Chapter 2

Random Variables

2.1 Basic Definitions

To take advantage of our measurements over the real numbers, in terms of its metric structure and arithmetic, we need to formally define this measurement process using the notion of a random variable.

Definition 2.1 (Random Variable)**.** *Let* $(\Omega, \mathcal{F}, \mathbb{P})$ *be some probability triple. Then, a* **Random Variable (RV)***, say* X*, is a function from the sample space* Ω *to the set of real numbers* $\mathbb{R}$

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

such that for every $x \in \mathbb{R}$*, the inverse image of the half-open real interval* $(-\infty, x]$ *is an element of the collection of events* $\mathcal{F}$*, i.e.:*

$$\text{for every } x \in \mathbb{R}, \qquad X^{[-1]}(\ (-\infty, x]\ ) := \{\omega : X(\omega) \leq x\} \in \mathcal{F}\ .$$

This definition can be summarised by the statement that a RV is an $\mathcal{F}$*-measurable map.* *We assign probability to the RV* X *as follows:*

$$\mathbb{P}(X \leq x) = \mathbb{P}(\ X^{[-1]}(\ (-\infty, x]\ )\ ) := \mathbb{P}(\ \{\omega : X(\omega) \leq x\}\ )\ . \qquad (2.1)$$

Definition 2.2 (Distribution Function)**.** *The* **Distribution Function (DF)** *or* **Cumulative Distribution Function (CDF)** *of any RV* X*, over a probability triple* $(\Omega, \mathcal{F}, \mathbb{P})$*, denoted by* F *is:*

$$F(x) := \mathbb{P}(X \leq x) = \mathbb{P}(\ \{\omega : X(\omega) \leq x\}\ ), \qquad \text{for any} \quad x \in \mathbb{R}\ . \qquad (2.2)$$

Thus, $F(x)$ *or simply* F *is a non-decreasing, right continuous,* $[0, 1]$*-valued function over* $\mathbb{R}$*. When a RV* X *has DF* F *we write* $X \sim F$*.*

Remark 2.3 (Notation)**.** *It is enough to understand the idea of random variables as explained above, and work with random variables using simplified notation like*

$$\mathbb{P}(2 \leq X \leq 3)$$

rather than

$$\mathbb{P}(\{\omega : 2 \leq X(\omega) \leq 3\})$$

but note that when learning or doing more advanced work this sample space notation is usually needed to clarify the true meaning of the simplified notation.

From the idea of a distribution function, we get something that resembles the fundamental theorem of calculus:

Proposition 2.4. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability triple and let X be a random variable with DF F. Then for $a < b$ we get*

$$\mathbb{P}(a < X \leq b) = F(b) - F(a). \tag{2.3}$$

Proof. For this proof we will be very formal for the sake of clarity, later on we will adopt the more relaxed notation. Define the sets

$$\begin{aligned} A &= \{\omega : \omega \in \Omega, X(\omega) \leq a\} \\ B &= \{\omega : \omega \in \Omega, X(\omega) \leq b\} \\ C &= \{\omega : \omega \in \Omega, a < X(\omega) \leq b\} \end{aligned}$$

Note that $A, B, C \in \mathcal{F}$, which follows from Definition 2.1. Furthermore, note that $B = A \cup C$ and that $A \cap C = \emptyset$. Now using Definition 2.2 and Lemma 1.11 and our above construction, we get

$$F(b) = \mathbb{P}(B) = \mathbb{P}(A \cup C) = \mathbb{P}(A) + \mathbb{P}(C) = F(a) + \mathbb{P}(C).$$

Rearranging the above,

$$F(b) - F(a) = \mathbb{P}(C) = \mathbb{P}(a < X \leq b).$$

which is (2.3). □

A special RV that often plays the role of 'building-block' in Probability and Statistics is the indicator function of an event A that tells us whether the event A has occurred or not. Recall that an event belongs to the collection of possible events $\mathcal{F}$ for our experiment.

Definition 2.5. *[Indicator Function] Given a probability triple* $(\Omega, \mathcal{F}, \mathbb{P})$*, the* **Indicator Function** *of an event* $A \in \{$ *which is denoted* $\mathbb{1}_A$ *is defined as follows:*

$$\mathbb{1}_A(\omega) := \begin{cases} 1 & \textit{if} \quad \omega \in A \\ 0 & \textit{if} \quad \omega \notin A \end{cases} \tag{2.4}$$

Lemma 2.6. *The indicator function* $\mathbb{1}_A$ *as in Definition 2.5 is a random variable.*

Proof. For $\mathbb{1}_A$ to be a RV, we need to verify that for any real number $x \in \mathbb{R}$, the inverse image $\mathbb{1}_A^{[-1]}(\ (-\infty, x]\ )$ is an event, ie :

$$\mathbb{1}_A^{[-1]}(\ (-\infty, x]\ ) := \{\omega : \mathbb{1}_A(\omega) \leq x\} \in \mathcal{F}\ .$$

Since the indicator function is either 0 or 1 we observe that we have to get the empty event if $x < 0$, furthermore we have to get the entire sample space if $x > 1$ (since it is always true). The last case is when $0 \leq x < 1$, which is only ok when $\mathbb{1}_A = 0$. Summarised below:

$$\mathbb{1}_A^{[-1]}(\ (-\infty, x]\ ) := \{\omega : \mathbb{1}_A(\omega) \leq x\} = \begin{cases} \emptyset & \text{if} \quad x < 0 \\ A^c & \text{if} \quad 0 \leq x < 1 \\ A \cup A^c = \Omega & \text{if} \quad 1 \leq x \end{cases}$$

Thus, $\mathbb{1}_A^{[-1]}(\ (-\infty, x]\ )$ is one of the following three sets that belong to $\mathcal{F}$; (1) $\emptyset$, (2) A^c and (3) Ω depending on the value taken by x relative to the interval $[0, 1]$. We have proved that $\mathbb{1}_A$ is indeed a RV. □

Model 2.7 (Indicator of an event as Bernoulli RV)**.** *Given a probability triple* $(\Omega, \mathcal{F}, \mathbb{P})$ *and an event* $A \in \mathcal{F}$*, the random variable* $\mathbb{1}_A$ *is called the* Bernoulli *RV for event* A *with a known probability* $\mathbb{P}(A)$*. We will adopt the notation* Bernoulli(θ) *for the RV by introducing a paramater* $\theta \in [0, 1]$ *for the typically unknown probability* $\mathbb{P}(A)$*.*

Lemma 2.8. *Given a probability triple* $(\Omega, \mathcal{F}, \mathbb{P})$ *and an event* $A \in \mathcal{F}$*, the following properties hold:*

1. $\mathbb{1}_A = 1 - \mathbb{1}_{A^c}$*, (complementation behaves like the probability)*
2. $\mathbb{1}_{A \cap B} = \mathbb{1}_A \mathbb{1}_B$ *(intersection becomes product)*
3. $\mathbb{1}_{A \cup B} = \mathbb{1}_A + \mathbb{1}_B - \mathbb{1}_A \mathbb{1}_B$ *(union becomes addition - intersection)*

As you can see, there is a lot of similarities between the properties of indicator function and the properties of the probability measure $\mathbb{P}$.

Proof. Exercise! □

We slightly abuse notation when A is a single element set by ignoring the curly braces.

2.2 Discrete Random Variables

When a RV takes at most countably many values from a discrete set $\mathbb{X}$, we call it a **discrete** RV. Recall that a set $\mathbb{X}$ is said to be discrete if we can enumerate its elements, i.e., find an enumerating or counting function $\mathbb{X} \ni x \mapsto i \in \mathbb{N}$ that associates each element $x \in \mathbb{X}$ to a natural number $i \in \mathbb{N}$. So, $\mathbb{X}$ is either finite with k elements in $\mathbb{X} = \{x_1, x_2, \ldots, x_k\}$ or countably infinite with the same cardinality as $\mathbb{N}$ with $\mathbb{X} = \{x_1, x_2, \ldots\}$.

Definition 2.9. *Let X be a $\mathbb{R}$-valued RV over a probability triple $(\Omega, \mathcal{F}, \mathbb{P})$. If X takes values in an enumerable set $\mathbb{X} \subset \mathbb{R}$ then we call X a $\mathbb{R}$-valued* **discrete** *random variable.*

The concept of distribution function (DF) does not differentiate between types of random variables, the next definition does:

Definition 2.10. *[probability mass function (PMF)] Let X be a $\mathbb{R}$-valued discrete RV over a probability triple $(\Omega, \mathcal{F}, \mathbb{P})$. We define the* **probability mass function** *(PMF) f of X to be the function $f : \mathbb{R} \to [0, 1]$ defined as follows:*

$$f(x) := \mathbb{P}(X = x) = \mathbb{P}(\ \{\omega : X(\omega) = x\}\ ) = \begin{cases} \theta_i & \text{if } x = x_i \in \mathbb{X}. \\ 0 & \text{otherwise.} \end{cases} \tag{2.5}$$

Theorem 2.11. *The relation between the DF F and PMF f for a discrete RV X is as follows:*

1. *For any $x \in \mathbb{R}$,*
$$F(x) = \sum_{x_i \leq x} f(x_i) = \sum_{x_i \leq x} \theta_i\ . \tag{2.6}$$

2. *For any $a, b \in \mathbb{R}$ with $a < b$,*
$$F(b) - F(a) = \sum_{a < x_i \leq b} \theta_i\ . \tag{2.7}$$

   This is just the sum of all probabilities θ_i for which x_i satisfies $a < x_i \leq b$.

