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*
$$P\left(\bigcup_{i \in \mathbb{N}^*} E_i\right) = \sum_{i \in \mathbb{N}^*} P(E_i)$$

Remark 1.1. There are more restrictions on the event space $\Sigma \subseteq \mathcal{P}(S)$. Technically, (S, Σ) should constitute a **measurable space**, and Σ should be a σ-algebra over S. These conditions exist to make our definitions make sense, because although intuitive, they might be undefined for general sets (e.g. picking $\Sigma = \mathcal{P}(S)$ may not work, certain sets might not be able to be given a consistent measure). These conditions are:

- $S \in \Sigma$
- $\forall E \in \Sigma : S \setminus E \in \Sigma$
- Let $\mathcal{S} \subseteq \Sigma$ be countable. Then $\bigcup \mathcal{S} \in \Sigma$. Equivalently, let $(E_i)_{i \in \mathbb{N}}$ be a sequence in Σ. Then $\bigcup_{i=0}^{\infty} E_i \in \Sigma$. This is equivalent, because countable means that there exists a bijection with $\mathbb{N}$, which in particular is a sequence.

This is why we can even find $P(S)$ and the countable unions in the above definition, or always find a complement event as we will see shortly.

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: $EF := E \cap F$)*
- $E \subseteq F \implies P(E) \leq P(F)$

Proof.
- Observe that $1 = P(S) = P(E \cup E^C) = P(E) + P(E^C)$, hence $P(E^C) = 1 - P(E)$.
- $P(\varnothing) = P(S^C) = 1 - P(S) = 1 - 1 = 0$
- Observe that $P(E) = 1 - P(E^C) \leq 1 - 0 = 1$. The other inequality is trivial.
- $P(E \cup F) = P(E \cup E^C F) = P(E) + P(E^C F) = P(E) + P(F) - P(EF)$, since
$$P(F) = P(EF \cup E^C F) = P(EF) + P(E^C F)$$
- $P(F) = P(E \cup E^C F) = P(E) + P(E^C F) \geq P(E)$, hence $P(E) \leq P(F)$

□

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**.

Proof. Observe that

- $P(E) \geq 0$ trivially for all $E \in \Sigma$
- $P(S) = \frac{|S|}{|S|} = 1$
- Let $E_1, \ldots, E_n \in \Sigma$. Suppose they are mutually disjoint. Then

$$\left| \bigcup_{i \in \mathbb{N}^*} E_i \right| = \sum_{i=1}^{n} |E_i|$$

  Hence the third axiom is also clearly satisfied.

□

Remark 1.2. 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.*

Proof.
- Clear
- $P(S|F) = \frac{P(SF)}{P(F)} = \frac{P(F)}{P(F)} = 1$
- Let $(E_i)_{i \in \mathbb{N}^*}$ be a sequence in Σ, where we again have mutually disjoint sets. Observe that

$$P\left(\bigcup_{i \in \mathbb{N}^*} E_i | F \right) := \frac{P(\bigcup_{i \in \mathbb{N}^*} FE_i)}{P(F)} = \sum_{i \in \mathbb{N}^*} \frac{P(FE_i)}{P(F)}$$

□

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 1.3. 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.

Proof.

- Trivial
- $P_X(S_X) = P(S) = 1$
- Let $(E_i)_{i \in \mathbb{N}^*}$ be a sequence in Σ_X. Suppose that for all $i, j \in \mathbb{N}^*$ where $i \neq j$ we have $E_i \cap E_j = \varnothing$. Then

$$\begin{aligned} P_X\left(\bigcup_{i \in \mathbb{N}^*} E_i\right) &= P\left(\left\{s \in S \mid X(S) \in \bigcup_{i \in \mathbb{N}^*} E_i\right\}\right) \\ &= P\left(\bigcup_{i \in \mathbb{N}^*} \{s \in S \mid X(S) \in E_i\}\right) \\ &= \sum_{i \in \mathbb{N}^*} P(\{s \in S \mid X(S) \in E_i\}) \\ &= \sum_{i \in \mathbb{N}^*} P_X(E_i) \end{aligned}$$

□

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

Proof. It suffices to show that $f_X = f_Z$, where those are the density functions of X and Z, respectively. We could similarly show that for all $z \in \mathbb{R}$ we have that

$$P(Z \leq z) = \Phi(z)$$

where

$$\Phi(z) := \int_{-\infty}^{z} \phi(t)\, dt$$

Observe that

$$\begin{aligned} P(Z \leq z) &= P\left(\frac{X-\mu}{\sigma} \leq z\right) \\ &= P(X \leq \sigma z + \mu) \\ &= \int_{-\infty}^{\sigma z+\mu} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right) \\ &= \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}t^2\right) dt \\ &= \Phi(z) \end{aligned}$$

□

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 $X \sim$ 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 $X \sim$ geom(p) for some $p \in (0, 1)$ then $EX = \frac{1}{p}$.*
- *If $X \sim$ Poisson(λ) for some $\lambda > 0$ then $EX = \lambda$.*

Proof. Let p be the mass function of X.

- Let $p \in (0, 1)$ and suppose $X \sim$ Bernoulli(p), that is, $S_X = \{0, 1\}$, $p(0) = 1 - p$, $p(1) = p$, then

$$EX = 0 \cdot (1 - p) + 1 \cdot p = p$$

- Let $n \in \mathbb{N}$, $p \in (0,1)$, and suppose $X \sim B(n,p)$, that is, $S_X = \{0, \dots, n\}$ and

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

Observe that

$$\begin{aligned} EX &= \sum_{i=0}^{n} i \binom{n}{i} p^i (1-p)^{n-i} \\ &= \sum_{i=1}^{n} i \frac{n!}{(n-i)!i!} p^i (1-p)^{n-i} \\ &= np \sum_{i=1}^{n} \frac{(n-1)!}{(n-i)!(i-1)!} p^{i-1} (1-p)^{n-i} \\ &= np \sum_{i=0}^{n} \frac{(n-1)!}{(n-1-i))!i!} p^i (1-p)^{n-1-i} \\ &= np(p+1-p)^n \\ &= np \end{aligned}$$

