P*(*X*), is a function that assigns a probability to each value it can take
- For example, if *X* is the number of heads in a 3 fair coin flips:

x	*P*(*X* = *x*)
0	1/8
1	3/8
2	3/8
3	1/8

$$\Omega = \{\mathrm{HHH}, \mathrm{HHT}, \mathrm{HTH}, \mathrm{THH}, \mathrm{HTT}, \mathrm{THT}, \mathrm{TTH}, \mathrm{TTT}\}$$

Discrete Random Variables

- A **discrete random variable** *X* can take only a finite number of possible values
  - or a countably infinite set of values
- A **probability mass function (PMF)** assigns a probability to each such value:

$$p_X(x) = P(X = x) = P(\{\omega \in \Omega \mid X(\omega) = x\})$$

- For a function to be a valid PMF, it must satisfy:
  - Each probability must lie between 0 and 1

$$0 \leq p_X(x) \leq 1, \quad \text{for all } x$$

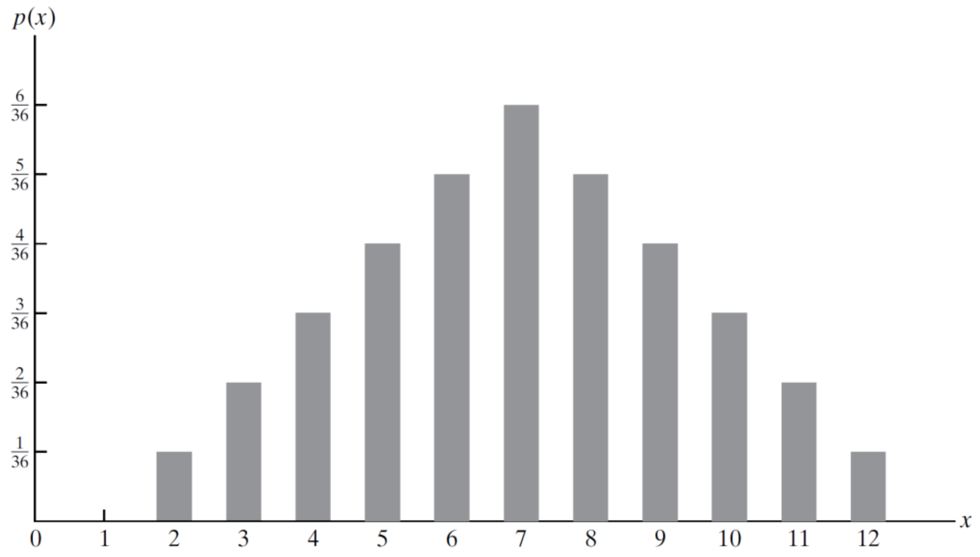
  - The total probability across all possible values of *X* must sum to 1:

$$\sum_{x \in \mathcal{R}_X} p_X(x) = 1$$

Discrete Random Variables

- For example, a PMF of a random variable representing the sum of two dice:

$$p_X(x) = \begin{cases} 1/36, & x = 2, \\ 2/36, & x = 3, \\ 3/36, & x = 4, \\ 4/36, & x = 5, \\ 5/36, & x = 6, \\ 6/36, & x = 7, \\ 5/36, & x = 8, \\ 4/36, & x = 9, \\ 3/36, & x = 10, \\ 2/36, & x = 11, \\ 1/36, & x = 12, \\ 0, & \text{otherwise.} \end{cases}$$

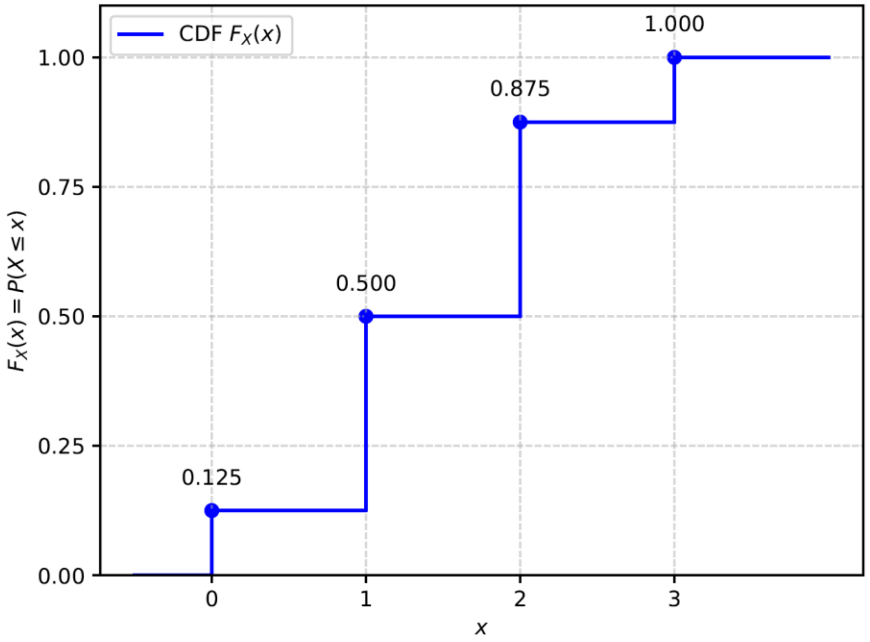


Cumulative Distribution Function

- A **cumulative distribution function** (CDF) of a random variable *X* gives the probability that X takes a values less or equal to a given number

$$F_X(x) = P(X \leq x) \quad \text{for all } x \in \mathbb{R}$$

- For example, the CDF of the number of heads in 3 fair coin flips:



Expectation

- The **expected value** (or **mean**) of *X* represents its average outcome over many trials
- Let *X* be a discrete RV with a finite set of values $x_1, x_2, \ldots, x_n$ and a PMF $p(x)$
- The expected value of *X* is a weighted average of its values:

$$\mathbb{E}[X] = \sum_{i=1}^{n} x_i p(x_i)$$

- For example, if *X* is the outcome of a fair die roll:

$$\mathbb{E}[X] = \sum_{i=1}^{6} iP(X = i) = \sum_{i=1}^{6} i \cdot \frac{1}{6} = \frac{1}{6}(1+2+3+4+5+6) = \frac{21}{6} = 3.5$$

Properties of Expectation

- Expectation of a constant $\mathbb{E}[c] = c$
- Scalar multiplication $\mathbb{E}[aX] = a\,\mathbb{E}[X]$
- Additivity $\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$
- Linearity of expectation:
  - For random variables $X_1, X_2, \ldots, X_n$ and constants $c_1, c_2, \ldots, c_n$:

$$\mathbb{E}\left[\sum_{i=1}^{n} c_i X_i\right] = \sum_{i=1}^{n} c_i \mathbb{E}[X_i]$$

Functions of Random Variables

- Often, we are interested in some function of the random variable
- Let X be a random variable and $Y = g(X)$ for some function g
  - Y is also a random variable
  - The PMF of Y is given by:

$$p_Y(y) = P(g(X) = y) = \sum_{\substack{x \in \mathcal{R}_X \\ g(x) = y}} p_X(x)$$

  - The expected value of Y is:

$$\mathbb{E}[Y] = \sum_{y \in \mathcal{R}_Y} y \cdot p_Y(y) = \sum_{x \in \mathcal{R}_X} g(x) \cdot p_X(x)$$

    - Known as the **Law of the Unconscious Statistician (LOTUS)**

Variance

- **Variance** is the expected squared deviation from the mean:

$$\mathrm{Var}(X) = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right] = \mathbb{E}[X^2] - \mathbb{E}[X]^2$$

- For example, if *X* is the outcome of a fair die, then

$$\mathbb{E}[X^2] = \sum_{i=1}^{6} i^2 \cdot \frac{1}{6} = \frac{1}{6}(1^2 + 2^2 + 3^2 + 4^2 + 5^2 + 6^2) = \frac{91}{6}$$

$$\mathrm{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \frac{91}{6} - \left(\frac{7}{2}\right)^2 = \frac{35}{12} = 2.917$$

- The square root of the variance is called **standard deviation**

$$\sigma_X = \sqrt{\mathrm{Var}(X)}$$

  - In the die example: $\sigma_X = \sqrt{\frac{35}{12}} \approx 1.71$

Properties of Variance

- Variance of a constant: $\mathrm{Var}(c) = 0$
- Scaling property: for any constants *a* and *b*:

$$\mathrm{Var}(aX + b) = a^2 \mathrm{Var}(X)$$

- Variance of the sum of two variables:

$$\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\,\mathrm{Cov}(X, Y)$$

- Variance of the sum of two independent variables:

$$\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$$

Bernoulli Distribution

- Consider an experiment that can either succeed or fail
  - The probability of success is given by *p* ($0 \leq p \leq 1$)
- Let *X* be a binary random variable defined as:

$$X = \begin{cases} 1, & \text{if the outcome is a success,} \\ 0, & \text{if the outcome is a failure.} \end{cases}$$

- Then, *X* follows a Bernoulli distribution with parameter *p* $X \sim \text{Bernoulli}(p)$
- PMF of *X*:

$$p(x) = \begin{cases} p & \text{if } x = 1 \\ 1 - p & \text{if } x = 0 \end{cases}$$

