ECE 30200 - Probabilistic Methods in Electrical and Computer Engineering

Zeke Ulrich

July 28, 2025

Contents

Background

The following formulas will be instrumental and may be familar.

Series

$$\sum_{k=0}^{n} r^k = \frac{1-r^{n+1}}{1-r} \tag{1}$$

$$\sum_{n=1}^{\infty} \frac{1}{n^2} = \frac{\pi^2}{6} \tag{2}$$

$$\sum_{k=1}^{\infty} k r^{k-1} = \frac{1}{(1-r)^2} \tag{3}$$

Combinatorics

$$\binom{n}{k} = \frac{n!}{k!(n-k)!} \tag{4}$$

$$(a+b)^n = \sum_{k=0}^{n} \binom{n}{k} a^{n-k} b^k \tag{5}$$

$$\binom{n}{k} + \binom{n}{k-1} = \binom{n+1}{k} \tag{6}$$

$$P(n,k) = \frac{n!}{(n-k)!} \tag{7}$$

where $P(n,k)$ is the number of ways to arrange k objects out of n (permutations).

$$C(n,k) = \binom{n}{k} = \frac{n!}{k!(n-k)!} \tag{8}$$

where $C(n,k)$ is the number of ways to choose k objects out of n (combinations).

Approximations

$$\begin{aligned} f(x) &= f(a) + f'(a)(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \dots && (9) \\ &= \sum_{n=0}^{\infty} \frac{f^{(n)}(a)}{n!}(x-a)^n && (10) \end{aligned}$$

$$\begin{aligned} 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \dots &= \sum_{k=0}^{\infty} \frac{x^k}{k!} && (11) \\ &= e^x && (12) \end{aligned}$$

$$\sin(x) = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \dots \tag{13}$$
$$= \sum_{n=0}^{\infty} (-1)^n \frac{x^{2n+1}}{(2n+1)!} \tag{14}$$

$$\cos(x) = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!} + \dots \tag{15}$$
$$= \sum_{n=0}^{\infty} (-1)^n \frac{x^{2n}}{(2n)!} \tag{16}$$

$$\ln(1+x) = x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + \dots \tag{17}$$
$$= \sum_{n=1}^{\infty} (-1)^{n+1} \frac{x^n}{n} \tag{18}$$

Calculus

$$\frac{d}{dx} \int_a^x f(t)\, dt = f(x) \tag{19}$$

$$\int_a^b f'(x)\, dx = f(b) - f(a) \tag{20}$$

$$\int f(g(x))g'(x)\, dx = \int f(u)\, du \tag{21}$$

$$\int u\, dv = uv - \int v\, du \tag{22}$$

$$\int \frac{1}{(x-a)(x-b)}\, dx = \frac{1}{b-a} \ln \left| \frac{x-a}{x-b} \right| + C \tag{23}$$

Linear Algebra

$$\vec{y} = \beta_1 \vec{x}_1 + \beta_2 \vec{x}_2 + \cdots + \beta_N \vec{x}_N \tag{24}$$

$$\langle \vec{a}, \vec{b} \rangle = \vec{a}\vec{b}^T \tag{25}$$
$$= \sum_{i=1}^{n} a_i b_i \tag{26}$$

where $\langle \vec{a}, \vec{b} \rangle$ denotes the inner product of vectors $\vec{a}$ and $\vec{b}$.

$$\|\vec{x}\|_p = \left(\sum_{i=1}^{n} |x_i|^p \right)^{1/p} \tag{27}$$

where $\|\vec{x}\|_p$ is the p-norm (or ℓ_p-norm) of vector $\vec{x}$.

$$\cos(\theta) = \frac{\langle \vec{a}, \vec{b} \rangle}{\|\vec{a}\|_2 \|\vec{b}\|_2} \tag{28}$$

where θ is the angle between vectors $\vec{a}$ and $\vec{b}$.

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\vec{y} \tag{29}$$

where $\hat{\beta}$ is the vector of least squares coefficients, $\mathbf{X}$ is the data matrix, and $\vec{y}$ is the target vector

Set Theory

The *set difference* $A \setminus B$ is the set of elements that are in A but not in B:

$$A \setminus B = \{x \mid x \in A \text{ and } x \notin B\} \tag{30}$$

Some important properties of set operations are:

- **Commutativity:**

$$A \cup B = B \cup A \tag{31}$$
$$A \cap B = B \cap A \tag{32}$$

- **Associativity:**

$$(A \cup B) \cup C = A \cup (B \cup C) \tag{33}$$
$$(A \cap B) \cap C = A \cap (B \cap C) \tag{34}$$

- **Distributivity:**

$$A \cup (B \cap C) = (A \cup B) \cap (A \cup C) \tag{35}$$
$$A \cap (B \cup C) = (A \cap B) \cup (A \cap C) \tag{36}$$

- **Identity:**

$$A \cup \emptyset = A \tag{37}$$
$$A \cap \Omega = A \tag{38}$$

- **Complement:**

$$A \cup A^c = \Omega \tag{39}$$
$$A \cap A^c = \emptyset \tag{40}$$

Probability Laws

A probability law must satisfy three axioms:

1. Non-negativity: $P(A) \geq 0 \forall A \in F$
2. Normalization: $P(\Omega) = 1$
3. Additivity: For any disjoint subsets $\{A_1, A_2, \dots\}$, it holds that

$$P\left[\bigcup_{n=1}^{\infty} A_n\right] = \sum_{n=1}^{\infty} P\left[A_n\right]$$

Probability Properties

$$P[A \cup B] = P[A] + P[B] - P[A \cap B] \tag{41}$$

$$P[A \cup B] \leq P[A] + P[B] \tag{42}$$

$$A \subseteq B \implies P[A] \leq P[B] \tag{43}$$

Formal Definitions

Outcomes

An *outcome* is the result of some *experiment*. If that experiment is flipping a coin, the outcome is either heads or tails. We could express the outcome of heads as H, and the outcome of tails as T. The set of all possible outcomes for an experiment is known as a sample space and is denoted by Ω. In this case $\Omega = \{H, T\}$.

Events

An *event* F is a subset of the sample space Ω. The formal definitions of probability are expressed with set notation. So the event where we have neither heads nor tails is written as $\{\}$. The event of heads could be expressed as $\{H\}$, and the event of tails could be expressed as $\{T\}$. The event of either heads or tails is $\{H, T\}$.

Probability Laws

A *probability law* is a function P that maps an event A to a real number in $[0, 1]$. For the coin example, the probability law might be $P(\{\}) = 0$, $P(\{H\}) = 0.5$, $P(\{T\}) = 0.5$, and $P(\{\Omega\}) = 1$. A probability law must satisfy three axioms:

1. Non-negativity: $P(A) \geq 0 \forall A \in F$
2. Normalization: $P(\Omega) = 1$
3. Additivity: For any disjoint subsets $\{A_1, A_2, \dots\}$, it holds that

$$P\left[\bigcup_{n=1}^{\infty} A_n\right] = \sum_{n=1}^{\infty} P\left[A_n\right]$$

Probability Space

A probability space is a triplet Ω, F, P.