- Let $p \in (0,1)$ and suppose $X \sim \text{geom}(p)$, that is, $S_X = \mathbb{N}^*$ and

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

Observe that EX is a power series:

$$EX = \sum_{i=1}^{\infty} i(1-p)^{i-1} p$$

Its coefficients are $a_k := k$, $k \in \mathbb{N}^*$. Well, its radius of convergence R is given by

$$R = \frac{1}{\overline{\lim}_{n \to \infty} \sqrt[n]{|a_n|}} = \frac{1}{1} = 1$$

Hence this series converges uniformly in $1-p$ if $|1-p| < 1$, which is assumed. Now observe that

$$\begin{aligned} EX &= \sum_{i=1}^{\infty} i(1-p)^{i-1} p \\ &= p \sum_{i=0}^{\infty} \frac{d}{dt} t^i \Big|_{t=1-p} \\ &= p \frac{d}{dt} \sum_{i=0}^{\infty} t^i \Big|_{t=1-p} \\ &= p \frac{d}{dt} \frac{1}{1-t} \Big|_{t=1-p} \\ &= p \frac{1}{(1-t)^2} \Big|_{t=1-p} \\ &= p \frac{1}{p^2} = p \end{aligned}$$

- Let $\lambda > 0$ and suppose $X \sim \text{Poisson}(\lambda)$, that is, $S_X = \mathbb{N}$ and

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

Observe that

$$\begin{aligned} EX &= \sum_{i=0}^{\infty} i \exp(-\lambda) \frac{\lambda^i}{i!} \\ &= \sum_{i=1}^{\infty} \exp(-\lambda) \frac{\lambda^i}{(i-1)!} \\ &= \lambda \exp(-\lambda) \sum_{i=1}^{\infty} \frac{\lambda(i-1)}{(i-1)!} \\ &= \lambda \exp(-\lambda) \sum_{i=0}^{\infty} \frac{\lambda^i}{i!} \\ &= \lambda \exp(-\lambda) \exp(\lambda) \\ &= \lambda \end{aligned}$$

□

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 $EX = \mu$.*

Proof. Let f be the probability density function of X.

- Observe that

$$\begin{aligned} EX &= \int_{\mathbb{R}} x f(x)\, dx \\ &= \int_{\alpha}^{\beta} \frac{x}{\beta - \alpha}\, dx \\ &= \frac{1}{\beta - \alpha} \frac{1}{2} x^2 \Big|_{\alpha}^{\beta} \\ &= \frac{1}{2} \frac{1}{\beta - \alpha} (\beta^2 - \alpha^2) \\ &= \frac{1}{2} \frac{1}{\beta - \alpha} (\alpha + \beta)(\beta - \alpha) \\ &= \frac{1}{2}(\alpha + \beta) \end{aligned}$$

- Recall that for $x \geq 0$ we have

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

Observe that

$$\begin{aligned} EX &= \int_{\mathbb{R}} xf(x)\,dx \\ &= \int_0^\infty x\lambda\exp(-\lambda x)\,dx \\ &= -x\exp(-\lambda x)\Big|_0^\infty + \int_0^\infty \exp(-\lambda x)\,dx \\ &= 0 - 0 - \frac{1}{\lambda}\exp(-\lambda x)\Big|_0^\infty \\ &= \frac{1}{\lambda} \end{aligned}$$

- Recall that

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

Observe that

$$\begin{aligned} EX &= \int_{\mathbb{R}} xf(x)\,dx \\ &= \int_{\mathbb{R}} x\frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right)dx \\ &= \int_{\mathbb{R}} (x+\mu)\frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{1}{2}\left(\frac{x}{\sigma}\right)^2\right)dx \\ &= \int_{\mathbb{R}} x\frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{1}{2}\left(\frac{x}{\sigma}\right)^2\right)dx + \mu\underbrace{\int_{\mathbb{R}}\frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{1}{2}\left(\frac{x}{\sigma}\right)^2\right)dx}_{1} \\ &= -\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right)\Big|_{-\infty}^{\infty} + \mu = \mu \end{aligned}$$

□

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 $EX = a$.

Proof. Let f be an even density function, that is, for all $x \in \mathbb{R}$, $f(x) = f(-x)$. Observe that

$$\begin{aligned} EX &= \int_{-\infty}^{\infty} xf(x)\,dx \\ &= \int_{-\infty}^{0} xf(x)\,dx + \int_0^\infty xf(x)\,dx \\ &= \int_0^\infty -xf(-x)\,dx + \int_0^\infty xf(x)\,dx \\ &= -\int_0^\infty xf(x)\,dx + \int_0^\infty xf(x)\,dx = 0 \end{aligned}$$

Observe now that if f is symmetric around a we have

$$\begin{aligned} &\int_{\mathbb{R}} xf(a-x)\,dx = \int_{\mathbb{R}} xg(-x) = 0 \\ \implies &\int_{\mathbb{R}} (a-t)f(t)\,dt = a\underbrace{\int_{\mathbb{R}} f(t)\,dt}_{1} - \int_{\mathbb{R}} tf(t)\,dt = 0 \\ \implies &EX = a \end{aligned}$$

□

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*

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

Proof.

$$\begin{aligned}
\sum_{x \in S_X : g(x) = y} P(X = x) &= \sum_{x \in S_X : g(x) = y} P_X(\{x\}) \\
&= P_X\Big(\bigcup \{\{x\} \in \mathcal{P}(S_X) \mid g(x) = y\}\Big) \\
&= P_X(\{x \in S_X \mid g(x) = y\}) \\
&= P(\{s \in S \mid X(s) \in \{x \in S_X \mid g(x) = y\}\}) \\
&= P(\{s \in S \mid g(X(s)) = y\}) \\
&= P(\{s \in S \mid Y(s) = y\}) \\
&= P(Y = y)
\end{aligned}$$