3. *From the fact that $\mathbb{P}(\Omega) = 1$, we get that the sum of all the probabilties is* 1*:*
$$\sum_i \theta_i = 1 \ . \tag{2.8}$$

Proof. Let us prove the first equality. First, recall that the definition of a discrete random variable required that X takes values in an enumerable $\mathbb{X}$, i.e. $\mathbb{X} = \{x_1, \ldots\}$, then for each $x_i \in \mathbb{X}$ define the sets

$$A_i = \{\omega : \omega \in \Omega, i - 1 < X(\omega) \leq i\} = \{X = i\},$$

and note that $A_i \in \mathcal{F}$, $\cup_i A_i = \Omega$ and they are mutually exclusive. Now using Definitions 1.22 and 2.2 we get

$$F(x) = \mathbb{P}(X \leq x) = \mathbb{P}\left(\cup_{i:x_i \leq x} A_i\right) = \sum_{i:x_i \leq x} \mathbb{P}(A_i) = \sum_{x_i \leq x} f(x_i).$$

The last equality of (2.6) is just the definition of θ_i in Definition 2.10.

The proof of the other properties is an Exercise! □

Remark 2.12. *When X only has finitely many possibilities, say k with $\mathbb{X} = \{x_1, x_2, \ldots, x_k\}$, then we may think of the probability $\mathbb{P}$ specified by $(\theta_1, \theta_2, \ldots, \theta_k)$ as a point in the* ***unit*** $(k-1)$ ***simplex****:*

$$\Delta^{k-1} := \{(\theta_1, \theta_2, \ldots, \theta_k) \in \mathbb{R}^k : \sum_i \theta_i = 1 \textit{ and } \theta_i \geq 0, \textit{ for all } i\} \tag{2.9}$$

In particular when X has only two possible values with $\mathbb{X} = \{x_1, x_2\}$ then $\theta_2 = 1 - \theta_1$, so we can avoid subscripts and take $\theta := \theta_1$ and realize that the probability $\mathbb{P}$ is now specified by the point $(\theta, 1 - \theta)$ in the ***unit*** 1 ***simplex****:*

$$\Delta^1 := \{(\theta, 1 - \theta) \in \mathbb{R}^2 : 0 \leq \theta \leq 1\} \ . \tag{2.10}$$

See https://en.wikipedia.org/wiki/Simplex.

Out of the class of discrete random variables we will define specific kinds as they arise often in applications. We classify discrete random variables into three types for convenience as follows:

- Discrete uniform random variables with finitely many possibilities
- Discrete non-uniform random variables with finitely many possibilities
- Discrete non-uniform random variables with (countably) infinitely many possibilities

Model 2.13 (Discrete Uniform)**.** *We say that a discrete random variable X is uniformly distributed over k possible values in $\mathbb{X} = \{x_1, x_2, \ldots, x_k\}$ if its probability mass function is:*

$$f(x) = \begin{cases} \theta_i = \frac{1}{k} & \textit{if } x = x_i, \quad \textit{where } i = 1, 2, \ldots, k \ , \\ 0 & \textit{otherwise} \ . \end{cases} \tag{2.11}$$

The distribution function for the discrete uniform random variable X is:

$$F(x) = \sum_{x_i \leq x} f(x_i) = \sum_{x_i \leq x} \theta_i = \begin{cases} 0 & \textit{if } -\infty < x < x_1 \ , \\ \frac{1}{k} & \textit{if } x_1 \leq x < x_2 \ , \\ \frac{2}{k} & \textit{if } x_2 \leq x < x_3 \ , \\ \vdots & \\ \frac{k-1}{k} & \textit{if } x_{k-1} \leq x < x_k \ , \\ 1 & \textit{if } x_k \leq x < \infty \ . \end{cases} \tag{2.12}$$

The discrete uniform RV with values in $\mathbb{X} = \{1, 2, \ldots, k\}$ is called the equi-probable de Moivre(k) *RV.*

Example 2.14. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability triple and let $A_1, \ldots, A_k \in \mathcal{F}$, then*

$$X = \sum_{i=1}^{k} \mathbb{1}_{A_i}$$

is a $\mathbb{R}$-valued discrete random variable, taking values in $\{0, \ldots, k\}$.

Can you write down the probability mass function and the distribution function for X when there is only two sets A_1, A_2?

2.3 Continuous Random Variables

If we have a random variable that does not take values in an enumerable set, then it is not discrete. However we often wish to allow the random variable to take any value in $\mathbb{R}$ or just every value in the interval $[0, 1]$. Let us define such a class of random variables.

Definition 2.15 (Continuous random variable)**.** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability triple and let X be a $\mathbb{R}$-valued random variable with distribution function F. We say that X is a* **continuous** *RV if there exists a piecewise-continuous function $f : \mathbb{R} \to [0, \infty]$, called the* **probability density function (PDF)** *of X, such that*

$$F(x) \ = \ \mathbb{P}(X \leq x) \ = \ \int_{-\infty}^{x} f(v) \, dv. \tag{2.13}$$

Remark 2.16. *see* https://en.wikipedia.org/wiki/Piecewise.

Remark 2.17. *There are actually random variables which are neither discrete or continuous, for instance, the product of a discrete and a continuous random variable!! We will deal with these later on.*

Theorem 2.18. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability triple and let X be a $\mathbb{R}$-valued continuous random variable, then the following holds:*

1. *For any $x \in \mathbb{R}$, the probability of observing a single value is zero:*

$$\mathbb{P}(X = x) = 0.$$

2. *The probability density function (density function) is the derivative of the distribution function. That is:*

$$f(x) = \frac{d}{dx}F(x) =: F'(x), \tag{2.14}$$

for every x at which $f(x)$ is continuous.

3. *For any $a, b \in \mathbb{R}$ with $a < b$,*

$$\begin{aligned} \mathbb{P}(a < X < b) = \mathbb{P}(a < X \leq b) = \mathbb{P}(a \leq X \leq b) \quad &= \mathbb{P}(a \leq X < b) \\ = F(b) - F(a) \quad &= \int_a^b f(v)dv \quad . \end{aligned} \tag{2.15}$$

4. *Finally:*

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

Proof. Fix x and let $\epsilon > 0$ be arbitrary, then from the properties of the probability measure, we get

$$F(x - \epsilon) \leq \mathbb{P}(X < x) \leq F(x)$$

However we know that integrals are continuous and as such $\lim_{\epsilon \to 0^+} F(x - \epsilon) = F(x)$, this proves that

$$\mathbb{P}(X < x) = \mathbb{P}(X \leq x). \tag{2.16}$$

Now to prove 1 we simply write

$$\mathbb{P}(X = x) = \mathbb{P}(x \leq X \leq x) = \mathbb{P}(X \leq x) - \mathbb{P}(X < x) = 0$$

where the last step follows from (2.16). Let $a \leq b$ and note that

$$\mathbb{P}(a \leq X \leq b) = \mathbb{P}(X \leq b) - \mathbb{P}(X < a)$$

Property 2 is the first fundamental theorem of calculus.

Property 3 follows from Proposition 2.4, (2.16), and Definition 2.15.

Finally propety 4 is an exercise!! □

The standard normal distribution is the most important continuous probability distribution. It was first described by De Moivre in 1733 and subsequently by C. F. Gauss (1777 - 1885). Many random variables have a normal distribution, or they are approximately normal, or can be transformed into normal random variables in a relatively simple fashion. Furthermore, the normal distribution is a useful approximation of more complicated distributions.

Model 2.19 (Normal$(0,1)$ or standard normal or Gaussian RV)**.** *A continuous random variable Z is called **standard normal** or* **standard Gaussian** *if its probability density function is*

$$\phi(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right) . \tag{2.17}$$

An exercise in calculus yields the first two derivatives of ϕ as follows:

$$\begin{aligned}\frac{d\phi}{dz} &= -\frac{1}{\sqrt{2\pi}} z \exp\left(-\frac{z^2}{2}\right) = -z\phi(z),\\ \frac{d^2\phi}{dz^2} &= \frac{1}{\sqrt{2\pi}}(z^2-1)\exp\left(-\frac{z^2}{2}\right) = (z^2-1)\phi(z) .\end{aligned}$$

Thus, ϕ has a global maximum at 0, it is concave down if $z \in (-1,1)$ and concave up if $z \in (-\infty,-1) \cup (1,\infty)$. This shows that the graph of ϕ is shaped like a smooth symmetric bell centred at the origin over the real line.

2.3.1 Viewing a deterministic real variable as a random variable

Consider the class of discrete RVs with distributions that place all probability mass on a single real number. This is the probability model for the deterministic real variable, which is often thought of as an unknown constant $\theta \in \mathbb{R}$.

Model 2.20 (Point Mass(θ))**.** *Given a specific point $\theta \in \mathbb{R}$, we say an RV X has point mass at θ or is* Point Mass(θ) *distributed if the DF is:*

$$F(x;\theta) = \begin{cases} 0 & \text{if } x < \theta \\ 1 & \text{if } x \geq \theta \end{cases} \tag{2.18}$$

and the PMF is:

$$f(x;\theta) = \begin{cases} 0 & \textit{if } x \neq \theta \\ 1 & \textit{if } x = \theta \end{cases} \tag{2.19}$$

Thus, Point Mass(θ) RV X is deterministic in the sense that every realisation of X is exactly equal to $\theta \in \mathbb{R}$. We will see that this distribution plays a central limiting role in asymptotic statistics.

