Probability Theory Notes (2023/2024)

Griffin Reimerink

Contents

1 Combinatorics

Definition *Falling factorial, binomial coefficient*

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

When the cardinality of a set is difficult to compute, a bijection can help.

Theorem *Inclusion-exclusion formula*

$$\left|\bigcup_{i=1}^{n} X_i\right| = \sum_{1\leq i_1 \leq n} |x_i| - \sum_{1 \leq i_1 < i_2 \leq n} |X_{i_1} \cap X_{i_2}| + \ldots + (-1)^{n+1} \sum_{1 \leq i_1 < i_2 < \ldots \leq n} |X_{i_1} \cap X_{i_2} \cap \ldots \cap X_{i_n}|$$

2 Probability

2.1 Probability spaces

Definition *Sample space*

A **sample space** is a set Ω.
Any element $\omega \in \Omega$ is called an **outcome**. Any subset $A \subset \Omega$ is called an **event**.

Definition *Probablity space*

A **probability space** is a triple $(\Omega, \mathscr{A}, \mathbb{P})$:

- Ω is a sample space
- $\mathscr{A}$ is a collection of events: $\begin{cases} \text{the set of all subsets of } \Omega \text{ if } \Omega \text{ is countable} \\ \text{a certain set of subsets of } \Omega \text{ if } \Omega \text{ is uncountable} \end{cases}$
- $\mathbb{P}$ is a probability function $\mathbb{P} : \mathscr{A} \to [0,1]$ satisfying:
  - $P(\Omega) = 1$
  - $\mathbb{P}(\bigcup_{i\geq 1} A_i) = \sum_{i \geq 1} \mathbb{P}(A_i)$ for any pairwise disjoint A_1, A_2, A_3...

Theorem

Let $\Omega = \{\omega_1, \omega_2, \ldots, \omega_n$ and $p_1, p_2, \ldots, p_n \geq 0$ with $\sum p_i = 1$.
For any $A \subset \Omega$ define $\mathbb{P}(A) = \sum_{i:\omega_i \in A} p_i$. Then $\mathbb{P}$ is a probability function on Ω.

Theorem *Properties of probabilities*

1. $\mathbb{P}(\varnothing) = 0$
2. If $B_1, B_2, \ldots, B_n$ are pairwise disjoint, then $\mathbb{P}(\bigcup_{i\leq n} B_i) = \sum_{i \leq n} \mathbb{P}(B_i)$
3. $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$
4. $0 \leq \mathbb{P}(A) \leq 1 \quad \forall A \in \mathscr{A}$
5. $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$
6. if $A \subset B$, then $P(A) \leq P(B)$

Theorem *σ-sub-additivity*

For any collection of events $A_1, A_2, \ldots, A_n$, we have $\quad \mathbb{P}(\bigcup_{i \leq n} B_i) \leq \sum_{i\leq n} \mathbb{P}(B_i)$

Theorem *Uniform probability of finite sample space*

Let $(\Omega, \mathscr{A}, \mathbb{P})$ be a probability space with $|\Omega| < \infty$.
If all outcomes $\omega \in \Omega$ have the same probability, then $\mathbb{P}(A) = \dfrac{|A|}{|\Omega|}$

2.2 Events' relations

Definition *Probability of A given B*

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

Theorem *Law of total probabilities*

Let $B \in \mathscr{A}$ with $\mathbb{P}(B) > 0$. Let $A_1, A_2, \ldots$ be a partition of $\mathscr{A}$. Then

$$\mathbb{P}(B) = \sum_{i \in \mathbb{N}} \mathbb{P}(B \mid A_i) \cdot \mathbb{P}(A_i)$$

Theorem *Bayes' formula*

Let $B \in \mathscr{A}$ with $\mathbb{P}(B) > 0$. Let $A_1, A_2, \ldots$ be a partition of $\mathscr{A}$. Then

$$\mathbb{P}(A_i \mid B) = \frac{\mathbb{P}(B \mid A_i) \cdot \mathbb{P}(A_i)}{\sum_{i \in \mathbb{N}} (\mathbb{P}(B \mid A_i)\mathbb{P}(A_i))}$$

Definition *Independence*

- Events A and B are **independent** if $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$
- Events $A_1, A_2, \ldots, A_n$ are **pairwise independent** if $\mathbb{P}(A_i \cap A_j) = \mathbb{P}(A_i)\mathbb{P}(A_j) \quad \forall i \neq j$
- Events $A_1, A_2, \ldots, A_n$ are **mutually independent** if $\mathbb{P}(A_{i_1} \cap \ldots \cap A_{i_k}) = \mathbb{P}(A_{i_1}) \cdot \ldots \cdot \mathbb{P}(A_{i_k})$ for any subcollection $A_{i_1}, A_{i_2}, \ldots, A_{i_k}$

Mutual independence implies pairwise independence.