□

Remark 2.4. A 'closed form' like this does not really exist for a continuous random variable. Rather, we can look at the CDF, and from there make statements about the PDF as well. Let X be a continuous random variable, let $g : \mathbb{R} \to \mathbb{R}$ be a function and let $Y = g(X)$ be a mapped random variable. Then Y is also a continuous random variable. Its CDF F_Y can be computed with:

$$F_Y(y) = P(Y \leq y) = P(g(X) \leq y)$$

Its PDF can then in turn be computed with

$$f_Y(y) = F'_Y(y) = \frac{d}{dy} F_Y(y)$$

Example 2.1. Let $X \sim N(0, 1)$, let $Y = |X|$. Observe that

$$\begin{aligned}
F_Y(y) &= P(Y \leq y) \\
&= P(|X| \leq y) \\
&= P(-y \leq X \leq y) && \text{(if } y \geq 0\text{)} \\
&= P(X \leq y) - P(X \leq -y) \\
&= \Phi(y) - \Phi(-y)
\end{aligned}$$

And for $y < 0$, we have that $P(|X| \leq y) = P(\varnothing) = 0$. Observe that furthermore

$$\begin{aligned}
\Phi(-y) &= \int_{-\infty}^{-y} \phi(t)\,dt \\
&= \int_{-\infty}^{\infty} \phi(t)\,dt - \int_{-y}^{\infty} \phi(t)\,dt \\
&= 1 - \int_{-\infty}^{y} \phi(-t)\,dt \\
&= 1 - \int_{-\infty}^{y} \phi(t)\,dt \\
&= 1 - \Phi(y)
\end{aligned}$$

Hence

$$\begin{aligned}
F_Y(y) &= \Phi(y) - \Phi(-y) \\
&= 2\Phi(y) - 1
\end{aligned}$$

And lastly

$$f_Y(y) = \begin{cases} 2\phi(y) & \text{if } y \geq 0 \\ 0 & \text{if } y < 0 \end{cases}$$

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*

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

Proof.

$$\begin{aligned}
EY &= \sum_{y \in S_Y} y p_Y(y) \\
&= \sum_{y \in S_Y} y \sum_{x \in S_X : g(x) = y} P(X = x) \\
&= \sum_{y \in S_Y} \sum_{x \in S_X : g(x) = y} g(x) P(X = x) \\
&= \sum_{x \in S_X} g(x) P(X = x)
\end{aligned}$$

□

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*

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

Proof.

$$\begin{aligned}
EY &= \int_{\mathbb{R}} y f_Y(y)\,dy \\
&= \int_{\mathbb{R}} g(x) f_Y(g(x)) g'(x)\,dx \\
&= \int_{\mathbb{R}} g(x) F_Y'(g(x)) g'(x)\,dx \\
&= \int_{\mathbb{R}} g(x) \frac{d}{dx} F_Y(g(x))\,dx \\
&= \int_{\mathbb{R}} g(x) \frac{d}{dx} P(g(X) \leq g(x))\,dx \\
&= \int_{\mathbb{R}} g(x) \frac{d}{dx} P(X \leq x)\,dx \\
&= \int_{\mathbb{R}} g(x) \frac{d}{dx} F_X(x)\,dx \\
&= \int_{\mathbb{R}} g(x) f_X(x)\,dx
\end{aligned}$$

□

Remark 2.5. This proof, however, only works if the inverse of g is increasing, which in turn means that since g is assumed to be continuous, we need that g is increasing, too. It turns out that this proof can be done in more generality, but it is a hassle. Even the bijectivity can be dropped. The proof rougly speaking goes by observing that for some (measurable) set $D \subseteq \mathbb{R}$

$$P_Y(D) = \int_D f_Y(x)\,dx = \int_{g^{-1}(D)} f_X(x)\,dx = \int_D (f_X \circ g^{-1})(y)(g^{-1})'(y)\,dy$$

Hence

$$f_Y(y) = (f \circ g^{-1})(y)(g^{-1})'(y)$$

And thus

$$\begin{aligned}
EY &= \int_{\mathbb{R}} y(f_X \circ g^{-1})(y)(g^{-1})'(y)\,dy \\
&= \int_{\mathbb{R}} g(x) f_X(x)(g^{-1})'(g(x)) g'(x)\,dx \\
&= \int_{\mathbb{R}} g(x) f_X(x) \frac{1}{g'(x)} g'(x)\,dx \\
&= \int_{\mathbb{R}} g(x) f_X(x)\,dx
\end{aligned}$$

This is still not the most general since we apparently also do not need bijectivity and differentiability, just surjectivity of g. The search term is the 'law of the unconscious statistician'.

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 $s \mapsto aX(S) + b$, $E(y) = aE(X) + b$.

Proof. By cases:

- Suppose X is a discrete random variable. Then clearly Y is also a discrete random variable, since $x \mapsto ax + b$ is bijective, and hence $S_X \cong S_Y$, so if S_X is finite, S_Y is too. Observe that

$$\begin{aligned} E(Y) &= \sum_{x \in S_X} (ax + b)P(X = x) \\ &= a \sum_{x \in S_X} xP(X = x) + b \sum_{x \in S_X} P(X = x) \\ &= aE(X) + b \end{aligned}$$

- Suppose X is a continuous random variable. By the logic above, we have that Y also must be continuous. Observe that

$$\begin{aligned} E(Y) &= \int_{\mathbb{R}} (ax + b) f_X(x)\, dx \\ &= a \int_{\mathbb{R}} x f_X(x)\, dx + b \int_{\mathbb{R}} f_X(x)\, dx \\ &= aE(X) + b \end{aligned}$$

□

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*

$$EZ = EX + EY$$

Proof. We will only prove the discrete case here, as we need a little measure theory to prove this in the continuous case (as far as I am aware). Observe that Z is then also discrete. We have that

$$EZ = \sum_{z \in S_Z} zP(Z = z)$$

Observe that

$$P(Z = z) = P(\{s \in S \mid Z(s) = z\})$$

So for a particular $z \in S_Z$ we have that

$$zP(Z = z) = \sum_{s \in S : Z(s) = z} Z(s)P(s)$$

Hence we have the following equivalent definition of discrete expectation:

$$EZ = \sum_{z \in S_Z} \sum_{s \in S : Z(s) = z} Z(s)P(s) = \sum_{s \in S} Z(s)P(s)$$