Probability Properties

$$P\left[A \cup B\right] = P\left[A\right] + P\left[B\right] - P\left[A \cap B\right] \tag{44}$$

$$P\left[A \cup B\right] \leq P\left[A\right] + P\left[B\right] \tag{45}$$

$$A \subseteq B \implies P\left[A\right] \leq P\left[B\right] \tag{46}$$

$$P\left[A|B\right] = \frac{P[A \cap B]}{P[B]} \tag{47}$$

Outcomes are statistically *independent* if $P(A|B) = P(A)$ (assuming P(B) > 0), or equivalently $P(A \cap B) = P(A)P(B)$.

Bayes Theorem states that for any two events A and B such that $P[A] > 0$ and $P[B] > 0$,

$$P[A|B] = \frac{P[B|A]P[A]}{P[B]} \tag{48}$$

The *Law of Total Probability* states that if $\{A_1, A_2, \ldots, A_n\}$ is a partition of Ω, then for any $B \subseteq \Omega$,

$$P[B] = \sum_{i=1}^{n} P[B|A_i]P[A_i] \tag{49}$$

Discrete Random Variables

A *random variable* X is a function $X : \Omega \implies \Re$ that maps an outcome $\epsilon \in \Omega$ to a number $X(\epsilon)$ on the real line. We call it a variable because it has multiple states.

The *expectation* of a random variable X is

$$E[X] = \sum_{x \in X(\Omega)} x p_X(x) \tag{50}$$

The difference between $E[X]$ and the mean is that $E[X]$ is computed from the ideal histogram, while mean is computed from the empirical histogram. In general for any functions g and h,

$$E[g(X)] = \sum_x g(x) p_X(x) \tag{51}$$

$$E[g(X) + h(X)] = E[g(X)] + E[h(X)] \tag{52}$$

$$E[cX] = cE[X] \tag{53}$$

$$E[X + c] = E[X] + c \tag{54}$$

The *variance* of a random variable X is

$$Var[X] = E\left[(X - \mu)^2\right] \tag{55}$$

or alternatively, the second moment minus the first moment squared.

$$E[X^2] - E[X]^2 \tag{56}$$

The *probability mass function* (PMF) $p_X(a)$ of a random variable X specifies the probability of obtaining a number $X(\epsilon) = a$. We denote a PMF as

$$p_X(a) = P[X = a] \tag{57}$$

PMFs are represented with histograms. A PMF should satisfy

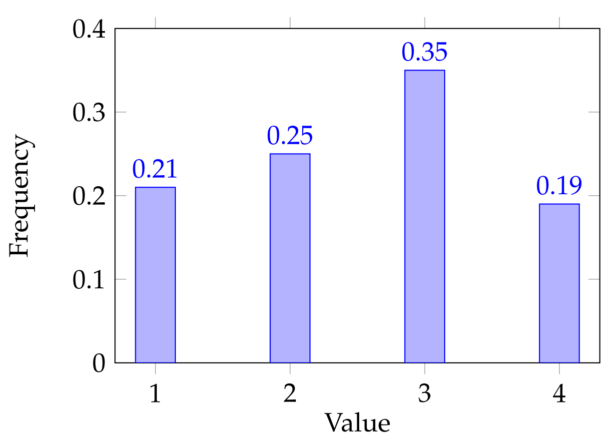

Figure 1: PMF

$$\sum_{x \in X(\Omega)} p_X(x) = 1 \tag{58}$$

The *cumulative distribution function* is given by

$$F_X(x) = P[X \leq x] \tag{59}$$
$$= \sum_{u \leq x} p_X(u) \tag{60}$$

and represents the sum of every impulse of the PMF up to x.

A *Bernoulli random variable* has a state of either 0 or 1. The probability of getting 1 is p and the probability of getting 0 is $1-p$. We write

$$X \sim Bernoulli(p) \tag{61}$$

or

$$X \sim B(p) \tag{62}$$

to say that X is drawn from a Bernoulli distribution with a parameter p. For a Bernoulli distribution,

$$E[X] = p \tag{63}$$
$$E[X^2] = p \tag{64}$$
$$Var[X] = p(1-p) \tag{65}$$

Say $S \sim B(1-p)$. Let

$$P(R=0|S=0) = 1-\epsilon_0 \tag{66}$$
$$P(R=1|S=0) = \epsilon_0 \tag{67}$$

then $R|S=0 \sim B(\epsilon_0)$. Let

$$P(R=0|S=1) = \epsilon_1 \tag{68}$$
$$P(R=1|S=0) = 1-\epsilon_1 \tag{69}$$

then $R|S=0 \sim B(1-\epsilon_1)$. Overall,

$$R|S \sim B(\epsilon_0^{1-S}(1-\epsilon_1)^S) \tag{70}$$

A *Rademacher random variable* has two states, -1 and 1. The probability of getting each is 0.5.

A *binomial random variable* has a PMF of

$$p_X(k) = \binom{n}{k} p^k (1-p)^{n-k}, k = 0, 1, \ldots n \tag{71}$$

where $0 < p < 1$ is the binomial parameter, and n is the total number of states. We write

$$X \sim Binomial(n, p) \tag{72}$$

to say that X is drawn from a binomial distribution with a parameter p of size n. If $X \sim Binomial(n, p)$, then

$$E[X] = np \tag{73}$$

$$E[X^2] = np(np + (1-p)) \tag{74}$$

$$Var[X] = np(1-p) \tag{75}$$

Let X be a *geometric random variable*. Then the PMF of X is

$$p_X(k) = (1-p)^{k-1}p, k = 1, 2, \ldots \tag{76}$$

We write

$$X \sim Geometric(p) \tag{77}$$

to say that X was drawn from a geometric distribution with a parameter p. If $X \sim Geometric(p)$ then

$$E[X] = \frac{1}{p} \tag{78}$$

$$E[X^2] = \frac{2}{p^2} - \frac{1}{p} \tag{79}$$

$$Var[X] = \frac{1-p}{p^2} \tag{80}$$

Let X be a *Poisson random variable*. Then the PMF of X is

$$p_X(k) = \frac{\lambda^k}{k!}e^{-\lambda}, k = 0, 1, 2, \ldots \tag{81}$$

where $\lambda > 0$ is the Poisson rate. We write $X \sim Poisson(\lambda)$ to say that X was drawn from a Poisson distribution with a parameter λ. If $X \sim Poisson(\lambda)$ then

$$E[X] = \lambda \tag{82}$$

$$E[X^2] = \lambda + \lambda^2 \tag{83}$$

$$Var[X] = \lambda \tag{84}$$

For small p and large n,

$$\binom{n}{k} p^k (1-p)^{n-k} \approx \frac{\lambda^k}{k!}e^{-\lambda} \tag{85}$$

Joint distributions are higher-dimensional PDFs, PMFs, or CDFs. We write

$$f_{X_1, X_2, \ldots, X_n}(x_1, x_2, \ldots, x_n) \equiv f_{\vec{X}}(\vec{x}) \tag{86}$$

The *joint PMF* of two random variables X and Y is notated by

$$p_{X,Y}(x, y) = P[X = x \text{ and } Y = y] \tag{87}$$

and represents the probability of both.

A *marginal PMF* is defined as

$$p_X(x) = \sum_{y \in \Omega_Y} p_X(x, y) \tag{88}$$

or w.l.o.g.

$$p_Y(y) = \sum_{x \in \Omega_X} p_X(x, y) \tag{89}$$

That is, it is the joint PMF summed over one of the variables.

