10. Discrete probability distributions

Let (Ω, p) be a probability space and $X : \Omega \to \mathbb{R}$ be a random variable. We define two objects associated to X.

Probability mass function (pmf). The range of X is a countable subset of $\mathbb{R}$, denote it by $\text{Range}(X)\{t_1, t_2, \dots\}$. Then, define $f_X : \mathbb{R} \to [0,1]$ as the function

$$f_X(t) = \begin{cases} \mathbf{P}\{\omega : X(\omega) = t\} & \text{if } t \in \text{Range}(X). \\ 0 & \text{if } t \notin \text{Range}(X). \end{cases}$$

One obvious property is that $\sum_{t\in\mathbb{R}} f_X(t) = 1$. Conversely, any non-negative function f that is non-zero on a countable set S and such that $\sum_{t\in\mathbb{R}} f(t) = 1$ is a pmf of some random variable.

Cumulative distribution function (CDF). Define $F_X : \mathbb{R} \to [0,1]$ by

$$F_X(t) = \mathbf{P}\{\omega : X(\omega) \le t\}.$$

Example 65. Let $\Omega = \{(i,j) : 1 \le i,j \le 6\}$ with $p_{(i,j)} = \frac{1}{36}$ for all $(i,j) \in \Omega$. Let $X : \Omega \to \mathbb{R}$ be the random variable defined by $X(i,j) = i + j$. Then, $\text{Range}(X) = \{2,3,\dots,12\}$. The pmf and CDF of X are given by

$$f_X(k) = \begin{cases} 1/36 & \text{if } k = 2. \\ 2/36 & \text{if } k = 3. \\ 3/36 & \text{if } k = 4. \\ 4/36 & \text{if } k = 5. \\ 5/36 & \text{if } k = 6. \\ 6/36 & \text{if } k = 7. \\ 5/36 & \text{if } k = 8. \\ 4/36 & \text{if } k = 9. \\ 3/36 & \text{if } k = 10. \\ 2/36 & \text{if } k = 11. \\ 1/36 & \text{if } k = 12. \end{cases} \qquad F_X(t) = \begin{cases} 0 & \text{if } t < 2. \\ 1/36 & \text{if } t \in [2,3). \\ 3/36 & \text{if } t \in [3,4). \\ 6/36 & \text{if } t \in [4,5). \\ 10/36 & \text{if } t \in [5,6). \\ 15/36 & \text{if } t \in [6,7). \\ 21/36 & \text{if } t \in [7,8). \\ 26/36 & \text{if } t \in [8,9). \\ 30/36 & \text{if } t \in [9,10). \\ 33/36 & \text{if } t \in [10,11). \\ 35/36 & \text{if } t \in [11,12). \\ 1 & \text{if } t \ge 12. \end{cases}$$

A picture of the pmf and CDF for a Binomial distribution are shown in Figure **??**.

Basic properties of a CDF: The following observations are easy to make.

(1) F is an increasing function on $\mathbb{R}$.
(2) $\lim_{t\to+\infty} F(t) = 1$ and $\lim\limits_{t\to-\infty} F(t) = 0$.
(3) F is right continuous, that is, $\lim\limits_{h\searrow 0} F(t+h) = F(t)$ for all $t \in \mathbb{R}$.
(4) F increases only in jumps. This means that if F has no jump discontinuities (an increasing function has no other kind of discontinuity anyway) in an interval $[a,b]$, then $F(a) = F(b)$.

Since $F(t)$ is the probability of a certain event, these statements can be proved using the basic rules of probability that we saw earlier.

PROOF. Let $t < s$. Define two events, $A = \{\omega : X(\omega) \le t\}$ and $B = \{\omega : X(\omega) \le s\}$. Clearly $A \subseteq B$ and hence $F(t) = \mathbf{P}(A) \le \mathbf{P}(B) = F(s)$. This proves the first property.

To prove the second property, let $A_n = \{\omega : X(\omega) \le n\}$ for $n \ge 1$. Then, A_n are increasing in n and $\bigcup_{n=1}^{\infty} A_n = \Omega$. Hence, $F(n) = \mathbf{P}(A_n) \to \mathbf{P}(\Omega) = 1$ as $n \to \infty$. Since F is increasing, it follows that $\lim_{t\to+\infty} F(t) = 1$. Similarly one can prove that $\lim_{t\to-\infty} F(t) = 0$.

Right continuity of F is also proved the same way, by considering the events $B_n = \{\omega : X(\omega) \le t + \frac{1}{n}\}$. We omit details. ■

Remark 66. It is easy to see that one can recover the pmf from the CDF and vice versa. For example, given the pmf f, we can write the CDF as $F(t) = \sum_{u:u\le t} f(u)$. Conversely, given the CDF, by looking at the locations of the jumps and the sizes of the jumps, we can recover the pmf.

The point is that probabilistic questions about X can be answered by knowing its CDF F_X. Therefore, in a sense, the probability space becomes irrelevant. For example, the expected value of a random variable can be computed using its CDF only. Hence, we shall often make statements like "X is a random variable with pmf f" or "X is a random variable with CDF F", without bothering to indicate the probability space. Some distributions (i.e., CDF or the associated pmf) occur frequently enough to merit a name.

Example 67. Let f and F be the pmf, CDF pair

$$f(t) = \begin{cases} p & \text{if } t = 1, \\ q & \text{if } t = 0, \end{cases} \qquad F_X(t) = \begin{cases} 1 & \text{if } t \ge 1, \\ q & \text{if } t \in [0,1), \\ 0 & \text{if } t < 0. \end{cases}$$

A random variable X having this pmf (or equivalently the CDF) is said to have *Bernoulli distribution* with parameter p and write $X \sim \text{Ber}(p)$. For example, if $\Omega = \{1,2,\dots,10\}$ with $p_i = 1/10$, and $X(\omega) = \mathbf{1}_{\omega\le 3}$, then $X \sim \text{Ber}(0.3)$. Any random variable taking only the values 0 and 1, has Bernoulli distribution.

Example 68. Fix $n \ge 1$ and $p \in [0,1]$. The pmf defined by $f(k) = \binom{n}{k} p^k q^{n-k}$ for $0 \le k \le n$ is called the *Binomial distribution* with parameters n and p and is denoted $\text{Bin}(n,p)$. The CDF is as usual defined by $F(t) = \sum_{u:\mathbf{u}\le t} f(u)$, but it does not have any particularly nice expression.

For example, if $\Omega = \{0,1\}^n$ with $p_{\underline{\omega}} = p^{\sum_i \omega_i} q^{n-\sum_i \omega_i}$, and $X(\underline{\omega}) = \omega_1 + \ldots + \omega_n$, then $X \sim \text{Bin}(n,p)$. In words, the number of heads in n tosses of a p-coin has $\text{Bin}(n,p)$ distribution.

Example 69. Fix $p \in (0,1]$ and let $f(k) = q^{k-1}p$ for $k \in \mathbb{N}_+$. This is called the *Geometric distribution* with parameter p and is denoted $\text{Geo}(p)$. The CDF is

$$F(t) = \begin{cases} 0 & \text{if } t < 1, \\ 1 - q^k & \text{if } k \le t < k+1, \text{ for some } k \ge 1. \end{cases}$$

For example, the number of tosses of a p-coin till the first head turns up, is a random variable with $\text{Geo}(p)$ distribution.

Example 70. Fix $\lambda > 0$ and define the pmf $f(k) = e^{-\lambda}\frac{\lambda^k}{k!}$. This is called the *Poisson distribution* with parameter λ and is denoted $\text{Pois}(\lambda)$.

In the problem of a psychic (randomly) guessing the cards in a deck, we have seen that the number of matches (correct guesses) had an *approximately* $\text{Pois}(1)$ distribution.

Example 71. Fix positive integers b,w and $m \le b+w$. Define the pmf $f(k) = \frac{\binom{b}{k}\binom{w}{m-k}}{\binom{b+w}{m}}$ where the binomial coefficient $\binom{x}{y}$ is interpreted to be zero if $y > x$ (thus $f(k) > 0$ only for $\max\{m-w,0\} \le k \le b$). This is called the *Hypergeometric distribution* with parameters b,w,m and we shall denote it by Hypergeo(b,w,m).

Consider a population with b men and w women. The number of men in a random sample (without replacement) of size m, is a random variable with the Hypergeo(b,w,m) distribution.