2.4 Transformations of random variables

Suppose we know the distribution of a random variable X. How do we find the distribution of a transformation of X, say $g(X)$?

Now, let us return to our question of determining the distribution of the transformation $g(X)$.

2.4.1 Transformations of discrete random variables

To answer this question we must first observe that the inverse image $g^{[-1]}$ satisfies the following properties:

- $g^{[-1]}(\mathbb{Y}) = \mathbb{X}$
- For any set A, $g^{[-1]}(A^c) = \left(g^{[-1]}(A)\right)^c$
- For any collection of sets $\{A_1, A_2, \ldots\}$,

$$g^{[-1]}\left(A_1 \cup A_2 \cup \cdots\right) = g^{[-1]}(A_1) \cup g^{[-1]}(A_2) \cup \cdots \ .$$

Let $\mathbb{X}$ be an enumerable subset of $\mathbb{R}$, then $\mathbb{Y} := g(\mathbb{X})$ is also enumerable and $(g(X))^{[-1]}((-\infty, x)) \in \mathcal{F}$. This is a subtle point, and I strongly encourage you to try to figure out why this is!!

Consequentially,

$$\mathbb{P}_g(A) = P\left(g(X) \in A\right) = P\left(X \in g^{[-1]}(A)\right) \tag{2.20}$$

satisfies the axioms of probability and gives the desired probability of the event A from the transformation $Y = g(X)$ in terms of the probability of the event given by the inverse image of A underpinned by the random variable X. It is crucial to understand this from the sample space Ω of the underlying experiment in the sense that (2.20) is just short-hand for its actual meaning:

$$\begin{aligned} P\left(\{\omega \in \Omega : g(X(\omega)) \in A\}\right) &= P\left(\left\{\omega \in \Omega : X(\omega) \in g^{[-1]}(A)\right\}\right) \\ &= P\left(X^{[-1]}(g^{[-1]}(A))\right) . \end{aligned}$$

Because we have more than one random variable to consider, namely, X and its transformation $Y = g(X)$ we will subscript the probability density or mass function and the distribution function by the random variable itself. For example we denote the distribution function of X by $F_X(x)$ and that of Y by $F_Y(y)$.

For a discrete random variable X with probability mass function f_X we can obtain the probability mass function f_Y of $Y = g(X)$ using (2.20) as follows:

$$\begin{aligned} f_Y(y) &= \mathbb{P}(Y = y) = \mathbb{P}(Y \in \{y\}) \\ &= P\left(g(X) \in \{y\}\right) = P\left(X \in g^{[-1]}(\{y\})\right) \\ &= P\left(X \in g^{[-1]}(y)\right) = \sum_{x \in g^{[-1]}(y)} f_X(x) = \sum_{x \in \{x : g(x) = y\}} f_X(x) \ . \end{aligned}$$

This gives the formula:

$$f_Y(y) = \mathbb{P}(Y = y) = \sum_{x \in g^{[-1]}(y)} f_X(x) = \sum_{x \in \{x : g(x) = y\}} f_X(x) \ . \qquad (2.21)$$

2.4.2 Transformations of continuous random variables

Suppose we know F_X and/or f_X of a continuous random variable X. Recall (2.20), in this formula it is essential that

$$\{\omega \in \Omega : X^{[-1]}(g^{[-1]}(A))\} \in \mathcal{F}.$$

That is, what we need to know is, when is $g(X)$ itself a random variable with respect to $(\Omega, \mathcal{F}, \mathbb{P})$?

Definition 2.21. *We say that a function $g : \mathbb{R} \to \mathbb{R}$ is Borel, if we denote Σ the Borel sigma algebra on $\mathbb{R}$ (see Section 1.2.2) and for every $A \in \Sigma$*

$$g^{[-1]}(A) \in \Sigma.$$

Remark 2.22. *The reason we need this seemingly abstract notion, is to guarantee that if $g : \mathbb{R} \to \mathbb{R}$ is Borel, then if X is any $\mathbb{R}$-valued RV, then $g(X)$ is an $\mathbb{R}$-valued RV.*

Recall that the definition of a $\mathbb{R}$-valued RV X in some probability triple $(\Omega, \mathcal{F}, \mathbb{P})$ was that

$$X^{[-1]}((-\infty, x]) \in \mathcal{F}, \quad \textit{for all } x \in \mathbb{R}.$$

The Borel sigma-algebra Σ is the smallest sigma-algebra that contains the half open intervals $(-\infty, x]$ and thus if we take $A \in \Sigma$ then we also get $X^{[-1]}(A) \in \mathcal{F}$ by the properties of inverses. Thus it is immediate that $(g(X))^{[-1]}(A) \in \mathcal{F}$.

Remark 2.23. *Recall that in the case when X is a $\mathbb{R}$-valued discrete random variable, we don't need to assume anything about g. This was because of the fact that all enumerable sets of $\mathbb{R}$ is Borel, but we only mentioned it in passing.*

Remark 2.24. *A small subset of Borel functions are the continuous, piecewise continuous and monotone functions, which will be the ones we use mostly.*

Our objective now, is to obtain F_Y and/or f_Y of Y from F_X and/or f_X.

One-to-one transformations

The easiest case for transformations of continuous random variables is when g is **one-to-one and monotone** (monotone implies Borel).

- First, let us consider the case when g is **increasing** (monotone) on the range of the random variable X. In this case g^{-1} is also an increasing function and we can obtain the distribution function of $Y = g(X)$ in terms of the distribution function of X as

  $$F_Y(y) = P\left(Y \le y\right) = P\left(g(X) \le y\right) = P\left(X \le g^{-1}(y)\right) = F_X(g^{-1}(y)) \ .$$

  Now, let us use the chainrule to compute the density of Y as follows:

  $$f_Y(y) = \frac{d}{dy}F_Y(y) = \frac{d}{dy}F_X\left(g^{-1}(y)\right) = f_X\left(g^{-1}(y)\right)\frac{d}{dy}\left(g^{-1}(y)\right) \ .$$

- Second, let us consider the case when g is **decreasing** (monotone) on the range of the random variable X. In this case g^{-1} is also a decreasing function and we can obtain the distribution function of $Y = g(X)$ in terms of the distribution function of X as

  $$F_Y(y) = P\left(Y \le y\right) = P\left(g(X) \le y\right) = P\left(X \ge g^{-1}(y)\right) = 1 - F_X(g^{-1}(y)) \ ,$$

  and the density of Y as

  $$f_Y(y) = \frac{d}{dy}F_Y(y) = \frac{d}{dy}\left(1 - F_X\left(g^{-1}(y)\right)\right) = -f_X\left(g^{-1}(y)\right)\frac{d}{dy}\left(g^{-1}(y)\right) \ .$$

  For a decreasing g, its inverse function g^{-1} is also decreasing and consequently the density f_Y is indeed positive because $\frac{d}{dy}\left(g^{-1}(y)\right)$ is negative.

We can combine the above two cases and obtain the following

Proposition 2.25 (Change of variable formula)**.** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability triple and let X be a $\mathbb{R}$-valued RV. If $g : \mathbb{R} \to \mathbb{R}$ is one-to-one and monotone (increasing or decreasing) on the range of X, i.e $X(\Omega) \subset \mathbb{R}$, then*

$$f_Y(y) = f_X\left(g^{-1}(y)\right)\left|\frac{d}{dy}g^{-1}(y)\right| \quad . \tag{2.22}$$

The next example yields the *l*ocation-scale family of normal random variables via a family of linear transformations of the standard normal random variable.

Example 2.26. *Let Z be the standard Gaussian or standard normal random variable with probability density function $\phi(z)$ given by Equation (2.17). For real numbers $\sigma > 0$ and μ consider the linear transformation of Z given by*

$$Y = g(Z) = \sigma Z + \mu \quad .$$

We are interested in the density of the tranformed random variable $Y = g(Z) = \sigma Z + \mu$. Once again, since g is a one-to-one monotone function let us follow the four steps and use the change of variable formula to obtain f_Y from $f_Z = \phi$ and g.

1. *$y = g(z) = \sigma z + \mu$ is a monotone increasing function over $-\infty < z < \infty$, the range of Z. So, we can apply the change of variable formula.*

2. *$z = g^{-1}(y) = (y - \mu)/\sigma$ is a monotone increasing function over the range of y given by, $-\infty < y < \infty$.*

3. *For $-\infty < y < \infty$,*

$$\left|\frac{d}{dy}g^{-1}(y)\right| = \left|\frac{d}{dy}\left(\frac{y-\mu}{\sigma}\right)\right| = \left|\frac{1}{\sigma}\right| = \frac{1}{\sigma} \quad .$$

4. *we can use (2.17) and (2.22) which gives*

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

to find the density of Y as follows:

$$f_Y(y) = f_Z\left(g^{-1}(y)\right)\left|\frac{d}{dy}g^{-1}(y)\right| = \phi\left(\frac{y-\mu}{\sigma}\right)\frac{1}{\sigma} = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2\right] \quad ,$$

for $-\infty < y < \infty$.

Thus, we have obtained the expression for the probability density function of the linear transformation $\sigma Z + \mu$ of the standard normal random variable Z. This analysis leads to the following definition.

