Lecture notes Probability Theory

Koen Oostveen

Contents

1 Probability spaces

Definition 1.1. A **sample space** is a nonempty set representing all possible outcomes of an experiment.

Axiom 1.1. *Let S be a sample space. Let $\Sigma \subseteq \mathcal{P}(S)$. Let $P : \Sigma \to [0,1]$. A triple (S, Σ, P) is then called a **probability space** (the Σ is commonly omitted) if the following axioms hold:*

1. $\forall E \in \Sigma : P(E) \geq 0$
2. $P(S) = 1$
3. *Let $(E_i)_{i \in \mathbb{N}^*}$ be a sequence in Σ. Suppose that for all $i, j \in \mathbb{N}^*$ where $i \neq j$ we have $E_i \cap E_j = \varnothing$. Then*
$$P\left(\bigcup_{i \in \mathbb{N}^*} E_i\right) = \sum_{i \in \mathbb{N}^*} P(E_i)$$

Definition 1.2. Let (S, Σ, P) be a probability space. Let $E \in \Sigma$ (E is called an **event**). We then define the set E^C, the so-called **complement** of S, by
$$E^C := S \setminus E$$

Theorem 1.1. *Let (S, Σ, P) be a probability space. Let $E, F \in \Sigma$. The following facts hold:*

- $P(E^C) = 1 - P(E)$
- $P(\varnothing) = 0$
- $0 \leq P(E) \leq 1$
- *(If $E \cap F \neq \varnothing$ then)* $P(E \cup F) = P(E) + P(F) - P(EF)$ *(notation: $EF := E \cap F$)*
- $E \subseteq F \implies P(E) \leq P(F)$

Definition 1.3. Let S be a sample space. Suppose S is finite. Let $\Sigma := \mathcal{P}(S)$. Let $P : \Sigma \to \mathbb{R}$ such that
$$P(E) := \frac{|E|}{|S|}$$
Then (S, Σ, P) is a probability space, called a **symmetric probability space**.

Remark 1.1. Symmetric probability spaces can also be interpreted using frequency: if one observes an event $E \in \Sigma$ n_e times out of n experiments, then one would like that
$$\frac{n_e}{n} \overset{n \to \infty}{\longrightarrow} P(E)$$

Definition 1.4. Let (S, Σ, P) be a probability space. Let $E, F \in \Sigma$ be events. We then define the **conditional probability** of E **given** F $P(E|F)$ by
$$P(E|F) := \frac{P(EF)}{P(F)}$$

Theorem 1.2. *The map $Q := P(\cdot|F)$ for some $F \in \Sigma$ makes the triple (S, Σ, Q) a probability space.*

Definition 1.5. Let (S, Σ, P) be a probability space. Let $E, F \in \Sigma$. We say that E and F are **independent** if
$$P(EF) = P(E)P(F)$$

Definition 1.6. Let (S, Σ, P) be a probability space. Let $E_1, \ldots, E_n \in \Sigma$, $n \in \mathbb{N}^*$. We say that these events are **pairwise independent** if for all $i, j = 1, \ldots, n$, where $i \neq j$, that E_i and E_j are independent. We say that these events are **independent** if
$$P\left(\bigcap_{i=1}^{n} E_i\right) = \prod_{i=1}^{n} P(E_i)$$

Remark 1.2. Let (S, Σ, P) be a probability space. Let $E, F \in \Sigma$. Observe that

$$\begin{aligned} P(E) &= P(EF \cup EF^C) \\ &= P(EF) + P(EF^C) \\ &= P(E|F)P(F) + P(E|F^C)P(F^C) \end{aligned}$$

This way of expressing probability in terms of conditional probability is called **conditioning**.

Theorem 1.3 (Bayes)**.** *Let (S, Σ, P) be a probability space. Let $E, F \in \Sigma$. Then*

$$P(F|E) = \frac{P(E|F)P(F)}{P(E)}$$

No proof since this a particularly easy consequence of the definitions.

2 Random variables

2.1 Definitions

Definition 2.1. Let (S, Σ, P) be a probability space. A map $X : S \to \mathbb{R}$ is called a **random variable**. We define the set $S_X := \text{im}(X) \subseteq \mathbb{R}$. Let $\Sigma_X \subseteq \mathcal{P}(S_X)$. Define the map $P_X : \Sigma_X \to \mathbb{R}$ by

$$P_X(D) := P(\{s \in S \mid X(S) \in D\})$$

Then also the triple (S_X, Σ_X, P_X) is a probability space.

Definition 2.2. Let (S, Σ, P) be a probability space. Let X be a random variable in said space. We define a map $F_X : \mathbb{R} \to \mathbb{R}$ by

$$a \mapsto P(X \leq a) := P_X((-\infty, a])$$

F_X is called the **cumulative distribution function** of X.