The proof is now trivial:

$$EZ = \sum_{s \in S} (X(s) + Y(s))P(s) = EX + EY$$

□

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 $\operatorname{var}(X)$, is the value

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

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

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

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

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

Proof. By a previous theorem, we have that

$$\begin{aligned}
\operatorname{var}(X) &= E((X - E(X))^2) \\
&= E(X^2 - 2XE(X) + E(x)^2) \\
&= E(X^2) - E(2XE(X)) + E(E(X)^2) \\
&= E(X^2) - 2E(x)^2 + E(X)^2 \\
&= E(X^2) - E(X)^2
\end{aligned}$$

□

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 $j \in S_Y$

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

Proof. We shall only prove the first statement, as the proof is analogous for the second. Observe that

$$\begin{aligned}
\sum_{j \in S_Y} P(X = i, Y = j) &= \sum_{j \in S_Y} P(\{s \in S \mid X(s) = i \wedge Y(s) = j\}) \\
&= P\left(\bigcup_{j \in S_Y} \{s \in S \mid X(s) = i \wedge Y(s) = j\} \right) \\
&= P(\{s \in S \mid X(s) = i\}) \\
&= P(X = i) && \text{(by definition)}
\end{aligned}$$

□

Remark 2.6. 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 2.7. 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 2.8. 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 2.2. 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 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 2.9. 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*

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

Proof. We only prove this if X and Y are discrete random variables. Observe that

$$\begin{aligned} E(g(X)h(Y)) &= \sum_x \sum_y g(x)h(y)P(X = x, Y = y) \\ &= \sum_x \sum_y g(x)h(y)P(X = x)P(Y = y) \\ &= \left(\sum_x g(x)P(X = x)\right)\left(\sum_y h(y)P(Y = y)\right) \\ &= E(g(x))E(h(Y)) \end{aligned}$$

□

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 2.10. 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 2.3. 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, $|\rho(X,Y)| \leq 1$.*

Proof.

$$\begin{aligned}\rho(X,Y) &= \frac{\operatorname{Cov}(X,Y)}{\sigma_X\sigma_Y}\\ &= \operatorname{Cov}\left(\frac{X}{\sigma_X}, \frac{Y}{\sigma_Y}\right)\\ &= \frac{1}{2}\left(\operatorname{var}\left(\frac{X}{\sigma_X} + \frac{Y}{\sigma_Y}\right) - \operatorname{var}\left(\frac{X}{\sigma_X}\right) - \operatorname{var}\left(\frac{Y}{\sigma_Y}\right)\right)\\ &= \frac{1}{2}\operatorname{var}\left(\frac{X}{\sigma_X} + \frac{Y}{\sigma_Y}\right) - 1\end{aligned}$$

Hence, $2 + 2\rho(X,Y) = \operatorname{var}\left(\frac{X}{\sigma_X} + \frac{Y}{\sigma_Y}\right) \geq 0$ which finishes the proof. □

Remark 2.11. Observe that the space of random variables on a probability space is an $\mathbb{R}$-vector space: we can define addition and scalar multiplication as expected, and verify the vector space axioms easily. For completeness, for two random variables X and Y, we define their sum $X+Y$ as the random variable $X+Y : S \to \mathbb{R}$ such that $(X+Y)(s) := X(s) + Y(s)$ for all $s \in S$. And for some $\lambda \in \mathbb{R}$ we define the random variable $\lambda X : S \to \mathbb{R}$ by $(\lambda X)(s) := \lambda X(s)$ for all $s \in S$. It is readily verified that this constitutes a vector space over $\mathbb{R}$. The zero element of this vector space is then simply the random variable that always assigns 0 to any element of S.
Now, observe that $\operatorname{Cov}(\cdot,\cdot)$ is an inner product on this space:

- Symmetry is trivial / already shown
- Linearity is trivial / already shown
- It is clear that we have $\operatorname{Cov}(X,X) \geq 0$, but not that we also have that for a nonzero random variable, we have a nonzero covariance with itself. Sadly, this cannot be proven under the current construction, unless we assume only discrete random variables. We could consider the quotient space instead over an equivalence relation that identifies random variables that are equal up to some constant. We will ignore this detail for now.

Hence, we have the Cauchy-Schwarz inequality:

$$|\operatorname{Cov}(X,Y)|^2 \leq \sigma_X^2 \sigma_Y^2$$

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)

Proof. Suppose first that there exist $a, b \in \mathbb{R}$ such that $X = aY + b$. Then

$$\begin{aligned}
\rho(X,Y) &= \frac{\operatorname{Cov}(X,Y)}{\sigma_X \sigma_Y} \\
&= \frac{\operatorname{Cov}(aY+b,Y)}{\sqrt{\operatorname{var}(aY+b)}\sigma_Y} \\
&= \frac{a\operatorname{Cov}(Y,Y)}{|a|\sigma_Y^2} \\
&= \operatorname{sgn}(a)
\end{aligned}$$

Now for the converse direction, suppose $|\rho(X,Y)| = 1$. Recall that under the above identification of quotient vector space and the properties of the Cauchy-Schwarz inequality, we have that if

$$|\operatorname{Cov}(X,Y)|^2 = \sigma_X^2 \sigma_Y^2$$

then there exists $a \in \mathbb{R}$ such that

$$a[X] = [Y]$$

(where the square brackets denote equivalence classes). Hence $[aX] = [Y]$, which means there exists $b \in \mathbb{R}$ such that $aX + b = Y$, which finishes the proof. Note that this also proves the converse direction, but that would be a quite boring proof, when an algebraic proof can be used here as well. □

Remark 2.12. 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 2.13. 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.*

Proof. By induction. □

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*

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

Proof.