Model 2.27 (Normal(μ, σ^2) RV)**.** *Given a location parameter $\mu \in (-\infty, +\infty)$ and a scale parameter $\sigma^2 > 0$, the* Normal(μ, σ^2) *or Gaussian(μ, σ^2) random variable X has probability density function:*

$$f(x; \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] \qquad (\sigma > 0) \ . \tag{2.23}$$

The normal distribution has the **distribution function**

$$F(x; \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} \exp\left[-\frac{1}{2}\left(\frac{v-\mu}{\sigma}\right)^2\right] dv \ . \tag{2.24}$$

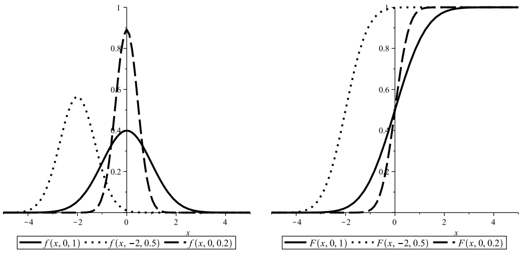

Figure 2.1: PDF and DF of a Normal(μ, σ^2) RV for different values of μ and σ^2

Direct method

If the transformation g in $Y = g(X)$ is not necessarily one-to-one then special care is needed to obtain the distribution function or density of Y. For a continuous random variable X with a known distribution function F_X we can obtain the distribution function F_Y of $Y = g(X)$ using (2.20) as follows:

$$\begin{aligned} F_Y(y) &= P\left(Y \leq y\right) = P\left(Y \in (-\infty, y]\right) \\ &= P\left(g(X) \in (-\infty, y]\right) = P\left(X \in g^{[-1]}((-\infty, y])\right) \\ &= P\left(X \in \{x : g(x) \in (-\infty, y]\}\right) \ . \end{aligned} \tag{2.25}$$

In words, the above equalities just mean that the probability that $Y \leq y$ is the probability that X takes a value x that satisfies $g(x) \leq y$. We can use this approach if it is reasonably easy to find the set $g^{[-1]}((-\infty, y]) = \{x : g(x) = (-\infty, y]\}$.

Example 2.28. *Let X be any random variable with distribution function F_X. Let $Y = g(X) = X^2$. Then we can find F_Y, the distribution function of Y from F_X as follows:*

- *Since $Y = X^2 \geq 0$, if $y < 0$ then $F_Y(y) = P\left(X \in \{x : x^2 < y\}\right) = \mathbb{P}(X \in \emptyset) = 0$.*
- *If $y \geq 0$ then*

$$\begin{aligned} F_Y(y) = P\left(Y \leq y\right) &= P\left(X^2 \leq y\right) \\ &= P\left(-\sqrt{y} \leq X \leq \sqrt{y}\right) \\ &= F_X(\sqrt{y}) - F_X(-\sqrt{y}) \ . \end{aligned}$$

By differentiation we get:

- *If $y < 0$ then $f_Y(y) = \frac{d}{dy}(F_Y(y)) = \frac{d}{dy}0 = 0$.*
- *If $y \geq 0$ then*

$$\begin{aligned} f_Y(y) = \frac{d}{dy}\left(F_Y(y)\right) &= \frac{d}{dy}\left(F_X(\sqrt{y}) - F_X(-\sqrt{y})\right) \\ &= \frac{d}{dy}\left(F_X(\sqrt{y})\right) - \frac{d}{dy}\left(F_X(-\sqrt{y})\right) \\ &= \frac{1}{2}y^{-\frac{1}{2}}f_X(\sqrt{y}) - \left(-\frac{1}{2}y^{-\frac{1}{2}}f_X(-\sqrt{y})\right) \\ &= \frac{1}{2\sqrt{y}}\left(f_X(\sqrt{y}) + f_X(-\sqrt{y})\right) \ . \end{aligned}$$

Therefore, the distribution function of $Y = X^2$ is:

$$F_Y(y) = \begin{cases} 0 & \text{if } y < 0 \\ F_X(\sqrt{y}) - F_X(-\sqrt{y}) & \text{if } y \geq 0 \ . \end{cases} \tag{2.26}$$

and the probability density function of $Y = X^2$ is:

$$f_Y(y) = \begin{cases} 0 & \text{if } y < 0 \\ \frac{1}{2\sqrt{y}}\left(f_X(\sqrt{y}) + f_X(-\sqrt{y})\right) & \text{if } y \geq 0 \ . \end{cases} \tag{2.27}$$

Using the direct method's (2.25), we can obtain the distribution function of the Normal(μ, σ^2) random variable from that of the tabulated distribution function of the Normal$(0, 1)$.

Proposition 2.29 (One Table to Rule Them All Gaussians)**.** *The distribution function $F_X(x;\mu,\sigma^2)$ of the* Normal(μ,σ^2) *random variable X and the distribution function $F_Z(z) = \Phi(z)$ of the standard normal random variable Z are related by:*

$$F_X(x;\mu,\sigma^2) \;=\; F_Z\left(\frac{x-\mu}{\sigma}\right) \;=\; \Phi\left(\frac{x-\mu}{\sigma}\right) \quad .$$

Proof. Let Z be a Normal$(0,1)$ random variable with distribution function $\Phi(z) = P(Z \leq z)$. We know that if $X = g(Z) = \sigma Z + \mu$ then X is the Normal(μ,σ^2) random variable. Therefore,

$$\begin{aligned} F_X(x;\mu,\sigma^2) &= P(X \leq x) = P\left(g(Z) \leq x\right) = P(\sigma Z + \mu \leq x) \\ &= P\left(Z \leq \frac{x-\mu}{\sigma}\right) \\ &= F_Z\left(\frac{x-\mu}{\sigma}\right) = \Phi\left(\frac{x-\mu}{\sigma}\right) \quad . \end{aligned}$$

□

Hence we often transform a general Normal(μ,σ^2) random variable, X, to a standardised Normal$(0,1)$ random variable, Z, by the substitution:

$$Z \;=\; \frac{X-\mu}{\sigma} \, .$$

2.5 Expectations and L^p spaces

Expectation is perhaps the most fundamental concept in probability theory. In fact, probability is itself an expectation as you will soon see!

Expectation is one of the fundamental concepts in probability. The expected value of a real-valued random variable gives the population mean, a measure of the centre of the distribution of the variable in some sense. Its variance measures its spread and so on.

Definition 2.30 (Expectation of a RV)**.** *The* **expectation***, or* **expected value***, or* **mean***, or* **first moment***, of a random variable X, with distribution function F and density f, is defined to be*

$$\mathbb{E}(X) := \int x \, dF(x) = \begin{cases} \sum_x x f(x) & \text{if } X \text{ is discrete} \\ \int x f(x) \, dx & \text{if } X \text{ is continuous} \ , \end{cases} \tag{2.28}$$

provided the sum or integral is well-defined. We say the expectation exists if

$$\int |x| \, dF(x) < \infty \ . \tag{2.29}$$

Sometimes, we denote $\mathbb{E}(X)$ by $\mathbb{E}\,X$ for brevity. Thus, the expectation is a single-number summary of the RV X and may be thought of as the average.

Definition 2.31. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be some probability triple, then for a random variable $X : \Omega \to \mathbb{R}$, we say that X is in $L^p(\mathbb{P})$ for some $1 \leq p < \infty$ if,*

$$\int |x|^p dF_X < \infty,$$

where F_X is the distribution function for X. If there is no fear of ambiguity we will simply write L^2 without referring to the measure $\mathbb{P}$.

So another way of saying that the expectation of the random variable X exists is the same as saying that $X \in L^1(\mathbb{P})$.

Definition 2.32 (Variance of a RV)**.** *Let X be a RV with mean or expectation $\mathbb{E}(X)$. The* **variance** *of X denoted by $\mathbb{V}(X)$ or simply $\mathbb{V}\,X$ is*

$$\mathbb{V}(X) := \mathbb{E}\left((X - \mathbb{E}(X))^2\right) = \int (x - \mathbb{E}(X))^2 \, dF(x) \ ,$$

provided this expectation exists. The **standard deviation** *denoted by* $\mathrm{sd}(X) := \sqrt{\mathbb{V}(X)}$*. Thus variance is a measure of "spread" of a distribution.*

Another way of saying that the variance exists is to say that $X \in L^2(\mathbb{P})$.

Definition 2.33 (k-th moment of a RV)**.** *Let $k = 1, \ldots,$ we call*

$$\mathbb{E}\Big(X^k\Big) = \int x^k \, dF(x)$$

as the k-th moment of the RV X and say that the k-th moment exists when $X \in L^k$. We call the following expectation as the k-th central moment:

$$\mathbb{E}\Big((X - \mathbb{E}(X))^k\Big) \; .$$

2.6 Multivariate Random Variables

Often, in experiments we are measuring two or more aspects simultaneously. For example, we may be measuring the diameters and lengths of cylindrical shafts manufactured in a plant or heights, weights and blood-sugar levels of individuals in a clinical trial. Thus, the underlying outcome $\omega \in \Omega$ needs to be mapped to measurements as realizations of random vectors in the real plane $\mathbb{R}^2 = (-\infty, \infty) \times (-\infty, \infty)$ or the real space $\mathbb{R}^3 = (-\infty, \infty) \times (-\infty, \infty) \times (-\infty, \infty)$:

$$\omega \mapsto (X(\omega), Y(\omega)) : \Omega \to \mathbb{R}^2 \qquad\qquad \omega \mapsto (X(\omega), Y(\omega), Z(\omega)) : \Omega \to \mathbb{R}^3$$

More generally, we may be interested in heights, weights, blood-sugar levels, family medical history, known allergies, etc. of individuals in the clinical trial and thus need to make m measurements of the outcome in $\mathbb{R}^m$ using a "measurable mapping" from $\Omega \to \mathbb{R}^m$. To deal with such multivariate measurements we need the notion of **random vectors** (R$\vec{\text{V}}$s), i.e. ordered pairs of random variables (X, Y), ordered triples of random variables (X, Y, Z), or more generally ordered m-tuples of random variables $(X_1, X_2, \ldots, X_m)$.