Theorem 2.1. *Properties of the CDF (let $a, b \in \mathbb{R}$, $a < b$):*

- F_X *is nondecreasing.*
- $\lim_{b \to -\infty} F_X(b) = 0$
- $\lim_{b \to \infty} F_X(b) = 1$
- $P(a < X \leq b) = F_X(b) - F_X(a)$
- $P(X < b) = \lim_{h \to 0^+} F_X(b - h)$
- $P(X = b) = P(X \leq b) - P(X < b) = F_X(b) - \lim_{h \to 0^+} F(b - h)$

Definition 2.3. Let (S, Σ, P) be a probability space. Let X be a random variable in said space. Suppose X is discrete, that is, S_X is finite. Then we define the **probability mass function** of X $p_x : \mathbb{R} \to \mathbb{R}$ such that

$$p_x(a) := P(X = a)$$

Theorem 2.2. *Properties of p_x:*

- *For all $a \in S_X$, $p(a) \geq 0$*
- $\sum_{x \in S_X} p(x) = 1$
- *For all $a, b \in \mathbb{R}$, $a < b$, we have* $P(a \leq X \leq b) = \sum_{x \in S_X \wedge a \leq x \leq b} p(x)$
- *For all $b \in \mathbb{R}$ we have* $F(b) := P(X \leq b) = \sum_{x \in S_X \wedge x \leq b} p(x)$
- *For all $x \in S_X$, $p(x) = F(x) - \lim_{h \to 0^+} F(x - h)$*

Note. From now on we might omit the declaration of a probability space, if it is not required for context. The following symbols will always mean the same thing from now on:

- S: any sample space
- Σ: any event space (algebra)
- P: any probability measure

2.2 Discrete random variables

Definition 2.4. Let X be a discrete random variable. Let $p \in [0,1]$. Suppose $S_X = \{0,1\}$ and $p(1) = p$, $p(0) = 1 - p$. Then X is said to be **Bernoulli distributed**.

Definition 2.5. Let X be a discrete random variable. Let $p \in [0,1]$, $n \in \mathbb{N}^*$ 'the amount of trials'. Suppose $S_X = \{0, \ldots, n\}$ and

$$p(i) = \binom{n}{i} p^i (1-p)^{n-i}$$

Then X is said to be **Binomially distributed**, and we write $X \sim B(n,p)$.

Definition 2.6. Let X be a discrete random variable. Let $p \in [0,1]$. Suppose $S_X = \mathbb{N}^*$ (X is the 'amount of experiments until 1 success') and

$$p(i) = p(1-p)^{i-1}$$

Then X is said to be **Geometrically distributed**, and we write $X \sim \text{Geo}(p)$. Some properties are:

- $P(X > i) = (1-p)^i$
- $P(X > j + i \mid X > j) = P(X > i) = (1-p)^i$

Definition 2.7. Let X be a discrete random variable. Let $\lambda > 0$ ('the amount of expected successful events'). Suppose $S_X = \mathbb{N}$ (X is the 'amount of experiments that are successful') and

$$p(i) = \exp(-\lambda)\frac{\lambda^i}{i!}$$

Then X is said to be **Poisson distributed**.
Suppose that $X \sim B(n,p)$. Let $\lambda := np$. Then

$$\lim_{n\to\infty} P(x = i) = p(i)$$

where p refers to the probability mass function of a Poisson distributed variable. That is, if you do arbitrarily many experiments, a binomially distributed variable behaves like a Poisson variable.

2.3 Continuous random variables

Definition 2.8. Let X be a **continuous** random variable, that is, S_X is not finite (and maybe also not countable). Let $f : \mathbb{R} \to \mathbb{R}$ be an (improperly) integrable function. Let $B \subseteq S_X$ be a closed interval. If we have the properties that:

- $P_X(B) = P(X \in B) = \int_B f(x)\,dx$
- $\forall x \in \mathbb{R} : f(x) \geq 0$
- $\int_{\mathbb{R}} f(x)\,dx = 1$

Then f is said to be a **probability density function** of X. If $B = [a,b]$ for some $a < b$ real numbers, it follows that

- $P(a \leq X \leq b) = \int_a^b f(x)\,dx$
- $F(b) := P(x \leq b) = \int_{-\infty}^b f(x)\,dx$
- $\forall x \in \mathbb{R} : f(x) = F'(x)$ (if f is continuous at x, which we can assume most of the time)

Definition 2.9. Let X be a continuous random variable and suppose it has probability density function f. The following will be a list of potential identities for f and associated 'labels'. Let $x \in \mathbb{R}$.

- Let $\alpha < \beta$ be real numbers. We say that X is **uniformly distributed** ($X \sim U(\alpha, \beta)$) if

$$f(x) = \begin{cases} \frac{1}{\beta - \alpha} & \text{if } \alpha < x < \beta \\ 0 & \text{otherwise} \end{cases}$$

- Let $\lambda > 0$. We say that X is **exponentially distributed** if for $x \geq 0$

$$f(x) = \lambda \exp(-\lambda x)$$

where we make the convention that if the density is unspecified somewhere, it is 0. Observe that for some $b > 0$ we have

$$\int_{-\infty}^{b} f(x)\,dx = \lambda \int_{0}^{b} \exp(-\lambda x) = \exp(-\lambda x)\Big|_{b}^{0} = 1 - \exp(-\lambda b)$$

Hence indeed

$$\int_{\mathbb{R}} f(x)\,dx = \lim_{b \to \infty} 1 - \exp(-\lambda b) = 1$$

- Let $a > 0$. Let $x \geq 0$. Let Γ be the analytic continuation of the factorial function (that is, Γ is analytic over its domain and for all $n \in \mathbb{N}$ we have $n! = \Gamma(n+1)$). Then we say that X is **Gamma distributed** (also **Erlang distributed** if $a \in \mathbb{N}^*$) if

$$f(x) = \frac{\lambda^a x^{a-1} \exp(-\lambda x)}{\Gamma(a)}$$

If we let $a \in \mathbb{N}^*$ we instead get

$$f(x) = \frac{\lambda^a x^{a-1} \exp(-\lambda x)}{(a-1)!} = \lambda \exp(-\lambda x) \frac{(\lambda x)^{a-1}}{(a-1)!}$$