Computing expectations from the pmf Let X be a random variable on (Ω, p) with pmf f. Then we claim that

$$\mathbf{E}[X] = \sum_{t\in\mathbb{R}} t f(t).$$

Indeed, let Range$(X) = \{x_1, x_2, \ldots\}$. Let $A_k = \{\omega : X(\omega) = x_k\}$. By definition of pmf we have $\mathbf{P}(A_k) = f(x_k)$. Further, A_k are pairwise disjoint and exhaustive. Hence

$$\mathbf{E}[X] = \sum_{\omega\in\Omega} X(\omega)p_\omega = \sum_k \sum_{\omega\in A_k} X(\omega)p_\omega = \sum_k x_k \mathbf{P}(A_k) = \sum_k x_k f(x_k).$$

Similarly, $\mathbf{E}[X^2] = \sum_k x_k^2 f(x_k)$. More generally, if $h : \mathbb{R} \to \mathbb{R}$ is any function, then the random variable $h(X)$ has expectation $\mathbf{E}[h(X)] = \sum_k h(x_k) f(x_k)$. Although this sounds trivial, there is a very useful point here. To calculate $\mathbf{E}[X^2]$ we do not have to compute the pmf of X^2 first, which can be done but would be more complicated. Instead, in the above formulas, $\mathbf{E}[h(X)]$ has been computed directly in terms of the pmf of X.

Exercise 72. Find $\mathbf{E}[X]$ and $\mathbf{E}[X^2]$ in each case.

(1) $X \sim \text{Bin}(n,p)$.
(2) $X \sim \text{Geo}(p)$.
(3) $X \sim \text{Pois}(\lambda)$.
(4) $X \sim \text{Hypergeo}(b,w,m)$.

11. Uncountable probability spaces - conceptual difficulties

The following two “random experiments” are easy to imagine, but difficult to fit into the framework of probability spaces.

(1) Toss a p-coin infinitely many times: Clearly the sample space is $\Omega = \{0,1\}^{\mathbb{N}}$. But what is $p_{\underline{\omega}}$ for any $\underline{\omega} \in \Omega$? The only reasonable answer is $p_{\underline{\omega}} = 0$ for all ω. But then how to define $\mathbf{P}(A)$ for any A? For example, if $A = \{\underline{\omega} : \omega_1 = 0, \omega_2 = 0, \omega_3 = 1\}$, then everyone agrees that $\mathbf{P}(A)$ “ought to be” $q^2 p$, but how does that come about? The basic problem is that Ω is uncountable, and probabilities of events are not got by summing probabilities of singletons.
(2) Draw a number at random from $[0,1]$: Again, it is clear that $\Omega = [0,1]$, but it also seems reasonable that $p_x = 0$ for all x. Again, Ω is uncountable, and probabilities of events are not got by summing probabilities of singletons. It is “clear” that if $A = [0.1, 0.4]$, then $\mathbf{P}(A)$ “ought to be” 0.3, but it gets confusing when one tries to derive this from something more basic!

The resolution: Let Ω be uncountable. There is a class of *basic subsets* (usually not singletons) of Ω for which we take the probabilities as given. We also take the rules of probability, namely, countable additivity, as axioms. Then we use the rules to compute the probabilities of more complex events (subsets of Ω) by expressing those events in terms of the basic sets using countable intersections, unions and complements and applying the rules of probability.

Example 73. In the example of infinite sequence of tosses, $\Omega = \{0,1\}^{\mathbb{N}}$. Any set of the form $A = \{\underline{\omega} : \omega_1 = \varepsilon_1, \dots, \omega_k = \varepsilon_k\}$ where $k \geq 1$ and $\varepsilon_i \in \{0,1\}$ will be called a basic set and its probability is defined to be $\mathbf{P}(A) = \prod_{j=1}^{k} p^{\varepsilon_j} q^{1-\varepsilon_j}$ where we assume that $p > 0$. Now consider a more complex event, for example, $B = \{\underline{\omega} : \omega_k = 1 \text{ for some } k\}$. We can write $B = A_1 \cup A_2 \cup A_3 \cup \dots$ where $A_k = \{\underline{\omega} : \omega_1 = 0, \dots, \omega_{k-1} = 0, \omega_k = 1\}$. Since A_k are pairwise disjoint, the rules of probability demand that $\mathbf{P}(B)$ should be $\sum_k \mathbf{P}(A_k) = \sum_k q^{k-1} p$ which is in fact equal to 1.

Example 74. In the example of drawing a number at random from $[0,1]$, $\Omega = [0,1]$. Any interval (a,b) with $0 \leq a < b \leq 1$ is called a basic set and its probability is defined as $\mathbf{P}(a,b) = b - a$. Now consider a non-basic event $B = [a,b]$. We can write $B = A_1 \cup A_2 \cup A_3 \dots$ where $A_k = (a + (1/k), b - (1/k))$. Then A_k is an increasing sequence of events and the rules of probability say that $\mathbf{P}(B)$ must be equal to $\lim_{k\to\infty} \mathbf{P}(A_k) = \lim_{k\to\infty}(b - a - (2/k)) = b - a$. Another example could be $C = [0.1, 0.2) \cup (0.3, 0.7]$. Similarly argue that $\mathbf{P}(\{x\}) = 0$ for any $x \in [0,1]$. A more interesting one is $D = \mathbb{Q} \cap [0,1]$. Since it is a countable union of singletons, it must have zero probability! Even more interesting is the $1/3$-Cantor set. Although uncountable, it has zero probability!

Consistency: Is this truly a solution to the question of uncountable spaces? Are we assured of never running into inconsistencies? Not always.

Example 75. Let $\Omega = [0,1]$ and let intervals (a,b) be open sets with their probabilities defined as $\mathbf{P}(a,b) = \sqrt{b-a}$. This quickly leads to problems. For example, $\mathbf{P}(0,1) = 1$ by definition. But $(0,1) = (0, 0.5) \cup (0.5, 1) \cup \{1/2\}$ from which the rules of probability would imply that $\mathbf{P}(0,1)$ must be at least $\mathbf{P}(0,1/2) + \mathbf{P}(1/2,1) = \frac{1}{\sqrt{2}} + \frac{1}{\sqrt{2}} = \sqrt{2}$ which is greater than 1. Inconsistency!

Exercise 76. Show that we run into inconsistencies if we define $\mathbf{P}(a,b) = (b-a)^2$ for $0 \leq a < b \leq 1$.

Thus, one cannot arbitrarily assign probabilities to basic events. In the next section we shall state a result on how to assign probabilities.

12. Probability distributions on the real line

Definition 77. A *cumulative distribution function* or CDF is a function $F : \mathbb{R} \to [0,1]$ be a function satisfying the following properties.

(1) F is increasing, $F(s) \leq F(t)$ for all $s < t$.
(2) F is right-continuous, $\lim\limits_{h \to 0} F(t+h) = F(t)$ for all $t \in \mathbb{R}$.
(3) $\lim\limits_{t\to-\infty} F(t) = 0$ and $\lim_{t\to+\infty} F(t) = 1$.

If (Ω, p) is a (discrete) probability space and $X : \Omega \to \mathbb{R}$ is a random variable, then the function $F_X : \mathbb{R} \to \mathbb{R}$, defined as $F_X(t) = \mathbf{P}\{\omega : X \leq t\}$ is a CDF. Does every CDF arise this way? [9] We shall give one quick example.

[9] The point of this definition is that a CDF can be used to define probabilities of intervals in a way that is not inconsistent with the rules of probability. We just state this result.

Example 79. Let

$$F(t) = \begin{cases} 0 & \text{if } t \le 0, \\ t & \text{if } t \in (0,1), \\ 1 & \text{if } t \ge 1. \end{cases}$$

Then F is a CDF. But since it has no jumps, it cannot possibly come from a discrete random variable. However, if we assign probabilities to intervals by setting $\mathbf{P}\{(a,b]\} = b-a$, what we get is precisely what we wanted in the experiment of "drawing a number at random from $[0,1]$". The theorem in the footnote assures us that we do not run into inconsistencies.

Working rules: Let F be any CDF. Then, there exists (a possibly uncountable) probability space and a random variable such that $F(t) = \mathbf{P}\{X \le t\}$ for all t. Then we say that X has distribution F. Since it takes a lot of technicalities to define what uncountable probability spaces look like and what random variables mean in this more general setting, we shall never define them. Instead we can use the following simple working rules to answer questions about the distribution of a random variable.

(1) For an $a < b$, we set $\mathbf{P}\{a < X \le b\} := F(b) - F(a)$.
(2) If $I_j = (a_j, b_j]$ are countably many pairwise disjoint intervals, and $I = \bigcup_j I_j$, then we define $\mathbf{P}\{X \in I\} := \sum_j F(b_j) - F(a_j)$.
(3) For a general set $A \subseteq \mathbb{R}$, here is a general scheme: Find countably many pairwise disjoint intervals $I_j = (a_j, b_j]$ such that $A \subseteq \cup_j I_j$. Then we define $\mathbf{P}\{X \in A\}$ as the infimum (over all such coverings by intervals) of the quantity $\sum_j F(b_j) - F(a_j)$.

All of probability in another line: Take an (interesting) random variable X with a given CDF F and an (interesting) set $A \subseteq \mathbb{R}$. Find $\mathbf{P}\{X \in A\}$.

There are loose threads here but they can be safely ignored for this course. We just remark about them for those who are curious to know.