We begin by defining what we mean

Definition 2.34 (Random Variable)**.** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be some probability triple. Then, a $\mathbb{R}^m$ valued* **Random Variable (RV)***, say X, is a function from the sample space Ω to the set vectors $\mathbb{R}^m$*

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

such that the inverse image of the half-open product intervals $(-\infty, x_1] \times \cdots \times (-\infty, x_m]$ for $x = (x_1, \ldots, x_m) \in \mathbb{R}^m$ is an element of the collection of events $\mathcal{F}$, i.e., for every $x \in \mathbb{R}^m$ we have

$$X^{[-1]}(\; (-\infty, x_1] \times \cdots \times (-\infty, x_m] \;) := \{\omega : X(\omega) \le x\} \in \mathcal{F},$$

where we interpret the inequality $X \le x$ to hold for all components. This definition can be summarised by the statement that a RV is an $\mathcal{F}$-measurable map from Ω to $\mathbb{R}^m$.

This definition looks remarkably similar to Definition 2.1, in fact it is the same. We can simply write it as follows.

Definition 2.35 (Abstract definition of a RV). *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability triple and let $(\mathbb{X}, \Sigma_X)$ be a space together with a sigma algebra Σ_X. Then we call a function $X : \Omega \to \mathbb{X}$ an $\mathbb{X}$-valued RV if*

$$X^{[-1]}(A) \in \mathcal{F}, \qquad \text{for all } A \in \Sigma_X.$$

If we in the above definition just set $\mathbb{X} = \mathbb{R}^m$ and Σ_X the Borel sigma algebra (or any subset that generates the Borel sigma algebra, like half spaces) as in Section 1.2.2, we obtain Definition 2.34.

Let us leave the abstract notion and go back to $\mathbb{R}^m$ valued RV's. Specifically, let us define the corresponding distribution function:

Definition 2.36 (JDF). *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability triple and let X be a $\mathbb{R}^m$ valued RV. Then the* **joint distribution function (JDF)** *or* **joint cumulative distribution function (JCDF)**, *$F_X(x) : \mathbb{R}^m \to [0,1]$ is defined as*

$$\begin{aligned} F_X(x) &= \mathbb{P}(\cap_{i=1}^m (X_i \le x_i)) \ = \ \mathbb{P}(X_1 \le x_1, \dots, X_m \le x_m) \\ &= P\left(\{\omega : X_1(\omega) \le x_1, \dots, X_m \le x_m\}\right), \end{aligned}$$

where $X = (X_1, \dots, X_m)$ and each $X_i \in \mathbb{R}$, and $x = (x_1, \dots, x_m) \in \mathbb{R}^m$.

Why do we need this? Well, lets say we do two measurements, X and Y. What does $\mathbb{P}(X \le x, Y \le y)$ mean? Let $Z = (X, Y)$, then consider this as a $\mathbb{R}^2$ valued RV, as such we can write

$$F_Z(z) = F_{X,Y}(x, y) = \mathbb{P}(X \le x, Y \le y).$$

The above extends to any finite number of random variables.

Lets now say that we have two random variables X, Y and we wish to compute something like

$$\mathbb{E}[X + Y], \quad \mathbb{E}[XY], \quad \mathbb{E}[X^r Y^s]$$

etc. then we need the joint distribution function to compute the above expecations.

From the above motivation, we would expect that each component of a $\mathbb{R}^m$ random variable is again a random variable:

Theorem 2.37. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability triple and let Z be a $\mathbb{R}^m$ valued RV, then for any $i = 1, \dots, m$, Z_i is a $\mathbb{R}$-valued RV. Its distribution function*

$$\begin{aligned} F_{Z_i}(z_i) &= \mathbb{P}(Z_i \le z_i) \\ &= \mathbb{P}(Z_1 \le \infty, \dots, Z_{i-1} \le \infty, Z_i \le z_i, Z_{i+1} \le \infty, \dots, Z_m \le \infty) \end{aligned}$$

is called the **marginal distribution**.

Proof. Let us prove this in the case when $m = 2$, that is, let $Z = (X, Y)$. Let us show that X is a $\mathbb{R}$ valued RV. We do this by showing that

$$X^{[-1]}((-\infty, x)) \in \mathcal{F}$$

What do we mean by $X^{[-1]}$? We mean,

$$\begin{aligned} X^{[-1]}((-\infty, x)) &= \{\omega \in \Omega : X(\omega) \leq x\} = \{\omega \in \Omega : X(\omega) \leq x, Y(\omega) < \infty\} \\ &= Z^{[-1]}((-\infty, x) \times (-\infty, \infty)) \in \mathcal{F} \end{aligned}$$

where the final step follows from the definition of Z being a $\mathbb{R}^2$-valued RV. □

We have seen the notion of independence of two events in Definition 1.18 or of a sequence of events in Definition 1.20. Recall that independence amounts to having the probability of the joint occurrence of the events to be given by the product of the probabilities of each of the events.

We can use the definition of independence of two events to define the independence of two random variables using their distribution functions.

Definition 2.38 (Independence of Two RVs)**.** *Consider an $\mathbb{R}^2$-valued RV $X := (X_1, X_2)$. Then the $\mathbb{R}$-valued RVs X_1 and X_2 are said to be independent or independently distributed if and only if*

$$\mathbb{P}(X_1 \leq x_1, X_2 \leq x_2) = \mathbb{P}(X_1 \leq x_1)\, \mathbb{P}(X_2 \leq x_2)$$

or equivalently,

$$F_{X_1,X_2}(x_1, x_2) = F_{X_1}(x_1) F_{X_2}(x_2) \ ,$$

for any pair of real numbers $(x_1, x_2) \in \mathbb{R}^2$.

The above definition can be extended to a sequence of random variables in the same way as in Definition 1.20.

2.6.1 Discrete random vectors

Let us specify the above fairly abstract concepts into something tangible, we start with discrete random vectors.

Definition 2.39. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability triple and let Z be a $\mathbb{R}^m$ valued RV, we say that Z is a discrete random variable if it takes values in an enumerable set $\mathbb{Z}$. The* **joint probability mass function** *f_Z of Z is the function $f_Z : \mathbb{R}^m \to [0, 1]$ defined as follows*

$$f_Z(z) := \mathbb{P}(Z = z) = \mathbb{P}(\{\omega : Z(\omega) = z\}) = \begin{cases} \theta_h & \textit{if } z = z_h \in \mathbb{Z} \\ 0 & \textit{otherwise} \end{cases}$$

The above definition is easier to grasp in the 2-d case. Let $Z = (X, Y) \in \mathbb{R}^2$ and let $\mathbb{X}$ and $\mathbb{Y}$ be two enumerable sets and consider their product $\mathbb{Z} = \{(x_i, y_j) : i = 1, \ldots, j = 1, \ldots\} \subset \mathbb{R}^2$ and we have $\theta_{i,j} = \mathbb{P}(X = x_i, Y = y_j) > 0$, then the JPMF can be written as

$$\begin{aligned} f_{X,Y}(x, y) &:= \mathbb{P}(X = x, Y = y) = \mathbb{P}(\{\omega : X(\omega) = x, Y(\omega) = y\}) \\ &= \begin{cases} p_{i,j} & \text{if } x = x_i, y = y_i, (x_i, y_i) \in \mathbb{Z} \\ 0 & \text{otherwise} \end{cases} \end{aligned}$$

In this case we can quite easily compute also the **marginal distribution** and the **marginal probability mass function** as follows: Say we have X and Y as above, and lets compute the marginal distribution for X, according to the definition it is

$$\begin{aligned} F_X(x) &= F_{X,Y}(x, \infty) = \mathbb{P}(X \leq x, Y \leq \infty) \\ &= \mathbb{P}(\{\omega : X(\omega) \leq x\} \cap \{\omega : Y(\omega) \leq \infty\}) \end{aligned}$$

now since Y is also discrete, i.e. taking values $y_1, y_2, \ldots$ we can write the above as

$$\{\omega : Y(\omega) \leq \infty\} = \bigcup_j \{\omega : Y(\omega) = y_j\} := \bigcup_j A_j$$

And clearly all A_j are mutually exclusive we can thus write

$$\begin{aligned} F_X(x) &= \mathbb{P}\left(\{\omega : X(\omega) \leq x\} \cap \left(\bigcup_j A_j\right)\right) \\ &= \sum_j \mathbb{P}(\{\omega : X(\omega) \leq x\} \cap A_j) \\ &= \sum_{x_i \leq x} \sum_j p_{i,j} \end{aligned}$$

if we now define $p_i = \sum_j p_{i,j}$ then we simply have

$$F_X(x) = \sum_{x_i \leq x} p_i$$

which is the same as if we had defined X alone as a discrete random variable. From the above it is also clear that the marginal probability mass function is given by

$$f_X(x) := \mathbb{P}(X = x) = \mathbb{P}(\{\omega : X(\omega) = x\}) = \begin{cases} p_i & \text{if } x = x_i \in \mathbb{X} \\ 0 & \text{otherwise} \end{cases}$$