- Let $\sigma > 0$. Let $\mu \in \mathbb{R}$. We say that X is **Normally distributed** ($X \sim N(\mu, \sigma^2)$) if

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

We also define the **standard normal distribution** ϕ according to $N(0,1)$, e.g.

$$\phi(x) := \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right)$$

Remark 2.1. Observe that if $X : S \to \mathbb{R}$ is a random variable such that $X \sim N(\mu, \sigma^2)$ for some $\sigma > 0$ and $\mu \in \mathbb{R}$, we can define a new random variable, Z according to $Z := \frac{X-\mu}{\sigma}$, that is, we define $Z : S \to \mathbb{R}$ such that

$$Z(s) := \frac{X(s) - \mu}{\sigma}$$

then $Z \sim N(0,1)$.

Remark 2.2. Let $a, b \in \mathbb{R}$, let X be a random variable under the above assumptions. Then $aX + b$ is also normally distributed.

2.4 Expectation

Definition 2.10. Let X be a discrete random variable. Let $S_X \subset \mathbb{R}$ be the range of X, which is assumed to be finite (by assumption of a discrete random variable). We define the **expected value** of X, EX, by

$$EX := \sum_{x \in S_X} x \underbrace{P(X = x)}_{p(x)}$$

Theorem 2.3. *Let X be a discrete random variable. Then the following assertions hold*

- *If $X \sim \text{Bernoulli}(p)$ for some $p \in (0,1)$ then $EX = p$.*
- *If $X \sim B(n,p)$ for some $n \in \mathbb{N}$ and $p \in (0,1)$ then $EX = np$.*
- *If $X \sim \text{geom}(p)$ for some $p \in (0,1)$ then $EX = \frac{1}{p}$.*
- *If $X \sim \text{Poisson}(\lambda)$ for some $\lambda > 0$ then $EX = \lambda$.*

Definition 2.11. Let X be a continuous random variable. Let $f : \mathbb{R} \to \mathbb{R}$ be the probability density function of X, which is assumed to exist. We define the **expected value** of X, EX, by

$$EX := \int_{\mathbb{R}} x f(x)\, dx$$

Theorem 2.4. *Let X be a continuous random variable. Then the following assertions hold*

- *Let $\alpha < \beta$ be real numbers and suppose $X \sim U(\alpha, \beta)$. Then $EX = \frac{1}{2}(\alpha + \beta)$.*
- *Let $\lambda > 0$ and suppose $X \sim E(\lambda)$. Then $EX = \frac{1}{\lambda}$.*
- *Let $\mu \in \mathbb{R}$ and $\sigma > 0$ and suppose $X \sim N(\mu, \sigma^2)$. Then $EX = \mu$.*

Remark 2.3. In general: if f is an even function, then $EX = 0$, or if f is symmetric around $a \in \mathbb{R}$ (that is, for all $x \in \mathbb{R}$ we have $f(a - x) = f(a + x)$) then $g(x) := f(x + a)$ is an even function $g(-x) = f(a - x) = f(a + x) = g(x)$ and hence $EX = a$.

2.4.1 Functions of random variables

Definition 2.12. Let X be a random variable. Let $S_X \subseteq \mathbb{R}$ be its range. Let $S_Y \subseteq \mathbb{R}$ be a set. Let $g : S_X \to S_Y$ be a surjective function. Let $Y : S \to \mathbb{R}$ be a random variable, such that $Y(s) := g(X(s))$. We say that such random variable Y is the **mapped random variable** of X under g. We write $Y = g(X)$ (by slight abuse of notation).

Theorem 2.5. *Let X be a discrete random variable. Let $g : S_X \to S_Y$ be a surjective function, let $Y := g(X)$. Let $y \in S_Y$. Then*

$$P(Y = y) = \sum_{x \in S_X : g(x) = y} P(X = x)$$

Theorem 2.6. *Let X be a discrete random variable. Let $g : S_X \to S_Y$ be a surjective function, let $Y := g(X)$. Then*

$$EY = \sum_{x \in S_X} g(x) p_X(x)$$

Theorem 2.7. *Let X be a continuous random variable. Let $g : \mathbb{R} \to \mathbb{R}$ be a bijective and differentiable function, let $Y := g(X)$. Then*

$$EY = \int_{\mathbb{R}} g(x) f_X(x)\, dx$$

Theorem 2.8. *Let X be a random variable. Let $a, b \in \mathbb{R}$. Then*

$$E(aX + b) = aE(X) + b$$

That is, if we define a random variable $Y : S \to \mathbb{R}$ such that $s \mapsto aX(S) + b$, $E(y) = aE(X) + b$.