The *conditional PMF* is given by

$$p_{X|Y}(x|y) = \frac{p_{X,Y}(x, y)}{p_Y(y)} \tag{90}$$

Let X and Y be two discrete random variables, and let A be an event. Then

$$P[X \in A | Y = y] = \sum_{x \in A} p_{X|Y}(x|y) \tag{91}$$

$$\begin{aligned} P[X \in A] &= \sum_{x \in A} \sum_{y \in \Omega_Y} p_{X|Y}(x|y) p_Y(y) && (92) \\ &= \sum_{y \in \Omega_Y} P[X \in A | Y = y] p_Y(y) && (93) \end{aligned}$$

The *conditional expectation* of X given $Y = y$ is

$$E[X|Y = y] = \sum_x x p_{X|Y}(x|y) \tag{94}$$

The *law of total expectation* is

$$E[X] = \sum_y E[X|Y = y] p_Y(y) \tag{95}$$

If two random variables X and Y are independent, then

$$\begin{aligned} p_{X,Y} &= p_X(x) p_Y(y) && (96) \\ f_{X,Y} &= f_X(x) f_Y(y) && (97) \end{aligned}$$

If a sequence of random variables $X_1, X_2, \ldots, X_N$ are independent, then their joint PDF (or joint PMF) can be factorized as

$$f_{X_1, X_2, \ldots, X_N}(x_1, x_2, \ldots, x_N) = \prod_{n=1}^{N} f_{X_n}(x_n) \tag{98}$$

The *joint CDF* of two random variables X and Y is the function $F_{X,Y}(x, y)$ such that

$$F_{X,Y}(x, y) = P[X \leq x \cap Y \leq y] \tag{99}$$

If X and Y are discrete, then

$$F_{X,Y}(x, y) = \sum_{y' \leq y} \sum_{x' \leq x} p_{X,Y}(x', y') \tag{100}$$

For two random variables X and Y, the *marginal CDF* is

$$F_X(x) = F_{X,Y}(x, \infty) \tag{101}$$

$$F_Y(y) = F_{X,Y}(\infty, y) \tag{102}$$

Let X and Y be two random variables. The *joint expectation* is

$$E[XY] = \sum_{y \in \Omega_Y} \sum_{x \in \Omega_X} xy \times p_{X,Y}(x, y) \tag{103}$$

If X and Y are discrete, then joint expectation is also called *correlation*. This can be written in matrix form as

$$\begin{bmatrix} p_{X,Y}(x_1, y_1) & p_{X,Y}(x_1, y_2) & \dots & p_{X,Y}(x_1, y_N) \\ p_{X,Y}(x_2, y_1) & p_{X,Y}(x_2, y_2) & \dots & p_{X,Y}(x_2, y_N) \\ \vdots & \vdots & \ddots & \vdots \\ p_{X,Y}(x_N, y_1) & p_{X,Y}(x_N, y_2) & \dots & p_{X,Y}(x_N, y_N) \end{bmatrix} \tag{104}$$

then the joint expectation is

$$E[XY] = \sum_{i=1}^{N} \sum_{j=1}^{N} x_i y_j \times p_{X,Y}(x_i, y_j) \tag{105}$$

Let the matrix in Equation 104 be $\mathbf{P}$. Let

$$\vec{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{bmatrix} \tag{106}$$

$$\vec{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix} \tag{107}$$

then

$$E[XY] = \begin{bmatrix} x_1 & x_2 & \dots & x_N \end{bmatrix} \begin{bmatrix} p_{X,Y}(x_1, y_1) & p_{X,Y}(x_1, y_2) & \dots & p_{X,Y}(x_1, y_N) \\ p_{X,Y}(x_2, y_1) & p_{X,Y}(x_2, y_2) & \dots & p_{X,Y}(x_2, y_N) \\ \vdots & \vdots & \ddots & \vdots \\ p_{X,Y}(x_N, y_1) & p_{X,Y}(x_N, y_2) & \dots & p_{X,Y}(x_N, y_N) \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix} \tag{108}$$

$$= \vec{x}^T \mathbf{P} \vec{y} \tag{109}$$

$E[XY]$ is a weighted inner product between the states. $\vec{x}$ and $\vec{y}$ are the states of the random variables X and Y. Recalling that the magnitude of the inner product of $\vec{a}$ and $\vec{b}$ is $|a||b|\cos(\theta)$ and that cosine is bounded, we have

$$-1 \le \frac{E[XY]}{\sqrt{E[X^2]}\sqrt{E[Y^2]}} \le 1 \tag{110}$$

Notice that the correlation of X, Y is proportional to the covariance.

The covariance of two random variables is

$$\text{Cov}(X, Y) = E[XY] - E[X]E[Y] \tag{111}$$

While ρ is

$$\rho = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}} \tag{112}$$

Continuous Random Variables

A *continuous random variable* is analogous to the discrete case. Recall that a probability is just a size of a set. It's easy to find the size of a discrete set because you can just count elements, but for an uncountable set new methods are needed. Luckily the intution for continuous random variables is intuitive, it's still just the size of a set A relative to Ω. Formally, if each event in A is equally likely, then

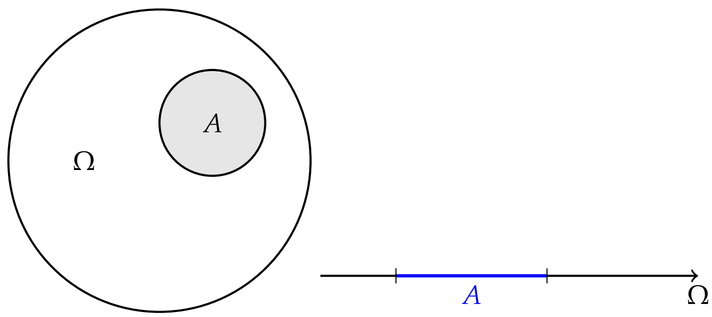

Figure 2: Continuous random variables

$$P[\{x \in A\}] = \frac{\int_A dx}{|\Omega|} \tag{113}$$

If we relax the assumption of equiprobability, then more generally

$$P[\{x \in A\}] = \int_A f_X(x)dx \tag{114}$$

$f_X(x)$ is called the *probability density function* (PDF). It is analogous to the probability mass function.

Formally, a probability density function is a mapping $f_X : \Omega \implies \Re$, with the following properties:

- Non-negativity: $f_X(x) \geq 0 \forall x \in \Omega$
- Unity: $\int_\Omega f_X(x)dx = 1$
- Measure of a set: $P[\{x \in A\}] = \int_A f_X(x)dx$

We can express a PDF in terms of a PMF with a train of delta functions like so:

$$f_X(x) = \sum_{x_k \in \Omega} p_X(x_k)\delta(x - x_k) \tag{115}$$

We can also define the probability density function as the derivative of the CDF, like so:

$$f_X(x) = \frac{d}{dx} p(X \leq x) \tag{116}$$

The expectation of a continuous random variable is

$$E[X] = \int_{\Omega} x f_X(x) dx \tag{117}$$

Properties of the expectation for continuous random variables:

- $E[aX] = aE[X]$
- $E[X + a] = E[X] + a$
- $E[aX + b] = aE[X] + b$

A random variable X has an expectation if it is absolutely integrable,

$$E[|X|] = \int_{\Omega} |x| f_X(x) dx < \infty \tag{118}$$

The variance of a continuous random variable X is

$$\begin{aligned} Var[X] &= E[(X - \mu)^2] &(119)\\ &= \int_{\Omega} (x - \mu)^2 f_X(x) dx &(120)\\ &= E[X^2] - \mu^2 &(121) \end{aligned}$$