2.6.2 Continuous random vectors

Arguably the continuous random variables are easier:

Definition 2.40. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability triple and let Z be a $\mathbb{R}^m$ valued RV. We say that Z is a continuous random variable if there exists a piecewise-continuous function $f_Z : \mathbb{R}^m \to [0, \infty)$, called the* **joint probability density function** *of Z such that*

$$F_Z(z) = \mathbb{P}(Z \leq z) = \int_{-\infty}^{z_1} \cdots \int_{-\infty}^{z_m} f_Z(v_1, \ldots, v_m) dv_1 \ldots dv_m.$$

The above definition is easier to grasp in the 2-d case. Consider $Z = (X, Y) \in \mathbb{R}^2$ then the joint density function satisfies

$$F_{X,Y}(x, y) = \mathbb{P}(X \leq x, Y \leq y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f_{X,Y}(u, v) du dv$$

In this case we can quite easily compute also the **marginal distribution** and the **marginal probability mass function** as follows: Say we have X and Y as above, and lets compute the marginal distribution for X, according to the definition it is

$$\begin{aligned} F_X(x) &= F_{X,Y}(x, \infty) = \mathbb{P}(X \leq x, Y \leq \infty) \\ &= \int_{-\infty}^{x} \int_{-\infty}^{\infty} f_{X,Y}(u, v) du dv \end{aligned}$$

Now let us define

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, v) dv$$

then

$$F_X(x) = \int_{-\infty}^{x} f_X(u) du.$$

This means that by Definition 2.15 X is a continuous random variable with density f_X.

For independent random variables we get

Theorem 2.41. *Consider two independent continuous $\mathbb{R}$ valued RVs, X, Y. Then*

$$f_{X,Y}(x, y) = f_X(x) f_Y(y).$$

Proof. Now using Definition 2.38 together with properties of interals we get

$$\begin{aligned} F_{X,Y}(x,y) = F_X(x)F_Y(y) &= \int_{-\infty}^{x} f_X(u)du \int_{-\infty}^{y} f_Y(v)dv \\ &= \int_{-\infty}^{x} f_X(u) \left(\int_{-\infty}^{y} f_Y(v)dv \right) du \\ &= \int_{-\infty}^{y} \int_{-\infty}^{x} f_X(u)f_Y(v)dudv. \end{aligned}$$

□

As we saw in the chapter about $\mathbb{R}$ valued RVs, we can look at functions of RVs. The first thing we need is an extended notion of Borel.

Definition 2.42. *We say that a function $g : \mathbb{R}^m \to \mathbb{R}^k$ is Borel, if we denote Σ_r the Borel sigma algebra on $\mathbb{R}^r$, $r = 1, \dots$ (see Section 1.2.2) and for every $A \in \Sigma_k$*

$$g^{[-1]}(A) \in \Sigma_m.$$

Lemma 2.43. *Let X be an $\mathbb{R}^m$ valued RV and let $g : \mathbb{R}^m \to \mathbb{R}^k$ be Borel. Then $g(X)$ is a $\mathbb{R}^k$ valued RV.*

Exercise 2.44. *Prove the above lemma in the case when $k = 1$.*

2.6.3 Properties of expectations

Now that we know what a joint distribution function is, we can make sense of for instance

$$\mathbb{E}[X + Y]$$

Let us list the immediate properties of the expectation here

Theorem 2.45. *Properties of the expectation*

1. *If $X \in L^1(\mathbb{P})$ is an $\mathbb{R}$ valued RV and $\alpha \in \mathbb{R}$, then*

$$\mathbb{E}[\alpha X] = \alpha \mathbb{E}[X]$$

2. *If $X, Y \in L^1(\mathbb{P})$ are $\mathbb{R}$ valued RV, then*

$$\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y].$$

3. *If $X, Y \in L^2(\mathbb{P})$ are independent $\mathbb{R}$ valued RV, then*

$$\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y].$$

4. *If $X, Y \in L^1(\mathbb{P})$ are $\mathbb{R}$-valued RVs then if $X \leq Y$ a.s., then*

$$\mathbb{E}[X] \leq \mathbb{E}[Y].$$

5. *Let X be an $\mathbb{R}$-valued RV then if $A \subset \mathbb{R}$ is Borel, then*

$$\mathbb{E}[\mathbb{1}_A(X)] = \mathbb{P}(X \in A)$$

Proof. We will only prove the above in the case of continuous random variables: For 1 we note that the linearity of integrals we have

$$\mathbb{E}[\alpha X] = \int_{-\infty}^{\infty} (\alpha x) dF(x) = \alpha \int_{-\infty}^{\infty} x dF(x) = \alpha \mathbb{E}[X].$$

Let us now consider 2, which requires the concept of marginals,

$$\begin{aligned} \mathbb{E}[X+Y] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (x+y) dF_{X,Y}(x,y) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x dF_{X,Y}(x,y) \\ &+ \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} y dF_{X,Y}(x,y) = I_X + I_Y \end{aligned}$$

Now

$$\begin{aligned} I_X &= \int_{-\infty}^{\infty} x dF_{X,Y}(x,y) = \int_{-\infty}^{\infty} x \left(\int_{-\infty}^{\infty} f_{X,Y}(x,y) dy \right) dx \\ &= \int_{-\infty}^{\infty} x f_X(x) dx = \mathbb{E}[X]. \end{aligned}$$

To prove 3 note that since X and Y are independent, then from Theorem 2.41 we have $f_{X,Y}(x,y) = f_X(x) f_Y(y)$ which gives us

$$\begin{aligned} \mathbb{E}[XY] &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xy dF_{X,Y}(x,y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xy f_X(x) f_Y(y) dx dy \\ &= \int_{-\infty}^{\infty} x f_X(x) \left(\int_{-\infty}^{\infty} y f_Y(y) dy \right) dx \\ &= \left(\int_{-\infty}^{\infty} x f_X(x) dx \right) \left(\int_{-\infty}^{\infty} y f_Y(y) dy \right) \\ &= \mathbb{E}[X] \mathbb{E}[Y]. \end{aligned}$$

□

2.6.4 L^p is a normed vector space

Definition 2.46. *We say that a set $\mathbb{X}$ is a vector space if there is a notion of addition, and a notion of multiplication with scalars, such that*

1. *For $X, Y \in \mathbb{X}$, we have $X + Y \in \mathbb{X}$,*

2. *For a scalar $\alpha \in \mathbb{R}$ we have $\alpha X \in \mathbb{X}$ if $X \in \mathbb{X}$.*

Definition 2.47. *We say that a set $\mathbb{X}$ is a normed vector space if it is a vector space and the following holds true: there is a function $\|\cdot\| : \mathbb{X} \to \mathbb{R}$ (called a* **norm***) such that*

1. *$\|X\| \geq 0$ for any $X \in \mathbb{X}$,*

2. *$\|X\| = 0$ if and only if $X = 0$,*

3. *For every vector $X \in \mathbb{X}$ and every $\alpha \in \mathbb{R}$, one has*
$$\|\alpha X\| = |\alpha|\|X\|$$

4. *The* triangle inequality *holds, that is, for $X \in \mathbb{X}$ and $Y \in \mathbb{X}$ we have*
$$\|X + Y\| \leq \|X\| + \|Y\|.$$

Theorem 2.48. *Let $(\Omega, \mathcal{F}, P)$ be a probability triple, then the set of random variables $L^p(\mathbb{P})$ for $1 \leq p < \infty$ is a normed vector space, with norm*
$$\|X\|_{L^p(\mathbb{P})} = (\mathbb{E}[|X|^p])^{\frac{1}{p}}$$
and with $X = Y$ in $L^p(\mathbb{P})$ if $X(\omega) = Y(\omega)$ for a.e $\omega \in \Omega$ with respect to $\mathbb{P}$. That is $\mathbb{P}(\{\omega \in \Omega, X(\omega) = Y(\Omega)) = 0$.

How would we prove such a theorem? Well, we basically need to verify all conditions in Definitions 2.46 and 2.47. Verifying Definition 2.46 and conditions 1,2,3 of Definition 2.47 is left to you to verify. We will however prove the triangle inequality using Hölders inequality:

Theorem 2.49. *Let $X \in L^p(\mathbb{P})$ and $Y \in L^q(\mathbb{P})$ with $\frac{1}{p} + \frac{1}{q} = 1$, for $1 < p < \infty$, then*
$$\mathbb{E}[XY] \leq \mathbb{E}[|X|^p]^{1/p}\,\mathbb{E}[|Y|^q]^{1/q}.$$

Lemma 2.50. *Let $\phi : \mathbb{R} \to \mathbb{R}$ be a convex function, and consider numbers x_1, x_2 and parameter $t \in [0, 1]$, then*
$$\phi(tx_1 + (1 - t)x_2) \leq t\phi(x_1) + (1 - t)\phi(x_2).$$

Lemma 2.51. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability triple and let $X \in L^1(\mathbb{P})$ be a $\mathbb{R}$-valued RV. Then if $\phi(X) \in L^1(\mathbb{P})$, where $\phi : \mathbb{R} \to \mathbb{R}$ is a convex function, we have*
$$\phi(\mathbb{E}[X]) \leq \mathbb{E}[\phi(X)].$$