3 Random variables

i.i.d stands for "Independent and identically distributed random variables"

Definition *Random variable*

Let $(\Omega, \mathscr{A}, \mathbb{P})$ be a probablity space. A **random variable** is a function $X : \Omega \to \mathbb{R}$.
The **distribution** of a random variable X is the function $A \subset \mathbb{R} \to \mathbb{P}(X \in A)$,
where $\mathbb{P}(X \in A) = \mathbb{P}(\{\omega \in \Omega : x(\omega) \in A\}$

Definition *Cumulative distribution function (cdf)*

The **cdf** of a random variable X is the function $F_X : \mathbb{R} \to [0, 1]$ given by $F_X(x) = \mathbb{P}(X \leq x)$

Lemma

$$\mathbb{P}(X = x) = F_x(x) - \lim_{y \nearrow x} F_X(y)$$

Theorem

F_x is nondecreasing, $\lim_{x \to -\infty} F_X(x) = 0$, and $\lim_{x \to \infty} F_X(x) = 1$

Definition *Discrete and continuous variables*

A random variable is **discrete** if it takes finitely or countably many values. (F_X is piecewise constant)
A random variable is **continuous** if its cdf is continuous.

Definition *Identically distributed variables*

X and Y are **identically distributed** if they have the same distribution.

3.1 Probability mass and density functions

Definition *Probability mass function (pmf)*

The **pmf** of a discrete random variable X is given by $f_X(x) = \mathbb{P}(X = x)$

Definition *Probability density function (pdf)*

The **pdf** of a continuous random variable X is a function $f_X : \mathbb{R} \to [0, \infty)$ that satisfies

$$F_x(x) = \int_{-\infty}^{x} f_X(y)\,\mathrm{d}y \qquad \forall x \in \mathbb{R}$$

Definition *Absolute continuity*

A random variable is **absolutely continuous** if it is continuous and has a probability density function.

Notation

$$X \sim F_X \text{ means } X \text{ has cdf } = F_X \qquad X \sim f_X \text{ means } \begin{cases} X \text{ has pmf } = f_X \text{ if } X \text{ is discrete} \\ X \text{ has pdf } = f_X \text{ if } X \text{ is continuous} \end{cases}$$

3.2 Functions of random variables

Theorem

1. If $Y = g(X)$ and g is strictly increasing, then $F_Y(y) = F_X(g^{-1}(y))$
2. If $Y = g(X)$, g is strictly decreasing and X is continuous, then $F_Y(y) = 1 - F_X(g^{-1}(y))$

Lemma

Assume X is discrete. Then, Y is also discrete and $f_Y(y) = \sum\limits_{x \in g^{-1}(y)} f_x(x)$.
If g is injective, then $f_Y(y) = f_X(g^{-1}(y))$

Theorem

Assume X has pdf f_X and $Y = g(X)$ with g differentiable and strictly increasing or strictly decreasing. Then,

$$f_Y(y) = f_X(g^{-1}(y)) \cdot \left|\frac{\mathrm{d}}{\mathrm{d}y} g^{-1}(y)\right|$$

3.3 Expectation

Definition *Expected value*

Let X be a random variable.
The **expected value** (or **mean** or **expectation**) of X is:

$$\mathbb{E}(X) = \begin{cases} \sum_x x \cdot f_x(x) & \text{if } X \text{ is discrete} \\ \int_{-\infty}^{\infty} x \cdot f_x(x)\,\mathrm{d}x & \text{if } X \text{ is absolutely continuous} \end{cases}$$

Theorem *Linearity of expectation*

Let $X_1, X_2, \ldots, X_n$ be random variables. Let $a_1, a_2, \ldots, a_n \in \mathbb{R}$.
Then $\mathbb{E}[a_1 X_1 + a_2 X_2 + \ldots + a_n X_n] = a_1 \mathbb{E} X_1 + a_2 \mathbb{E} X_2 + \ldots + a_n \mathbb{E} X_n$

Theorem

Let X be a discrete random variable and $g : \mathbb{R} \to \mathbb{R}$. Then $\mathbb{E}(g(X)) = \sum_x g(x) f_X(x)$.