Remark 80. The above method starts from a CDF F and defines $\mathbf{P}\{X \in A\}$ for all subsets $A \subseteq \mathbb{R}$. However, for most choices of F, the countable additivity property turns out to be violated! However, the sets which do violate them rarely arise in practice and hence we ignore them for the present.

Exercise 81. Let X be a random variable with distribution F. Use the working rules to find the following probabilities.

(1) Write $\mathbf{P}\{a < X < b\}$, $\mathbf{P}\{a \le X < b\}$, $\mathbf{P}\{a \le X \le b\}$ in terms of F.
(2) Show that $\mathbf{P}\{X = a\} = F(a) - F(a-)$. In particular, this probability is zero unless F has a jump at a.

We now illustrate how to calculate the probabilities of rather non-trivial sets in a special case. It is not always possible to get an explicit answer as here.

Example 82. Let F be the CDF defined in example 79. We calculate $\mathbf{P}\{X \in A\}$ for two sets A.

Theorem 78. *Let $\Omega = \mathbb{R}$ and let intervals of the form $(a,b]$ with $a < b$ be called basic sets. Let F be any distribution function.* Define *the probabilities of basic sets as $\mathbf{P}\{(a,b]\} = F(b) - F(a)$. Then, applying the rules of probability to compute probabilities of more complex sets (got by taking countable intersections, unions and complements) will never lead to inconsistency.*

1. $A = \mathbb{Q} \cap [0,1]$. Since A is countable, we may write $A = \cup_n \{r_n\}$ and hence $A \subseteq \cup_n I_n$ where $I_n = (r_n, r_n + \delta 2^{-n}]$ for any fixed $\delta > 0$. Hence $\mathbf{P}\{X \in A\} \leq \sum_n F(r_n + \delta 2^{-n}) - F(r_n) \leq 2\delta$. Since this is true for every $\delta > 0$, we must have $\mathbf{P}\{X \in A\} = 0$. (We stuck to the letter of the recipe described earlier. It would have been simpler to say that any countable set is a countable union of singletons, and by the countable additivity of probability, must have probability zero. Here we used the fact that singletons have zero probability since F is continuous).

2. $A =$ Cantor's set[10] How to find $\mathbf{P}\{X \in A\}$? Let A_n be the set of all $x \in [0,1]$ which do not have 1 in the first n digits of their ternary expansion. Then $A \subseteq A_n$. Further, it is not hard to see that $A_n = I_1 \cup I_2 \cup \ldots \cup I_{2^n}$ where each of the intervals I_j has length equal to 3^{-n}. Therefore, $\mathbf{P}\{X \in A\} \leq \mathbf{P}\{X \in A_n\} = 2^n 3^{-n}$ which goes to 0 as $n \to \infty$. Hence, $\mathbf{P}\{X \in A\} = 0$.

13. Examples of distributions

Cumulative distributions will also be referred to as simply distribution functions or distributions. We start by giving two large classes of CDFs. There are CDFs that do not belong to either of these classes, but for practical purposes they may be ignored (for now).

(1) (CDFs with pmf). Let f be a pmf, i.e., let $t_1, t_2, \ldots$ be a countable subset of reals and let $f(t_i)$ be non-negative numbers such that $\sum_i f(t_i) = 1$. Then, define $F : \mathbb{R} \to \mathbb{R}$ by
$$F(t) := \sum_{i : t_i \leq t} f(t_i).$$
Then, F is a CDF. Indeed, we have seen that it is the CDF of a discrete random variable. A special feature of this CDF is that it increases only in jumps (in more precise language, if F is continuous on an interval $[s,t]$, then $F(s) = F(t)$).

(2) (CDFs with pdf). Let $f : \mathbb{R} \to \mathbb{R}_+$ be a function such that $\int_{-\infty}^{+\infty} f(u)du = 1$ (if it worries you that the integral may not exist, assume that f is continuous or piecewise continuous). Such a function is called a *probability density function* or pdf for short. Then, define $F : \mathbb{R} \to \mathbb{R}$ by
$$F(t) := \int_{-\infty}^{t} f(u)du.$$
Again, F is a CDF. Indeed, it is clear that F has the increasing property (if $t > s$, then $F(t) - F(s) = \int_s^t f(u)du$ which is non-negative because $f(u)$ is non-negative for all u), and its limits at $\pm\infty$ are as they should be (why?). As for right-continuity, F is in-fact continuous. Actually F is differentiable except at points where f is discontinuous and $F'(t) = f(t)$.

Remark 83. We understand the pmf. For example if X has pmf f, then $f(t_i)$ is just the probability that X takes the value t_i. How to interpret the pdf? If X has pdf f, then as we already remarked, the CDF is continuous and hence $\mathbf{P}\{X = t\} = 0$. Therefore $f(t)$ cannot

[10] To define the Cantor set, recall that any $x \in [0,1]$ may be written in ternary expansion as $x = 0.u_1u_2 \ldots := \sum_{n=1}^{\infty} u_n 3^{-n}$ where $u_n \in \{0,1,2\}$. This expansion is unique except if x is a rational number of the form $p/3^m$ for some integers p, m (these are called triadic rationals). For triadic rationals, there are two possible ternary expansions, a terminating one and a non-terminating one (for example, $x = 1/3$ can be written as $0.100\ldots$ or as $0.0222\ldots$). For definiteness, for triadic rationals we shall always take the non-terminating ternary expansion. With this preparation, the Cantor set is defined as the set of all x which do not have the digit 1 in their ternary expansion.

be interpreted as $\mathbf{P}\{X = t\}$ (in fact, pdf can take values greater than 1, so it cannot be a probability!).

To interpret $f(a)$, take a small positive number δ and look at

$$F(a+\delta) - F(a) = \int_a^{a+\delta} f(u)du \approx \delta f(a).$$

In other words, $f(a)$ measures the chance of the random variable taking values near a. Higher the pdf, greater the chance of taking values near that point.

Among distributions with pmf, we have seen the Binomial, Poisson, Geometric and Hypergeometric families of distributions. Now we give many important examples of distributions (CDFs) with densities.

Example 84. Uniform distribution on the interval $[a,b]$**:**, denoted Unif($[a,b]$) where $a < b$ is the distribution with density and distribution given by

$$\text{PDF: } f(t) = \begin{cases} \frac{1}{b-a} & \text{if } t \in (a,b) \\ 0 & \text{otherwise} \end{cases} \qquad \text{CDF: } F(t) = \begin{cases} 0 & \text{if } t \le a \\ \frac{t-a}{b-a} & \text{if } t \in (a,b) \\ 1 & \text{if } t \ge b. \end{cases}$$

Example 85. Exponential distribution with parameter λ**:**, denoted Exp(λ) where $\lambda > 0$ is the distribution with density and distribution given by

$$\text{PDF: } f(t) = \begin{cases} \lambda e^{-\lambda t} & \text{if } t > 0 \\ 0 & \text{otherwise} \end{cases} \qquad \text{CDF: } F(t) = \begin{cases} 0 & \text{if } t \le 0 \\ 1 - e^{-\lambda t} & \text{if } t > 0. \end{cases}$$

Example 86. Normal distribution with parameters μ, σ^2**:**, denoted N(μ, σ^2) where $\mu \in \mathbb{R}$ and $\sigma^2 > 0$ is the distribution with density and distribution given by

$$\text{PDF: } \varphi_{\mu,\sigma^2}(t) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(t-\mu)^2} \qquad \text{CDF: } \Phi_{\mu,\sigma^2}(t) = \int_{-\infty}^{t} \varphi_{\mu,\sigma^2}(u)du.$$

There is no closed form expression for the CDF. It is standard notation to write φ and Φ to denote the normal density and CDF when $\mu = 0$ and $\sigma^2 = 1$. N$(0,1)$ is called the standard normal distribution. By a change of variable one can check that $\Phi_{\mu,\sigma^2}(t) = \Phi(\frac{t-\mu}{\sigma})$.

We said that the normal CDF has no simple expression, but is it even clear that it is a CDF?! In other words, is the proposed density a true pdf? Clearly $\varphi(t) = \frac{1}{\sqrt{2\pi}} e^{-t^2/2}$ is non-negative. We need to check that its integral is 1.

Lemma 87. *Fix* $\mu \in \mathbb{R}$ *and* $\sigma > 0$ *and let* $\varphi(t) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(t-\mu)^2}$. *Then,* $\int\limits_{-\infty}^{\infty} \varphi(t)dt = 1$.