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

- Expected value: $\mathbb{E}[X] = p$
- Variance: $\text{Var}(X) = p(1-p)$

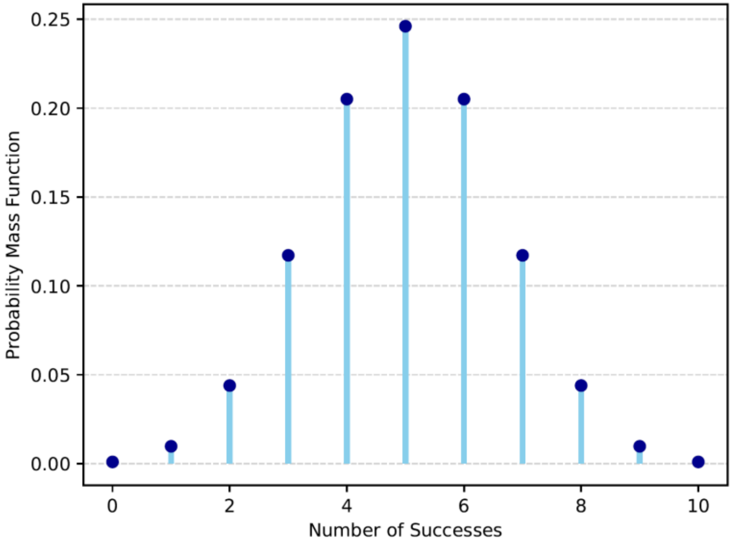
Binomial Distribution

- Models the number of successes in a sequence of *n* independent Bernoulli trials
  - each with a probability of success *p*
- A random variable $X \sim \text{Binomial}(n, p)$ counts the number of successes in *n* trials
- PMF of *X*:

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

  - e.g., let *X* be the number of heads in 10 fair coin flips

$$P(X = 5) = \binom{10}{5} (0.5)^5 (1-0.5)^5 = \frac{10!}{5!5!} \cdot (0.5)^{10} \approx 0.246$$



- Expected value: $\mathbb{E}[X] = np$
- Variance: $\text{Var}(X) = np(1-p)$

Geometric Distribution

- Models the number of Bernoulli trials needed to obtain the first success
- A random variable $X \sim$ Geometric(p) represents the trial with the first success
- PMF of X:

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

  - e.g., if X denotes the number of coin flips until the first heads appears

$$P(X = k) = (0.5)^{k-1} \cdot 0.5 = 0.5^k$$

- Expected value: $$\mathbb{E}[X] = \frac{1}{p}$$
- Variance: $$\mathrm{Var}(X) = \frac{1 - p}{p^2}$$

Poisson Distribution

- Models the number of events occurring randomly in a fixed interval of time
  - A parameter λ represents the expected number of events in that interval
- A **Poisson random variable** counts the number of events in that interval

$$X \sim \text{Poisson}(\lambda)$$

- PMF of X: $$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}, \quad k = 0, 1, 2, \ldots$$
- Expected value and variance: $\mathbb{E}[X] = \lambda, \quad \text{Var}(X) = \lambda$
- Suppose you are waiting at a bus stop where buses arrive at an average rate of 4 buses per hour. What is the probability that exactly 3 buses arrive in 30 minutes?
  - If *X* is the number of buses arriving in 30 minutes, then $X \sim \text{Poisson}(2)$

$$P(X = 3) = \frac{2^3 e^{-2}}{3!} \approx 0.18$$

Categorical Distribution

- Generalizes the Bernoulli distribution to *k* possible outcomes (categories)
  - Each category *i* occurs with probability p_i
  - The probabilities sum to 1 $$\sum_{i=1}^{k} p_i = 1$$
- Let *X* be a categorical random variable $X \sim \text{Categorical}(p_1, \dots, p_k)$
- PMF of *X*:

$$P(X = i) = p_i, \quad i \in \{1, 2, \dots, k\}$$

- Example: consider a chatbot that predicts the next word in a sentence from a vocabulary of *n* words
  - Define a random variable *X* representing the next word
  - *X* follows a categorical distribution over the *n* possible words
  - p_i is the likelihood of choosing word *i* in the vocabulary

Common Discrete Probability Distributions

X	*X Counts*	*p(x)*	*Values of X*	*E(x)*	*V(x)*
Discrete uniform	Outcomes that are equally likely (finite)	$\frac{1}{b-a+1}$	$a \le x \le b$	$\frac{b+a}{2}$	$\frac{(b-a+2)(b-a)}{12}$
Binomial	Number of sucesses in n fixed trials	$\binom{n}{x} p^x (1-p)^{n-x}$	$x = 0, 1, \ldots, n$	np	$np(1-p)$
Poisson	Number of arrivals in a fixed time period	$\frac{e^{-\lambda}\lambda^x}{x!}$	$x = 0, 1, 2, \ldots$	λ	λ
Geometric	Number of trials up through 1st success	$(1-p)^{x-1}p$	$x = 1, 2, 3, \ldots$	$\frac{1}{p}$	$\frac{1-p}{p^2}$
Negative Binomial	Number of trials up through kth success	$\binom{x-1}{k-1}(1-p)^{x-k}p^k$	$x = k, k+1, \ldots$	$\frac{k}{p}$	$\frac{k(1-p)}{p^2}$
Hyper - geometric	Number of marked individuals in sample taken without replacement	$\frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}}$	$\max(0, M+n-N) \le x \le \min(M, n)$	$n * \frac{M}{N}$	$\frac{nM(N-M)(N-n)}{N^2(N-1)}$

Continuous Random Variables

- A **continuous random variable** *X* can take an infinite number of values
- The probability of observing any exact value of *X* is zero

$$P(X = x) = 0, \quad \text{for all } x \in \mathbb{R}$$

- Therefore, we consider the probability that *X* falls within an interval

$$P(a \leq X \leq b)$$

Probability Density Function

- **Probability density function (PDF)** of *X* is a nonnegative function *f*(x) such that the probability that *X* lies within an interval [*a*, *b*] is given by the integral:

$$P(a \leq X \leq b) = \int_a^b f(x)dx$$

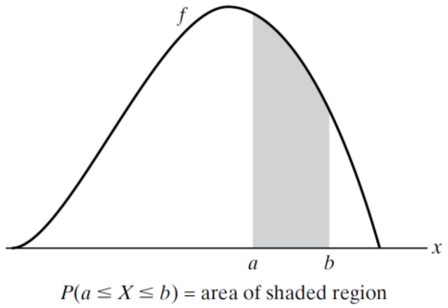


FIGURE 5.1: Probability density function *f*.

- For a function *f* to be a valid PDF it needs to satisfy:
  - It must be nonnegative for all real numbers

$$f(x) \geq 0, \quad \text{for all } x \in \mathbb{R}$$

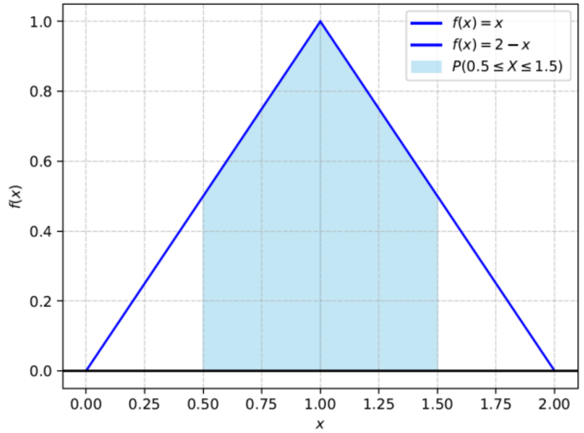
  - The total probability over the entire real line must equal 1:

$$\int_{-\infty}^{\infty} f(x)dx = 1$$

Probability Density Function

- For example, consider the function:

$$f(x) = \begin{cases} x & 0 \leq x \leq 1 \\ 2 - x & 1 < x \leq 2 \\ 0 & \text{otherwise} \end{cases}$$



- To verify that it is valid PDF, we check that it integrates to 1:

$$\int_0^1 x\,dx + \int_1^2 (2-x)\,dx = \left[\frac{x^2}{2}\right]_0^1 + \left[2x - \frac{x^2}{2}\right]_1^2 = \frac{1}{2} + \frac{1}{2} = 1$$

- The probability that *X* falls within the interval [0.5, 1.5] is:

$$\begin{aligned} P(0.5 \leq X \leq 1.5) &= \int_{0.5}^1 x\,dx + \int_1^{1.5} (2-x)\,dx \\ &= \left[\frac{x^2}{2}\right]_{0.5}^1 + \left[2x - \frac{x^2}{2}\right]_1^{1.5} = \left(\frac{1}{2} - \frac{(0.5)^2}{2}\right) + \left(3 - \frac{2.25}{2} - 2 + \frac{1}{2}\right) \\ &= 0.375 + 0.375 = 0.75 \end{aligned}$$

Cumulative Distribution Function

- The cumulative distribution function (CDF) of a continuous variable X is:

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

- If the CDF is differentiable at a point *x*, then the PDF is its derivative at that point:

$$f_X(x) = \frac{d}{dx} F_X(x)$$

Expectation

- If X is a continuous random variable with PDF $f(x)$, then its expected value is:

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x f(x) dx$$

- For example, given the PDF

$$f(x) = \begin{cases} x, & 0 \le x \le 1, \\ 2 - x, & 1 < x \le 2, \\ 0, & \text{otherwise} \end{cases}$$

- The expected value is:

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x f(x)\, dx = \int_0^1 x^2\, dx + \int_1^2 x(2-x)\, dx$$

$$= \left[\frac{x^3}{3}\right]_0^1 + \left[x^2 - \frac{x^3}{3}\right]_1^2 = \frac{1}{3} + \left(\frac{4}{3} - \frac{2}{3}\right) = 1$$

Variance

- The variance is defined as in the discrete case

$$\mathrm{Var}(X) = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right] = \mathbb{E}[X^2] - \mathbb{E}[X]^2$$

- In our example,

$$\mathbb{E}[X^2] = \int_{-\infty}^{\infty} x^2 f(x)\,dx = \int_0^1 x^3\,dx + \int_1^2 x^2(2-x)\,dx$$

$$= \left[\frac{x^4}{4}\right]_0^1 + \left[\frac{2x^3}{3} - \frac{x^4}{4}\right]_1^2 = \frac{1}{4} + \left(\frac{4}{3} - \frac{5}{12}\right) = \frac{7}{6}$$

$$\mathrm{Var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 = \frac{7}{6} - 1^2 = \frac{1}{6}$$

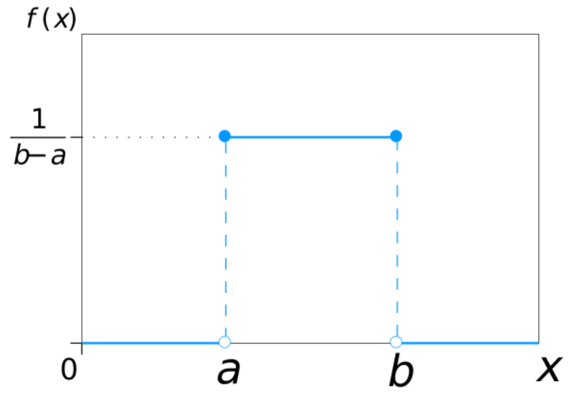
Continuous Uniform Distribution

- A **continuous uniform** variable *X* has a constant density over an interval [*a*, *b*]

$$X \sim \mathcal{U}(a, b)$$

- PDF of *X*:

$$f(x) = \begin{cases} \dfrac{1}{b-a} & \text{if } a \leq x \leq b \\ 0 & \text{otherwise} \end{cases}$$



- Expected value:

$$\mathbb{E}[X] = \frac{a+b}{2}$$

- Variance:

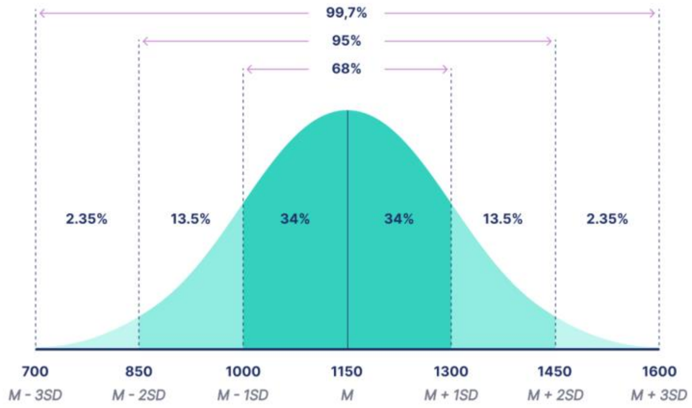
$$\text{Var}(X) = \frac{b-a}{12}$$

Normal (Gaussian) Distribution

- One of the most fundamental and widespread probability distributions
  - Arises naturally in many situations due to the Central Limit Theorem (CLT)
- X is a **normal random variable** with mean μ and variance σ^2, written $X \sim N(\mu, \sigma^2)$, if its PDF is:

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

- The density function is a bell-shaped curve symmetric about μ
- The **68-95-99.7 rule**:
  - 68% of the values lie within one standard deviation of the mean
  - 95% lie within two standard deviations of the mean
  - 99.7% lie within three standard deviations of the mean



Properties of the Normal Distribution

- Linear transformation: If $X \sim N(\mu, \sigma^2)$, then for any constants a and b

$$Y = aX + b \sim \mathcal{N}(a\mu + b, a^2\sigma^2)$$

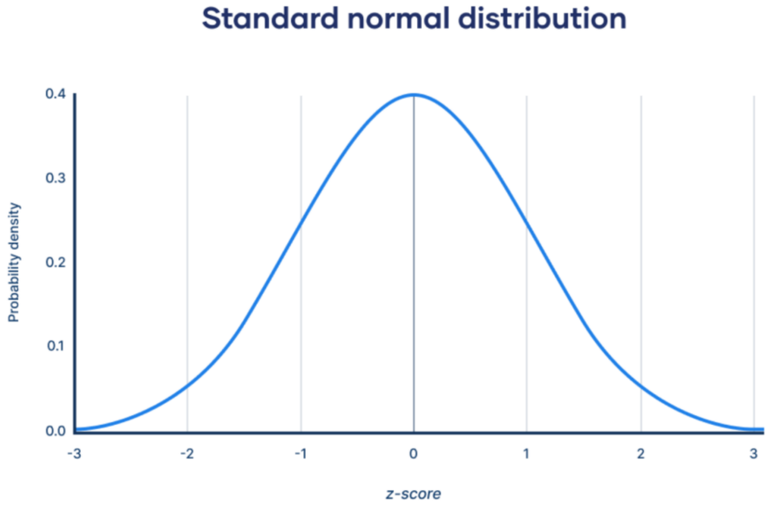
- Sum of independent normal variables: If $X_1, \ldots, X_n$ are independent normal variables with means μ_i and standard deviations σ_i, then their sum is also normally distributed

$$\sum_{i=1}^{n} X_i \sim \mathcal{N}\left(\sum_{i=1}^{n} \mu_i, \sum_{i=1}^{n} \sigma_i^2\right)$$

Standard Normal Distribution

- A **standard normal variable** $Z \sim N(0, 1)$ is a normal variable with $\mu = 0$ and $\sigma = 1$
- Its PDF is:

$$f_Z(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right)$$



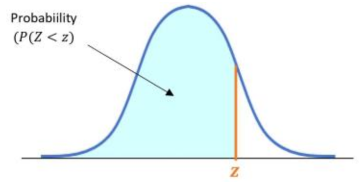
- If X is normally distributed with μ, σ^2 then the following variable is standard normal

$$Z = \frac{X - \mu}{\sigma}$$

Standard Normal Distribution

- To compute probabilities, we use the CDF of a standard normal random variable is:

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \exp\left(\frac{-t^2}{2}\right) dt$$



- There is no closed-form solution to this integral
- Instead, we compute these values using tables / software
  - In Python you can use [scipy.stats.norm.cdf]
- Due to symmetry $\Phi(-x) = 1 - \Phi(x)$
- For a normally distributed variable $X \sim N(\mu, \sigma^2)$

$$P(a \le X \le b) = \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)$$

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5754
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7258	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7518	0.7549
0.7	0.7580	0.7612	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7996	0.8023	0.8051	0.8079	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9430	0.9441
1.6	0.9452	0.9463	0.9474	0.9485	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9700	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9762	0.9767
2.0	0.9773	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9865	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9980	0.9980	0.9981
2.9	0.9981	0.9982	0.9983	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998	0.9998

Standard Normal Distribution

- Example: assuming that the height of adult males is normally distributed with mean 175 cm and standard deviation 7.5 cm, what is the probability that a randomly selected male is shorter than 170 cm?
- Define a random variable $X \sim \mathcal{N}(175, 7.5^2)$

$$P(X \leq 170) = P\left(Z \leq \frac{170 - 175}{7.5}\right) = P(Z \leq -0.667) = \Phi(-0.667) \approx 0.2525$$