A continuous *uniform random variable* has a PDF of

$$f_X(x) = \begin{cases} \frac{1}{b-a} & a \leq x \leq b \\ 0 & \text{else} \end{cases} \tag{122}$$

We write

$$X \sim Uniform(a, b) \tag{123}$$

to mean that X is drawn from a uniform distribution on an interval $[a, b]$. It has a CDF given by

$$F_X(x) = \begin{cases} 0 & x < a \\ \frac{x-a}{b-a} & a \leq x \leq b \\ 1 & x > b \end{cases} \tag{124}$$

If $X \sim Uniform(a, b)$ then

$$\begin{aligned} E[X] &= \frac{a+b}{2} &(125)\\ Var[X] &= \frac{(b-a)^2}{12} &(126) \end{aligned}$$

A continuous *exponential random variable* has a PDF of

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \\ 0 & \text{else} \end{cases} \tag{127}$$

An exponential random variable is the interarrival time between two consecutive Poisson events

We write

$$X \sim Exponential(\lambda) \tag{128}$$

to mean that X is drawn from an exponential distribution of parameter λ. It has a CDF given by

$$F_X(x) = 1 - e^{-\lambda x} \tag{129}$$

If $X \sim Exponential(\lambda)$, then

$$E[X] = \frac{1}{\lambda} \tag{130}$$

$$Var[X] = \frac{1}{\lambda^2} \tag{131}$$

A *Gaussian random variable* is a random variable X such that its PDF is

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \tag{132}$$

We write

$$X \sim Gaussian(\mu, \sigma^2) \tag{133}$$

or

$$X \sim \mathcal{N}\left(\mu, \sigma^2\right) \tag{134}$$

to mean that X is drawn from a Gaussian of parameter (μ, σ^2). If $X \sim \mathcal{N}(\mu, \sigma^2)$, then

$$E[X] = \mu \tag{135}$$

$$Var[X] = \sigma^2 \tag{136}$$

The *standard Gaussian* random variable has a PDF given by

$$f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \tag{137}$$

The CDF of the standard Gaussian is defined as the Φ function.

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{t^2}{2}} dt \tag{138}$$

The CDF of the standard Gaussian is related to the *error function*, which is defined as

$$\text{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt \tag{139}$$

by the relation

$$\Phi(x) = \frac{1}{2}\left(1 + \text{erf}\left(\frac{x}{\sqrt{2}}\right)\right) \tag{140}$$

The CDF of an arbitrary Gaussian is related via the transformation

$$F_X(x) = \phi\left(\frac{x-\mu}{\sigma}\right) \tag{141}$$

Let $X \sim \mathcal{N}(\mu, \sigma^2)$, then

- $\Phi(y) = 1 - \Phi(-y)$
- $P[X \geq b] = 1 - \Phi(\frac{b-\mu}{\sigma})$
- $P[|X| \geq b] = 1 - \Phi(\frac{b-\mu}{\sigma}) + \Phi(\frac{-b-\mu}{\sigma})$

In addition to mean and variance, we introduce two more useful quantities, *skewness* and *kurtosis*.

$$E[X] = \mu \tag{142}$$
$$E\left[(X-\mu)^2\right] = \sigma^2 \tag{143}$$
$$E\left[\left(\frac{X-\mu}{\sigma}\right)^3\right] = \gamma \tag{144}$$
$$E\left[\left(\frac{X-\mu}{\sigma}\right)^4\right] = \kappa \tag{145}$$

Excess kurtosis is defined as $\kappa - 3$

Skewness measures the asymmetry of a distribution. A Gaussian distribution has skewness 0. Kurtosis measures how heavy-tailed the distribution is. If the kurtosis is positive, then the tails decay faster than a Gaussian. If the kurtosis is negative, then the distribution has a tail that decays more slowly than a Gaussian.

The definition of a CDF is

$$F_X(x) = P[X \leq x] \tag{146}$$

Let X be a continuous random variable. if the CDF F_X is continuous

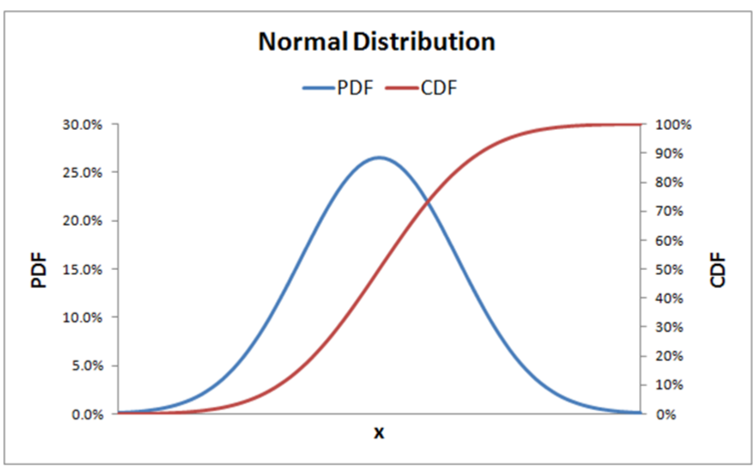

Figure 3: PDF and CDF

at any $a \leq x \leq b$, then

$$P[a \leq X \leq b] = F_X(b) - F_X(a) \tag{147}$$

A function $F_X(x)$ is said to be left continuous if at $x = b$

$$F_X(b) = \lim_{h \Longrightarrow 0} F_X(b-h) \tag{148}$$

and right continuous if

$$F_X(b) = \lim_{h \Longrightarrow 0} F_X(b+h) \tag{149}$$

and continuous if $F_X(x)$ is both left and right continuous. All CDFs are right continuous.

For any random variable X, discrete or continuous,

$$P[X = b] = \begin{cases} F_X(b) - F_X(b^-) & \text{if } F_X \text{ is discontinuous at } x = b \\ 0 & \text{else} \end{cases} \tag{150}$$

The PDf is the derivative of the CDF.

$$f_X(x) = \frac{d}{dx} \int_{-\infty}^{x} f_X(t)dt \tag{151}$$

provided F_X is differentiable at x. If not, then

$$f_X(x) = F_X(x) - \lim_{h \Longrightarrow 0} F_X(x-h) \tag{152}$$

Let X be a continuous random variable with PDF f_X. The median of X is a point $c \in \Re$ such that

$$\int_{-\infty}^{c} f_X(x)dx = \int_{c}^{\infty} f_X(x)dx \tag{153}$$

Let X be a continuous random variable. The mode is the point c such that $f_X(x)$ attains the maximum.

$$x = \text{argmax}_{x \in \Omega} f_X(x) \tag{154}$$

The mean $E[X]$ can be computed from F_X as

$$E[X] = \int_{0}^{\infty} (1 - F_X(t))dt \tag{155}$$

Recall that joint distributions are higher-dimensional PDFs, PMFs, or CDFs.

$$f_{\mathbf{X}}(\vec{x}) = f_{X_1,\ldots,X_N}(x_1, \ldots, x_n) \tag{156}$$

Let X and Y be two continuous random variables.