Theorem *Properties of expected value of functions*

Let $X, Y : \Omega \to \mathbb{R}$ be random variables, $a, b, c \in \mathbb{R}$
and $g_1, g_2 : \mathbb{R} \to \mathbb{R}$ s.t. $\mathbb{E}(g_1(X)), \mathbb{E}(g_2(X)), \mathbb{E}(g_2(Y))$ are defined. Then

Linearity:

1. $\mathbb{E}[ag_1(X) + bg_2(Y) + c] = a\mathbb{E}[g_1(X)] + b\mathbb{E}[g_2(Y)] + c$

Monotonicity:

2. If $g_1 \geq 0$, then $\mathbb{E}[g_1(X)] \geq 0$
3. If $g_1 \geq g_2$, then $\mathbb{E}[g_1(X)] \geq \mathbb{E}[g_2(X)]$
4. If $a \leq g_1 \leq b$, then $a \leq \mathbb{E}[g_1(X)] \leq b$

Theorem

1. If X is a discrete random variable taking values in $\{0, 1, \ldots\}$, then $$\mathbb{E}(X) = \sum_{n=0}^{\infty}(1 - F_X(n))$$
2. If X is a continuous nonnegative random variable, then $$\mathbb{E}(X) = \int_0^{\infty}(1 - F_X(n))\,\mathrm{d}x$$

3.4 Variance

Definition *Variance and standard deviation*

n-th moment: $\mathbb{E}(X)^n$
n-th central moment: $\mathbb{E}(X - \mathbb{E}(X))^n$
Variance: $\mathrm{Var}(X) = \mathbb{E}((X - \mathbb{E}(X))^2) = \mathbb{E}(X^2) - (\mathbb{E}(X))^2$
Standard deviation: $\sigma = \sqrt{\mathrm{Var}(X)}$

Theorem

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

3.5 Classical distributions

Definition *Classical discrete distributions*

Discrete uniform distribution:	$X \sim \mathrm{Unif}(a, b)$ if $f_x(x) = \dfrac{1}{b - a + 1}$	$a, b \in \mathbb{Z},\ a < b$
Bernoulli distribution:	$X \sim \mathrm{Ber}(p)$ if $f_x(1) = p$ and $f_x(0) = 1 - p$	$p \in [0, 1]$
Binomial distribution:	$X \sim \mathrm{Bin}(n, p)$ if $f_x(x) = \binom{n}{x} p^x (1 - p)^{n-x}$	$x = 0, 1, \ldots, n$
Geometric distribution:	$X \sim \mathrm{Geo}(p)$ if $f_x(x) = (1 - p)^{x-1} p$	$x = 1, 2, \ldots$
Poisson distribution:	$X \sim \mathrm{Poi}(\lambda)$ if $f_x(x) = \dfrac{\lambda^x}{x!} e^{-\lambda}$	$\lambda > 0,\ x = 0, 1, 2, \ldots$

For the binomial distribution, we perform n Bernoulli trials with probability p of success, where X is the number of successes. For the geometric distribution, we perform Bernoulli trials with probability p of success until the first success is obtained, where X is the number of trials.

Expectation and variance of classical discrete distributions

Discrete uniform distribution	$\mathbb{E}(X) = \frac{a+b}{2}$	$\mathrm{Var}(X) = \frac{(b-a+1)^2 - 1}{12}$
Bernoulli distribution	$\mathbb{E}(X) = p$	$\mathrm{Var}(X) = p(1 - p)$
Binomial distribution	$\mathbb{E}(X) = np$	$\mathrm{Var}(X) = np(1 - p)$
Geometric distribution	$\mathbb{E}(X) = \frac{1}{p}$	$\mathrm{Var}(X) = \frac{1-p}{p^2}$
Poisson distribution	$\mathbb{E}(X) = \lambda$	$\mathrm{Var}(X) = \lambda$

Theorem *Poisson limit theorem*

Assume $(p_n)_{n\geq 1}$ is a sequence such that $p_n \in [0,1]$ for each n and $\lim_{n\to\infty} n \cdot p_n = \lambda > 0$.
Then for each $k \geq 1$,

$$\lim_{n\to\infty} \underbrace{\binom{n}{k}(p_n)^k(1-p_n)^{n-k}}_{f_X(k) \text{ for } X\sim \text{Bin}(n,p_n)} = \underbrace{\frac{\lambda^k}{k!}e^{-\lambda}}_{f_X(k) \text{ for } X\sim \text{Poi}(\lambda)}$$

Definition *Gamma function*

$$\Gamma(a) \int_0^\infty t^{a-1} \cdot e^{-t}\,\mathrm{d}t$$

Properties of the gamma function

- $\Gamma(a+1) = a\Gamma(a)$ for any $a > 0$.
- $\Gamma(n) = (n-1)!$ for any positive integer n.