- Recall that if $X \sim B(n,p)$ then its mass function is (for $i = 0, \ldots, n$)

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

Hence for $t \in \mathbb{R}$

$$\begin{aligned}\phi_X(t) &= \sum_{i=0}^{n} \exp(ti)p(i) \\ &= \sum_{i=0}^{n} \binom{n}{i} \exp(t)^i p^i (1-p)^{n-i} \\ &= \sum_{i=0}^{n} \binom{n}{i} (p\exp(t))^i (1-p)^{n-i} \\ &= (p\exp(t) + 1 - p)^n\end{aligned}$$

- Recall that we have for $i \in \mathbb{N}$ that

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

Hence for $t \in \mathbb{R}$

$$\begin{aligned}\phi_X(t) &= \sum_{i=0}^{n} \exp(ti)p(i) \\ &= \exp(-\mu) \sum_{i=0}^{n} \frac{(\exp(t)\mu)^i}{i!} \\ &= \exp(-\mu)\exp(\exp(t)\mu) \\ &= \exp(\mu(\exp(t) - 1))\end{aligned}$$

- Recall that if f is the pdf. of X, then for $x \geq 0$ we have

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

Hence for $t \in \mathbb{R}$

$$\begin{aligned}\phi_X(t) &= \int_{\mathbb{R}} \exp(tx)f(x)\,dx \\ &= \lambda \int_0^{\infty} \exp((t-\lambda)x)\,dx \\ &= \frac{\lambda}{t-\lambda} \exp((t-\lambda)x)\Big|_0^{\infty} \\ &= \frac{\lambda}{\lambda - t} \qquad \text{(if } t < \lambda\text{, does not exist otherwise)}\end{aligned}$$

- Recall that for $x \in \mathbb{R}$

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

Hence for $t \in \mathbb{R}$ we have

$$\begin{aligned}
\phi_X(t) &= \int_{\mathbb{R}} \exp(tx) f(x)\, dx \\
&= \frac{1}{\sigma\sqrt{2\pi}} \int_{\mathbb{R}} \exp(tx) \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right) dx \\
&= \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} \exp(t(x\sigma+\mu)) \exp\left(-\frac{1}{2}x^2\right) dx \\
&= \exp(\mu t) \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} \exp\left(tx\sigma - \frac{1}{2}x^2\right) dx \\
&= \exp(\mu t) \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} \exp\left(-\frac{1}{2}((x-t\sigma)^2 - t^2\sigma^2)\right) dx \\
&= \exp\left(\frac{1}{2}(t\sigma)^2 + \mu t\right) \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} \exp\left(-\frac{1}{2}(x-t\sigma)^2\right) dx \\
&= \exp\left(\frac{1}{2}(t\sigma)^2 + \mu t\right)
\end{aligned}$$

□

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

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

Proof. Let $Y := aX + b$ for some $a, b \in \mathbb{R}$, $a \neq 0$. We show only the right implication, as the converse amounts to showing the same fact in the same way, mutatis mutandis for '$X = \frac{Y-b}{a}$'.

$$\begin{aligned}
\phi_Y(t) &= E(\exp(tY)) \\
&= E(\exp(t(aX+b))) \\
&= \exp(bt) E(\exp(atX)) \\
&= \exp\left(\frac{1}{2}(at\sigma)^2 + \mu at + bt\right) \\
&= \exp\left(\frac{1}{2}(a\sigma)^2 t^2 + (a\mu + b)t\right)
\end{aligned}$$

□

Remark 2.14. 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 $t \in \mathbb{R}$.

Proof. (of the last property). Observe that

$$\begin{aligned}
\phi_{X+Y}(t) &= E(\exp(t(X+Y))) \\
&= E(\exp(tX)\exp(tY)) \\
&= E(\exp(tX))E(\exp(tY)) \\
&= \phi_X(t)\phi_Y(t)
\end{aligned}$$

□

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 $X+Y \sim N(\mu_1+\mu_2,\sigma_1^2+\sigma_2^2)$.*

Proof. Proof by staring at the formula's, please. □

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*

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

Proof. Two cases:

- Suppose that X is discrete with range S_X. Then

$$\begin{aligned} EX &= \sum_{x \in S_X} xP(X=x) \\ &= \sum_{x \in S_X \wedge x < a} xP(X=x) + \sum_{x \in S_X \wedge x \geq a} xP(X=x) \\ &\geq \sum_{x \in S_X \wedge x \geq a} xP(X=x) \\ &\geq a \sum_{x \in S_X \wedge x \geq a} P(X=x) \\ &= aP(X \geq a) \end{aligned}$$

- Suppose that X is continuous. Then

$$\begin{aligned} EX &= \int_{\mathbb{R}} xf(x)\,dx \\ &= \int_0^\infty xf(x)\,dx \\ &= \int_0^a xf(x)\,dx + \int_a^\infty xf(x)\,dx \\ &\geq \int_a^\infty xf(x)\,dx \\ &\geq a\int_a^\infty f(x)\,dx \\ &= aP(X \geq a) \end{aligned}$$

□

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*

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

Proof. We already know from the previous theorem that

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

□

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

Proof. Consider that for some $\varepsilon > 0$ we have

$$1 - P(|\overline{X}_n - \mu| < \varepsilon) = P(|\overline{X}_n - \mu| \geq \varepsilon) \leq \frac{\sigma^2}{n\varepsilon^2} \to 0$$

Hence

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

□

Remark 2.16. Consider a sequence of random variables that are all a single independent Bernoulli trial that an event $E \in \Sigma$ occurs with probability $p \in (0, 1)$. Observe that $\mu = EX = p$. Then the sum of the random variables is the total amount of successes, denoted by n_E. Then indeed, as we have hinted at before, we have that $\frac{n_E}{n} \to p$.