PROOF. It suffices to check the case $\mu = 0$ and $\sigma^2 = 1$ (why?). To find its integral is quite non-trivial. Let $I = \int_{-\infty}^{\infty} \varphi(t)dt$. We introduce the two-variable function $h(t,s) := \varphi(t)\varphi(s) = (2\pi)^{-1}e^{-(t^2+s^2)/2}$. On the one hand,

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h(t,s)dtds = \left(\int_{-\infty}^{+\infty} \varphi(t)dt\right)\left(\int_{-\infty}^{+\infty} \varphi(s)ds\right) = I^2.$$

On the other hand, using polar co-ordinates $t = r\cos\theta$, $s = r\sin\theta$, we see that

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h(t,s)dtds = \int_0^{\infty}\int_0^{2\pi}(2\pi)^{-1}e^{-r^2/2}rd\theta dr = \int_0^{\infty} re^{-r^2/2}dr = 1$$

since $\frac{d}{dr}e^{-r^2/2} = -re^{-r^2/2}$. Thus $I^2 = 1$ and hence $I = 1$. ■

Example 88. Gamma distribution with shape parameter ν and scaler parameter λ:, where $\nu > 0$ and $\lambda > 0$, denoted Gamma(ν,λ) is the distribution with density and distribution given by -

$$\text{PDF: } f(t) = \begin{cases} \frac{1}{\Gamma(\nu)}\lambda^{\nu}t^{\nu-1}e^{-\lambda t} & \text{if } t > 0 \\ 0 & \text{otherwise} \end{cases} \qquad \text{CDF: } F(t) = \begin{cases} 0 & \text{if } t \le 0 \\ \int_0^t f(u)du & \text{if } t > 0. \end{cases}$$

Here $\Gamma(\nu) := \int_0^{\infty} t^{\nu-1}e^{-t}dt$. Several clarifications are needed.

Firstly, f is a density. To see this, make the change of variable $\lambda t = u$ to see that

$$\int_0^{\infty} \lambda^{\nu}e^{-\lambda t}t^{\nu-1}dt = \int_0^{\infty} e^{-u}u^{\nu-1}d\nu = \Gamma(\nu).$$

Thus, $\int_0^{\infty} f(t)dt = 1$.

Note that $\nu = 1$ gives the exponential distribution. Thus, the Gamma family subsumes the exponential distributions. For positive integer values of ν, one can actually write an expression for the CDF of Gamma(ν,λ) as (this is a homework problem)

$$F_{\nu,\lambda}(t) = 1 - e^{-\lambda t}\sum_{k=0}^{\nu-1}\frac{(\lambda t)^k}{k!}.$$

Once the expression is given, it is easy to check it by induction (and integration by parts). A curious observation is that the right hand side is exactly $\mathbf{P}(N \ge \nu)$ where $N \sim \text{Pois}(\lambda t)$. This is in fact indicating a deep connection between Poisson distribution and the Gamma distributions.

The Gamma function: The function $\Gamma : (0,\infty) \to \mathbb{R}$ defined by $\Gamma(\nu) = \int_0^{\infty} e^{-t}t^{\nu-1}dt$ is a very important function that often occurs in mathematics and physics. There is no simpler expression for it, although one can find it explicitly for special values of ν. One of its most important properties is that $\Gamma(\nu+1) = \nu\Gamma(\nu)$. To see this, consider

$$\Gamma(\nu+1) = \int_0^{\infty} e^{-t}t^{\nu}dt = -e^{-t}t^{\nu}\Big|_0^{\infty} + \nu\int_0^{\infty} e^{-t}t^{\nu-1}dt = \nu\Gamma(\nu).$$

Starting with $\Gamma(1) = 1$ (direct computation) and using the above relationship repeatedly one sees that $\Gamma(\nu) = (\nu-1)!$ for positive integer values of ν. Thus, the Gamma function interpolates the factorial function (which is defined only for positive integers). It is also possible to prove that $\Gamma(1/2) = \sqrt{\pi}$ and from that we can compute $\Gamma(3/2) = \sqrt{\pi}/2$, $\Gamma(5/2) = 3\sqrt{\pi}/4$ etc. Yet another interesting fact about the Gamma function is its asymptotics.

Stirling's approximation: $\frac{\Gamma(\nu+1)}{\nu^{\nu+\frac{1}{2}}e^{-\nu}\sqrt{2\pi}} \to 1$ as $\nu \to \infty$.

Example 89. Beta distribution with parameters α, β**:**, where $\alpha, \beta > 0$, denoted Beta(α, β) is the distribution with density and distribution given by -

$$\text{PDF: } f(t) = \begin{cases} \frac{1}{B(\alpha,\beta)} t^{\alpha-1}(1-t)^{\beta-1} & \text{if } t \in (0,1) \\ 0 & \text{otherwise} \end{cases} \qquad \text{CDF: } F(t) = \begin{cases} 0 & \text{if } t \leq 0 \\ \int_0^t f(u)du & \text{if } t \in (0,1) \\ 0 & \text{if } t \geq 1. \end{cases}$$

Here $B(\alpha, \beta) := \int_0^1 t^{\alpha-1}(1-t)^{\beta-1} dt$. Again, for special values of α, β (eg., positive integers), one can find the value of $B(\alpha, \beta)$, but in general there is no simple expression. However, it can be expressed in terms of the Gamma function!

Proposition 90. *For any* $\alpha, \beta > 0$*, we have* $B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$.

PROOF. For $\beta = 1$ we see that $B(\alpha, 1) = \int_0^1 t^{\alpha-1} = \frac{1}{\alpha}$ which is also equal to $\frac{\Gamma(\alpha)\Gamma(1)}{\Gamma(\alpha+1)}$ as required. Similarly (or by the symmetry relation $B(\alpha, \beta) = B(\beta, \alpha)$), we see that $B(1, \beta)$ also has the desired expression.

Now for any other *positive integer* value of α and real $\beta > 0$ we can integrate by parts and get

$$\begin{aligned} B(\alpha, \beta) &= \int_0^1 t^{\alpha-1}(1-t)^{\beta-1} dt \\ &= -\frac{1}{\beta} t^{\alpha-1}(1-t)^{\beta} \Big|_0^1 + \frac{\alpha-1}{\beta} \int_0^1 t^{\alpha-2}(1-t)^{\beta} dt \\ &= \frac{\alpha-1}{\beta} B(\alpha-1, \beta+1). \end{aligned}$$

Note that the first term vanishes because $\alpha > 1$ and $\beta > 0$. When α is an integer, we repeat this for α times and get

$$B(\alpha, \beta) = \frac{(\alpha-1)(\alpha-2)\ldots 1}{\beta(\beta+1)\ldots(\beta+\alpha-2)} B(1, \beta+\alpha-1).$$

But we already checked that $B(1, \beta+\alpha-1) = \frac{\Gamma(1)\Gamma(\alpha+\beta-1)}{\Gamma(\alpha+\beta)}$ from which we get

$$B(\alpha, \beta) = \frac{(\alpha-1)(\alpha-2)\ldots 1}{\beta(\beta+1)\ldots(\beta+\alpha-2)} \frac{\Gamma(1)\Gamma(\alpha+\beta-1)}{\Gamma(\alpha+\beta)} = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

by the recursion property of the Gamma function. Thus we have proved the proposition when α is a positive integer. By symmetry the same is true when β is a positive integer (and α can take any value). We do not bother to prove the proposition for general $\alpha, \beta > 0$ here. ■

Example 91. The standard Cauchy distribution: is the distribution with density and distribution given by

$$\text{PDF: } f(t) = \frac{1}{\pi(1+t^2)} \qquad \text{CDF: } F(t) = \frac{1}{2} + \frac{1}{\pi} \tan^{-1} t.$$

One can also make a parametric family of Cauchy distributions with parameters $\lambda > 0$ and $a \in \mathbb{R}$ denoted Cauchy(a, λ) and having density and CDF

$$f(t) = \frac{\lambda}{\pi(\lambda^2 + (t-a)^2)} \qquad F(t) = \frac{1}{2} + \frac{1}{\pi} \tan^{-1}\left(\frac{t-a}{\lambda}\right).$$

Remark 92. Does every CDF come from a pdf? Not necessarily. For example any CDF that is not continuous (for example, CDFs of discrete distributions such as Binomial, Poisson, Geometric etc.). In fact even continuous CDFs may not have densities (there is a good example manufactured out of the $1/3$-Cantor set, but that would take us out of the topic now). However, suppose F is a *continuous* CDF and suppose F is differentiable except at finitely many points and that the derivative is a continuous function. Then $f(t) := F'(t)$ defines a pdf which by the fundamental theorm of Calculus satisfies $F(t) = \int_{-\infty}^{t} f(u)du$.

14. Simulation

As we have emphasized, probability is applicable to many situations in the real world. As such one may conduct experiments to verify the extent to which theorems are actually valid. For this we need to be able to draw numbers at random from any given distribution.