Theorem 2.9. *Let (S, Σ, P) be a probability space. Let X and Y be random variables in said space. Suppose that X is continuous if and only if Y is too. Let $Z : S \to \mathbb{R}$ be defined by $s \mapsto X(s) + Y(s)$ (in other words, $Z := X + Y$). Then*

$$EZ = EX + EY$$

Definition 2.13. Let X be a random variable. We say that for some $n \in \mathbb{N}$ that its n-th **moment** is the expected value of X^n, that is, the expected value of a random variable $S \to \mathbb{R}$ such that $s \mapsto X(s)^n$. We say that for some $n \in \mathbb{N}$ that its n-th **central moment** is the value

$$E((X - E(X))^n)$$

We say that the **variance** of X, denoted by $\text{var}(X)$, is the value

$$\text{var}(X) := E((X - E(X))^2)$$

That is, the variance of X is its 2-nd central moment.
We say that the **standard deviation** σ_X of X is

$$\sigma_X := \sqrt{\text{var}(X)}$$

Theorem 2.10. *Let X be a random variable. Then*

$$\text{var}(X) = E(X^2) - (E(X))^2$$

2.5 Joint distributions

2.5.1 Joint distribution functions

Definition 2.14. Let (S, Σ, P) be a probability space. Let X and Y be discrete random variables in said space. We introduce the following notation: for $i \in S_X$ and $j \in S_Y$,

$$P(X = i, Y = j) := P(\{s \in S \mid X(s) = i \wedge Y(s) = j\})$$

(likewise for multiple variables). The function $p : S_X \times S_Y \to [0, 1]$ such that $i, j \mapsto P(X = i, Y = j)$ is called the **joint probability mass function** of X and Y.

Theorem 2.11. *Let (S, Σ, P) be a probability space. Let X and Y be discrete random variables in said space. Let $p : S_X \times S_Y$ be the joint probability mass function of X and Y. Then for all $i \in S_X$*

$$P(X = i) = \sum_{j \in S_Y} P(X = i, Y = j)$$

and for all $j \in S_Y$

$$P(Y = j) = \sum_{i \in S_X} P(X = i, Y = j)$$

Remark 2.4. Observe that we can now make statements about conditions on random variables: we can interpret the expression $P(X = i, Y = j)$ as the probability of the set $\{s \in S \mid X(s) = i \wedge Y(s) = j\}$. Hence, we can either 'define' or observe that

$$P(X = i \mid Y = j) = \frac{P(X = i, Y = j)}{P(Y = j)}$$

For a fixed $j \in S_Y$, this generates another probability mass function, called the **conditional probability mass function** of X **given** $Y = j$.
Because we have such a mass function, we can define the **conditional expectation** of X **given** $Y = j$, as follows:

$$E(X \mid Y = j) := \sum_{x \in S_X} xP(X = x \mid Y = j)$$

Similarly, we can define a **joint probability distribution function** $F : \mathbb{R} \times \mathbb{R} \to [0, 1]$ such that

$$F(x, y) := P(X \leq x, Y \leq y)$$

If X and Y are discrete, then

$$F_X(x) = F(x, b) \quad , \quad F_Y(y) = F(a, y)$$

where $a := \max S_X$, $b := \max S_Y$. If X and Y are continuous then

$$F_X(x) = \lim_{b \to \infty} F(x, b) \quad , \quad F_Y(y) = \lim_{a \to \infty} F(a, y)$$

Definition 2.15. Let (S, Σ, P) be a probability space. Let X and Y be continuous random variables in said space. A function $f : \mathbb{R}^2 \to \mathbb{R}$ is said to be the **joint probability density function** of X and Y if for all (measurable) sets $A \subseteq S_X$ and $B \subseteq S_Y$, we have

$$P(X \in A, Y \in B) = \int_B \int_A f(x, y)\, dx\, dy$$

and for all $x, y \in \mathbb{R}^2$, $f(x, y) \geq 0$.

Theorem 2.12. *Under the above assumptions for a joint probability density function f, and density functions f_X and f_Y for the variables X and Y, respectively, we have the following properties:*

- $P(X \in A) = \int_A \int_{\mathbb{R}} f(x, y)\, dy\, dx$

- $P(Y \in B) = \int_B \int_{\mathbb{R}} f(x,y)\,dx\,dy$
- $f_X(x) = \int_{\mathbb{R}} f(x,y)\,dy$
- $f_Y(y) = \int_{\mathbb{R}} f(x,y)\,dx$
- $\int_{\mathbb{R}} \int_{\mathbb{R}} f(x,y) = 1$

The proof is omitted, is trivial.

Remark 2.5. Let (S, Σ, P) be a probability space. Let X and Y be discrete random variables in said space. Let $S_Z \subseteq \mathbb{R}$ be a set and let $g : S_X \times S_Y \to S_Z$ be a surjective function. We define a new random variable $Z : S \to S_Z$, denoted by $Z := g(X,Y)$, by $s \mapsto g(X(s), Y(s))$. Observe that the probability mass function for some $z \in S_Z$ is given by (where $g^{-1}(z)$ denotes the set of preimages of z)

$$\begin{aligned} P(Z = z) &= P(g(X,Y) = z) \\ &= P((X,Y) \in g^{-1}(z)) \\ &= \sum_{(x,y) \in g^{-1}(z)} P(X = x, Y = y) \end{aligned}$$

Observe that the expectation becomes

$$\begin{aligned} EZ &= \sum_{z \in S_Z} zP(Z = z) \\ &= \sum_{z \in S_Z} \sum_{(x,y) \in g^{-1}(z)} g(x,y) P(X = x, Y = y) \\ &= \sum_{x \in S_X} \sum_{y \in S_Y} g(x,y) P(X = x, Y = y) \end{aligned}$$

Theorem 2.13. *Let (S, Σ, P) be a probability space. Let X and Y be continuous random variables in said space. Let $f : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ be the density function of X and Y. Let $S_Z \subseteq \mathbb{R}$ be a set and let $g : S_X \times S_Y \to S_Z$ be a surjective and integrable function. We define a new random variable Z as above. Then*

$$EZ = \int_{\mathbb{R}} \int_{\mathbb{R}} g(x,y) f(x,y)\,dx\,dy$$

Result does not seem that far fetched, but it is problematic to prove, as we do not have a way of expressing the density of Z, we also need the multivariable substitution law.