  - Approximately 25.25% of adult males in the population are shorter than 170 cm

The Quantile Function

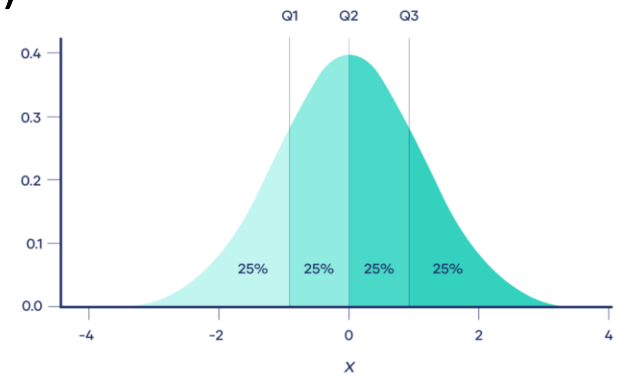
- **Quantile** is a value below which a specified proportion of the data falls
  - **Percentile** is similar but uses percentages (0-100%) instead of proportions
- The **quantile function** is the inverse of the CDF

$$F_X^{-1}(p) = x \quad \Leftrightarrow \quad F_X(x) = p$$

- For standard normal distribution, it gives the value *z* such that $\phi(z) = p$
  - e.g., the 95th percentile of the standard normal distribution

$$\Phi^{-1}(0.95) \approx 1.645$$

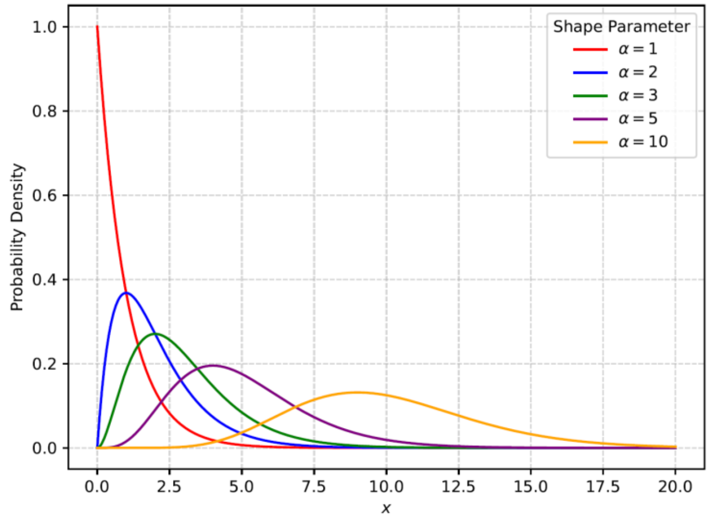
  - In Python, you can compute this value using **scipy.stats.norm.ppf**()



Gamma Distribution

- A flexible distribution defined by a **shape parameter** α and **rate parameter** λ
- Models the waiting time until α events occur in a Poisson process with rate λ
- Commonly used in Bayesian statistics
- Its PDF is given by

$$f(x) = \begin{cases} \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x} & x \geq 0 \\ 0 & x < 0 \end{cases}$$



  - The gamma function generalizes the factorial to non-integers

$$\Gamma(x) = \int_0^\infty t^{x-1} e^{-t}\, dt, \quad x > 0$$

Common Continuous Distributions

Distribution	PDF	Mean	Variance
Uniform(a,b)	$\frac{1}{b-a}, \quad a \le x \le b$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
Normal(μ,σ^2)	$\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$	μ	σ^2
Student's $t(\nu)$	$\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}$	$0 \ (\nu > 1)$	$\frac{\nu}{\nu-2} \ (\nu > 2)$
Cauchy(x_0,γ)	$\frac{1}{\pi\gamma\left[1+\left(\frac{x-x_0}{\gamma}\right)^2\right]}$	–	–
Log-Normal(μ,σ^2)	$\frac{1}{x\sigma\sqrt{2\pi}}\exp\left(-\frac{(\log x-\mu)^2}{2\sigma^2}\right), \ x > 0$	$e^{\mu+\sigma^2/2}$	$(e^{\sigma^2}-1)e^{2\mu+\sigma^2}$
Exponential(λ)	$\lambda e^{-\lambda x}, \ x \ge 0$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$
Gamma(k,θ)	$\frac{x^{k-1}e^{-x/\theta}}{\theta^k\Gamma(k)}, \ x > 0$	$k\theta$	$k\theta^2$
Chi-Squared(ν)	$\frac{x^{\nu/2-1}e^{-x/2}}{2^{\nu/2}\Gamma(\nu/2)}, \ x \ge 0$	ν	2ν
Beta(α,β)	$\frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)}, \ 0 \le x \le 1$	$\frac{\alpha}{\alpha+\beta}$	$\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$

Conditional Probabilities

- The likelihood of an event occurring given that another event has already occurred
- The conditional probability of an event *A* given another event *B* is defined as:

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

  - assuming that $P(B) \neq 0$
- Example: suppose a class has 30 students. Out of these, 18 students take mathematics, and 10 students take both mathematics and physics. If a student is known to take mathematics, what is the probability that they also take physics?
  - Define *A* as the event that a student takes physics
  - Define *B* as the event that a student takes math

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{10/30}{18/30} = \frac{10}{18} \approx 0.556$$

The Product Rule

- Expresses the joint probability of two events in terms of their conditional probability

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

  - Follows directly from the definition of conditional probability

- Example: suppose a box contains 5 red balls and 3 blue balls. Two balls are drawn at random, one after the other, without replacement. What is the probability that both balls are red?
  - Let *A* be the event that the first ball drawn is red
  - Let *B* be the event that the second ball drawn is red

$$P(A \cap B) = P(A)P(B|A) = \frac{5}{8} \cdot \frac{4}{7} = \frac{5}{14}$$

The Chain Rule

- Generalizes the product rule to any number of events
- For n events A_1, …, A_n, their joint probability can be expressed as

$$P(A_1, A_2, \ldots, A_n) = P(A_1)P(A_2|A_1)P(A_3|A_1, A_2) \cdots P(A_n|A_1, \ldots, A_{n-1})$$
$$= \prod_{i=1}^{n} P(A_i|A_1, \ldots, A_{i-1})$$

Law of Total Probability

- Let B_1, …, B_n be n disjoint events whose union is the entire sample space
- Then, for any event A, $$P(A) = \sum_{i=1}^{n} P(A \cap B_i) = \sum_{i=1}^{n} P(B_i)P(A|B_i)$$
- Suppose you have two boxes:
  - **Box A** contains 3 red balls and 2 blue balls
  - **Box B** contains 1 red ball and 4 blue balls
- You randomly choose a box with equal probability and then you randomly draw a ball from that box, what is the probability that the ball you draw is red?
  - Let R be the event that the ball is red
  - Let A/B be the event that you choose box A/B, respectively

$$P(R) = P(A) \cdot P(R|A) + P(B) \cdot P(R|B) = \frac{1}{2} \cdot \frac{3}{5} + \frac{1}{2} \cdot \frac{1}{5} = \frac{4}{10} = 0.4$$

Bayes' Rule

- Allows us to update the probability of an event in light of new evidence
- For any two events *A*, *B*, such that $P(B) \neq 0$

Likelihood

Prior probability

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

Posterior probability

Evidence (marginal probability)

- Can be used to infer the probability of a cause given the observed effect:

$$P(cause \mid effect) = \frac{P(effect \mid cause)P(cause)}{P(effect)}$$

It's hard to estimate this

but often easier to estimate this

Bayes' Rule Example

- Meningitis causes a stiff neck
  - as do lots of other things
- A doctor knows
  - Meningitis causes a stiff neck 70% of time
  - Prior probability of a patient having meningitis is 1 in 50,000
  - Prior probability of a stiff neck is 1%

- How likely is a patient with stiff neck to have meningitis?

Bayes' Rule Example

- Denote by S the event of having stiff neck
- Denote by M the event of having meningitis
- From the data we have:
  - $P(S|M) = 0.7$
  - $P(M) = 0.00002$
  - $P(S) = 0.01$
- Using Bayes' rule the posterior probability of having meningitis is:

$$P(M|S) = \frac{P(S|M)P(M)}{P(S)} = \frac{0.7 \cdot 0.00002}{0.01} = 0.0014$$

Independence

- Two events A and B are **independent**, denoted by $A \perp B$, if

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

- In this case,

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

  - i.e., observing one event doesn't change the probability that the other event occurs
- For example, suppose you toss two fair coins
  - Let A be the event that the first coin shows heads
  - Let B be the event that the second coin shows heads
  - These two events are independent
  - Therefore, their joint probability is:

$$P(A \cap B) = P(A)P(B) = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$$

Conditional Independence

- A more limited form of independence that is more common than full independence
- Events *A* and *B* are **conditionally independent** given event *C* if, once we know that *C* occurs, the probability that *A* occurs is unaffected by whether or not *B* occurs

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

  - or, equivalently:

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

- For example, if I have a cavity, the probability that the dentist probe catches it is independent of whether I have a toothache:

$$P(\text{toothache}, \text{catch} \mid \text{cavity}) = P(\text{toothache} \mid \text{cavity})P(\text{catch} \mid \text{cavity})$$

  - but in general toothache and catch are not independent, because a toothache may be more likely when catch is true

$$P(\text{toothache}, \text{catch}) \neq P(\text{toothache})P(\text{catch})$$

Joint Probability Distributions

- The **joint probability distribution** of X and Y specifies the probability for all combinations of values for X and Y
- For discrete random variables X and Y, we define their **joint PMF** as

$$p_{XY}(x, y) = P(X = x, Y = y), \quad \text{for all } x \in \mathcal{R}_X, \; y \in \mathcal{R}_Y$$

- More generally, the joint PMF of n discrete random variables X_1, …, X_n is:

$$p_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = P(X_1 = x_1, \ldots, X_n = x_n)$$

- A valid joint PMF must satisfy:
  - The probability of any combination of values must be nonnegative

$$p_{X_1,\ldots,X_n}(x_1, \ldots, x_n) \geq 0, \quad \text{for all } x_1 \in \mathcal{R}_{X_1}, \ldots, x_n \in \mathcal{R}_{X_n}$$