We will prove the triangle inequality in the case that $p = 2$ (we leave $p \neq 2$ as an exercise), that is in Theorem 2.49 we have $p = q = 2$, hence

$$\begin{aligned}\mathbb{E}\left[|X+Y|^2\right] &= \mathbb{E}\left[|X+Y||X+Y|\right] \leq \mathbb{E}\left[(|X|+|Y|)|X+Y|\right] \\ &\leq \left(\mathbb{E}\left[|X|^2\right]^{1/2} + \mathbb{E}\left[|Y|^2\right]^{1/2}\right)\mathbb{E}\left[|X+Y|^2\right]^{1/2}\end{aligned}$$

in the first inequality we used the triangle inequality for the absolute value and in the last inequality we used Theorem 2.49 with $p = q = 2$. Dividing both sides by $\mathbb{E}\left[|X+Y|^2\right]^{1/2}$ gives

$$\mathbb{E}\left[|X+Y|^2\right]^{1/2} \leq \mathbb{E}\left[|X|^2\right]^{1/2} + \mathbb{E}\left[|Y|^2\right]^{1/2}$$

or equivalently

$$\|X+Y\|_{L^2(\mathbb{P})} \leq \|X\|_{L^2(\mathbb{P})} + \|Y\|_{L^2(\mathbb{P})},$$

which proves the triangle inequality. For the case of $p \neq 2$ one does

$$\mathbb{E}\left[|X+Y|^p\right] = \mathbb{E}\left[|X+Y||X+Y|^{p-1}\right] \leq \mathbb{E}\left[(|X|+|Y|)|X+Y|^{p-1}\right]$$

and then apply Hölders inequality with $q = \frac{p-1}{p}$.

Theorem 2.52. *The following are consequences of Hölders inequality*

1. *If $X \in L^p(\mathbb{P})$ and $Y \in L^q(\mathbb{P})$ then $XY \in L^1(\mathbb{P})$ if $\frac{1}{p} + \frac{1}{q} = 1$.*
2. *If $X \in L^r(\mathbb{P})$ then $X \in L^s(\mathbb{P})$ for $1 \leq s \leq r < \infty$.*
3. *If $X \in L^p(\mathbb{P})$ and $Y \in L^q(\mathbb{P})$ then $XY \in L^r(\mathbb{P})$ if $\frac{1}{p} + \frac{1}{q} = \frac{1}{r}$.*

Proof. Statement 1 follows immediately from Theorem 2.49.

To prove 2. First apply Hölders inequality with p to be chosen

$$\mathbb{E}\left[|X|^s\right] = \mathbb{E}\left[|X|^s \times 1\right] \leq \mathbb{E}\left[|X|^{sp}\right]^{1/p}\mathbb{E}\left[1^q\right]^{1/q}.$$

Now choose $p = r/s \geq 1$ which gives

$$\mathbb{E}\left[|X|^s\right] \leq \mathbb{E}\left[|X|^r\right]^{\frac{s}{r}}$$

or equivalently

$$\|X\|_{L^s(\mathbb{P})} \leq \|X\|_{L^r(\mathbb{P})}.$$

To prove 3. Apply Hölders inequality with the pair (s, h) and get

$$\mathbb{E}\left[|XY|^r\right] = \mathbb{E}\left[|X|^r|Y|^r\right] \leq \mathbb{E}\left[|X|^{rs}\right]^{1/s}\mathbb{E}\left[|Y|^{rh}\right]^{1/h}$$

now choose $h = \frac{p}{p-r}$ and $s = \frac{s-1}{s} = \frac{p}{r}$ which gives

$$\mathbb{E}\left[|XY|^r\right] \leq \mathbb{E}\left[|Y|^{\frac{pr}{p-r}}\right]^{(p-r)/p} \mathbb{E}\left[|X|^p\right]^{r/p}$$

the last term is finite but what about the expectation of Y, well note that

$$\frac{pr}{p-r} = q.$$

□

The same ideas as in Theorem 2.52 can be extended to a product of k random variables. Here the Hölder exponents become $\sum_{i=1}^{k} \frac{1}{p_i} = 1/r$. Try this out for yourself!

Functions of random variables

What we saw above with moments, i.e. powers of random variables is a special case of functions of random variables. Often we are interested in functions of random variables, like correlation etc. Let us define what we mean with the expectation of a random variable.

Definition 2.53 (Expectation of a function of a RV)**.** *The* ***Expectation*** *of a function $g(X)$ of a random variable X is defined as:*

$$\mathbb{E}(g(X)) := \int g(x)dF(x) = \begin{cases} \sum_x g(x)f(x) & \text{if } X \text{ is a discrete RV} \\ \int_{-\infty}^{\infty} g(x)f(x)dx & \text{if } X \text{ is a continuous RV} \end{cases}$$

provided $\mathbb{E}(g(X))$ exists, i.e., $\int |g(x)|dF(x) < \infty$.

Remark 2.54. *Notational convenience: Note in the above how we wrote*

$$\mathbb{E}\left[g(X)\right] = \int g(x)dF(x)$$

even if X was discrete. This can be made rigorous by the introduction of point measures (dirac measures or atomic measure), we will however skip that part in this course and instead work with the understanding that the integral is a sum in the discrete case.

The above can be taken as a defintion, but why does it make sense? It is actually something we can prove, and its called the law of the unconcious statistician. We will only prove this in the case of one-to-one monotone transformations. Lets do it for increasing functions f. If we think of the transformation part we note that using the direct method that if X is a random variable and f is a function then for $Y = f(X)$ the distribution is

$$F_Y(y) = \mathbb{P}\left[X \in \{x : f(x) \in (-\infty, y]\}\right]$$

and $f_Y = \frac{d}{dy}F_Y(y)$, so by a change of variables we get $y = f(x)$

$$\mathbb{E}\left[Y\right] = \int y f_Y(y)dy = \int f(x)f_Y(f(x))\frac{df}{dx}dx$$

now, $\frac{d}{dx}F_Y(f(x)) = f_Y(f(x))\frac{df}{dx}$, but since f is monotone and increasing we get $F_Y(f(x)) = P(f(X) \leq f(x)) = \mathbb{P}\left[X \leq x\right] = F_X(x)$. This gives that

$$\mathbb{E}\left[Y\right] = \int y f_Y(y)dy = \int f(x)f_Y(f(x))\frac{df}{dx}dx = \int f(x)f_X(x)dx$$

Now all of this mean that $f(X)$ is a random variable and one can question in what $L^p(\mathbb{P})$ it lies, i.e. what value of p? Well, the existence of the expectation is equivalent to saying that $f(X) \in L^1(\mathbb{P})$. But actually, since if $X \in \mathbb{R}$ is a real valued random variable with density f_X where $f_X : \mathbb{R} \to \mathbb{R}$ then we can actually view everything from the perspective of integrability in $\mathbb{R}$. That is, we can define

$$L^p(F) = \{f \mid f : \mathbb{R} \to \mathbb{R}, \int |f(x)|^p dF(x) < \infty\}$$

this space of functions from $\mathbb{R}$ to $\mathbb{R}$ inherits all properties of $L^p(\mathbb{P})$ and as such it is a normed vector space.

2.6.5 Conditional Random Variables

Often we will have a condition where one of the two random variables that make up a random vector (X_1, X_2) already occurs and takes a value. And we might want to compute the probability of the occurrence of the other random variable given this conditional information. For this all we need to do is extend the idea of conditional probabilities to $\mathbb{R}^2$-valued random variables.

Definition 2.55. *Let (X, Y) be a $\mathbb{R}^2$ valued random variable, and let $A \subset \mathbb{R}$ be a Borel set such that $\mathbb{P}(Y \in A) > 0$ then define the conditional distribution function of X given that $Y \in A$ as*

$$F_{X|Y}(x \mid A) := \frac{\mathbb{P}(X \leq x, Y \in A)}{\mathbb{P}(Y \in A)}.$$

Lemma 2.56. *Let (X, Y) be a $\mathbb{R}^2$ valued random variable, then*

$$F_{X|Y}(x \mid (-\infty, y))F_Y(y) = \frac{F_{X,Y}(x, y)}{F_Y(y)}F_Y(y) = F_{X,Y}(x, y).$$

If Y is a discrete random variable and $f_Y(y) > 0$ the above definition is well defined for $A = \{y\}$ and we can write

$$F_{X|Y}(x \mid y) := \frac{\mathbb{P}(X \leq x, Y = y)}{\mathbb{P}(Y = y)}.$$

If however Y is continuous we know from Theorem 2.18 that $\mathbb{P}(Y = y) = 0$ and as such Definition 2.55 does not apply and we need another definition. Fix a point $y \in \mathbb{R}$ such that $f_Y(y) > 0$ and let $\epsilon > 0$, then define $U_\epsilon = [y - \epsilon, y + \epsilon]$, hence $\mathbb{P}(Y \in U_\epsilon) > 0$ since f_Y is piecewise continuous. We can now compute

$$\begin{aligned} F_{X|Y}(x|U_\epsilon) = \frac{\mathbb{P}(X \leq x, Y \in U_\epsilon)}{\mathbb{P}(Y \in U_\epsilon)} &= \frac{\int_{-\infty}^{x} \left(\int_{y-\epsilon}^{y+\epsilon} f_{X,Y}(u,v) dv \right) dv}{\int_{y-\epsilon}^{y+\epsilon} f_Y(v) dv} \\ &= \frac{\int_{-\infty}^{x} \left(\frac{1}{2\epsilon} \int_{y-\epsilon}^{y+\epsilon} f_{X,Y}(u,v) dv \right) dv}{\frac{1}{2\epsilon} \int_{y-\epsilon}^{y+\epsilon} f_Y(v) dv} \end{aligned}$$

from this we can "define"

$$F_{X|Y}(x \mid y) := \lim_{\epsilon \to 0} F_{X|Y}(x|U_\epsilon) = \int_{-\infty}^{x} \frac{f_{X,Y}(u,y)}{f_Y(y)} du.$$

Why "define", well basically we need to make sure what happens if we are at points of discontinuity of $f_{X,Y}(u,v)$ and $f_Y(y)$. This does not turn into any problems as there are only distinct points of discontinuity for $f_Y(y)$ and $f_{X,Y}(x,y)$ which all have probability zero and dont contribute to the integral.