For example, take the case of Bernoulli($1/2$) distribution. One experiment that can give this is that of physically tossing a coin. This is not entirely satisfactory for several reasons. Firstly, are real coins fair? Secondly, what if we change slightly and want to generate from Ber(0.45)? In this section, we describe how to draw random numbers from various distributions on a computer. We do not fully answer this question. Instead what we shall show is

If one can generate random numbers from Unif($[0,1]$*) distribution, then one can draw random numbers from any other distribution. More precisely, suppose U is a random variable with Unif(*$[0,1]$*) distribution. We want to simulate random numbers from a given distribution F. Then, we shall find a function $\psi : [0,1] \to \mathbb{R}$ so that the random variable $X := \psi(U)$ has the given distribution F.*

The question of how to draw random numbers from Unif($[0,1]$) distribution is a very difficult one and we shall just make a few superficial remarks about that.

Drawing random numbers from a discrete pmf: First start with an example.

Example 93. Suppose we want to draw random numbers from Ber(0.4) distribution. Let $\psi : [0,1] \to \mathbb{R}$ be defined as $\psi(t) = \mathbf{1}_{t \le 0.4}$. Let $X = \psi(U)$, i.e., $X = 1$ if $U \le 0.4$ and $X = 0$ otherwise. Then

$$\mathbf{P}\{X = 1\} = \mathbf{P}\{U \le 0.4\} = 0.4, \qquad \mathbf{P}\{X = 0\} = \mathbf{P}\{U > 0.4\} = 0.6.$$

Thus, X has Ber(0.4) distribution.

It is clear how to generalize this.

General rule: Suppose we are given a pmf f

$$\begin{pmatrix} t_1 & t_2 & t_3 & \dots \\ f(t_1) & f(t_2) & f(t_3) & \dots \end{pmatrix}.$$

Then, define $\psi : [0,1] \to \mathbb{R}$ as

$$\psi(u) = \begin{cases} t_1 & \text{if } u \in [0, f(t_1)] \\ t_2 & \text{if } u \in (f(t_1), f(t_1) + f(t_2)] \\ t_3 & \text{if } u \in (f(t_1) + f(t_2), f(t_1) + f(t_2) + f(t_3)] \\ \vdots & \vdots \end{cases}.$$

Then define $X = f(U)$. Clearly X takes the values $t_1, t_2, \dots$ and

$$\mathbf{P}\{X = t_k\} = \mathbf{P}\left\{\sum_{j=1}^{k-1} f(t_j) < U \le \sum_{j=1}^{k} f(t_j)\right\} = f(t_k).$$

Thus X has pmf f.

Exercise 94. Draw 100 random numbers from each of the following distributions and draw the histograms. Compare with the pmf.

(1) Bin(n,p) for $n = 10, 20, 40$ and $p = 0.5, 0.3, 0.9$.
(2) Geo(p) for $p = 0.9, 0.5, 0.3$.
(3) Pois(λ) with $\lambda = 1, 4, 10$.
(4) Hypergeo(N_1, N_2, m) with $N_1 = 100, N_2 = 50, m = 20, N_1 = 1000, N_2 = 1000, m = 40$.

Drawing random numbers from a pdf: Clearly the procedure used for generating from a pmf is inapplicable here. First start with two examples. As before U is a Unif($[0,1]$) random variable.

Example 95. Suppose we want to draw from the Unif($[3,7]$) distribution. Set $X = 4U + 3$. Clearly X

$$\mathbf{P}\{X \le t\} = \mathbf{P}\{U \le \frac{t-3}{4}\} = \begin{cases} 0 & \text{if } t < 0 \\ (t-3)/4 & \text{if } 3 \le t \le 7 \\ 1 & \text{if } t > 7 \end{cases}.$$

This is precisely the CDF of Unif($[3,7]$) distribution.

Example 96. Here let us do the opposite, just take some function of a uniform variable and see what CDF we get. Let $\psi(t) = t^3$ and let $X = \varphi(U) = U^3$. Then,

$$F(t) := \mathbf{P}\{X \le t\} = \mathbf{P}\{U \le t^{1/3}\} = \begin{cases} 0 & \text{if } t < 0 \\ t^{1/3} & \text{if } 0 \le t \le 1 \\ 1 & \text{if } t > 1 \end{cases}.$$

Differentiating the CDF, we get the density

$$f(t) = F'(t) = \begin{cases} \frac{1}{3}t^{-2/3} & \text{if } 0 < t < 1 \\ 0 & \text{otherwise.} \end{cases}$$

The derivative does not exist at 0 and 1, but as remarked earlier, it does not matter if we change the value of the density at finitely many points (as the integral over any interval will remain the same). Anyway, we notice that the density is that of Beta$(1/3, 1)$. Hence $X \sim \text{Beta}(1/3, 1)$.

This gives us the idea that to generate random number from a CDF F, we should find a function $\psi : [0,1] \to \mathbb{R}$ such that $X := \psi(U)$ has the distribution F. How to find the distribution of X?

Lemma 97. *Let $\psi : (0,1) \to \mathbb{R}$ be a strictly increasing function with $a = \psi(0+)$ and $b = \psi(1-)$. Let $X = \psi(U)$. Then X has CDF*

$$F(t) = \begin{cases} 0 & \text{if } t \le a \\ \psi^{-1}(t) & \text{if } a < t < b \\ 1 & \text{if } t \ge b. \end{cases}$$

If is ψ also differentiable and the derivative does not vanish anywhere (or vanishes at finitely many points only), then X has pdf

$$f(t) = \begin{cases} \left(\psi^{-1}\right)'(t) & \text{if } a < t < b \\ 0 & \text{if } t \notin (a,b). \end{cases}$$

PROOF. Since ψ is strictly increasing, $\psi(u) \le t$ if and only if $u \le \psi^{-1}(t)$. Hence,

$$F(t) = \mathbf{P}\{X \le t\} = \mathbf{P}\{U \le \psi^{-1}(t)\} = \begin{cases} 0 & \text{if } t \le a \\ \psi^{-1}(t) & \text{if } a < t < b \\ 1 & \text{if } t \ge b. \end{cases}$$

If ψ is differentiable at and $\psi(u) \neq 0$, then ψ^{-1} is differentiable at $t = \psi(u)$ (and indeed, $(\psi^{-1})'(t) = \frac{1}{\psi'(u)}$). Thus we get the formula for the density. ■

From this lemma, we immediately get the following rule for generating random numbers from a density.

How to simulate from a CDF: Let F be a CDF that is strictly increasing on an interval $[A,B]$ where $F(A) = 0$ and $F(B) = 1$ (it is allowed to take $A = -\infty$ and/or $B = +\infty$). Then define $\psi : (0,1) \to (A,B)$ as $\psi(u) = F^{-1}(u)$. Let $U \sim \text{Unif}([0,1])$ and let $X = \psi(U)$. Then X has CDF equal to F.

This follows from the lemma because ψ is define as the inverse of F and hence F (restricted to (A,B)) is the inverse of ψ. Further, as the inverse of a strictly increasing function, the function ψ is also strictly increasing.

Example 98. Consider the Exponential distribution with parameter λ whose CDF is

$$F(t) = \begin{cases} 0 & \text{if } t \le 0 \\ 1 - e^{-\lambda t} & \text{if } t > 0 \end{cases}$$

Take $A = 0$ and $B = +\infty$. Then F is increasing on $(0,\infty)$ and its inverse is the function $\psi(u) = -\frac{1}{\lambda}\log(1-u)$. Thus to simulate a random number from Exp(λ) distribution, we set $X = -\frac{1}{\lambda}\log(1-U)$.

When the CDF is not explicitly available as a function we can still adopt the above procedure but only numerically. Consider an example.

Example 99. Suppose $F = \Phi$, the CDF of $N(0,1)$ distribution. Then we do not have an explicit form for either Φ or for its inverse Φ^{-1}. With a computer we can do the following. Pick a large number of closely placed points, for example divide the interval $[-5,5]$ into 1000 equal intervals of length 0.01 each. Let the endpoints of these intervals be labelled $t_0 < t_1 < \ldots < t_{1000}$. For each i, calculate $\Phi(t_i) = \int_{-\infty}^{t_i} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx$ using numerical methods for integration, say the numerical value obtained is w_i. This is done only once and create the table of values

$$\begin{array}{cccccc} t_0 & t_1 & t_2 & \ldots & \ldots & t_{1000} \\ w_0 & w_1 & w_2 & \ldots & \ldots & w_{1000} \end{array}.$$

Now draw a uniform random number U. Look up the table and find the value of i for which $w_i < U < w_{i+1}$. Then set $X = t_i$. If it so happens that $U < w_0$, set $X = t_0 = -5$ and if $U > w_{1000}$ set $X = t_{1000} = 5$. But since $\Phi(-5) < 0.00001$ and $\Phi(5) > 0.99999$, it is highly unlikely that the last two cases will occur. The random variable X has a distribution close to $N(0,1)$.

Exercise 100. Give an explicit method to draw random numbers from the following densities.

(1) Cauchy distribution with density $\frac{1}{\pi(1+x^2)}$.