Definition *Classical continuous distributions*

Uniform distribution: $X \sim \text{ContUnif}(a,b)$ if $f_X(x) = \begin{cases} \frac{1}{b-a} \text{ if } x \in (a,b) \\ 0 \text{ otherwise.} \end{cases}$ $a < b$

Exponential distribution: $X \sim \text{Exp}(\lambda)$ if $f_X(x) = \begin{cases} \lambda e^{-\lambda x} \text{ if } x > 0 \\ 0 \text{ otherwise.} \end{cases}$ $\lambda > 0$

Gamma distribution: $X \sim \text{Gamma}(\alpha,\beta)$ if $f_X(x) = \begin{cases} \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1}e^{-\beta x} \text{ if } x > 0 \\ 0 \text{ otherwise.} \end{cases}$ $\alpha, \beta > 0$

Normal distribution: $X \sim \mathcal{N}(\mu,\sigma^2)$ if $f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$ $\mu \in \mathbb{R}, \sigma > 0$

We call $\mathcal{N}(0,1)$ the **standard normal distribution**.

Expectation and variance of classical continuous distributions

Uniform distribution	$\mathbb{E}(X) = \frac{b+a}{2}$	$\text{Var}(X) = \frac{(b-a)^2}{12}$
Exponential distribution	$\mathbb{E}(X) = \frac{1}{\lambda}$	$\text{Var}(X) = \frac{1}{\lambda^2}$
Gamma distribution	$\mathbb{E}(X) = \frac{\alpha}{\beta}$	$\text{Var}(X) = \frac{\alpha}{\beta^2}$
Normal distribution	$\mathbb{E}(X) = \mu$	$\text{Var}(X) = \sigma^2$

Theorem

If $X \sim \mathcal{N}(\mu,\sigma^2)$ and $Y = aX + b$ with $a \neq 0$, then $Y \sim \mathcal{N}(a\mu + b, a^2\sigma^2)$
If $X \sim \mathcal{N}(\mu,\sigma^2)$ and $Z = \frac{X-\mu}{\sigma} \sim \mathcal{N}(0,1)$

4 Random vectors

Definition *Random vector*

Let $n \in \mathbb{N}$. An n-dimensional **random vector** is a function from a sample space Ω into $\mathbb{R}^n$.
A random variable is a 1-dimensional random vector.
The **distribution** of a random vector is the function $A \mapsto \mathbb{P}((X_1, X_2, \ldots, X_n) \in A)$

4.1 Joint and marginal distributions

Definition *Joint cumulative distribution function*

The **joint cdf** of $(X_1,\ldots,X_n)$ is $F_{X_1,\ldots,X_n}(x_1,\ldots,x_n) = \mathbb{P}(X_1 \leq x_1, \ldots, X_n \leq x_n)$

Definition *Discrete random vectors*

A random vector $(X_1, \dots, X_n)$ is **discrete** if it takes countably many values.
The **joint pmf** of $(X_1, \dots, X_n)$ is $f_{X_1,\dots,X_n}(x_1, \dots, x_n) = \mathbb{P}(X_1 = x_1, \dots, X_n = x_n)$

Definition *Continuous random vectors*

A random vector $(X_1, \dots, X_n)$ is **continuous** if there exists a function $f_{X_1,\dots,X_n} : \mathbb{R}^n \to [0, \infty)$ such that

$$\mathbb{P}((X_1, X_2, \dots, X_n) \in A) = \int \cdots \int_A f_{X_1,\dots,X_n}(x_1, \dots, x_n)\, \mathrm{d}x_1 \dots \mathrm{d}x_n$$

This function is the **joint pdf** of $(X_1, \dots, X_n)$.

Definition *Marginals*

Let $(X_1, \dots, X_n)$ be a random vector with joint pdf $f_{X_1,\dots,X_n}$.
The pmf's $f_{X_1}, \dots, f_{X_n}$ of the (univariate) random variables $X_1, \dots, X_n$ are called the **marginal pmf's**.
We similarly define **marginal pdf's** for continuous random vectors.