Remark 2.6. Using the above results we can once again easily reprove Theorem 2.9. We can also see that for random variables $X_1, \ldots, X_n$, for some $n \in \mathbb{N}$, and $a_1, \ldots, a_n \in \mathbb{R}$, we have

$$E(a_1 X_1 + \ldots + a_n X_n) = a_1 E(X_1) + \ldots a_n E(X_n)$$

Example 2.1. Let $X \sim B(n,p)$ for some $n \in \mathbb{N}$ and $p \in [0,1]$. Observe that we can define random variables $X_1, \ldots X_n$ such that for all $i = 1, \ldots, n$ we have $X_i \sim B(1,p)$, which means X_i is a single Bernoulli trial and $EX_i = p$. We also have that $X = X_1 + \ldots + X_n$. Hence

$$EX = \sum_{i=1}^{n} EX_n = np$$

2.5.2 Independent random variables

Definition 2.16. Let X and Y be discrete random variables. We say that X and Y are **independent** if for all $x \in S_X$ and $y \in S_Y$ we have

$$P(X = x, Y = y) = P(X = x) P(Y = y)$$

Definition 2.17. Let X and Y be random variables. We say that X and Y are **(mutually) independent** if for all $a, b \in \mathbb{R}$ we have

$$P(X \leq a, Y \leq b) = P(X \leq a)P(Y \leq b)$$

Remark 2.7. If X, Y are discrete random variables then the above definition is equivalent to: for all $x \in S_X$, $y \in S_Y$, we have

$$P(X = x, Y = y) = P(X = x)P(Y = y)$$

and if X, Y are continuous, then: if $f : \mathbb{R}^2 \to \mathbb{R}$ is the density function of X and Y, then for all x, y we have

$$f(x, y) = f_X(x)f_Y(y)$$

where f_X and f_Y are the density functions of X and Y, respectively.

Theorem 2.14. *Let X and Y be random variables. If X and Y are independent, then for functions $g : S_X \to \mathbb{R}$ and $h : S_Y \to \mathbb{R}$ we have*

$$E(g(X)h(Y)) = E(g(X))E(h(Y))$$

Corollary. *If X, Y are independent then*

$$E(XY) = E(X)E(Y)$$

The converse is not necessarily true.

2.5.3 Covariance

Definition 2.18. Let X and Y be random variables. We define a quantity called the **covariance** of X with Y by

$$\operatorname{Cov}(X,Y) := E((X-EX)(Y-EY))$$

Also observe that this definition is equivalent to

$$\operatorname{Cov}(X,Y) := E(XY) - E(X)E(Y)$$

A consequence of that is that if X and Y are independent, then $\operatorname{Cov}(X,Y) = 0$.

Remark 2.8. The following properties hold for random variables X, Y, Z:

- $\operatorname{Cov}(X,X) = \operatorname{var}(X)$
- $\operatorname{Cov}(X,Y) = \operatorname{Cov}(Y,X)$
- For all $c \in \mathbb{R}$, $\operatorname{Cov}(cX,Y) = c\operatorname{Cov}(X,Y)$
- $\operatorname{Cov}(X,Y+Z) = \operatorname{Cov}(X,Y) + \operatorname{Cov}(X,Z)$
- $$\begin{aligned}\operatorname{var}(X+Y) &= \operatorname{Cov}(X+Y,X+Y)\\ &= \operatorname{Cov}(X,X) + 2\operatorname{Cov}(X,Y) + \operatorname{Cov}(Y,Y)\\ &= \operatorname{var}(X) + \operatorname{var}(Y) + 2\operatorname{Cov}(X,Y)\end{aligned}$$
- $\operatorname{var}(X-Y) = \operatorname{var}(X) + \operatorname{var}(Y) - 2\operatorname{Cov}(X,Y)$

A more general result: if $X_1, \ldots, X_n$ are random variables then

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

Clearly, if those random variables are all independent, then var is linear.

Example 2.2. Let $X \sim B(n,p)$ for some $n \in \mathbb{N}$ and $p \in [0,1]$. Observe that we can define random variables $X_1, \ldots X_n$ such that for all $i = 1, \ldots, n$ we have $X_i \sim B(1,p)$, which means X_i is a single Bernoulli trial, which is independent of all others. We also have that $X = \sum_{i=1}^{n} X_i$. Then

$$\operatorname{var}(X) = \sum_{i=1}^{n} \operatorname{var}(X_i) = np(1-p)$$

Definition 2.19. Let X and Y be random variables. We define a quantity, the **correlation coefficient** of X and Y, $\rho(X,Y)$, by

$$\rho(X,Y) := \frac{\operatorname{Cov}(X,Y)}{\sigma_X \sigma_Y}$$

where $\sigma_X := \sqrt{\operatorname{var}(X)}$ and similarly for Y.