Definition 2.57 (Conditional PDF or PMF). *Let (X, Y) be a $\mathbb{R}^2$ valued RV. Then the* **conditional probability mass / density function** *is defined as*

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

if $f_Y(y) > 0$

Lemma 2.58. *Let (X, Y) be a $\mathbb{R}^2$ valued RV. Then*

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

where the left hand side is interpreted as 0 if $f_Y(y) = 0$.

Exercise 2.59. *Consider two independent fair coin tosses (2 sided), i.e. $X, Y \sim \text{Bernoulli}(1/2)$, and let 1 be heads and 0 is tails. Let $Z = X + Y$, what is the PMF of Z given X. Once, you have this, what is the joint PMF of (Z, X)?*

These definitions allows us to construct conditional expectations, like

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

But it also allows for the following very useful property

Theorem 2.60 (The tower property)**.** *Let (X, Y) be a $\mathbb{R}^2$ valued RV where $\mathbb{E}[X]$ is well defined. Then*

$$\mathbb{E}[\mathbb{E}[X \mid Y]] = \mathbb{E}[X].$$

In the above we introduced a new notation, namely $\mathbb{E}[X \mid Y]$, what is this? Denote $g(y) = \mathbb{E}[X \mid Y = y]$, then define

$$\mathbb{E}[X \mid Y] := g(Y).$$

Proof. We will also prove this only for continuous RVs, the discrete is an exercise!!

Lets begin by writing down the LHS

$$\begin{aligned}\mathbb{E}[\mathbb{E}[X \mid Y]] = \mathbb{E}[g(Y)] = \int_{-\infty}^{\infty} g(y) f_Y(y) dy = \int_{-\infty}^{\infty} \mathbb{E}[X \mid Y = y] f_Y(y) dy \\ = \int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} x f_{X|Y}(x \mid y) dx \right) f_Y(y) dy \\ = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x f_{X|Y}(x \mid y) f_Y(y) dx dy \\ = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x f_{X,Y}(x, y) dx dy = \mathbb{E}[X].\end{aligned}$$

We switched the order of integration which follows by Fubinis theorem and our assumption that $\mathbb{E}[X]$ is well defined, i.e. $\mathbb{E}[|X|] < \infty$. In the last stages we used Lemma 2.58 and the definition of marginal density. □

2.6.6 Mixed random variables

We have dealt with both discrete and continuous random variables, but what about mixed random variables? For instance, if X is continuous and Y is discrete, what is (X, Y)? Actually it is neither, but we have all the tools to deal with it, we just have to mix the two concepts.

2.7 Examples Of Modeling

Now that we have developed the language of

1. Probability
2. Random variables
3. Dependence and independence
4. Conditional distribution etc.

We can take some common problems encountered in data science and model it.

2.7.1 Email spam filtering

Lets say that you wish to construct an email filter that takes as input an email and predicts if it is spam or not.

- The experiment, the next incoming email.
- $\Omega = \{\text{All strings of length 1000}\}$, that is, an outcome is one email (We limit to the first 1000 characters, but that is arbitrary).
- Ω is finite, so the σ-algebra is just all subsets of Ω, i.e. the power set.
- The probability measure $\mathbb{P}$ is unknown, but it is there, not all emails are equally probable. This is crucially the case, because we want to estimate probabilities based on the data.
- To each email, there is a function that tells us if it is a spam or not. We could take this as X and $X : \Omega \to \{0, 1\}$, 0 would be not spam and 1 would be spam.

This is our first setup of the problem. We are recieving an email, this is the experiment. It is either spam or not spam, and this is the value of the RV. X. However our initial problem was to predict if the email was a spam or not, i.e. we wish to use some information in the email and construct a decision function, which takes the email and outputs if it is a 1 or a 0. We pretty much would like to know how the unknown X works using a set of observations.

We can use many things, but one of the simplest would be. Let $Z : \Omega \to \{0, 1\}$ be a function that is 1 if the email contains the word “donate” and hope that Z is a good predictor of X. But how do we phrase that?

$$\mathbb{P}(X = 1 \mid Z = 1) > \mathbb{P}(X = 1)$$

if the above inequality holds, we know that if "donate" is in the email, then it is more likely to be spam. It is thus a valuable predictor. If $\mathbb{P}(X = 1 \mid Z = 1) = 1$ then it is a perfect predictor, since if you know $Z = 1$ then you know that $X = 1$.

If however

$$\mathbb{P}(X = 1 \mid Z = 1) = \mathbb{P}(X = 1)$$

then X and Z are independent, i.e. knowing Z gives nothing about X.

You have just seen the simplest case of a word based prediction model, in practice however you would use several words and in that case its called a **bag of words** model. One particularly successful model is **Naive Bayes** which is very close to what we did above, but with more words.

2.7.2 Number of website requests during a day

Lets say that you are monitoring the number of website requests, and let us assume that each request requires some processing work to be commissioned. That is, what you want to know is, how much resources should I give?

In this particular problem, the goal would be to predict the number of requests on a given day in advance so that there is time to add resources.

- The experiment, recording the website requests during a day. For each website request you log a bunch of information (where it came from, what day, what the request was, etc.)
- $\Omega = \{\text{all sequences of all valid website requests}\}$. Since we are recording during a day, we can have any number of requests. Often, however this set is enumerable. The choice here is fairly arbitrary of what Ω should be, we could just as well define Ω to be an abstract set which contains all possible outcomes of the experiment. With this we mean, all possible information that can be recorded is recorded in ω.
- Here we can as σ-algebra, $\mathcal{F}$ choose all possible subsets of Ω, i.e. the power set. If we would go the route of taking Ω to be the abstract set, then $\mathcal{F}$ is unknown, but it is here.
- Again, the probability measure $\mathbb{P}$ is here, but unknown. In this case, very very very hard to estimate, so we will not even try to.
- $X : \Omega \to \{0, 1, 2, \ldots\}$, a function that takes a sequence of valid website requests and outputs the number of requests. With the simpler Ω we know that X is measurable, however for the abstract Ω we have to assume it is. (This is crucial, we assume that what we observe is a random variable!!! This is a modeling assumption.)

Our goal here is to find a good way to predict X beforehand. Specifically, since it can take on many values we want to estimate the distribution function

$$F_X(x) = \mathbb{P}(X \leq x).$$

Lets say that if $X > 10$ we need to commission extra resources, that is we would then like to know

$$\mathbb{P}(X > 10) = 1 - F_X(10).$$

Let's say that some information about the outcome can be known, like the day of the week. Call this random variable $Z : \Omega \to \{1, 2, 3, 4, 5, 6, 7\}$, where the number is the day of week. It is likely that X and Z are not independent and that the conditional distribution of X given Z, which is

$$F_{X|Z}(x \mid z)$$

can instead be estimated. It is fairly reasonable to assume that the amount of traffic is different for different days, (compare wednesday to saturday for a work related website). Here we can take decisions based on

$$\mathbb{P}(X > 10 \mid Z = z) = 1 - F_{X|Z}(x \mid z).$$

In the case of a more abstract Ω we could envision other random variables Y which contains some information, say its 1 if today is a holiday. We could then try to estimate

$$\mathbb{P}(X > 10 \mid Z = z, Y = y) = 1 - F_{X|Z,Y}(x \mid z, y).$$

2.7.3 Summary

You have now seen to common estimation examples where we defined what could be known and left the unknown. We assumed that the unknown quantities existed (this is a modeling assumption).

Here is a modeling procedure which is always defined

- Lets say you are doing some experiment, say, looking at an image.
- Let Ω be all possible outcomes where you do not specify what an outcome is, it can be seen as unknown.
- Assume that there are unknown $\mathcal{F}$ and $\mathbb{P}$, that is we assume that underlying our problem there is a $(\Omega, \mathcal{F}, \mathbb{P})$, i.e. a probability triple, which is unknown.

- We assume that some information about the image, say if it contains a cat or not (1 or 0) is knowable, call that value X. We further assume that X is a random variable with respect to the probability triple $(\Omega, \mathcal{F}, \mathbb{P})$.

Now, since everything is unknown but assumed to be there, and all we care about is whether or not there is a cat in the image. We could try to estimate

$$F_X(x) = \mathbb{P}(X \leq x)$$

and since X only takes values 0 and 1, it would be enough to estimate

$$\mathbb{P}(X = 1).$$

If we do independent repeats of our experiment and record if its a cat or not, then it is reasonable to expect that the average would be a good estimate for $\mathbb{P}(X = 1)$. Point being, we did not need to know $(\Omega, \mathcal{F}, \mathbb{P})$ in order to estimate $\mathbb{P}(X = 1)$.

This idea is underlying many models, we assume that our repeated experiments are independent and that the value we are looking at is a random variable. This means we can “ignore” what $(\Omega, \mathcal{F}, \mathbb{P})$ actually is, but know that we assume its well defined and there.