Construction of marginals from a joint pmf

$$f_X(x) = \sum_y f_{X,Y}(x, y) \qquad f_Y(y) = \sum_x f_{X,Y}(x, y)$$

Construction of marginals from a joint pdf

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)\, \mathrm{d}y \qquad f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)\, \mathrm{d}x$$

Theorem *Expectation of random vectors*

Let $g : \mathbb{R}^n \to \mathbb{R}$.

$$\mathbb{E}[g(x_1, \dots, x_n)] = \begin{cases} \sum_{x_1} \cdots \sum_{x_n} g(x_1, \dots, x_n) f_{X_1,\dots,X_n}(x_1, \dots, x_n) & \text{(Discrete)} \\ \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} g(x_1, \dots, x_n) f_{X_1,\dots,X_n}(x_1, \dots, x_n)\, \mathrm{d}x_1 \dots \mathrm{d}x_n & \text{(Continuous)} \end{cases}$$

Theorem

1. If X and Y are random variables and $a, b \in \mathbb{R}$, then $\mathbb{E}[aX + bY] = a\mathbb{E}[X] + b\mathbb{E}[Y]$
2. If $\mathbb{P}(X \geq Y) = 1$, then $\mathbb{E}(X) \geq \mathbb{E}(Y)$.

4.2 Conditional distribution

Definition *Conditional pmf*

Let (X, Y) be a discrete random vector with joint pmf $f_{X,Y}$ and marginals f_X and f_Y.
The conditional pmf of X given Y is the function

$$f_{X|Y}(x \mid y) = \mathbb{P}(X = x \mid Y = y) = \frac{\mathbb{P}(X = x, Y = y)}{\mathbb{P}(Y = y)}$$

defined for all x and for all y such that $f_Y(y) > 0$.

Definition *Conditional pdf*

Let (X, Y) be a continuous random vector with joint density $f_{X,Y}(x, y)$. The conditional probability density function of X given Y is

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

defined for all x and for all y such that $f_Y(y) > 0$.

Definition *Independent random variables*

Random variables $X_1, \dots, X_n$ (defined on the same probability space) are independent if

$$\mathbb{P}(X_1 \in A_1, \dots, X_n \in A_n) = P(X_1 \in A_1) \dots \mathbb{P}(X_n \in A_n) \text{ for all } A_1, \dots A_n \subset \mathbb{R}.$$

Lemma

Random variables $X_1, \dots, X_n$ are independent if and only if

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

Lemma

Assume that there exist non-negative functions g, h such that we can write $f_{X,Y}(x, y) = g(x) \cdot h(y)$. Then X and Y are independent and

$$f_X(x) = \frac{g(x)}{\int_{-\infty}^{\infty} g(s)\,\mathrm{d}s} \qquad f_Y(y) = \frac{h(y)}{\int_{-\infty}^{\infty} h(t)\,\mathrm{d}t}$$

Definition *Conditional expectation and variance*

Let X, Y be random variables and E be an event with $\mathbb{P}(E) > 0$.

$$\mathbb{E}[X \mid E] = \frac{\mathbb{E}[X \mathbb{1}_E]}{\mathbb{P}(E)} \qquad \mathrm{Var}[X \mid E] = \mathbb{E}[X^2 \mid E] - (\mathbb{E}[X \mid E]^2)$$

Theorem

If X and Y are independent, then for any g, h we have

$$\mathbb{E}(g(X)h(Y)) = \mathbb{E}(g(X))E(h(Y)) \qquad E(XY) = E(X)E(Y) \qquad \mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$$

4.3 Transformation of vectors

Theorem

Assume X, Y are independent. Let $U = g_1(X)$ and $V = g_2(Y)$ with $g_1, g_2 : \mathbb{R} \to \mathbb{R}$
Then U and V are independent.