The *joint PDF* of X and Y is a function $f_{X,Y}(x,y)$ that can be integrated to yield a probability

$$P[A] = \int_{A} f_{X,Y}(x,y)dxdy \tag{157}$$

for any event $A \subseteq \Omega_X \times \Omega_Y$.

The *marginal PDF* is defined as

$$f_X(x) = \int_{\Omega_Y} f_{X,Y}(x,y)dy \tag{158}$$

and

$$f_Y(y) = \int_{\Omega_X} f_{X,Y}(x,y)dx \tag{159}$$

The *marginal CDF is*

$$F_X(x) = F_{X,Y}(x,\infty) \tag{160}$$
$$F_Y(y) = F_{X,Y}(\infty,y) \tag{161}$$

Let $F_{X,Y}(x,y)$ be the joint CDF of X and Y. Then the joint PDF can be obtained through

$$f_{X,Y}(x,y) = \frac{\partial^2}{\partial_y \partial_x} F_{X,Y}(x,y) \tag{162}$$

If two random variables are *independent*, then

$$p_{X,Y}(x,y) = p_X(x)p_Y(y) \tag{163}$$

and

$$f_{X,Y}(x,y) = f_X(x)f_Y(y) \tag{164}$$

If a sequence of random variables $X_1, \ldots, X_N$ are independent, then their joint PDF can be factorized.

$$f_{X_1,\ldots,X_N}(x_1,\ldots,x_n) = \prod_{n=1}^{N} f_{X_n}(x)n \tag{165}$$

A collection of random variables $X_1, \ldots, X_N$ are called *independent and identically distributed* (i.i.d.) if all are independent and have the same distribution, i.e. $f_{X_1}(x) = \cdots = f_{X_N}(x)$.

The *joint expectation* is

$$E[XY] = \int_{y\in\Omega_y}\int_{x\in\Omega_x} xyf_{X,Y}(x,y)dxdy \tag{166}$$

For an arbitrary $g(X,Y)$,

$$E[g(X,Y)] = \int_{y\in\Omega_y}\int_{x\in\Omega_x} g(x,y)f_{X,Y}(x,y)dxdy \tag{167}$$

Recall

$$\text{Cov}(X,Y) = E[XY] - E[X]E[Y] \tag{168}$$

and now we also state that

$$\text{Var}[X+Y] = \text{Var}[X] + 2\text{Cov}(X,Y) + \text{Var}[Y] \tag{169}$$

We also state that covariance is zero, then so is the correlation. However if the correlation is zero, the covariance is not necesarily zero.

$$\text{Cov}(X,Y) = 0 \implies \text{Corr}(X,y) = 0 \tag{170}$$

If X and Y are independent, then

$$E[XY] = E[X]E[Y] \tag{171}$$

This implies that X and Y are uncorrelated (i.e. $\text{Cov}(X, Y) = 0$), but the converse is not true.

Let X and Y be two continuous random variables. The *conditional PDF* of X given Y is

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)} \tag{172}$$

Let X and Y be continuous random variables and A be an event. Then

$$P[X \in A|Y = y] = \int_A f_{X|Y}(x|y)dx \tag{173}$$

$$P[X \in A] = \int_{\Omega_Y} P[X \in A|Y = y] f_Y(y)dy \tag{174}$$

The *conditional expectation* of X given $Y = y$ is

$$E[X|Y = y] = \int_{-\infty}^{\infty} x f_{X|Y}(x|y)dx \tag{175}$$

The *law of total expectation* is

$$E[X] = \int_{-\infty}^{\infty} E[X|Y = y] f_Y(y)dy \tag{176}$$

The theorem is sometimes also written

$$E[X] = E_Y[E_{X|Y}[X|Y]] \tag{177}$$

Functions

Functions of Random Variables

In general, given some random variable X, we may wish to know the properties of $Y = g(X)$, where g is a function.

To find the PDF of $Y = g(X)$, the first step is to find the CDF

$$F_Y(y) = F_X(g^{-1}(y)) \tag{178}$$

The next step is to find the PDF, given by

$$f_Y(y) = \left(\frac{d}{dy}g^{-1}(y)\right) f_X(g^{-1}(y)) \tag{179}$$

Suppose X is an exponential random variable with parameter λ, and let $Y = aX + b$. Then the CDF and PDF of Y are respectively

$$F_Y(y) = 1 - e^{-\frac{\lambda}{a}(y-b)}, y \geq b \tag{180}$$

$$f_Y(y) = \frac{\lambda}{a}e^{-\frac{\lambda}{a}(y-b)}, y \geq b \tag{181}$$

Suppose X is a uniform random variable in $[a, b]$, $a > 0$, and let $Y = X^2$. Then the CDF and PDF of Y are respectively

$$F_Y(y) = \frac{\sqrt{y} - a}{b - a}, a^2 \leq y \leq b^2 \tag{182}$$

$$f_Y(y) = \frac{1}{\sqrt{y}(b-a)}, a^2 \leq y \leq b^2 \tag{183}$$

To generate random numbers from an arbitrary distribution F_X, first generate a random number $U \sim Uniform(0, 1)$, then let $Y = F_X^{-1}(U)$. The distribution of Y is F_X.

Given two random variables X and Y, the PDF of $Z = X + Y$ is given by

$$f_Z(z) = f_X(x) * f_Y(y) \tag{184}$$

$$= \int_{-\infty}^{\infty} f_X(z-y)f_Y(y)dy \tag{185}$$

As more random variablaes are summed, their distribution (no matter the distribution) of each individual variable) approaches a Gaussian.

Let $X_1 \sim \text{Gauss}(\mu_1, \sigma_1^2)$ and $X_2 \sim \text{Gauss}(\mu_2, \sigma_2^2)$, then

$$X_1 + X_2 \sim \text{Gauss}\left(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2\right) \tag{186}$$

Given two random variables X and Y, the PDF of $Z = XY$ is given by

$$f_Z(z) = \int_{-\infty}^{\infty} \frac{1}{|y|} f_X(\frac{z}{y}) f_Y(y)dy \tag{187}$$

The PDF of $Z = X - Y$ is given by

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(z+y) f_Y(y)\, dy \tag{188}$$

The PDF of $Z = \frac{X}{Y}$ is given by

$$f_Z(z) = \int_{-\infty}^{\infty} |y|\, f_X(zy) f_Y(y)\, dy \tag{189}$$

For variables $X_1, X_2, \ldots, X_n$, all independent, let

$$Z = \prod_{i=1}^{n} X_i \tag{190}$$

The density is given recursively

$$f_Z(z) = \int_{-\infty}^{\infty} \frac{1}{|y|} f_W(\frac{z}{y}) f_{X_n}(y) dy \tag{191}$$

where $W = \prod_{i=1}^{n-1}$ and f_W is the density of the product of the first $n-1$ variables.

Moment Generating Functions

For any random variable X, the *moment generating function* (MGF) is

$$M_X(s) = E\left[e^{sX}\right] \tag{192}$$

For discrete X

$$M_X(s) = \sum_{x \in \Omega} e^{sx} p_X(x) \tag{193}$$

For continuous X

$$M_X(s) = \int_{-\infty}^{\infty} e^{sx} f_X(x) dx \tag{194}$$

MGFs have the following properties:

- $M_X(0) = 1$
- $\frac{d^k}{ds^k} M_X(s)\Big|_{s=0} = E\left[X^k\right], \quad k \in \mathcal{Z}^+$

The "moment generating" title comes from the ability to determine any order moment by evaluating the derivative at $s = 0$.