  - The total probability of all combinations must equal 1:

$$\sum_{x_1 \in \mathcal{R}_{X_1}} \cdots \sum_{x_n \in \mathcal{R}_{X_n}} p_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = 1$$

Joint Probability Distribution Example

- **Example**: Dental visit
- We have 3 binary random variables: Cavity, Catch, and Toothache

	toothache		¬toothache	
	catch	¬catch	catch	¬catch
cavity	0.108	0.012	0.072	0.008
¬cavity	0.016	0.064	0.144	0.576

- The full joint distribution can be used as the “knowledge base” from which all probabilities of events in the system may be derived

Joint Probability Distribution Example

- For example, what is the probability of cavity given that we have a toothache?

	toothache		¬toothache	
	catch	¬catch	catch	¬catch
cavity	0.108	0.012	0.072	0.008
¬cavity	0.016	0.064	0.144	0.576

$$P(\text{cavity} \mid \text{toothache}) = \frac{P(\text{cavity}, \text{toothache})}{P(\text{toothache})} = \frac{0.108 + 0.012}{0.108 + 0.012 + 0.016 + 0.064} = 0.6$$

Marginal Distributions

- The **marginal distribution** of a random variable is obtained by summing the joint probabilities over all the other variables (this process is called **marginalization**)
- For two discrete random variables *X* and *Y*, their marginal PMFs are:

$$p_X(x) = P(X = x) = \sum_y P(X = x, Y = y)$$

$$p_Y(y) = P(Y = y) = \sum_x P(X = x, Y = y)$$

- For *n* variables, the marginal probability of a subset of variables is obtained by summing over the remaining ones:

$$P(X_1 = x_1, X_2 = x_2) = \sum_{x_3} \sum_{x_4} \cdots \sum_{x_n} p_{X_1, \ldots, X_n}(x_1, x_2, \ldots, x_n)$$

Marginal Distribution Example

- For example, we can extract the marginal distribution of Cavity from the joint distribution by summing up all possible values of the other variables:

	toothache		¬toothache	
	catch	¬catch	catch	¬catch
cavity	0.108	0.012	0.072	0.008
¬cavity	0.016	0.064	0.144	0.576

$$P(\text{cavity}) = \sum_{c \in \text{Catch},\, t \in \text{Toothache}} P(\text{cavity}, c, t) = 0.108 + 0.012 + 0.072 + 0.008 = 0.2$$

$$P(\neg\text{cavity}) = 1 - P(\text{cavity}) = 0.8$$

Joint Probability Density Function

- For two continuous random variables X and Y, their **joint PDF** is defined as:

$$P(a \le X \le b,\ c \le Y \le d) = \int_a^b \int_c^d f_{X,Y}(x,y)\,dy\,dx$$

- For *n* continuous variables:

$$P(a_1 \le X_1 \le b_1,\ \ldots,\ a_n \le X_n \le b_n) = \int_{a_1}^{b_1} \cdots \int_{a_n}^{b_n} f_{X_1,\ldots,X_n}(x_1,\ldots,x_n)\,dx_n \cdots dx_1$$

- For *f* to be a valid joint PDF, it must satisfy:
  - The function must be nonnegative for all values of the variables:

$$f(x_1,\ldots,x_n) \ge 0, \quad \text{for all } x_1,\ldots,x_n \in \mathbb{R}$$

  - The total probability over the entire space must equal to 1:

$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(x_1,\ldots,x_n)\,dx_1 \ldots dx_n = 1$$

Joint Probability Density Function

- Example: consider random variables *X* and *Y* with a joint PDF:

$$f(x,y) = \begin{cases} x+y & 0 \le x \le 1,\ 0 \le y \le 1 \\ 0 & \text{otherwise} \end{cases}$$

- To verify that this is a valid joint PDF we compute the integral:

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)\,dy\,dx = \int_0^1\int_0^1 (x+y)\,dy\,dx = \int_0^1 \left[xy + \frac{y^2}{2}\right]_0^1 dx$$
$$= \int_0^1 \left(x + \frac{1}{2}\right) dx = \left[\frac{x^2}{2} + \frac{x}{2}\right]_0^1 = 1$$

- The probability that both variables are less than 0.5 is:

$$P(X \le 0.5, Y \le 0.5) = \int_0^{0.5}\int_0^{0.5} (x+y)\,dy\,dx = \int_0^{0.5} \left[xy + \frac{y^2}{2}\right]_0^{0.5} dx$$
$$= \int_0^{0.5} \left(0.5x + \frac{0.25}{2}\right) dx = \left[\frac{0.5x^2}{2} + 0.125x\right]_0^{0.5} = 0.125$$

Marginal and Conditional Density Functions

- The **marginal density function** of variable *X* is obtained by integrating the joint PDF over the other variables in the function

$$f_X(x) = \int_{-\infty}^{\infty} f_{XY}(x, y)\, dy$$
$$f_Y(y) = \int_{-\infty}^{\infty} f_{XY}(x, y)\, dx$$

- The **conditional density function** of *Y* given *X* is:

$$f_{Y|X}(y|x) = \frac{f_{XY}(x, y)}{f_X(x)}, \quad \text{for } f_X(x) > 0$$

  - where $f_X(x)$ is the marginal density function of *X*

Expectation in Joint Distributions

- We cannot define directly an expectation of joint distribution since expected value needs to return a single value
- Instead, we define expectation of scalar-valued functions of the variables
  - For discrete random variables

$$\mathbb{E}[g(X_1, \ldots, X_n)] = \sum_{x_1} \cdots \sum_{x_n} g(x_1, \ldots, x_n) P(X_1 = x_1, \ldots, X_n = x_n)$$

  - For continuous random variables

$$\mathbb{E}[g(X_1, \ldots, X_n)] = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} g(x_1, \ldots, x_n) f_{X_1, \ldots, X_n}(x_1, \ldots, x_n)\, dx_1 \ldots dx_n$$

- As in the single-variable case, the expectation satisfies the linearity property:

$$\mathbb{E}\left[a g_1(X_1, \ldots, X_n) + b g_2(X_1, \ldots, X_n)\right] = a\, \mathbb{E}\left[g_1(X_1, \ldots, X_n)\right] + b\, \mathbb{E}\left[g_2(X_1, \ldots, X_n)\right]$$

Independent Random Variables

- Two random variables *X* and *Y* are independent if knowing the value of one variable doesn't provide information on the value of the other
- Their joint probability is equal to the product of their marginal probabilities
  - Discrete random variables *X* and *Y* are independent if **for all values** of *x* and *y*:

$$p_{XY}(x, y) = p_X(x)p_Y(y)$$

  - Continuous random variables *X* and *Y* are independent if **for all values** of *x* and *y*:

$$f_{XY}(x, y) = f_X(x)f_Y(y)$$

Independent Random Variables

- Let T be a binary random variable representing the temperature (hot/cold)
- Let W be a binary random variable representing the weather (sun/rain)
- Their marginal and joint probabilities are given in the following tables:

$P(T)$

T	P
hot	0.5
cold	0.5

$P(W)$

W	P
sun	0.6
rain	0.4

$P(T, W)$

T	W	P
hot	sun	0.3
hot	rain	0.2
cold	sun	0.2
cold	rain	0.3

- Are the variable independent?
  - No, because $P(T = \text{cold}, W = \text{sun}) = 0.2 \neq P(T = \text{cold}) \cdot P(W = \text{sun}) = 0.5 \cdot 0.6 = 0.3$

Independent and Identically Distributed Variables

- Random variables $X_1, \ldots, X_n$ are **i.i.d.** (independent and identically distributed) if:
  - They are mutually independent

$$f_{X_1,\ldots,X_n}(x_1,\ldots,x_n) = f_{X_1}(x_1)\cdots f_{X_n}(x_n)$$

  - Have the same probability distribution

$$f_{X_1}(x) = f_{X_2}(x) = \cdots = f_{X_n}(x), \quad \text{for all } x \in \mathbb{R}$$

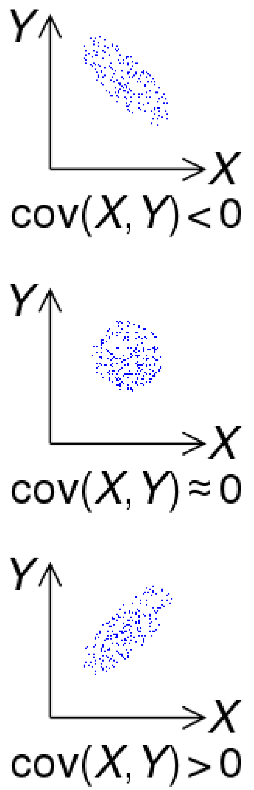
  - f can be a PMF (for discrete variables) or PDF (for continuous variables)
- For example, suppose $X_1, \ldots, X_n$ represent outcomes of n independent tosses of a coin
  - Then $X_1, \ldots, X_n$ are i.i.d. Bernoulli random variables
- The i.i.d. assumption is common in statistics and machine learning
  - e.g., we assume that all observations in the dataset are i.i.d. (sampled independently from the same underlying distribution)