Theorem 2.15. *For two random variables X and Y, $|\rho(X,Y)| \leq 1$.*

Theorem 2.16. *Let X and Y be random variables. Then*

$$\exists a, b \in \mathbb{R} : X = aY + b \iff |\rho(X,Y)| = 1$$

(that is, X and Y are linearly correlated if their correlation coefficient is maximal in absolute value)

Remark 2.9. Consider two continuous random variables X, Y, let $a \in \mathbb{R}$, and let f be the joint probability density function. If we want to find the cumulative density of $X + Y$ at a, we compute

$$\begin{aligned} F_{X+Y}(a) &= P(X + Y \leq a) \\ &= P((X, Y) \in A := \{(x, y) \in \mathbb{R}^2 \mid x + y \leq a\}) \end{aligned}$$

Observe that such set A is a Jordan region that can also be decomposed, as such, for fixed $y \in \mathbb{R}$, we integrate over the set $(-\infty, a - y]$, hence

$$P(X + Y \leq a) = \int_{\mathbb{R}} \int_{-\infty}^{a-y} f(x, y)\, dx\, dy$$

Now if we suppose that X and Y are independent, we can simplify further. We can suppose we have marginal density functions f_X and f_Y and a cumulative density function F_X, then

$$\begin{aligned} P(X + Y \leq a) &= \int_{\mathbb{R}} \int_{-\infty}^{a-y} f_X(x) f_Y(y)\, dx\, dy \\ &= \int_{\mathbb{R}} F_X(a - y) f_Y(y)\, dy \end{aligned}$$

We can now find the density of $X + Y$:

$$\begin{aligned} f_{X+Y}(a) &= F'_{X+Y}(a) \\ &= \frac{d}{da} \int_{\mathbb{R}} F_X(a - y) f_Y(y)\, dy \\ &= \int_{\mathbb{R}} \frac{\partial}{\partial a} F_X(a - y) f_Y(y)\, dy \\ &= \int_{\mathbb{R}} f_X(a - y) f_Y(y)\, dy \\ &= (f_X * f_Y)(a) \end{aligned}$$

Written more succinctly, $f_{X+Y} = f_X * f_Y$.

2.6 Moment generating functions

Definition 2.20. Let X be a random variable. We define its **moment generating function**, $\phi_X : \mathbb{R} \to \mathbb{R}$, by

$$\phi_X(t) := E(\exp(tX))$$

Remark 2.10. Recall the notation of the n-th derivative from Analysis II: if $f : I \to \mathbb{R}$ is an n-times differentiable function at $t \in I$, where I is an open set, then we write $f^{(n)}(t)$ for its n-th derivative at t.

Theorem 2.17. *For all $n \in \mathbb{N}^*$, $\phi_X^{(n)}(0) = E(X^n)$, if it exists.*

Theorem 2.18. *Several forms of MGF:*

- *If $X \sim B(n, p)$ for some $n \in \mathbb{N}^*$ and $p \in [0, 1]$, then*

$$\phi_X(t) = (p \exp(t) + 1 - p)^n$$

- *If $X \sim \text{Poisson}(\mu)$ for $\mu > 0$ then*

$$\phi_X(t) = \exp(\mu(\exp(t) - 1))$$

- *If X is exponentially distributed with parameter $\lambda > 0$ then*

$$\phi_X(t) = \frac{\lambda}{\lambda - t} \qquad (t < \lambda)$$

- *If $X \sim N(\mu, \sigma^2)$ for some $\sigma \in [0, \infty)$ and $\mu \in \mathbb{R}$, then*

$$\phi_X(t) = \exp\left(\frac{1}{2}(\sigma t)^2 + \mu t\right)$$

Corollary. *We now also have the useful fact that for all $a, b, \mu \in \mathbb{R}$, $a \neq 0$ and $\sigma > 0$*

$$X \sim N(\mu, \sigma) \iff aX + b \sim N(a\mu + b, a^2\sigma^2)$$

Remark 2.11. Summarizing, we have the following properties for a moment generating function ϕ_X of a variable X:

- $\phi_X^{(n)}(0) = E(X^n)$
- If for another variable Y and almost every $t \in \mathbb{R}$ we have that $\phi_X(t) = \phi_Y(t)$, then $X \sim Y$, that is, the moment generating function uniquely determines the distribution if it is smooth on an open interval around 0.
- If X, Y are independent then $\phi_{X+Y}(t) = \phi_X(t)\phi_Y(t)$ for all $t \in \mathbb{R}$.

Theorem 2.19. *Let X and Y be independent random variables.*

- *Suppose $X \sim B(n, p)$ and $Y \sim B(m, p)$. Then $X + Y \sim B(n + m, p)$.*
- *Suppose $X \sim \text{Poisson}(\lambda)$ and $Y \sim \text{Poisson}(\mu)$. Then $X + Y \sim \text{Poisson}(\lambda + \mu)$.*
- *Suppose $X \sim N(\mu_1, \sigma_1^2)$ and $Y \sim N(\mu_2, \sigma_2^2)$. Then $X + Y \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$.*