Let X and Y be independent random variables. Let $Z = X + Y$. Then

$$M_Z(s) = M_X(s) M_Y(s) \tag{195}$$

In general, let $Z = \sum_{n=1}^{N} X_n$. Then the MGF of Z is

$$M_Z(s) = \prod_{n=1}^{N} M_{X_n}(s) \tag{196}$$

Let $X_1, \ldots, X_N$ be a sequence of i.i.d. Bernoulli random variables with parameter p, then $Z = \sum_{i=1}^{N} X_i$ is a binomial random variable with parameters (N, p).

Characteristic Functions

For this course, the *characteristic function* of a random variable X is

$$\Phi_X(j\omega) = E\left[e^{-j\omega X}\right] \tag{197}$$

Law of Large Numbers

The law of large numbers is a probabilistic statement about the sample average. Suppose that we have a collection of i.i.d. random variables $X_1, \ldots, X_N$. The sample average of these N random variables is defined as follows:

$$\bar{X}_N = \frac{1}{N} \sum_{n=1}^{N} X_n \tag{198}$$

If the random variables $X_1, \ldots, X_N$ are i.i.d. so that they have the same population mean $E[X_n] = \mu$ then

$$E\left[\bar{X}_N\right] = \frac{1}{N} \sum_{n=1}^{N} E\left[X_n\right] \tag{199}$$

$$= \mu \tag{200}$$

Therefore the mean of $\bar{X}_N$ is the population mean.

If $X_1, \ldots, X_N$ have the same variance $\mathrm{Var}(\bar{X}_N)$ then

$$\mathrm{Var}(\bar{X}_N) = \frac{1}{N^2} \sum_{n=1}^{N} \mathrm{Var}(X_N) \tag{201}$$

$$= \frac{1}{N^2} \sum_{n=1}^{N} \sigma^2 \tag{202}$$

$$= \frac{\sigma^2}{N} \tag{203}$$

Therefore the variance shinks to 0 as N grows.

The *weak law of large numbers* says that if $X_1, \ldots, X_N$ is a set of i.i.d. random variables with mean μ and variance σ^2 and $E\left[X^2\right] < \infty$, then if we let

$$\bar{X}_N = \frac{1}{N} \sum_{n=1}^{N} X_n \tag{204}$$

for any $\epsilon > 0$

$$\lim_{N \to \infty} P\left[|\bar{X_N} - \mu| > \epsilon\right] = 0 \tag{205}$$

We say that a sequence of random variables $A_1, \ldots, A_N$ *converges in probability* to a deterministic number α for every $\epsilon > 0$. That is,

$$\lim_{N \to \infty} P\left[|A_N - \alpha| > \epsilon\right] = \tag{206}$$

We write $A_N \xrightarrow{P} \alpha$ to denote convergence in probability.

Central Limit Theorem

Let $\bar{X}_N$ be the sample average, and let

$$Z_N = \sqrt{N}\left(\frac{\bar{X}_N - \mu}{\sigma}\right) \tag{207}$$

be the normalized variable. The *central limit theorem* is: the CDF of Z_N converges pointwise to the CDF of Gaussian(0,1). The choice of language is extremely careful here. We are not saying that the PDF of Z_N converges to the PDF of a Gaussian, nor that the random variable Z_N converges to a Gaussian random variable. Formally, we write

$$\lim_{N\to\infty} F_{\bar{Z_N(z)}} = F_Z(z) \tag{208}$$

Consider $X_n \sim \text{Exponential}(\lambda)$, and let $X_1, \ldots, X_N$ be i.i.d. copies. Define $Z_N = \sum_{n=1}^{N} X_n$. Then

$$E[Z_N] = \frac{N}{\lambda} \tag{209}$$

and

$$\text{Var}[Z_N] = \frac{N}{\lambda^2} \tag{210}$$

Maximum Likelihood Estimation

Estimation seeks to recover an unknown parameter θ of a distribution $f_X(x;\theta)$ from observed samples $X_1,\dots,X_N$. Formally, if the forward model generates samples

$$X_1,\dots,X_N \sim f_X(\,\cdot\,;\theta), \tag{211}$$

then estimation inverts this to find θ given realizations $x_1,\dots,x_N$.

As an example,

- **Bernoulli:**

$$X_n \sim \text{Bernoulli}(\theta), \qquad p_X(x;\theta) = \theta^x(1-\theta)^{1-x},\ x \in \{0,1\}. \tag{212}$$

- **Gaussian:**

$$X_n \sim \mathcal{N}(\mu,\sigma^2), \qquad f_X(x;(\mu,\sigma^2)) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big(-\tfrac{(x-\mu)^2}{2\sigma^2}\Big). \tag{213}$$

One may treat $\theta = (\mu,\sigma^2)$ or fix one parameter and infer the other.

Likelihood and Log-Likelihood

Given i.i.d. samples $X_1,\dots,X_N$ with joint density

$$f(x_1,\dots,x_N;\theta), \tag{214}$$

the *likelihood* of θ is

$$L(\theta \mid x_1,\dots,x_N) \ = \ \prod_{n=1}^{N} f_X(x_n;\theta)\,. \tag{215}$$

The *log-likelihood* is

$$\ell(\theta) \ = \ \log L(\theta) = \sum_{n=1}^{N} \log f_X(x_n;\theta)\,. \tag{216}$$

For $X_n \sim \mathcal{N}(\mu,\sigma^2)$,

$$L(\mu,\sigma^2) = \prod_{n=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big(-\tfrac{(x_n-\mu)^2}{2\sigma^2}\Big), \tag{217}$$

so

$$\ell(\mu,\sigma^2) = -\frac{N}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2. \tag{218}$$

For $X_n \sim \text{Bernoulli}(\theta)$, let $S = \sum_{n=1}^{N} x_n$. Then

$$\ell(\theta) = S\log\theta + (N-S)\log(1-\theta). \tag{219}$$

The *maximum-likelihood* (ML) estimate maximizes the likelihood:

$$\widehat{\theta}_{\text{ML}} = \arg\max_{\theta} L(\theta \mid x_1,\dots,x_N) \tag{220}$$

$$= \arg\max_{\theta}\ \ell(\theta). \tag{221}$$

Closed-Form Solutions

- **Bernoulli:**
$$\frac{d\ell}{d\theta} = 0 \implies \widehat{\theta}_{\text{ML}} = \frac{1}{N}\sum_{n=1}^{N} x_n. \tag{222}$$

- **Gaussian Mean (known σ^2):**
$$\widehat{\mu}_{\text{ML}} = \frac{1}{N}\sum_{n=1}^{N} x_n. \tag{223}$$

- **Gaussian Variance (known μ):**
$$\widehat{\sigma}^2_{\text{ML}} = \frac{1}{N}\sum_{n=1}^{N} (x_n - \mu)^2. \tag{224}$$

Erdős-Rényi Social Network

In the single-membership Erdős-Rényi graph on N nodes, each edge indicator $X_{ij} \sim \text{Bernoulli}(p)$ independently. Let

$$S = \sum_{i=1}^{N}\sum_{j=1}^{N} x_{ij}. \tag{225}$$

The log-likelihood is

$$\ell(p) = S\log p + (N^2 - S)\log(1-p), \tag{226}$$

so the ML estimate is
$$\widehat{p}_{\text{ML}} = \frac{S}{N^2}. \tag{227}$$

Single-Photon Imaging

A 1-bit photon sensor reports

$$Y_n = \begin{cases} 1, & X_n \geq 1, \\ 0, & X_n = 0, \end{cases} \quad X_n \sim \text{Poisson}(\lambda). \tag{228}$$