Covariance

- Covariance measures the extent to which two variables vary together
  - i.e., whether they tend to increase or decrease in tandem
- The covariance between random variables *X* and *Y* is defined as:

$$\mathrm{Cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] = \mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y]$$

  - Cov(*X*, *Y*) > 0 indicates that *X* and *Y* tend to increase or decrease together
  - Cov(*X*, *Y*) < 0 indicates that *X* and *Y* tend to move in opposite directions
  - Cov(*X*, *Y*) = 0 means that X and Y are uncorrelated (linearly)
    - It doesn't imply that they are independent
    - However, if *X* and *Y* are independent then Cov(*X*, *Y*) = 0



Covariance

- Assume we have two discrete variables *X* and *Y* with the following joint distribution

x_i	y_i	p_i
2	3	1/5
4	7	1/5
6	5	1/5
8	10	1/5
10	15	1/5

- We first compute the expectations:

$$\mathbb{E}[X] = \sum_{i=1}^{5} x_i p_i = \frac{2+4+6+8+10}{5} = \frac{30}{5} = 6 \qquad \mathbb{E}[Y] = \sum_{i=1}^{5} y_i p_i = \frac{3+7+5+10+15}{5} = \frac{40}{5} = 8$$

$$\mathbb{E}[XY] = \sum_{i=1}^{5} x_i y_i p_i = \frac{2\cdot 3 + 4\cdot 7 + 6\cdot 5 + 8\cdot 10 + 10\cdot 15}{5} = \frac{294}{5} = 58.8$$

- Therefore: $\text{Cov}(X,Y) = \mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y] = 58.8 - 6\cdot 8 = 10.8$

Properties of the Covariance

- Symmetry:

$$\mathrm{Cov}(X,Y) = \mathrm{Cov}(Y,X)$$

- Linearity:

$$\mathrm{Cov}\left(\sum_{i=1}^{n} a_i X_i,\ \sum_{j=1}^{m} b_j Y_j\right) = \sum_{i=1}^{n}\sum_{j=1}^{m} a_i b_j \,\mathrm{Cov}(X_i, Y_j)$$

- Variance as a special case of covariance:

$$\mathrm{Var}(X) = \mathrm{Cov}(X,X)$$

- Variance of a sum of variables:

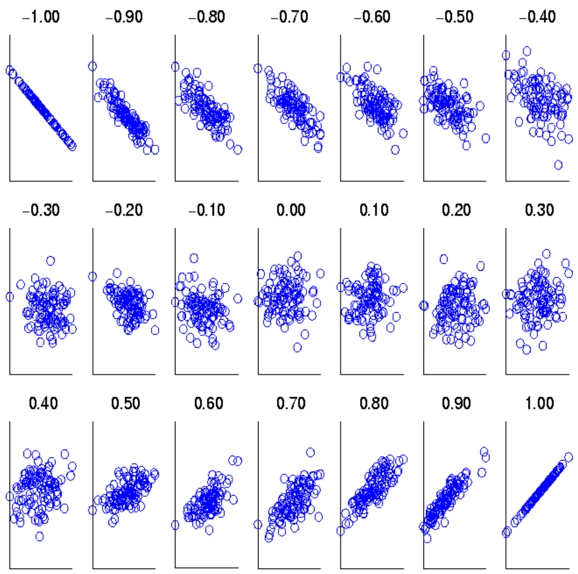
$$\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\,\mathrm{Cov}(X,Y)$$

$$\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \mathrm{Var}(X_i) + 2\sum_{i<j} \mathrm{Cov}(X_i, X_j)$$

Correlation Coefficient

- Covariance depends on the scale of the variables and can take any real value
- Correlation coefficients are normalized between -1 and +1
  - Allowing to quantify the strength of the correlation and not only the direction
- The most widely used is **Pearson's correlation coefficient** defined as:

$$\rho_{XY} = \frac{\mathrm{Cov}(X,Y)}{\sigma_X \sigma_Y}$$



  - $\rho_{XY} \approx 1$ indicates strong positive linear relationship
  - $\rho_{XY} \approx 0$ indicates no linear relationship
  - $\rho_{XY} \approx -1$ indicates strong negative linear relationship
- It is scale invariant: for any constants a, $c > 0$

$$\rho_{aX+b,\, cY+d} = \rho_{XY}$$

Non-Linear Relationships

- Pearson correlation coefficient measures the linear relationship between objects
- If the coefficient is 0, non-linear relationships may still exist
- For example, if

$$X = (\ -3,\ -2,\ -1,\ 0,\ 1,\ 2,\ 3)$$
$$Y = (\ 9,\ 4,\ 1,\ 0,\ 1,\ 4,\ 9)$$

  - Then $Y = X^2$, but their Pearson correlation coefficient is 0
- Other correlation coefficients measure other types of relationships
  - **Spearman's rank correlation coefficient** measures the strength of monotonic relationship
    - Defined as the Pearson correlation coefficient applied to the ranked values of X and Y

$$\rho_s = \frac{\text{Cov}(R(X), R(Y))}{\sigma_{R(X)}\sigma_{R(Y)}}$$

Random Vectors

- A **random vector** is a function **X**: $\Omega \to R^n$ whose components are random variables
  - Can simplify computations when working with multiple random variables

$$\mathbf{X} = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{pmatrix}$$

- The definitions of PMD, PDF, and CDF of **X** are based on the corresponding definitions of jointly distributed random variables:
  - PMF of **X**

$$p_{\mathbf{X}}(\mathbf{x}) = p_{X_1,\ldots,X_n}(x_1,\ldots,x_n) = P(X_1 = x_1,\ldots,X_n = x_n)$$

  - PDF of **X**

$$f_{\mathbf{X}}(\mathbf{x}) = f_{X_1,\ldots,X_n}(x_1,\ldots,x_n)$$

The Mean Vector

- The **mean vector** of a random vector **X** contains the means of its components:

$$\boldsymbol{\mu} = \mathbb{E}[\mathbf{X}] = \begin{pmatrix} \mathbb{E}[X_1] \\ \mathbb{E}[X_2] \\ \vdots \\ \mathbb{E}[X_n] \end{pmatrix}$$

- Satisfies the following properties
  - Linearity with respect to dot products: for any constant vector **a** ∈ R^n

$$\mathbb{E}[\mathbf{a}^T\mathbf{X}] = \mathbf{a}^T\,\mathbb{E}[\mathbf{X}]$$

  - Expectation of sums of vectors:

$$\mathbb{E}[\mathbf{X}_1 + \mathbf{X}_2 + \cdots + \mathbf{X}_k] = \mathbb{E}[\mathbf{X}_1] + \mathbb{E}[\mathbf{X}_2] + \cdots + \mathbb{E}[\mathbf{X}_k]$$

  - Linearity under matrix transformation: for any constant matrix A ∈ R$^{m\times n}$

$$\mathbb{E}[A\mathbf{X}] = A\,\mathbb{E}[\mathbf{X}]$$

Covariance Matrix

- Describes how the components of a random vector vary together
  - Extends the concept of variance to higher dimensions
- The covariance matrix of a random vector **X**: $\Omega \to$ R^n is defined as:

$$\Sigma = \mathrm{Cov}(\mathbf{X}) = \mathbb{E}\left[(\mathbf{X} - \mathbb{E}[\mathbf{X}])(\mathbf{X} - \mathbb{E}[\mathbf{X}])^T\right]$$

- Expanding the definition yields:

$$\Sigma = \begin{pmatrix} \mathrm{Var}(X_1) & \mathrm{Cov}(X_1, X_2) & \cdots & \mathrm{Cov}(X_1, X_n) \\ \mathrm{Cov}(X_2, X_1) & \mathrm{Var}(X_2) & \cdots & \mathrm{Cov}(X_2, X_n) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}(X_n, X_1) & \mathrm{Cov}(X_n, X_2) & \cdots & \mathrm{Var}(X_n) \end{pmatrix}$$

Properties of the Covariance Matrix

- The matrix is symmetric:

$$\mathrm{Cov}(X_i, X_j) = \mathrm{Cov}(X_j, X_i) \quad \text{for all } i, j$$

- The matrix is positive semidefinite: for any constant vector **a** $\in$ R^n

$$\mathbf{a}^T \Sigma \mathbf{a} = \mathrm{Var}(\mathbf{a}^T \mathbf{X}) \geq 0$$

- Covariance of linear transformations: for any constant matrix $A \in$ R$^{m \times n}$

$$\mathrm{Cov}(A\mathbf{X}) = A\, \mathrm{Cov}(\mathbf{X}) A^T$$

Correlation Matrix

- Contains the Pearson correlation coefficients between the components of **X**

$$R_{ij} = \rho(X_i, X_j) = \frac{\text{Cov}(X_i, X_j)}{\sigma_{X_i}\sigma_{X_j}}$$

  - *R* is symmetric positive semidefinite with all diagonal entries equal to 1
- In NumPy, you can compute it using the function **np.corrcoef**():

```
A = np.random.random((3, 3))
A
```

```
array([[0.37454012, 0.95071431, 0.73199394],
       [0.59865848, 0.15601864, 0.15599452],
       [0.05808361, 0.86617615, 0.60111501]])
```

```
R = np.corrcoef(A)
R
```

```
array([[ 1.        , -0.92660504,  0.99832331],
       [-0.92660504,  1.        , -0.94681794],
       [ 0.99832331, -0.94681794,  1.        ]])
```