2.7 Limit theorems

Theorem 2.20. *Let (S, Σ, P) be a probability space. Let X be a random variable in said space. If $X \geq 0$ for all outcomes in S, then for all $a > 0$ we have*

$$P(X \geq a) \leq \frac{1}{a} EX$$

Theorem 2.21. *Let X be a random variable. If $EX = \mu$ and $\text{var}(X) = \sigma^2$, then for all $k > 0$ we have*

$$P(|X - \mu| \geq k) \leq \frac{\sigma^2}{k^2}$$

Theorem 2.22. *Let (S, Σ, P) be a probability space. Let $(X_k)_{k \in \mathbb{N}^*}$ be a sequence of random variables in that space, independent and identically distributed (iid). Suppose for all $k \in \mathbb{N}^*$ we have $EX_k = \mu$ for some $\mu \in \mathbb{R}$. We define a new sequence of random variables, for $n \in \mathbb{N}^*$,*

$$\overline{X}_n := \frac{1}{n} \sum_{k=1}^{n} X_k$$

Then, with probability (wp) 1

$$\lim_{n \to \infty} \overline{X}_n = \mu$$

(whatever that means, to be honest).
Instead of proving this result, we will consider a more tangible / defined result: for all $\varepsilon > 0$ we have that

$$\lim_{n \to \infty} P(|\overline{X}_n - \mu| < \varepsilon) = 1$$

Remark 2.12. The former result is called the **strong law of large numbers**, while the bottom result is the **weak law of large numbers**.

Theorem 2.23 (Central limit)**.** *Let $(X_k)_{k \in \mathbb{N}^*}$ be a sequence of random variables, independent and identically distributed (iid). Suppose for all $k \in \mathbb{N}^*$ we have $EX_k = \mu$ for some $\mu \in \mathbb{R}$ and $\text{var}(X_k) = \sigma^2$ for $\sigma > 0$. Then*

$$P\left(\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \leq a\right) \to \Phi(a) \qquad (n \to \infty)$$

Remark 2.13. I do not know how to remove the assumption of analyticity here.

3 Conditional probability

3.2 Discrete random variables

Remark 3.1. Recall the definition of conditional probability: if (S, Σ, P) is a probability space and $E, F \in \Sigma$ are events, then

$$P(E|F) = \frac{P(EF)}{P(F)}$$

Recall furthermore the definition of conditional probability for some random variables X, Y and $x \in S_X$, $y \in S_Y$:

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

Definition 3.1. Let X and Y be discrete random variables. The **conditional probability mass function** of X given Y, $p_{X|Y} : S_X \times S_Y \to \mathbb{R}$ is defined by

$$p_{X|Y}(x|y) := P(X = x|Y = y) \overset{\text{recall}}{=} \frac{P(X = x, Y = y)}{P(Y = y)}$$

The **conditional probability distribution function** of X given Y is defined by

$$F_{X|Y}(x|y) := P(X \leq x|Y = y) \overset{\text{recall}}{=} \sum_{a \in S_X \wedge a \leq x} p_{X|Y}(a|y)$$

The **conditional expectation** of X given Y, $E(X|Y = y)$ for some $y \in S_Y$ is defined by

$$E(X|Y = y) := \sum_{x \in S_X} x\, p_{X|Y}(x|y)$$

3.3 Continuous random variables

Definition 3.2. Let X and Y be continuous random variables. Let $f_{X,Y} : \mathbb{R}^2 \to \mathbb{R}$ be the joint probability density function of X and Y and let $f_Y : \mathbb{R} \to \mathbb{R}$ be the probability density function of Y.

The **conditional probability density function** of X given Y, $f_{X|Y} : \mathbb{R}^2 \to \mathbb{R}$ is defined by

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

The **conditional probability distribution function** of X given Y is defined by

$$F_{X|Y}(x|y) := \int_{-\infty}^{x} f_{X|Y}(t|y)\, dt$$

The **conditional expectation** of X given Y, $E(X|Y = y)$ for some $y \in \mathbb{R}$ is defined by

$$E(X|Y = y) := \int_{-\infty}^{\infty} x f_{X|Y}(x|y)\, dx$$

3.4 Conditional expectation

Theorem 3.1. *Let X and Y be random variables. Then*

$$EX = E(E(X|Y))$$

That is, if X and Y are discrete,

$$EX = \sum_{y \in S_Y} E(X|Y = y)P(Y = y)$$

And if X and Y are continuous and f_Y is the density function of Y,

$$EX = \int_{\mathbb{R}} E(X|Y = y) f_Y(y)\, dy$$

We have already shown the discrete case.

Remark 3.2. Recall that

$$\operatorname{var}(X) := E((X - EX)^2) = E(X^2) - (EX)^2$$

Analogously,

$$\operatorname{var}(X|Y = y) := E((X - E(X|Y = y))^2|Y = y) = E(X^2|Y = y) - (E(X|Y = y))^2$$

Now we can play the same tricks as earlier, define a mapping $g : S_Y \to \mathbb{R}$ such that $y \mapsto \operatorname{var}(X|Y = y)$, then define $\operatorname{var}(X|Y) := g(Y)$. That is, we map the random variable Y to a new random variable, valued with the variance of X conditioned on Y. Observe that we then also have that

$$\operatorname{var}(X|Y) = E(X^2|Y) - (E(X|Y))^2$$

Theorem 3.2 (Law of total variance)**.** *Let X and Y be random variables. Then*

$$\operatorname{var}(X) = E(\operatorname{var}(X|Y)) + \operatorname{var}(E(X|Y))$$

3.5 Computations using conditional probability

Going back to the law of total probability, and using the same partition Q that can be summarized by $Q \ni F_y := \{Y = y\}$, but now taking an arbitrary event $E \in \Sigma$, we find that

$$P(E) = \sum_{y \in S_Y} P(E|Y = y)P(Y = y)$$

Now if Y is instead a continuous random variable, we can consider an indicator random variable, I, which is defined by

$$I(s) := \begin{cases} 1 & \text{if } s \in E, \text{ that is, } E \text{ 'occurs'} \\ 0 & \text{otherwise} \end{cases}$$