Thus
$$P[Y_n = 1] = 1 - e^{-\lambda}, \quad P[Y_n = 0] = e^{-\lambda}, \tag{229}$$

and for measurements $y_1, \ldots, y_N$ with $S = \sum_n y_n$,

$$\ell(\lambda) = S\log(1 - e^{-\lambda}) - (N - S)\,\lambda. \tag{230}$$

Setting $d\ell/d\lambda = 0$ yields

$$\widehat{\lambda}_{\text{ML}} = -\ln\Big(1 - \tfrac{S}{N}\Big). \tag{231}$$

Kullback-Leibler Divergence

The *Kullback-Leibler* (KL) divergence measures how much information is lost when a model distribution Q is used to approximate a true distribution P. It is defined in the discrete case as

$$KL(P||Q) = \sum_{x \in \mathcal{X}} P(x) \log \frac{P(x)}{Q(x)} \tag{232}$$

and in the continuous case as

$$KL(P||Q) = \int_{-\infty}^{\infty} p(x) \log \frac{p(x)}{q(x)} dx \tag{233}$$

Poisson Processes

If

$$S_n = \sum_{i=1}^{n} X_i \tag{234}$$

where $X_i \sim \text{Exponential}(\lambda)$ then

$$S_n \sim \text{Gamma}(n, \lambda) \tag{235}$$

and

$$f_{S_n}(s) = \frac{\lambda^n s^{n-1} e^{-\lambda s}}{(n-1)!} \tag{236}$$

If

$$S_n = \sum_{i=1}^{n} X_i \tag{237}$$

where $X_i \sim \text{Poisson}(\lambda_i)$ then

$$S_n \sim \text{Poisson}(\sum_{i=1}^{n}, \lambda_i) \tag{238}$$

When we say that exponential random variables are memoryless, we mean that if

$$X \sim \text{Exponential}(\lambda) \tag{239}$$

then

$$P(X > t + s | X > s) = P(X > t) \tag{240}$$

A *Poisson process* can be defined thus: If $X_n \sim \text{Exponential}(\lambda)$ for $n = 1, 2, 3, \ldots$ then

$$N(t) = \max \left\{ n : \sum_{i=1}^{n} X_i \leq t \right\} \tag{241}$$

is called a Poisson process with rate λ. In more words, if the interarrival times follow an exponential distribution with rate λ, then the number of arrivals by time t is called a Poisson process with rate λ.

The number of arrivals by time s follows a Poisson distribution

$$N(s) \sim \text{Poisson}(\lambda s) \tag{242}$$

A Poisson process has two important properties:

$$N(t + s) - N(s) \sim \text{Poisson}(\lambda t) \tag{243}$$

and $N(t_1) - N(t_0), \ldots, N(t_n) - N(t_{n-1})$ are independent.

This is rather surprising. It is surprising that if $N(t)$ is a Poisson process with rate λ then $\tilde{N}(t) = N(t + r) - N(r)$ is a Poisson process with rate λ and is independent of $N(r)$.

Reference

Series

$$\sum_{k=0}^{n} r^k = \frac{1-r^{n+1}}{1-r} \tag{1}$$

$$\sum_{n=1}^{\infty} \frac{1}{n^2} = \frac{\pi^2}{6} \tag{2}$$

$$\sum_{k=1}^{\infty} k r^{k-1} = \frac{1}{(1-r)^2} \tag{3}$$

Combinatorics

$$\binom{n}{k} = \frac{n!}{k!(n-k)!} \tag{4}$$

$$(a+b)^n = \sum_{k=0}^{n} \binom{n}{k} a^{n-k} b^k \tag{5}$$

$$\binom{n}{k} + \binom{n}{k-1} = \binom{n+1}{k} \tag{6}$$

$$P(n,k) = \frac{n!}{(n-k)!} \tag{7}$$

where $P(n,k)$ is the number of ways to arrange k objects out of n (permutations).

$$C(n,k) = \binom{n}{k} = \frac{n!}{k!(n-k)!} \tag{8}$$

where $C(n,k)$ is the number of ways to choose k objects out of n (combinations).

Approximations

$$\begin{aligned} f(x) &= f(a) + f'(a)(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \dots && (9) \\ &= \sum_{n=0}^{\infty} \frac{f^{(n)}(a)}{n!}(x-a)^n && (10) \end{aligned}$$

$$\begin{aligned} 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \dots &= \sum_{k=0}^{\infty} \frac{x^k}{k!} && (11) \\ &= e^x && (12) \end{aligned}$$

$$\begin{aligned} \sin(x) &= x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \dots && (13) \\ &= \sum_{n=0}^{\infty} (-1)^n \frac{x^{2n+1}}{(2n+1)!} && (14) \end{aligned}$$

$$\cos(x) = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!} + \dots \tag{15}$$
$$= \sum_{n=0}^{\infty} (-1)^n \frac{x^{2n}}{(2n)!} \tag{16}$$

$$\ln(1+x) = x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + \dots \tag{17}$$
$$= \sum_{n=1}^{\infty} (-1)^{n+1} \frac{x^n}{n} \tag{18}$$

Calculus

$$\frac{d}{dx} \int_a^x f(t)\, dt = f(x) \tag{19}$$

$$\int_a^b f'(x)\, dx = f(b) - f(a) \tag{20}$$

$$\int f(g(x))g'(x)\, dx = \int f(u)\, du \tag{21}$$

$$\int u\, dv = uv - \int v\, du \tag{22}$$

$$\int \frac{1}{(x-a)(x-b)}\, dx = \frac{1}{b-a} \ln\left|\frac{x-a}{x-b}\right| + C \tag{23}$$

Linear Algebra

$$\vec{y} = \beta_1 \vec{x}_1 + \beta_2 \vec{x}_2 + \cdots + \beta_N \vec{x}_N \tag{24}$$

$$\langle \vec{a}, \vec{b} \rangle = \vec{a}\vec{b}^T \tag{25}$$
$$= \sum_{i=1}^{n} a_i b_i \tag{26}$$

where $\langle \vec{a}, \vec{b} \rangle$ denotes the inner product of vectors $\vec{a}$ and $\vec{b}$.

$$\|\vec{x}\|_p = \left(\sum_{i=1}^{n} |x_i|^p \right)^{1/p} \tag{27}$$

where $\|\vec{x}\|_p$ is the p-norm (or ℓ_p-norm) of vector $\vec{x}$.

$$\cos(\theta) = \frac{\langle \vec{a}, \vec{b} \rangle}{\|\vec{a}\|_2 \|\vec{b}\|_2} \tag{28}$$

where θ is the angle between vectors $\vec{a}$ and $\vec{b}$.

$$\hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \vec{y} \tag{29}$$

where $\hat{\beta}$ is the vector of least squares coefficients, $\mathbf{X}$ is the data matrix, and $\vec{y}$ is the target vector