Lemma

If (X_1, X_2) is discrete and $(Y_1, Y_2) = g(X_1, X_2)$, then $f_{Y_1,Y_2}(y_1, y_2) = \sum_{(x_1,x_2)\in g^{-1}(y_1,y_2)} f_{X_1,X_2}(x_1, x_2)$.
If g is injective, then $f_Y(y) = f_X(g^{-1}(y))$

Theorem

Let (X_1, X_2) be a continuous random vector and let $g : D \to \mathbb{R}$ injective and differentiable, where

$$D := \{(x_1, x_2) \in \mathbb{R}^2 : f_{X_1,X_2}(x_1, x_2) \neq 0\} \qquad R = g(D) \subset \mathbb{R}^2.$$

If $h = g^{-1}$ and $(Y_1, Y_2) = g(X_1, X_2)$, then

$$f_{Y_1,Y_2}(y_1, y_2) = \begin{cases} f_{X_1,X_2}(g^{-1}(y_1, y_2)) \cdot \left| \dfrac{\partial(h_1, h_2)}{\partial(y_1, y_2)} \right| & \text{if } (y_1, y_2) \in R \\ 0 & \text{otherwise} \end{cases}$$

This theorem can be expanded to higher dimensions.

Definition *Convolution*

The distribution of $X + Y$ is called the **convolution** of (the distributions of) X and Y.

Theorem *Convolution formula*

Suppose that X and Y are independent random variables.

1. If X and Y are discrete, then $Z = X + Y$ is discrete and its pmf is
$$f_{X+Y}(z) = \sum_x f_X(x) f_y(z - x)$$
2. If X and Y are continuous, then $Z = X + Y$ is continuous and its pdf is
$$f_{X+Y}(z) = \int_{-\infty}^{\infty} f_X(x) f_y(z - x)\,\mathrm{d}x$$

4.4 Covariance and correlation

Definition *Covariance and correlation*

Let X, Y be random variables with $\mathbb{E}[X] = \mu_X$ and $\mathbb{E}[Y] = \mu_Y$.

- **Covariance** between X and Y: $\mathrm{Cov}(X, Y) = \mathbb{E}[(X - \mu_X)(Y - \mu_Y)]$
- **Correlation** between X and Y: $\rho_{X,Y} = \dfrac{\mathrm{Cov}(X, Y)}{\sigma_X \sigma_Y}$

If $\mathrm{Cov}(X, Y) = 0$ then X, Y are **uncorrelated**.

Lemma

For any random variable X, $\mathbb{P}(X = 0) = 1$ if and only if $\mathbb{E}[X^2] = 0$.

Theorem *Properties of covariance*

1. $\mathrm{Cov}(X, Y) = \mathrm{Cov}(Y, X)$
2. $\mathrm{Cov}(X, X) = \mathrm{Var}(X)$
3. $\mathrm{Cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$
4. If X, Y are independent, they are uncorrelated. (the converse is not true!)
5. $\mathrm{Cov}(aX + bY, Z) = a \,\mathrm{Cov}(X, Z) + b \,\mathrm{Cov}(Y, Z)$
6. If either of X or Y is constant, then $\mathrm{Cov}(X, Y) = 0$.
7. $|\mathrm{Cov}(X, Y)| \leq \sigma_X \sigma_Y$ (and $\rho_{X,Y} \in [-1, 1]$)
8. Assume $\sigma_X \sigma_Y > 0$. Then,
   $\mathrm{Cov}(X, Y) = \sigma_X \sigma_Y$ (and $\rho_{X,Y} = 1$) if and only if $X = aY + b$ for some $a > 0, b \in \mathbb{R}$
   $\mathrm{Cov}(X, Y) = -\sigma_X \sigma_Y$ (and $\rho_{X,Y} = -1$) if and only if $X = aY + b$ for some $a < 0, b \in \mathbb{R}$

Theorem

- $\mathrm{Cov}\left(\sum_{i \leq m} a_i X_i, \sum_{j \leq n} b_j Y_j\right) = \sum_{i \leq m} \sum_{j \leq n} a_i b_j \,\mathrm{Cov}(X_i, Y_j)$
- $\mathrm{Var}\left(\sum_{i \leq n} X_i\right) = \sum_{i \leq n} (X_i) + 2 \sum_{i \leq i < j \leq n} \mathrm{Cov}(X_i, X_j)$
- If $X_1, \ldots, X_n$ are independent, then $\mathrm{Var}\left(\sum_{i \leq n} X_i\right) = \sum_{i \leq n} \mathrm{Var}(X_i)$

5 Moment generating function

Definition *Moment generating function*

The **mgf** of a random variable is the function $M_X(t) = \mathbb{E}[e^{tX}] = \begin{cases} \sum_x e^t x f_X(x) & \text{if } X \text{ is discrete} \\ \int_{-\infty}^{\infty} e^t x f_X(x)\,\mathrm{d}x & \text{if } X \text{ is continuous} \end{cases}$