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 $\operatorname{var}(X_k) = \sigma^2$ for $\sigma > 0$. Then*

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

Proof. Attempt using MGFS. Consider that we should define a sequence

$$Y_n := \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} = \frac{\sum_{k=1}^{n} X_k - n\mu}{\sigma\sqrt{n}} = \sum_{k=1}^{n} \frac{1}{\sqrt{n}} \underbrace{\frac{X_k - \mu}{\sigma}}_{:=Z_n}$$

Now all of the Z_n's are independent, so for all $t \in \mathbb{R}$

$$\phi_{Y_n}(t) = \prod_{n\in\mathbb{N}^*} \phi_{Z_n}\left(\frac{t}{\sqrt{n}}\right) = \phi_{Z_1}\left(\frac{t}{\sqrt{n}}\right)^n$$

If we suppose that the MGF is analytic, we know that its Taylor series converges, which is to say

$$\phi_{Z_1}(t) = \sum_{k=0}^{\infty} \frac{\phi_{Z_1}^{(k)}(0)}{n!} x^k = \sum_{k=0}^{\infty} \frac{E(Z_1^k)}{k!} t^k$$

Now observe that $EZ_1 = 0$ and $EZ_1^2 = 1$, and observe that

$$\phi_{Z_1}\left(\frac{t}{\sqrt{n}}\right) = 1 + \frac{t^2}{2n} + \sum_{k=3}^{\infty} \frac{E(Z_1^k)}{k!}\left(\frac{t}{\sqrt{n}}\right)^k$$

Hence $\phi_{Y_n}(t) \to \exp(t^2/2)$. Therefore, the PDF and hence also the CDF of Y_n converges pointwise (and uniformly due to continuity on all compact intervals) to that of a variable with a standard normal distribution, which implies the theorem. □

Remark 2.17. 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

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

Remark 3.2. Recall that if we have a random variable $Z := \sum_i X_i$, then

$$\begin{aligned}
E(Z|Y = y) &= \sum_{z \in S_Z} zP(Z = z|Y = y) \\
&= \sum_{s \in S} Z(s)P(s|Y = y) \\
&= \sum_i \sum_{s \in S} X_i(s)P(s|Y = y) \\
&= \sum_i \sum_{x \in S_{X_i}} xP(X_i = x|Y = y) \\
&= \sum_i E(X_i|Y = y)
\end{aligned}$$

Hence,

$$E\left(\sum_i X_i|Y = y\right) = \sum_i E(X_i|Y = y)$$

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

Remark 3.3. Recall that if (S, Σ, P) is a probability space, $E \in \Sigma$ is an event and $Q \subseteq \Sigma$ such that for all $A, B \in Q$ we have $A = B \vee A \cap B = \varnothing$ (disjoint) and $\bigcup Q = S$, then (if it converges)

$$P(E) = \sum_{F \in Q} P(E|F)P(F)$$

This is the **law of total probability**. We can define a partition Q of S as follows: let X and Y be random variables, then for some $x \in S_X$

$$E := \{s \in S \mid X(s) = x\}$$

$$Q := \{F \in \Sigma \mid \exists y \in S_Y : \forall s \in F : Y(s) = y\}$$

Using this we find the useful result that

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

Observe furthermore that

$$\begin{aligned} EX &= \sum_{x \in S_X} xP(X=x) \\ &= \sum_{x \in S_X} \sum_{y \in S_Y} xP(X=x|Y=y)P(Y=y) \\ &= \sum_{y \in S_Y} E(X|Y=y)P(Y=y) \end{aligned}$$

We can define a function $g : S_Y \to \mathbb{R}$ by $g(y) := E(X|Y=y)$, such that

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

Now we introduce some crazy notation: we let $E(X|Y) := E(g(Y))$ be a random variable. Then

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

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 3.4. 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*

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

Proof. Firstly,

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

Then

$$\begin{aligned} \operatorname{var}(E(X|Y)) &= E((E(X|Y))^2) - (E(E(X|Y)))^2 \\ &= E((E(X|Y))^2) - (EX)^2 \end{aligned}$$

Summing then gives the desired result. □

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.

Example 3.2. A practical example: let $X :=$ 'number of reports', $Y :=$ 'number of thefts'. Obviously, $S_X = S_Y = \mathbb{N}$. Some information is given: for $n, k \in \mathbb{N}$ we have

$$P(Y = n) = \exp(-8)\frac{8^n}{n!}$$
$$P(X = k|Y = n) = \binom{n}{k}\frac{1}{2^n} \qquad \text{(if } k \leq n\text{, 0 otherwise)}$$

Intuitively, this means: the amount of thefts is Poisson distributed with $\lambda = 8$, and if we know that n thefts are committed, then the probability of k thefts being reported is binomial, e.g. there is a $1/2$ chance that a theft is reported. Consider that then $X|Y = n$ has expectation np where $p = 1/2$. But what about $E(Y|X = k)$?

$$\begin{aligned} E(Y|X = k) &= \sum_{n=0}^{\infty} nP(Y = n|X = k) \\ &= \sum_{n=0}^{\infty} n\frac{P(X = k|Y = n)P(Y = n)}{P(X = k)} \end{aligned}$$

$$\begin{aligned} P(X = k) &= \sum_{n=0}^{\infty} P(X = k|Y = n)P(Y = n) \\ &= \sum_{n=k}^{\infty} \binom{n}{k}\frac{1}{2^n}\exp(-8)\frac{8^n}{n!} \\ &= \frac{\exp(-8)4^k}{k!}\sum_{n=k}^{\infty}\frac{4^n}{(n-k)!} \\ &= \frac{\exp(-8)4^k}{k!}\sum_{n=0}^{\infty}\frac{4^n}{n!} \\ &= \frac{\exp(-8)\exp(4)4^k}{k!} = \frac{\exp(-4)4^k}{k!} \end{aligned}$$

$$\begin{aligned} E(Y|X = k) &= \sum_{n=0}^{\infty} n\frac{P(X = k|Y = n)P(Y = n)}{P(X = k)} \\ &= \sum_{n=k}^{\infty} n\binom{n}{k}\frac{1}{2^n}\exp(-8)\frac{8^n}{n!}\frac{k!}{4^k}\exp(4) \\ &= \exp(-4)\sum_{n=k}^{\infty} n\frac{4^{n-k}}{(n-k)!} \\ &= \exp(-4)\sum_{n=0}^{\infty}(n+k)\frac{4^n}{n!} \\ &= \exp(-4)\left(\sum_{n=1}^{\infty}\frac{4^n}{(n-1)!} + k\sum_{n=0}^{\infty}\frac{4^n}{n!}\right) \\ &= \exp(-4)(4\exp(4) + k\exp(4)) \\ &= 4 + k \end{aligned}$$

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*

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

Proof. Observe that

$$\begin{aligned} P(X > s + t | X > t) &= \frac{P(X > s + t, X > t)}{P(X > t)} \\ &= \frac{P(X > s + t)}{P(X > t)} \\ &= \frac{1 - P(X \leq s + t)}{1 - P(X \leq t)} \\ &= \frac{\exp(-\lambda(s + t))}{\exp(-\lambda t)} \\ &= \exp(-\lambda s) \\ &= 1 - P(X \leq s) = P(X > s) \end{aligned}$$

□

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.