Set Theory

$$A \setminus B = \{x \mid x \in A \text{ and } x \notin B\} \tag{30}$$

- **Commutativity:**

$$A \cup B = B \cup A \tag{31}$$
$$A \cap B = B \cap A \tag{32}$$

- **Associativity:**

$$(A \cup B) \cup C = A \cup (B \cup C) \tag{33}$$
$$(A \cap B) \cap C = A \cap (B \cap C) \tag{34}$$

- **Distributivity:**

$$A \cup (B \cap C) = (A \cup B) \cap (A \cup C) \tag{35}$$
$$A \cap (B \cup C) = (A \cap B) \cup (A \cap C) \tag{36}$$

- **Identity:**

$$A \cup \emptyset = A \tag{37}$$
$$A \cap \Omega = A \tag{38}$$

- **Complement:**

$$A \cup A^c = \Omega \tag{39}$$
$$A \cap A^c = \emptyset \tag{40}$$

Probability Laws

1. Non-negativity: $P(A) \geq 0 \forall A \in F$
2. Normalization: $P(\Omega) = 1$
3. Additivity: For any disjoint subsets $\{A_1, A_2, \ldots\}$, it holds that

$$P\left[\bigcup_{n=1}^{\infty} A_n\right] = \sum_{n=1}^{\infty} P\left[A_n\right]$$

Probability Properties

$$P\left[A \cup B\right] = P\left[A\right] + P\left[B\right] - P\left[A \cap B\right] \tag{41}$$
$$P\left[A \cup B\right] \leq P\left[A\right] + P\left[B\right] \tag{42}$$
$$A \subseteq B \implies P\left[A\right] \leq P\left[B\right] \tag{43}$$

Discrete Random Variables

$$E[g(X)] = \sum_x g(x) p_X(x) \tag{44}$$
$$E[g(X) + h(X)] = E[g(X)] + E[h(X)] \tag{45}$$
$$E[cX] = cE[X] \tag{46}$$
$$E[X + c] = E[X] + c \tag{47}$$

$$\begin{aligned} \text{Var}[X] &= E\left[(X-\mu)^2\right] && (48) \\ &= E[X^2] - (E[X])^2 && (49) \end{aligned}$$

$$P[X \in A | Y = y] = \sum_{x \in A} p_{X|Y}(x|y) \tag{50}$$

$$\begin{aligned} P[X \in A] &= \sum_{x \in A} \sum_{y \in \Omega_Y} p_{X|Y}(x|y) p_Y(y) && (51) \\ &= \sum_{y \in \Omega_Y} P[X \in A | Y = y] p_Y(y) && (52) \end{aligned}$$

$$E[X|Y = y] = \sum_x x p_{X|Y}(x|y) \tag{53}$$
$$E[X] = \sum_y E[X|Y = y] p_Y(y) \tag{54}$$
$$F_{X,Y}(x, y) = P\left[X \leq x \cap Y \leq y\right] \tag{55}$$
$$F_{X,Y}(x, y) = \sum_{y' \leq y} \sum_{x' \leq x} p_{X,Y}(x', y') \tag{56}$$
$$\text{Cov}(X, Y) = E[XY] - E[X]E[Y] \tag{57}$$
$$\rho = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}} \tag{58}$$

Continuous Random Variables

Conditions of a PDF f_X:

- Non-negativity: $f_X(x) \geq 0 \forall x \in \Omega$
- Unity: $\int_\Omega f_X(x) dx = 1$
- Measure of a set: $P[\{x \in A\}] = \int_A f_X(x) dx$

$$f_X(x) = \frac{d}{dx} p(X \leq x) \tag{59}$$

$$E[g(X)] = \int_\Omega g(x) p_X(x) \tag{60}$$
$$E[g(X) + h(X)] = E[g(X)] + E[h(X)] \tag{61}$$
$$E[cX] = cE[X] \tag{62}$$
$$E[X + c] = E[X] + c \tag{63}$$

$$\text{Var}[X] = E\left[(X-\mu)^2\right] \tag{64}$$
$$= E[X^2] - (E[X])^2 \tag{65}$$

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{t^2}{2}} dt \tag{66}$$

All CDFs are monotonically increasing, and additionally right continuous. That is,

$$F_X(b) = \lim_{h \implies 0} F_X(b+h) \tag{67}$$

$$P[X=b] = \begin{cases} F_X(b) - F_X(b^-) & \text{if } F_X \text{ is discontinuous at } x=b \\ 0 & \text{else} \end{cases} \tag{68}$$

$$f_X(x) = \frac{d}{dx}\int_{-\infty}^{x} f_X(t)dt \tag{69}$$

provided F_X is differentiable at x. If not, then

$$f_X(x) = F_X(x) - \lim_{h \implies 0} F_X(x-h) \tag{70}$$

Let X be a continuous random variable with PDF f_X. The median of X is a point $c \in \Re$ such that

$$\int_{-\infty}^{c} f_X(x)dx = \int_{c}^{\infty} f_X(x)dx \tag{71}$$

Let X be a continuous random variable. The mode is the point c such that $f_X(x)$ attains the maximum.

$$x = \text{argmax}_{x\in\Omega} f_X(x) \tag{72}$$

The mean $E[X]$ can be computed from F_X as

$$E[X] = \int_0^{\infty} (1 - F_X(t))dt \tag{73}$$

$$E[g(X,Y)] = \int_{y\in\Omega_y}\int_{x\in\Omega_x} g(x,y) f_{X,Y}(x,y)dxdy \tag{74}$$

$$\text{Var}[X+Y] = \text{Var}[X] + 2\text{Cov}(X,Y) + \text{Var}[Y] \tag{75}$$

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)} \tag{76}$$

$$P[X\in A|Y=y] = \int_A f_{X|Y}(x|y)dx \tag{77}$$

$$P[X\in A] = \int_{\Omega_Y} P[X\in A|Y=y] f_Y(y)dy \tag{78}$$

$$E[X|Y=y] = \int_{-\infty}^{\infty} x f_{X|Y}(x|y)dx \tag{79}$$

$$E[X] = \int_{-\infty}^{\infty} E[X|Y=y] f_Y(y)dy \tag{80}$$
$$= E_Y[E_{X|Y}[X|Y]] \tag{81}$$

$$\text{Cov}(X,Y) = 0 \implies \text{Corr}(X,y) = 0 \tag{82}$$

Functions of Random Variables

To find the PDF of $Y = g(X)$, the first step is to find the CDF

$$F_Y(y) = F_X(g^{-1}(y)) \tag{83}$$

The next step is to find the PDF, given by

$$f_Y(y) = \left(\frac{d}{dy}g^{-1}(y)\right) f_X(g^{-1}(y)) \tag{84}$$

Given two random variables X and Y, the PDF of $Z = XY$ is given by

$$f_Z(z) = \int_{-\infty}^{\infty} \frac{1}{|y|} f_X(\frac{z}{y}) f_Y(y) dy \tag{85}$$

The PDF of $Z = X - Y$ is given by

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(z+y) f_Y(y)\, dy \tag{86}$$

The PDF of $Z = \frac{X}{Y}$ is given by

$$f_Z(z) = \int_{-\infty}^{\infty} |y|\, f_X(zy) f_Y(y)\, dy \tag{87}$$

The PDF of $Z = X + Y$ is given by

$$\begin{aligned} f_Z(z) &= f_X(x) * f_Y(y) & (88) \\ &= \int_{-\infty}^{\infty} f_X(z-y) f_Y(y) dy & (89) \end{aligned}$$