(2) Beta$(\frac{1}{2},\frac{1}{2})$ density $\frac{1}{\pi}\frac{1}{\sqrt{x(1-x)}}$ on $[0,1]$ (and zero elsewhere).

(3) Pareto(α) distribution which by definition has the density

$$f(t) = \begin{cases} \alpha t^{-\alpha-1} & \text{if } t \ge 1, \\ 0 & \text{if } t < 1. \end{cases}$$

We have described a general principle. When we do more computations with random variables and understand the relationships between different distributions, better tricks can be found. For example, we shall see later that we can generate two $N(0,1)$ random numbers as follows: Pick two uniform random numbers U,V and set $X = \sqrt{-2\log(1-U)}\cos(2\pi V)$ and $Y = \sqrt{-2\log(1-U)}\sin(2\pi V)$. Then it turns out that X and Y have exactly $N(0,1)$ distribution! As another example, suppose we need to generate from Gamma$(3,1)$ distribution, we can first generate three uniforms U_1,U_2,U_3 and set $\xi_i = -\log(1-U_i)$ (so ξ_i have exponential distribution) and then define $X = \xi_1+\xi_2+\xi_3$. It turns out that X has Gamma$(3,1)$ distribution!

Remark 101. We have conveniently skipped the question of how to draw random numbers from uniform distribution in the first place. This is a difficult topic and various results, proved and unproved, are used in generating such numbers. For example,

15. Joint distributions

In many situations we study several random variables at once. In such a case, knowing the individual distributions is not sufficient to answer all relevant questions. This is like saying that knowing $\mathbf{P}(A)$ and $\mathbf{P}(B)$ is insufficient to calculate $\mathbf{P}(A\cap B)$ or $\mathbf{P}(A\cup B)$ etc.

Definition 102 (Joint distribution)**.** Let $X_1,X_2,\ldots,X_m$ be random variables on the same probability space. We call $\mathbf{X} = (X_1,\ldots,X_m)$ a *random vector*, as it is just a vector of random variables. The CDF of $\mathbf{X}$, also called the joint CDF of $X_1,\ldots,X_m$ is the function $F : \mathbb{R}^m \to \mathbb{R}$ defined as

$$F(t_1,\ldots,t_m) = \mathbf{P}\{X_1 \le t_1,\ldots,X_m \le t_m\} = \mathbf{P}\left\{\bigcap_{i=1}^{m}\{X_i \le t_i\}\right\}.$$

.

Example 103. Consider two events A and B in the probability space and let $X = \mathbf{1}_A$ and $Y = \mathbf{1}_B$ be their indicator random variables. Their joint CDF is given by

$$F(s,t) = \begin{cases} 0 & \text{if } s<0 \text{ or } t<0 \\ \mathbf{P}(A^c\cap B^c) & \text{if } s\ge 0,\ t<1 \text{ or } t\ge 0,\ s<1 \\ \mathbf{P}(A) & \text{if } 0\le s<1 \text{ and } t\ge 1 \\ \mathbf{P}(B) & \text{if } 0\le t<1 \text{ and } s\ge 1 \\ \mathbf{P}(A\cap B) & \text{if } s\ge 1, t\ge 1 \end{cases}$$

Properties of joint CDFs: The following properties of the joint CDF $F : \mathbb{R}^m \to [0,1]$ are analogous to those of the 1-dimensional CDF and the proofs are similar.

(1) F is increasing in each co-ordinate. That is, if $s_1 \le t_1, \dots, s_m \le t_m$, then $F(s_1,\dots,s_m) \le F(t_1,\dots,t_m)$.
(2) $\lim F(t_1,\dots,t_m) = 0$ if $\max\{t_1,\dots,t_m\} \to -\infty$ (i.e., one of the t_i goes to $-\infty$).
(3) $\lim F(t_1,\dots,t_m) = 1$ if $\min\{t_1,\dots,t_m\} \to +\infty$ (i.e., all of the t_i goes to $+\infty$).
(4) F is right continuous in each co-ordinate. That is $F(t_1+h_1,\dots,t_m+h_m) \to F(t_1,\dots,t_m)$ as $h_i \to 0+$.

Conversely any function having these four properties is the joint CDF of some random variables.

From the joint CDF, it is easy to recover the individual CDFs. Indeed, if $F : \mathbb{R}^m \to \mathbb{R}$ is the CDF of $\mathbf{X} = (X_1,\dots,X_m)$, then the CDF of X_1 is given by $F_1(t) := F(t,+\infty,\dots,+\infty) := \lim F(t,s_2,\dots,s_m)$ as $s_i \to +\infty$ for each $i = 2,\dots,m$. This is true because if $A_n := \{X_1 \le t\} \cap \{X_2 \le n\} \cap \dots \cap \{X_m \le n\}$, then as $n \to \infty$, the events A_n increase to the event $A = \{X_1 \le t\}$. Hence $\mathbf{P}(A_n) \to \mathbf{P}(A)$. But $\mathbf{P}(A_n) = F(t,n,n,\dots,n)$ and $\mathbf{P}(A) = F_1(t)$. Thus we see that $F_1(t) := F(t,+\infty,\dots,+\infty)$.

More generally, we can recover the joint CDF of any subset of $X_1,\dots,X_n$, for example, the joint CDF of $X_1,\dots,X_k$ is just $F(t_1,\dots,t_k,+\infty,\dots,+\infty)$.

Joint pmf and pdf: Just like in the case of one random variable, we can consider the following two classes of random variables.

(1) Distributions with a pmf. These are CDFs for which there exist points $\mathbf{t}_1,\mathbf{t}_2,\dots$ in $\mathbb{R}^m$ and non-negative numbers w_i such that $\sum_i w_i = 1$ (often we write $f(t_i)$ in place of w_i) and such that for every $\mathbf{t} \in \mathbb{R}^m$ we have
$$F(\mathbf{t}) = \sum_{i\,:\,\mathbf{t}_i \le \mathbf{t}} w_i$$
where $\mathbf{s} \le \mathbf{t}$ means that each co-ordinate of s is less than or equal to the corresponding co-ordinate of $\mathbf{t}$.
(2) Distributions with a pdf. These are CDFs for which there is a non-negative function (may assume piecewise continuous for convenience) $f : \mathbb{R}^m \to \mathbb{R}_+$ such that for every $\mathbf{t} \in \mathbb{R}^m$ we have
$$F(\mathbf{t}) = \int_{-\infty}^{t_1} \dots \int_{-\infty}^{t_m} f(u_1,\dots,u_m)du_1\dots du_m.$$

We give two examples, one of each kind.

Example 104. (Multinomial distribution). Fix parameters r,m (two positive integers) and $p_1,\dots,p_m$ (positive numbers that add to 1). The *multinomial pmf* with these parameters is given by
$$f(k_1,\dots,k_{m-1}) = \frac{r!}{k_1!k_2!\dots k_{m-1}!(r-\sum_{i=1}^{m-1}k_i)!} p_1^{k_1}\dots p_{m-1}^{k_{m-1}} p_m^{r-\sum_{i=1}^{m-1}k_i},$$
if $k_i \ge 0$ are integers such that $k_1+\dots+k_{m-1} \le r$. One situation where this distribution arises is when r balls are randomly placed in m bins, with each ball going into the jth bin with probability p_j, and we look at the random vector $(X_1,\dots,X_{m-1})$ where X_k is the number of balls that fell into the kth bin. This random vector has the multinomial pmf[11]

[11] In some books, the distribution of $(X_1,\dots,X_m)$ is called the multinomial distribution. This has the pmf
$$g(k_1,\dots,k_m)\frac{r!}{k_1!k_2!\dots k_{m-1}!k_m!} p_1^{k_1}\dots p_{m-1}^{k_{m-1}} p_m^{k_m}$$

In this case, the marginal distribution of X_k is $\mathrm{Bin}(r, p_k)$. More generally, $(X_1, \ldots, X_\ell)$ has multinomial distribution with parameters $r, \ell, p_1, \ldots, p_\ell, p_0$ where $p_0 = 1 - (p_1 + \ldots + p_\ell)$. This is easy to prove, but even easier to see from the balls in bins interpretation (just think of the last $n - \ell$ bins as one).

Example 105. (Bivariate normal distribution). This is the density on $\mathbb{R}^2$ given by

$$f(x,y) = \frac{\sqrt{ab-c^2}}{2\pi} e^{-\frac{1}{2}\left[a(x-\mu)^2+b(y-\nu)^2+2c(x-\mu)(y-\nu)\right]},$$

where μ, ν, a, b, c are real parameters. We shall impose the conditions that $a > 0$, $b > 0$ and $ab - c^2 > 0$ (otherwise the above does not give a density, as we shall see).