We know that $EI = P(E)$ and hence $E(I|Y = y) = P(E|Y = y)$. By the law of total expectation for continuous random variables, we have

$$EI = \int_{\mathbb{R}} E(I|Y = y) f_Y(y)\, dy$$

which is equivalent to

$$P(E) = \int_{\mathbb{R}} P(E|Y = y) f_Y(y)\, dy$$

Example 3.1. Let X and Y be continuous random variables with densities f_X and f_Y, respectively. Then

$$\begin{aligned} P(X < Y) &= \int_{\mathbb{R}} P(X < Y|Y = y) f_Y(y)\, dy \\ &= \int_{\mathbb{R}} P(X < y) f_Y(y)\, dy \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{y} f_X(x) f_Y(y)\, dx\, dy \end{aligned}$$

This result is quite intuitive as well.

5 The exponential distribution

Remark 5.1. Recall the definition of the exponential distribution. Let X be a random variable. X is said to be exponentially distributed if it is continuous and there exists some $\lambda \in \mathbb{R}$ such that its density function $f : \mathbb{R} \to \mathbb{R}$ is given for all $x \in \mathbb{R}$ by

$$f(x) = \begin{cases} \lambda \exp(-\lambda x) & \text{if } x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

Observe that its distribution function $F(x) := P(X \leq x)$ can be computed by

$$\begin{aligned} F(x) &= \int_{-\infty}^{x} f(t)\, dt \\ &= \lambda \int_{0}^{x} \exp(-\lambda t)\, dt && \text{(if } x \geq 0\text{)} \\ &= -\exp(-\lambda t)\Big|_{0}^{x} \\ &= \begin{cases} 1 - \exp(-\lambda x) & \text{if } x \geq 0 \\ 0 & \text{otherwise} \end{cases} \end{aligned}$$

Observe that its expectation is

$$EX = \lambda \int_{0}^{\infty} x \exp(-\lambda x) = \ldots = \frac{1}{\lambda} \qquad \text{(use IBP)}$$

Its variance is

$$\operatorname{var}(X) = E(X^2) - (EX)^2 = \ldots = \frac{1}{\lambda^2} \qquad \text{(use IBP)}$$

And as we have shown earlier, its MGF $\phi : (-\infty, \lambda) \to \mathbb{R}$ is (undefined for $t \geq \lambda$)

$$\phi(t) = \frac{\lambda}{\lambda - t}$$

Theorem 5.1. *Let X be an exponentially distributed variable. Then X is **memoryless**, that is, for all $s, t \geq 0$ real numbers, we have*

$$P(X > s + t | X > t) = P(X > s)$$

Remark 5.2. This also holds for geometrically distributed variables, this is readily verified.

Theorem 5.2. *Let X be an memoryless random variable. Then*

$$X \text{ continuous} \implies X \text{ exponentially distibuted}$$

$$X \text{ discrete} \implies X \text{ geometrically distibuted}$$

Since the converses have also already been shown, these are equivalences.

Definition 5.1. Let X be a continuous random variable. Suppose $F(t) := P(X \leq t)$ is the distribution function of X and suppose it is differentiable everywhere, such that its density f is indeed the derivative of F, e.g. X has continuous density. Furthermore suppose that $P(X \geq 0) = 1$.
We define the **hazard rate** r of X near a value $t \geq 0$ by

$$r(t) := \frac{f(t)}{1 - F(t)}$$

Remark 5.3. If $X \sim \operatorname{Exp}(\lambda)$ for some $\lambda > 0$ we have that $r(t) = \lambda$ for all $t \geq 0$.

Theorem 5.3. *Let X be a continuous random variable such that the hazard rate is defined. Then its distribution is uniquely determined by its hazard rate. In other words, if Y is another such random variable, then*

$$X \sim Y \iff r_X = r_Y$$

Definition 5.2. A continuous random variable X is said to be **hyperexponentially distributed** if its density is a linear combination of finitely many exponential densities, that is, there exists $\lambda_1, \ldots, \lambda_n \in \mathbb{R}$ and $p_1, \ldots, p_n \in [0,1]$ for some $n \in \mathbb{N}^*$ such that $\sum_{i=1}^{n} p_i = 1$, and the density f of X is given for all $x \geq 0$ by

$$f(x) := \sum_{i=1}^{n} p_i \lambda_i \exp(-\lambda_i x)$$

Note that we can omit the inclusion of these p_i values, this is a convention to stress the meaning: that p_i is the probability that some total variable takes the i-th exponential distribution.

Theorem 5.4. *Let (S, Σ, P) be a probability space. Let $n \in \mathbb{N}$. Let $X_1, \ldots, X_n$ be random variables in that space, i.i.d., such that there exists a $\lambda > 0$ such that for all $i = 1, \ldots, n$, we have $X_i \sim \mathrm{Exp}(\lambda)$. Let $S_n := X_1 + \ldots + X_n$. Then $S_n \sim \mathrm{Gamma}(n, \lambda)$, that is, its density function f_{S_n} is given by (for all $t \in \mathbb{R}$)*

$$f_{S_n}(t) = \begin{cases} \lambda \exp(-\lambda t) \frac{(\lambda t)^{n-1}}{(n-1)!} & \text{if } t \geq 0 \\ 0 & \text{otherwise} \end{cases}$$