Multivariate Distributions

- Multivariate distribution gives the full probabilistic model of a random vector
- It includes:
  - A joint probability distribution of the random variables in the vector
  - Dependency or correlation structure between variables (e.g., covariance matrix)

The Multivariate Normal Distribution

- A random vector **X** follows a **multivariate normal distribution** with mean vector $\boldsymbol{\mu} \in \mathbb{R}^n$ and covariance matrix $\Sigma \in \mathbb{R}^{n\times n}$, denoted by $\mathbf{X} \sim N(\boldsymbol{\mu}, \Sigma)$, if its PDF is given by:

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)$$

  - $|\Sigma|$ denotes the determinant of the covariance matrix
  - Σ must be invertible ($|\Sigma| \neq 0$) for the PDF to be well-defined
- When Σ is **diagonal**, the PDF factorizes into a product of univariate normal densities:

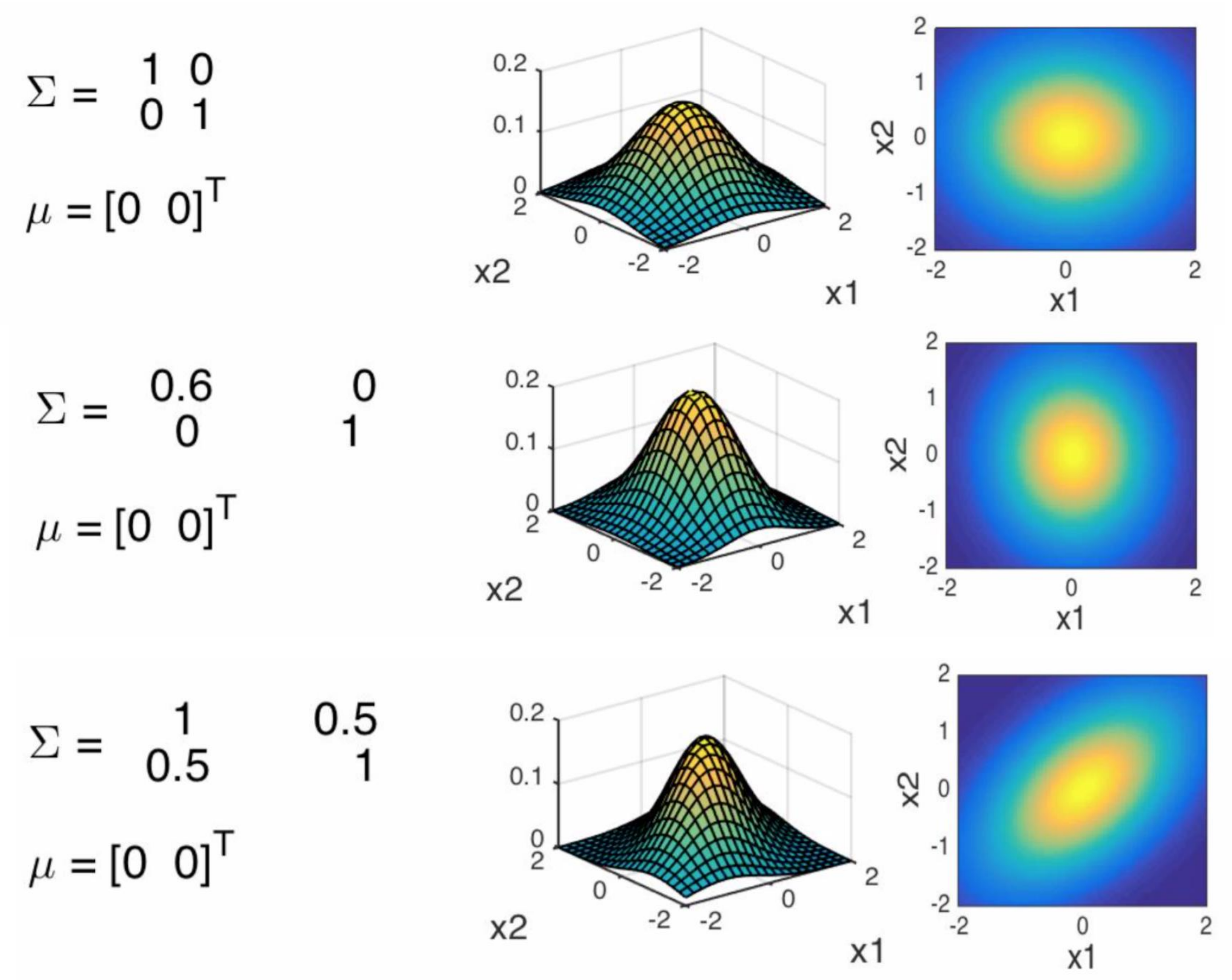
$$f_{\mathbf{X}}(\mathbf{x}) = \prod_{i=1}^{n}\left(\frac{1}{\sqrt{2\pi}\sigma_i}\exp\left(-\frac{1}{2}\cdot\frac{(x_i-\mu_i)^2}{\sigma_i^2}\right)\right) = \prod_{i=1}^{n} f_{X_i}(x_i), \quad X_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$$

  - In this case, all the variable X_i are independent
- The standard multivariate normal distribution has $\boldsymbol{\mu} = \mathbf{0}$ and $\Sigma = I$

$$\mathbf{Z} \sim \mathcal{N}(\mathbf{0}, I)$$

The Multivariate Normal Distribution

- Examples:



Properties of Multivariate Normal Distributions

- If **X** ~ $N(\boldsymbol{\mu}, \Sigma)$ then each component of **X** is univariate normal $X_i \sim \mathcal{N}(\mu_i, \Sigma_{ii})$
- Closure under linear transformations: If A is a constant matrix and **b** constant vector

$$A\mathbf{X} + \mathbf{b} \sim \mathcal{N}(A\mu + \mathbf{b}, A\Sigma A^T)$$

- Closure under summation: $\mathbf{X} + \mathbf{Y} \sim \mathcal{N}(\mu_X + \mu_Y, \Sigma_X + \Sigma_Y)$
- Closure under marginalization and conditioning: if

$$\mathbf{X} = \begin{pmatrix} \mathbf{X}_A \\ \mathbf{X}_B \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu_A \\ \mu_B \end{pmatrix}, \begin{pmatrix} \Sigma_{AA} & \Sigma_{AB} \\ \Sigma_{BA} & \Sigma_{BB} \end{pmatrix}\right)$$

  - Then the marginal and conditional distributions are also multivariate normal:

$$\mathbf{X}_A \sim \mathcal{N}(\mu_A, \Sigma_{AA}), \quad \mathbf{X}_B \sim \mathcal{N}(\mu_B, \Sigma_{BB})$$

$$\mathbf{X}_A | \mathbf{X}_B \sim \mathcal{N}\left(\mu_A + \Sigma_{AB}\Sigma_{BB}^{-1}(\mathbf{X}_B - \mu_B), \Sigma_{AA} - \Sigma_{AB}\Sigma_{BB}^{-1}\Sigma_{BA}\right)$$

$$\mathbf{X}_B | \mathbf{X}_A \sim \mathcal{N}\left(\mu_B + \Sigma_{BA}\Sigma_{AA}^{-1}(\mathbf{X}_A - \mu_A), \Sigma_{BB} - \Sigma_{BA}\Sigma_{AA}^{-1}\Sigma_{AB}\right)$$

The Multinomial Distribution

- Models repeated trials where each trials results in one of *k* possible categories
  - Counts how many times each category has occurred in *n* independent trials
  - When $k = 2$, this reduces to the binomial distribution
- Let the probabilities of the *k* outcomes be $p_1, \dots, p_k$
- Let X_i represent the number of times category *i* has occurred in *n* trials
- Then the random vector $\mathbf{X} = (X_1, \dots, X_k)$ follows a multinomial distribution
- The PMF of **X** is:

$$P(X_1 = x_1, \dots, X_k = x_k) = \begin{cases} \frac{n!}{x_1! \cdots x_k!} \prod_{i=1}^k p_i^{x_i} & \text{if } \sum_{i=1}^k x_i = n \\ 0 & \text{otherwise} \end{cases}$$

- For example, suppose that a fair die is rolled 5 times
  - The probability that 2 and 3 appear twice each and 6 appears once is:

$$P(X_2 = 2, X_3 = 2, X_6 = 1) = \frac{5!}{2! \cdot 2! \cdot 1!} \left(\frac{1}{6}\right)^2 \left(\frac{1}{6}\right)^2 \left(\frac{1}{6}\right)^1 = 30 \cdot \left(\frac{1}{6}\right)^5 \approx 0.00386$$

The Law of Large Numbers

- The sample average of i.i.d. random variables converges to the expected value of their underlying distribution
- Let $X_1, \ldots, X_n$ be i.i.d. random variables with expected value $\mu = E[X_i]$
- Define their **sample mean** as: $$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$
- Then for any $\varepsilon > 0$, $$\lim_{n\to\infty} P(|\overline{X}_n - \mu| > \varepsilon) = 0$$