The first thing is to check that this is indeed a density. We recall the one-dimensional Gaussian integral

$$\int_{-\infty}^{+\infty} e^{-\frac{\tau}{2}(x-a)^2} dx = \sqrt{2\pi}\frac{1}{\sqrt{\tau}} \text{ for any } \tau > 0 \text{ and any } a \in \mathbb{R}. \tag{1}$$

We shall take $\mu = \nu = 0$ (how do you compute the integral if they are not?). Then, the exponent in the density has the form

$$ax^2 + by^2 + 2cxy = b\left(y + \frac{c}{b}\right)^2 + \left(a - \frac{c^2}{b}\right)x^2.$$

Therefore,

$$\begin{aligned}\int_{-\infty}^{\infty} e^{-\frac{1}{2}\left[ax^2+by^2+2cxy\right]}dy &= e^{-\frac{1}{2}(a-\frac{c^2}{b})x^2}\int_{-\infty}^{\infty} e^{-\frac{b}{2}(y+\frac{c}{b})^2} \\ &= e^{-\frac{1}{2}(a-\frac{c^2}{b})x^2}\frac{\sqrt{2\pi}}{\sqrt{b}}\end{aligned}$$

by (1) but ony if $b > 0$. Now we integrate over x and use (1) again (and the fact that $a - \frac{c^2}{b} > 0$) to get

$$\begin{aligned}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-\frac{1}{2}\left[a(x-\mu)^2+b(y-\nu)^2+2c(x-\mu)(y-\nu)\right]}dydx &= \frac{\sqrt{2\pi}}{\sqrt{b}}\int_{-\infty}^{\infty} e^{-\frac{1}{2}(a-\frac{c^2}{b})x^2}dx \\ &= \frac{\sqrt{2\pi}}{\sqrt{b}}\frac{\sqrt{2\pi}}{\sqrt{a-\frac{c^2}{b}}} = \frac{2\pi}{ab-c^2}.\end{aligned}$$

This completes the proof that $f(x,y)$ is indeed a density. Note that $b > 0$ and $ab - c^2 > 0$ also implies that $a > 0$.

Matrix form of writing the density: Let $\Sigma^{-1} = \begin{bmatrix} a & c \\ c & b \end{bmatrix}$. Then, $\det(\Sigma) = \frac{1}{\det(\Sigma^{-1})} = \frac{1}{ab-c^2}$. Hence, we may re-write the density above as (let $\mathbf{u}$ be the column vector with co-ordinates x, y)

$$f(x,y) = \frac{1}{2\pi\sqrt{\det(\Sigma)}} e^{-\frac{1}{2}\mathbf{u}^t\Sigma^{-1}\mathbf{u}}.$$

where k_i are non-negative integers such that $k_1 + \ldots + k_m = r$. We have chosen our convention so that the binomial distribution is a special case of the multinomial...

This is precisely in the form in which we wrote for general n in the example earlier. The conditions $a > 0, b > 0, ab - c^2 > 0$ translate precisely to what is called positive-definiteness. One way to say it is that Σ is a symmetric matrix and all its eigenvalues are strictly positive.

Final form: We can now introduce an extra pair of parameters μ_1, μ_2 and define a density

$$f(x,y) = \frac{1}{2\pi\sqrt{\det(\Sigma)}} e^{-\frac{1}{2}(\mathbf{u}-\mu)^t \Sigma^{-1} (\mathbf{u}-\mu)}.$$

where μ is a column vector with co-ordinates μ_1, μ_2. This is the full bi-variate normal density.

Example 106. (A class of examples). Let $f_1, f_2, \ldots, f_m$ be one-variable densities. In other words, $f_i : \mathbb{R} \to \mathbb{R}_+$ and $\int_{-\infty}^{\infty} f_i(x)dx = 1$. Then, we can make a multivariate density as follows. Define $f : \mathbb{R}^m \to \mathbb{R}_+^m$ by $f(x_1, \ldots, x_m) = f_1(x_1) \ldots f_m(x_m)$. Then f is a density.

If X_i are random variables on a common probability space and the joint density of $(X_1, \ldots, X_m)$ if $f(x_1, \ldots, x_m)$, then we say that X_i are *independent random variables*. It is easy to see that the marginal density of X_i if f_i. It is also the case that the joint CDF factors as $F_X(x_1, \ldots, x_m) = F_{X_1}(x_1) \ldots F_{X_m}(x_m)$.

16. Change of variable formula

Let $\mathbf{X} = (X_1, \ldots, X_m)$ be a random vector with density $f(t_1, \ldots, t_m)$. Let $T : \mathbb{R}^m \to \mathbb{R}^m$ be a one-one function which is continuously differentiable (many exceptions can be made as remarked later).

Let $\mathbf{Y} = T(\mathbf{X})$. In co-ordinates we may write $\mathbf{Y} = (Y_1, \ldots, Y_m)$ and $Y_1 = T_1(X_1, \ldots, X_m) \ldots Y_m = T_m(X_1, \ldots, X_m)$ where $T_i : \mathbb{R}^m \to \mathbb{R}$ are the components of T.

Question: What is the joint density of $Y_1, \ldots, Y_m$?

The change of variable formula: In the setting described above, the joint density of $Y_1, \ldots, Y_m$ is given by

$$g(\mathbf{y}) = f\left(T^{-1}\mathbf{y}\right) \left| J[T^{-1}](\mathbf{y}) \right|$$

where $J[T^{-1}](\mathbf{y})$ is the Jacobian determinant of the function T^{-1} at the point $\mathbf{y} = (y_1, \ldots, y_m)$.

Justification: We shall not prove this formula, but give a imprecise but convincing justification that can be made into a proof. There are two factors on the right. The first one, $f(T^{-1}\mathbf{y})$ is easy to understand - if $\mathbf{Y}$ is to be close to $\mathbf{y}$, then $\mathbf{X}$ must be close to $T^{-1}\mathbf{y}$. The second factor involving the Jacobian determinant comes from the volume change. Let us explain with analogy with mass density which is a more familiar quantity.

Consider a solid cube with non-uniform density. If you rotate it, the density at any point now is the same as the original density, but at a different point (the one which came to the current position). Instead of rotating, suppose we uniformly expand the cube so that the center stays where it is and the side of the cube becomes twice what it is. What happens to the density at the center? It goes down by a factor of 8. This is simply because of volume change - the same mass spreads over a larger volume. More generally, we can have non-uniform expansion, we may cool some parts of the cube, heat some parts and to varying degrees. What happens to the density? At each point, the density changes by a factor given by the Jacobian determinant.

Now for a slightly more mathematical justification. We use the language for two variables ($m = 2$) but the same reasoning works for any m. Fix twp point $\mathbf{x} = (x_1, x_2)$ and $\mathbf{y} = (y_1, y_2)$ such that $\mathbf{y} = T(\mathbf{x})$ (and hence $\mathbf{x} = T^{-1}(\mathbf{y})$). The density of $\mathbf{Y}$ at $\mathbf{y}$ is given by

$$g(\mathbf{y}) \approx \frac{1}{\text{area}(\mathcal{N})}\mathbf{P}\{\mathbf{Y} \in \mathcal{N}\}$$

where $\mathcal{N}$ is a small neighbourhood of the point $\mathbf{y}$ (for example a disk of small radius δ centered at $\mathbf{y}$). By the one-one nature of T and the relationship $\mathbf{Y} = T(\mathbf{X})$, we see that

$$\mathbf{P}\{\mathbf{Y} \in \mathcal{N}\} = \mathbf{P}\{\mathbf{X} \in T^{-1}(\mathcal{N})\}$$

where $T^{-1}(\mathcal{N})$ is the image of $\mathcal{N}$ after mapping by T^{-1}. Now, $T^{-1}(\mathcal{N})$ is a small neighbourhood of $\mathbf{x}$ (if $\mathcal{N}$ is a disk, then $T^{-1}(\mathcal{N})$ would be an approximate ellipse) and hence, by the same interpretation of density we see that

$$\mathbf{P}\{\mathbf{X} \in T^{-1}(\mathcal{N})\} \approx \text{area}(T^{-1}(\mathcal{N}))f(\mathbf{x})$$

Putting the three displayed equations together, we arrive at the formula

$$g(\mathbf{y}) \approx f(\mathbf{x})\frac{\text{area}(T^{-1}(\mathcal{N}))}{\text{area}(\mathcal{N})}$$

Thus the problem boils down to how areas change under transformations. A linear map $S(\mathbf{y}) = A\mathbf{y}$ where A is a 2×2 matrix changes area of any region by a factor of $|\det(A)|$, i.e., $\text{area}(S(\mathcal{R})) = |\det(A)|\text{area}(\mathcal{R})$.

The differentiability of T means that in a small neighbourhood of $\mathbf{y}$, the mapping T^{-1} looks like a linear map, $T^{-1}(\mathbf{y}+\mathbf{h}) \approx \mathbf{x} + DT^{-1}(\mathbf{y})\mathbf{h}$. Therefore, the areas of small neighbourhoods of $\mathbf{y}$ change by a factor equal to $|\det(DT^{-1}(\mathbf{y}))|$ which is the Jacobian determinant. In other words, $\text{area}(T^{-1}(\mathcal{N})) \approx |JT^{-1}(\mathbf{y})|\text{area}(\mathcal{N})$. Consequently $g(\mathbf{y}) = f(T^{-1}\mathbf{y})|JT^{-1}(\mathbf{y})|$.