Proof. Observe that according to the previous steps, we need that

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

Let us now suppose that X is continuous. Observe that the function $P(X > \cdot)$ satisfies a certain functional equation. We know that this function is right continuous, since it is equal to $1 - F(\cdot)$. It is even differentiable almost everywhere, being the integral of a piecewise continuous function. Suppose f is a solution to the functional equation. Let $s = 0$ in the equation and we have in particular that for all $t \in \mathbb{R}$, $f(t) = f(0)f(t)$. We can rule out that f is constant zero as its limit at $-\infty$ must be 1 and its limit at ∞ must be 0. By the intermediate value theorem, there exists at least one $t \geq 0$ such that $f(t) \neq 0$, so we conclude that $f(0) = 1$. Furthermore, because f must be monotonically decreasing, we know that also for all $t < 0$ we have $f(t) = 1$.
Now let $t > 0$ and not at a potential point of discontinuity. f is differentiable at t, e.g. in particular

$$f'(t) = \lim_{h \to 0^+} \frac{f(t+h) - f(t)}{h} = f(t) \lim_{h \to 0^+} \frac{f(h) - 1}{h} = f(t) \lim_{h \to 0^+} \frac{f(h) - f(0)}{h} = f(t)f'^{+}(0)$$

Since f is monotonically decreasing and positive, we know that $f'(t) \leq 0$, so $f'^{+}(0) < 0$ (since equal to zero would mean the derivative is zero everywhere, which is a contradiction). We conclude that f is strictly decreasing at $(0, \infty)$, and hence continuous everywhere except at possibly 0. But by right-continuity we figure out that f is actually continuous everywhere, and differentiable everywhere except at 0. Well then, let $t > 0$ and let g be another solution to the functional equation and observe that

$$\begin{aligned} \frac{d}{dt}\frac{f(t)}{g(t)} &= \frac{f'(t)g(t) - f(t)g'(t)}{(g(t))^2} \\ &= \frac{f(t)}{g(t)}(f'^{+}(0) - g'^{+}(0)) \end{aligned}$$

Hence we get that f/g is also an exponential function by the uniqueness of the solution to the above differential equation. But since one possibility for f is an exponential function, we see that g must also be exponential. Hence any solution to the functional equation is an exponential function if and only if at least one solution is exponential, which it is. Hence there must exist $C, \lambda \in \mathbb{R}$ such that for all $t \in \mathbb{R}$

$$g(t) = \begin{cases} C\exp(-\lambda t) & \text{if } t \geq 0 \\ 1 & \text{otherwise} \end{cases}$$

By the necessity of continuity we find that $C = 1$. But then, for all $t \geq 0$, $F(t) = 1 - \exp(\lambda t)$ and for all $t < 0$, $F(t) = 0$, hence X is exponentially distributed.
Lastly, if X is instead discrete, we can make the fair assumption that its range is $\mathbb{N}^*$, if not we can redefine such that it is by relabelling, which does not change the distribution. Let $p := 1 - P(X > 1) = P(X \leq 1)$. We would like to show that for all $n \in \mathbb{N}^*$, we have

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

e.g. X is geometrically distributed with parameter p. Consider that

$$P(X = 1) = P(X \leq 1) - P(X < 1) = p$$

so the base case is proven. By induction, suppose the above statement holds for a particular $n \in \mathbb{N}^*$, then

$$P(X = n + 1) = P(X > n) - P(X > n + 1) = P(X > n) - P(X > n)(1 - p) = P(X > n)p$$

By an earlier proof we know that

$$P(X > n) = (1 - p)^n$$

□

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. Let $t, \varepsilon > 0$. Observe that

$$\begin{aligned} \frac{1}{\varepsilon} P(X \in (t, t+\varepsilon) | X > t) &= \frac{1}{\varepsilon} \frac{P(X \in (t, t+\varepsilon))}{P(X > t)} \\ &= \frac{1}{\varepsilon(1 - F(t))} \int_t^{t+\varepsilon} f(x)\, dx \\ &\overset{\varepsilon \to 0}{\to} \frac{f(t)}{1 - F(t)} = r(t) \end{aligned} \tag{FTC}$$

Hence, the hazard rate of X at a value $t \geq 0$ is the change in probability near t given X is already greater than t.
The interpretation is that if X represents for example the timestamp at which a component of a system fails, then the hazard rate is the change in probability of the component breaking at time t, given it has survived until at least time t.

Remark 5.4. If $X \sim \text{Exp}(\lambda)$ 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*

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

Proof. The implication is trivial, the direction the other way is more tricky. Consider that $-r$ is the derivative of $t \mapsto \ln(1 - F(t))$, e.g.

$$\ln(1 - F(t)) = \int_0^t \frac{d}{dx} \ln(1 - F(x))\, dx = -\int_0^t r(x)\, dx$$

But this clearly implies

$$F(t) = 1 - \exp\left(-\int_0^t r(x)\, dx \right)$$

Hence the distribution is uniquely determined by the hazard rate, and in other words, if $r_X = r_Y$, then $F_X = F_Y$ and $X \sim Y$. □

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.