Theorem

$\mathbb{E}(X^n) = \frac{\mathrm{d}^n}{\mathrm{d}t^n} M_X(t)\Big|_{t=0}$

Theorem

$M_{aX+b}(t) = e^{bt} M_X(at)$

Theorem

If X, Y are such that $M_X(t) = M_Y(t)$ for all t in some neighborhood of 0, then $F_X = F_Y$

Theorem

If X, Y are independent, then $M_{X+Y}(t) = M_X(t) M_Y(t)$ for all $T \geq 0$

Theorem

Let $X \sim \mathcal{N}(\mu_X \sigma_X^2), Y \sim \mathcal{N}(\mu_Y \sigma_Y^2)$ be independent. Let $a, b, c \in \mathbb{R}$.
Then $aX + bY + c \sim \mathcal{N}(a\mu_X + b\mu_Y + c, a^2\sigma_1^2 + b^2\sigma_2^2$

Definition *Joint moment generating function*

The **joint mgf** of a random vector $X_1, \ldots, X_n$ is the function $M_{X_1,\ldots,X_n}(t_1, \ldots, t_n) = \mathbb{E}(e^{t_1 X_1 + \ldots + t_n X^n})$

6 Statistics

Definition *Random sample, parameter, statistic*

A **random sample** of size n is simply a sequence $X_1, \ldots, X_n$ of independent random variables, all with the same pdf or pmf $f(x)$. We thus have $f_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = \prod_{i \leq n} f(x_i)$

A **parameter** is a constant that defines the pdf/pmf.
A **statistic** is a function of a random sample, $Y = T(X_1, \ldots, X_n), T : \mathbb{R}^n \to \mathbb{R}$

Definition *Sample mean and variance*

$$\overline{X}_n = \frac{X_1 + \ldots + X_n}{n} \qquad S_n^2 = \frac{1}{n-1} \sum_{i \leq n} (X_i - \overline{X}_n)^2$$

Lemma

$$S_n^2 = \frac{1}{n-1} \sum_{i \leq n} X_i^2 - \frac{n}{n-1} \overline{X}_n^2$$

Theorem

Let $X_1, \ldots, X_n$ be independent and identically distributed with mean μ and variance σ^2. Then,

1. $\mathbb{E}(\overline{X}_n) = \mu$
2. $\mathrm{Var}(\overline{X}_n) = \frac{\mathrm{Var}\, X_1}{n}$
3. $\mathbb{E}(S_n^2) = \sigma^2$

6.1 Law of large numbers

Definition *Convergence in probability*

A sequence of r.v. $X_1, X_2, \ldots$ **converges in probability** to $c \in \mathbb{R}$ (notation: $X_n \xrightarrow[\mathbb{P}]{n\to\infty} c$) if

$$\forall \varepsilon > 0 \quad \lim_{n\to\infty} \mathbb{P}(|X_n - c| > \varepsilon) = 0$$

Definition *Consistent and unbiased estimators*

Let $X_1, \dots, X_n$ be a random sample from a pmf/pdf $f(x)$ with parameter θ.
Assume Y_n is a statistic associated to $X_1, \dots, X_n$.
We say Y_n is an **unbiased estimator** for θ if $\mathbb{E}(Y) = \theta$
We say Y_n is a **consistent estimator** for θ if Y_n converges to θ in probability as $n \to \infty$

Theorem *Weak law of large numbers*

Let $X_1, X_2, \dots$ be independent and identically distributed random variables with $\mathbb{E}(X_i) = \mu$ and $\text{Var}(X_i) = \sigma^2 \leq \infty$. Then $\overline{X}_n$ is a consistent estimator for μ.

Theorem *Markov inequality*

Let $a > 0$ and Y be any non-negative random variable. Then, $\mathbb{P}(Y \geq a) \leq \dfrac{\mathbb{E}[Y]}{a}$

Theorem *Chebyschev's inequality*

If X is any random variable and $a > 0$ then $\mathbb{P}(|X - \mathbb{E}[X]| \geq a) \leq \dfrac{\text{Var}(X)}{a^2}$

6.2 Central limit theorem

Definition *Convergence in distribution*

A sequence of r.v. $X_1, X_2, \dots$ **converges in distribution** to X (notation: $X_n \xrightarrow[(d)]{n\to\infty} X$) if

$$\lim_{n\to\infty} F_{X_n}(x) = F_X(x) \text{ for every } x \in \mathbb{R} \text{ at which } F_X \text{ is continuous}$$