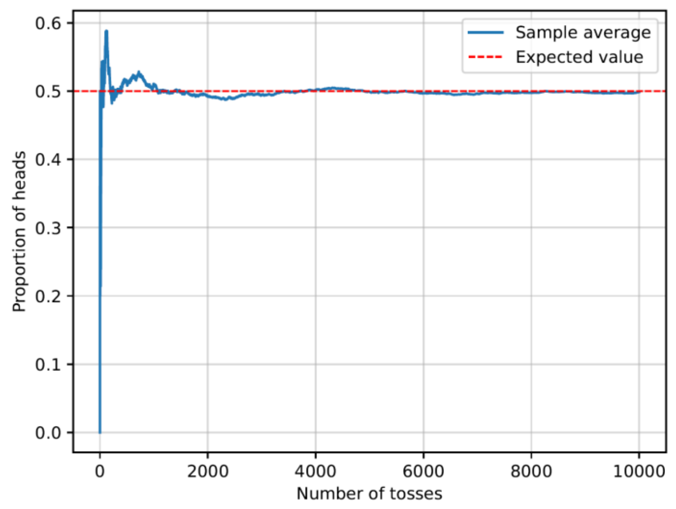


Figure A.24: Illustration of the Law of Large Numbers: The sample average of outcomes from repeated fair coin tosses converges to the expected value $\mu = 0.5$.

The Central Limit Theorem (CLT)

- The sum (or average) of a large number of i.i.d. variables is approximately normally distributed, regardless of their original distribution
- Let $X_1, \ldots, X_n$ be i.i.d. random variables with mean μ and variance σ^2
- Their **standardized sample mean** is:

$$Z_n = \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}}$$

- Then Z_n converges in distribution to the standard normal distribution

$$\lim_{n\to\infty} P(Z_n \le x) = \Phi(x), \quad \text{for all } x \in \mathbb{R}$$

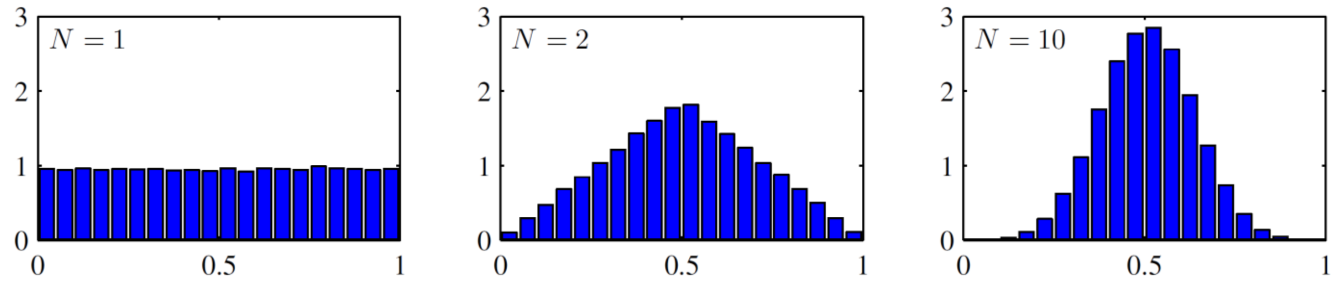


Figure 2.6 Histogram plots of the mean of N uniformly distributed numbers for various values of N. We observe that as N increases, the distribution tends towards a Gaussian.

Example for Using the CLT

- An airline is designing seating for a small aircraft that can carry 100 passengers
- The aircraft cannot safely carry more than a total of 8,500 kg
- Assume the average weight of a passenger (+baggage) is 82 kg with std 15 kg
- What is the probability that the total weight of 100 passengers exceeds the limit?

Solution

- Let X_i ($1 \leq i \leq 100$) denote the weight of a single passenger
  - The X_i s are assumed to be i.i.d.
- Let W be the total of the weights of all 100 passengers: $W = \sum_{i=1}^{100} X_i$
- Expected value and variance of W:

$$\mathbb{E}[W] = 100 \cdot 82 = 8200, \quad \mathrm{Var}(W) = 100 \cdot 15^2 = 22500$$

- The standardized W is:

$$Z = \frac{W - 8200}{\sqrt{22500}} = \frac{W - 8200}{150}$$

- By the CLT, Z is approximately standard normal ($n = 100 > 30$), therefore

$$P(W > 8500) = P\left(Z > \frac{8500 - 8200}{150}\right) = P(Z > 2) = 1 - \Phi(2) \approx 1 - 0.9772 = 0.0228$$

  - There is only 2.28% chance of exceeding the limit

Maximum Likelihood Estimation

- A widely used method for estimating distribution parameters from observed data
- Idea: Choose the parameter values that make the observed data most probable
- Assume that we have a set of data points $\{x_1, ..., x_n\}$ drawn independently from some probability distribution $P(X; \theta)$ with unknown parameter(s) θ
  - e.g., θ could be the mean and variance of a normal distribution
- The **likelihood function** of θ as the probability of observing the data under θ

$$\mathcal{L}(\theta|x_1, \ldots, x_n) = p(x_1, \ldots, x_n|\theta)$$

- The goal is to fine the value of that maximizes the likelihood
  - Called the **maximum likelihood estimator (MLE)**

$$\hat{\theta}_{\text{MLE}} = \underset{\theta}{\operatorname{argmax}}\, \mathcal{L}(\theta)$$

Maximum Likelihood Estimation

- Under i.i.d. assumption of the observed values we can write

$$\mathcal{L}(\theta) = p(x_1|\theta)p(x_2|\theta)\cdots p(x_n|\theta) = \prod_{i=1}^{n} p(x_i|\theta)$$

- To simplify computations, we typically work with the **log-likelihood** function:

$$\ell(\theta) = \log \mathcal{L}(\theta) = \sum_{i=1}^{n} \log p(x_i|\theta)$$

- To find the MLE we take its derivative and set it equal to zero:

$$\frac{d\ell(\theta)}{d\theta} = 0$$

MLE Example (1)

- There are 10 balls in a bag. Each ball is either red or green.
- Let θ be the number of red balls
- We draw 5 balls **with replacement** out of the bag getting:
  - "red", "red", “green", "red", "green” (in that order)
- What is the maximum likelihood estimate for θ?
- The likelihood function is:

$$\mathcal{L}(\theta) = P(\text{red, red, green, red, green}|\theta) = \left(\frac{\theta}{10}\right)^3 \left(\frac{10-\theta}{10}\right)^2$$

- The log-likelihood function is:

$$\begin{aligned}\ell(\theta) &= \log \mathcal{L}(\theta) = 3(\log\theta - \log 10) + 2(\log(10-\theta) - \log 10)\\ &= 3\log\theta + 2\log(10-\theta) - 5\log 10\end{aligned}$$

- To find the MLE we compute:

$$\frac{\partial \ell}{\partial \theta} = \frac{3}{\theta} - \frac{2}{10-\theta} = 0 \;\Rightarrow\; 3(10-\theta) = 2\theta \;\Rightarrow\; \theta = 6$$

$$\hat{\theta}_{\text{MLE}} = 6$$

MLE Example (2)

- Given n points $x_1, \ldots, x_n$ drawn from univariate normal distribution $N(\mu, \sigma)$
- Find the MLEs for μ and σ
- The likelihood function of the parameters is:

$$\mathcal{L}(\mu, \sigma | X) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x_i - \mu)^2}{2\sigma^2}\right)$$

- Thus, the log-likelihood is:

$$\begin{aligned}
\ell(\mu, \sigma | X) &= \sum_{i=1}^{n} \log\left[\frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x_i - \mu)^2}{2\sigma^2}\right)\right] \\
&= n \log \frac{1}{\sqrt{2\pi}\sigma} - \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{2\sigma^2} \\
&= -\sum_{i=1}^{n} \frac{(x_i - \mu)^2}{2\sigma^2} - n \log \sigma - \frac{n}{2} \log 2\pi
\end{aligned}$$

MLE Example (2)

- Taking partial derivatives of the log-likelihood w.r.t. μ, σ and setting them to 0:

$$\frac{\partial \ell}{\partial \mu} = -\sum_{i=1}^{n} \frac{-2(x_i - \mu)}{2\sigma^2} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - \mu) = 0$$

$$\Rightarrow \sum_{i=1}^{n} x_i - n\mu = 0$$

$$\Rightarrow \mu = \frac{\sum_{i=1}^{n} x_i}{n}$$

$$\frac{\partial \ell}{\partial \sigma} = -\sum_{i=1}^{n} \frac{-2(x_i - \mu)^2}{2\sigma^3} - \frac{n}{\sigma} = \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{\sigma^3} - \frac{n}{\sigma} = 0$$

$$\Rightarrow \frac{1}{\sigma^3}\sum_{i=1}^{n}(x_i - \mu)^2 = \frac{n}{\sigma}$$

$$\Rightarrow \sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n}}$$

- **Conclusion**: the MLE of the mean is the sample mean and the MLE of the standard deviation is the sample standard deviation

Additional Resources

- For further reference consult:
  - David Blei's probability review
  - The book Sheldon Ross: A First Course in Probability