Enlarging the applicability of the change of variable formula: The change of variable formula is applicable in greater generality than we stated above.

(1) Firstly, T does not have to be defined on all of $\mathbb{R}^m$. It is sufficient if it is defined on the range of $\mathbf{X}$ (i.e., if $f(t_1, \ldots, t_m) = 0$ for $(t_1, \ldots, t_m) = \mathbb{R}^m \setminus A$, then it is enough if T is defined on A.

(2) Even within the range of $\mathbf{X}$, we can allow T to be undefined, but $\mathbf{X}$ must have zero probability to fall in the set where it is undefined. For example, it can happen at finitely many points, or on a line (if $m \geq 2$) or on a plane (if $m \geq 3$) etc.

(3) Similarly, the differentiability of T is required only on a subset outside of which $\mathbf{X}$ has probability 0 of falling.

(4) One-one property of T is important, but there are special cases which can be dealt with by a slight modification. For example, if $T(x) = x^2$ or $T(x_1, x_2) = (x_1^2, x_2^2)$ where we can split the space into parts on each of which T is one-one.

Example 107. Let X_1, X_2 be independent Exp(λ) random variables. Let $T(x_1, x_2) = (x_1 + x_2, \frac{x_1}{x_1+x_2})$. This is well-defined on $\mathbb{R}_+^2$ (and note that $\mathbf{P}\{(X_1, X_2) \in \mathbb{R}_+^2\} = 1$) and its range is $\mathbb{R}_+ \times (0,1)$. The inverse function is $T^{-1}(y_1, y_2) = (y_1 y_2, y_1(1-y_2))$. Its Jacobian determinant is

$$J[T^{-1}](y_1, y_2) = \det\begin{bmatrix} y_2 & y_1 \\ 1-y_2 & -y_1 \end{bmatrix} = -y_1.$$

(X_1, X_2) has density $f(x_1,x_2) = \lambda^2 e^{-\lambda(x_1+x_2)}$ for $x_1, x_2 > 0$ (henceforth it will be a convention that the density is zero except where we specify it). Hence, the random variables $Y_1 = X_1 + X_2$ and $Y_2 = \frac{X_1}{X_1+X_2}$ have joint density

$$g(y_1,y_2) = f(y_1y_2, y_1(1-y_2))|J[T^{-1}](y_1,y_2)| = \lambda^2 e^{-\lambda(y_1y_2+y_1(1-y_2))}y_1 = \lambda^2 y_1 e^{-\lambda y_1}$$

for $y_1 > 0$ and $y_2 \in (0,1)$.

In particular, we see that $Y_1 = X_1 + X_2$ has density $h_1(t) = \int_0^1 \lambda^2 t e^{-\lambda t} ds = \lambda^2 t e^{-\lambda t}$ (for $t > 0$) which means that $Y_1 \sim \text{Gamma}(2, \lambda)$. Similarly, $Y_2 = \frac{X_1}{X_1+X_2}$ has density $h_2(s) = \int_0^\infty \lambda^2 t e^{-\lambda t} dt = 1$ (for $s \in (0,1)$) which means that Y_2 has $\text{Unif}(0,1)$ distribution. In fact, Y_1 and Y_2 are also independent since $g(u,v) = h_1(u)h_2(v)$.

Exercise 108. Let $X_1 \sim \text{Gamma}(\nu_1, \lambda)$ and $X_2 \sim \text{Gamma}(\nu_2, \lambda)$ (note that the shape parameter is the same) and assume that they are independent. Find the joint distribution of $X_1 + X_2$ and $\frac{X_1}{X_1+X_2}$.

Example 109. Suppose we are given that X_1 and X_2 are independent and each has $\text{Exp}(\lambda)$ distribution. What is the distribution of the random variable $X_1 + X_2$?

The change of variable formula works for transformations from $\mathbb{R}^m$ to $\mathbb{R}^m$ whereas here we have two random variables X_1, X_2 and our interest is in one random variable $X_1 + X_2$. To use the change of variable formula, we must introduce an *auxiliary* variable. For example, we take $Y_1 = X_1 + X_2$ and $Y_2 = X_1/(X_1 + X_2)$. Then as in the first example, we find the joint density of (Y_1, Y_2) using the change of variable formula and then integrate out the second variable to get the density of Y_1.

Let us emphasize the point that if our interest is only in Y_1, then we have a lot of freedom in choosing the auxiliary variable. The only condition is that from Y_1 and Y_2 we should be able to recover X_1 and X_2. Let us repeat the same using $Y_1 = X_1 + X_2$ and $Y_2 = X_2$. Then, $T(x_1, x_2) = (x_1 + x_2, x_2)$ maps $\mathbb{R}^2_+$ onto $Q := \{(y_1, y_2) : y_1 > y_2 > 0\}$ in a one-one manner. The inverse function is $T^{-1}(y_1, y_2) = (y_1 - y_2, y_2)$. It is easy to see that $JT^{-1}(y_1, y_2) = 1$ (check!). Hence, by the change of variable formula, the density of (Y_1, Y_2) is given by

$$\begin{aligned} g(y_1,y_2) &= f(y_1-y_2, y_2) \cdot 1 \\ &= \lambda^2 e^{-\lambda(y_1-y_2)} e^{-\lambda y_2} \quad (\text{if } y_1 > y_2 > 0) \\ &= \lambda^2 e^{-\lambda y_1} \mathbf{1}_{y_1 > y_2 > 0}. \end{aligned}$$

To get the density of Y_1, we integrate out the second variable. The density of Y_1 is

$$\begin{aligned} h(u) &= \int_{-\infty}^{\infty} \lambda^2 e^{-\lambda y_1} \mathbf{1}_{y_1 > y_2 > 0} dy_2 \\ &= \lambda^2 e^{-\lambda y_1} \int_0^{y_1} dy_2 \\ &= \lambda^2 y_1 e^{-\lambda y_1} \end{aligned}$$

which agrees with what we found before.

Example 110. Suppose $R \sim \text{Exp}(\lambda)$ and $\Theta \sim \text{Unif}(0, 2\pi)$ and the two are independent. Define $X = \sqrt{R}\cos(\Theta)$ and $Y = \sqrt{R}\sin(\Theta)$. We want to find the distribution of (X,Y). For this, we first write the joint density of (R,Θ) which is given by

$$f(r,\theta) = \frac{1}{2\pi}\lambda e^{-\lambda r} \quad \text{for } r > 0, \theta \in (0,2\pi).$$

Define the transformation $T : \mathbb{R}_+ \times (0,2\pi) \to \mathbb{R}^2$ by $T(r,\theta) = (\sqrt{r}\cos\theta, \sqrt{r}\sin\theta)$. The image of T consists of all $(x,y) \in \mathbb{R}^2$ with $y \neq 0$. The inverse is $T^{-1}(x,y) = (x^2 + y^2, \arctan(y/x))$ where $\arctan(y/x)$ is defined so as to take values in $(0,\pi)$ when $y > 0$ and to take values in $(\pi, 2\pi)$ when $y < 0$. Thus

$$JT^{-1}(x,y) = \det \begin{bmatrix} 2x & 2y \\ \frac{-y}{x^2+y^2} & \frac{x}{x^2+y^2} \end{bmatrix} = 2.$$

Therefore, (X,Y) has joint density

$$g(x,y) = 2f(x^2+y^2, \arctan(y/x)) = \frac{\lambda}{\pi} e^{-\lambda(x^2+y^2)}.$$

This is for $(x,y) \in \mathbb{R}^2$ with $y \neq 0$, but as we have remarked earlier, the value of a pdf in $\mathbb{R}^2$ on a line does not matter, we may define $g(x,y)$ as above for all (x,y) (main point is that the CDF does not change). Since $g(x,y)$ separates into a function of x and a function of y, X,Y are independent $N(0, \frac{1}{2\lambda})$.

Remark 111. Relationships between random variables derived by the change of variable formulas can be used for simulation too. For instance, the CDF of $N(0,1)$ is not explicit and hence simulating from that distribution is difficult (must resort to numerical methods). However, we can easily simulate it as follows. Simulate an $\text{Exp}(1/2)$ random variable R (easy, as the distribution function can be inverted) and simulate an independent $\text{Unif}(0,2\pi)$ random variable Θ. Then set $X = \sqrt{R}\cos(\Theta)$ and $Y = \sqrt{R}\sin(\Theta)$. These are two independent $N(0,1)$ random numbers. Here it should be noted that the random numbers in $(0,1)$ given by a random number generator are supposed to be independent uniform random numbers (otherwise, it is not acceptable as a random number generator).