Remark 5.5. If $X_1, \ldots, X_n$ are random variables that are each exponentially distributed, with parameter λ_i for variable X_i, $i = 1, \ldots, n$, and another discrete random variable T is independent of each, and has range $\{1, \ldots, n\}$, then X_T is also hyperexponentially distributed. X_T, by the way, maps the random variable T to the value of its corresponding X_i if $T = i$ for a certain event. That is, we of course have our probabilty space (S, Σ, P), and we then define $X_T : S \to \mathbb{R}$ by

$$s \mapsto X_{T(s)}(s)$$

We can see that X_T is hyperexponentially distributed by conditioning (let F be the CDF of X_T and f its density):

$$\begin{aligned}
F(x) &= P(X_T \leq x) = 1 - P(X_T > x) \\
&= 1 - \sum_{i=1}^{n} P(X_T > x | T = i) P(T = i) \\
&= 1 - \sum_{i=1}^{n} P(X_i > x) P(T = i) \\
&= \sum_{i=1}^{n} (1 - \exp(-\lambda_i x)) P(T = i) \\
f(x) &= \sum_{i=1}^{n} \lambda_i \exp(-\lambda_i x) P(T = i)
\end{aligned}$$

So indeed we have a hyperexponential distribution. Observe that the failure rate thus is

$$r(t) = \frac{\sum_{i=1}^{n} \lambda_i \exp(-\lambda_i t) P(T = i)}{\sum_{i=1}^{n} \exp(-\lambda_i t) P(T = i)}$$

Suppose $\min_i \lambda_i = \lambda_j$ for some $j = 1, \ldots, n$. That is, for all $i = 1, \ldots, n$, $\lambda_i \geq \lambda_j$. For convenience, relabel such that $j = 1$. If we assume there are no duplicate distributions, which would be fair in the normal case, then also $\lambda_i > \lambda_1$. Then

$$r(t) = \frac{\lambda_1 P(T = 1) + \sum_{i=2}^{n} \lambda_i \exp(-(\lambda_i - \lambda_1) t) P(T = i)}{P(T = 1) + \sum_{i=2}^{n} \exp(-\lambda_i t) P(T = i)} \overset{t \to \infty}{\to} \frac{\lambda_1 P(T = 1)}{P(T = 1)} = \lambda_1$$

In full generality:

$$r(t) = \min_i \lambda_i$$

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 $X_i \sim \text{Exp}(\lambda)$. Let $S_n := X_1 + \ldots + X_n$. Then $S_n \sim \text{Gamma}(n, \lambda)$, that is, its density function f_{S_n} is given by (for all $t \in \mathbb{R}$)*

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

Proof. By induction on n. The base case is trivial. Consider now that the variables are independent, so $f_{S_{n+1}} = f_{S_n} * f_{X_{n+1}}$. Let $t \in \mathbb{R}$. Consider that if $t < 0$ then $f_{S_{n+1}}(t) = 0$ because we would be integrating over $\varnothing$. Hence, let $t \geq 0$. Observe that

$$\begin{aligned}
f_{S_{n+1}}(t) &= \int_{-\infty}^{\infty} f_{S_n}(\tau) f_{X_{n+1}}(t - \tau)\, d\tau \\
&= \int_0^t \lambda \exp(-\lambda \tau) \frac{(\lambda \tau)^{n-1}}{(n-1)!} \lambda \exp(-\lambda(t - \tau))\, d\tau \\
&= \lambda^{n+1} \exp(-\lambda t) \int_0^t \frac{\tau^{n-1}}{(n-1)!}\, d\tau \\
&= \lambda^{n+1} \exp(-\lambda t) \frac{\tau^n}{n!} \Big|_0^t \\
&= \lambda \exp(-\lambda t) \frac{(\lambda t)^n}{n!}
\end{aligned}$$

□

Remark 5.6. Let X, Y be independent exponential random variables with parameters λ and μ respectively. We might be interested in the probability that $X < Y$. Well:

$$\begin{aligned}
P(X < Y) &= \int_{\mathbb{R}} P(X < Y | Y = t) f_Y(t)\, dt \\
&= \int_0^\infty P(X < x) \mu \exp(-\mu x)\, dx \\
&= \int_0^\infty (1 - \exp(-\lambda x)) \mu \exp(-\mu x)\, dx \\
&= 1 - \mu \int_0^\infty \exp(-(\lambda + \mu)x)\, dx \\
&= 1 - \frac{\mu}{\lambda + \mu} \exp(-(\lambda + \mu)x) \Big|_0^\infty \\
&= \frac{\lambda}{\lambda + \mu}
\end{aligned}$$

Now if we let $Z := \min\{X, Y\}$ be a new random variable, we can find its distribution:

$$\begin{aligned}
F_Z(z) &= 1 - P(Z > z) \\
&= 1 - P(X > z, Y > z) \\
&= 1 - P(X > z) P(Y > z) \\
&= 1 - \exp(-\lambda z) \exp(-\mu z) \\
&= 1 - \exp(-(\lambda + \mu) z)
\end{aligned}$$

Hence $Z \sim \text{Exp}(\lambda + \mu)$. Note that for $t \in \mathbb{R}$, the probability that $Z > t$ is independent of $X_1 < X_2$. We can see this by direct computation:

$$\begin{aligned}
P(X_1 < X_2 | Z > t) &= P(X_1 < X_2 | X_1 > t, X_2 > t) \\
&= P(X_2 > X_1 - t | X_1 > t) && \text{(memoryless)} \\
&= P(X_1 - t < X_2 - t) && \text{(reverse memoryless)} \\
&= P(X_1 < X_2) && \text{(which is independent of choice of } t\text{)}
\end{aligned}$$

In more generality, let $X_1, \ldots, X_n$ be independent exponential random variables with parameters $\lambda_1, \ldots, \lambda_n$, respectively. Let $Z := \min\{X_1, \ldots, X_n\}$. Then

$$Z \sim \text{Exp}(\lambda_1 + \ldots + \lambda_n)$$

We can see this easily by induction. Furthermore, we can let $i = 1, \ldots, n$ and consider when $X_i < X_j$ for all $j = 1, \ldots, n$, $j \neq i$. Well:

$$\begin{aligned}
&P(\{s \in S \mid \forall j = 1, \ldots, n : j \neq i \implies X_i(s) < X_j(s)\}) \\
&\quad = P\left(X_i < \min_{i \neq j} X_j \right) \\
&\quad = P(X_i < Z) && (Z := \min_{i \neq j} X_j) \\
&\quad = \frac{\lambda_i}{\sum_{j=1}^n \lambda_j} && (Z \sim \text{Exp}(\textstyle\sum_{j \neq i} \lambda_j))
\end{aligned}$$