Lemma

If X continuous and $X_n \xrightarrow[(d)]{n\to\infty} X$, then $\mathbb{P}(X_n = x) \xrightarrow{n\to\infty} 0 \quad \forall x \in \mathbb{R}$

Theorem

If X continuous and $X_n \xrightarrow[(d)]{n\to\infty} X$, then for every interval $I \subset \mathbb{R}$, $\lim_{n\to\infty} \mathbb{P}(X_n \in I) = \mathbb{P}(X \in I)$

Theorem *Central limit theorem*

Let $X_1, X_2, \dots$ i.i.d. with mean μ and variance σ^2 (both finite). Then,

$$\sqrt{n}\frac{\overline{X}_n - \mu}{\sigma} \xrightarrow[(d)]{n\to\infty} Z \quad \text{where } Z \sim \mathcal{N}(0,1).$$

Theorem

Assume that $X_1, X_2, \dots$ and X are such that $\lim_{n\to\infty} M_{X_n}(t) = M_X(t)$ for all t in a neighborhood of 0. Then $X_n \xrightarrow[(d)]{n\to\infty} X$.

7 Random walks

Definition *Random walk*

Let X_t denote the step a particle makes at time t.

Then $X_1, X_2, \dots$ are i.i.d. with $\begin{cases} +1 & \text{with probability } p \\ -1 & \text{with probability } 1-p \end{cases}$

The position at time t will be $S_t = \sum_{s=1}^{t} X_s$. If $p = \frac{1}{2}$, then we have a **symmetric random walk**.

Expectation and variance of random walks

$$\mathbb{E}X_i = 2p-1 \qquad \operatorname{Var} X_i = 4p(1-p) \qquad \mathbb{E}S_t = t(2p-1) \qquad \operatorname{Var} S_t = t \cdot 4p(1-p)$$

Theorem

$$\mathbb{P}(\text{The random walk will revisit the origin}) = \mathbb{P}(\exists t > 0 : S_t = 0) = \begin{cases} 1 & \text{if } p = \frac{1}{2} \\ 2(1-p) & \text{if } p > \frac{1}{2} \\ 2p & \text{if } p < \frac{1}{2} \end{cases}$$

Definition *Gambler's ruin*

The **gambler's ruin** is similar to a random walk,
but the gambler starts at i€ and ends at 0€ (☹) or a target N€ (☺)

Lemma

In the gambler's ruin, the probability that the gambler keeps playing indefinitely,
without ever reaching 0€ or N€, equals zero.

Theorem

$$\mathbb{P}(\text{the walker visits the origin infinitely many times}) = \begin{cases} 1 & \text{if } p = \frac{1}{2} \\ 0 & \text{otherwise} \end{cases}$$

8 Bivariate normal distribution

Definition *Bivariate normal distribution*

A random vector (X, Y) is said to follow a **bivariate normal distribution** $\mathcal{N}(\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \rho)$
with parameters $\mu_X \in \mathbb{R}, \mu Y \in \mathbb{R}, \sigma_X^2 > 0, \sigma_Y^2 > 0$ and $\rho \in (-1, 1)$ if

$$f_{X,Y}(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \exp\left\{ -\frac{1}{2(1-\rho^2)} \left(\left(\frac{x-\mu_X}{\sigma_X} \right)^2 + \left(\frac{y-\mu_Y}{\sigma_Y} \right)^2 - 2\rho \frac{x-\mu_X}{\sigma_X} \frac{y-\mu_Y}{\sigma_Y} \right) \right\}$$

Lemma *Linear combination*

If $Z_1 \sim \mathcal{N}(0,1)$ and $Z_2 \sim \mathcal{N}$ are independent, and $U = \sigma_1 Z_1 + \mu_1, V = \rho\sigma_2 Z_1 + \sqrt{1-\rho^2}\sigma_2 Z_2 + \mu_2$,
then $(U, V) \sim \mathcal{N}(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$

Theorem

If $(X, Y) \sim \mathcal{N}(\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \rho)$, then $X \sim \mathcal{N}(\mu_X, \sigma_X)$ $\quad Y \sim \mathcal{N}(\mu_Y, \sigma_Y)$ $\quad \rho_{X,Y} = \rho$

Corollary

If $(X, Y) \sim \mathcal{N}(\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \rho)$, then

$$aX + bY \sim \mathcal{N}(a\mu_X + b\mu_Y, a^2\sigma_X^2 + b^2\sigma_Y^2 + 2ab\rho\sigma_X\sigma_Y)$$