6. Distributions of random variables.

The *distribution* or *law* of a random variable is defined as follows.

Definition 6.0.1. *Given a random variable X on a probability triple $(\Omega, \mathcal{F}, \mathbf{P})$, its distribution (or law) is the function μ defined on $\mathcal{B}$, the Borel subsets of $\mathbf{R}$, by*

$$\mu(B) \;=\; \mathbf{P}(X \in B) \;=\; \mathbf{P}\left(X^{-1}(B)\right), \qquad B \in \mathcal{B}.$$

If μ is the law of a random variable, then $(\mathbf{R}, \mathcal{B}, \mu)$ is a valid probability triple. We shall sometimes write μ as $\mathcal{L}(X)$ or as $\mathbf{P}\,X^{-1}$. We shall also write $X \sim \mu$ to indicate that μ is the distribution of X.

We define the *cumulative distribution function* of a random variable X by $F_X(x) = \mathbf{P}(X \le x)$, for $x \in \mathbf{R}$. By continuity of probabilities, the function F_X is *right-continuous*, i.e. if $\{x_n\} \searrow x$ then $F_X(x_n) \to F_X(x)$. It is also clearly a non-decreasing function of x, with $\lim_{x\to\infty} F_X(x) = 1$ and $\lim_{x\to-\infty} F_X(x) = 0$. We note the following.

Proposition 6.0.2. *Let X and Y be two random variables (possibly defined on different probability triples). Then $\mathcal{L}(X) = \mathcal{L}(Y)$ if and only if $F_X(x) = F_Y(x)$ for all $x \in \mathbf{R}$.*

Proof. The "if" part follows from Corollary 2.5.9. The "only if" part is immediate upon setting $B = (-\infty, x]$. ∎

6.1. Change of variable theorem.

The following result shows that distributions specify completely the expected values of random variables (and functions of them).

Theorem 6.1.1. *(Change of variable theorem.) Given a probability triple $(\Omega, \mathcal{F}, \mathbf{P})$, let X be a random variable having distribution μ. Then for any Borel-measurable function $f : \mathbf{R} \to \mathbf{R}$, we have*

$$\int_\Omega f\left(X(\omega)\right) \mathbf{P}(d\omega) \;=\; \int_{-\infty}^{\infty} f(t)\, \mu(dt)\,, \tag{6.1.2}$$

i.e. $\mathbf{E}_{\mathbf{P}}[f(X)] = \mathbf{E}_\mu(f)$, provided that either side is well-defined. In words, the expected value of the random variable $f(X)$ with respect to the probability measure $\mathbf{P}$ on Ω is equal to the expected value of the function f with respect to the measure μ on $\mathbf{R}$.

Proof. Suppose first that $f = \mathbf{1}_B$ is an indicator function of a Borel set $B \subseteq \mathbf{R}$. Then $\int_\Omega f(X(\omega))\,\mathbf{P}(d\omega) = \int_\Omega \mathbf{1}_{\{X(\omega)\in B\}}\mathbf{P}(d\omega) = \mathbf{P}(X \in B)$, while $\int_{-\infty}^{\infty} f(t)\mu(dt) = \int_{-\infty}^{\infty} \mathbf{1}_{\{t\in B\}}\mu(dt) = \mu(B) = \mathbf{P}(X \in B)$, so equality holds in this case.

Now suppose that f is a non-negative simple function. Then f is a finite positive linear combination of indicator functions. But since both sides of (6.1.2) are linear functions of f, we see that equality still holds in this case.

Next suppose that f is a general non-negative Borel-measurable function. Then by Proposition 4.2.5, we can find a sequence $\{f_n\}$ of non-negative simple functions such that $\{f_n\} \nearrow f$. We know that (6.1.2) holds when f is replaced by f_n. But then by letting $n \to \infty$ and using the Monotone Convergence Theorem (Theorem 4.2.2), we see that (6.1.2) holds for f as well.

Finally, for general Borel-measurable f, we can write $f = f^+ - f^-$. Since (6.1.2) holds for f^+ and for f^- separately, and since it is linear, therefore it must also hold for f. ∎

Remark. The method of proof used in Theorem 6.1.1 (namely considering first indicator functions, then non-negative simple functions, then general non-negative functions, and finally general functions) is quite widely applicable; we shall use it again in the next subsection.

Corollary 6.1.3. *Let X and Y be two random variables (possibly defined on different probability triples). Then $\mathcal{L}(X) = \mathcal{L}(Y)$ if and only if $\mathbf{E}[f(X)] = \mathbf{E}[f(Y)]$ for all Borel-measurable $f : \mathbf{R} \to \mathbf{R}$ for which either expectation is well-defined. (Compare Proposition 6.0.2 and Remark 5.4.2.)*

Proof. If $\mathcal{L}(X) = \mathcal{L}(Y) = \mu$ (say), then Theorem 6.1.1 says that $\mathbf{E}[f(X)] = \mathbf{E}[f(Y)] = \int_{\mathbf{R}} f\,d\mu$.

Conversely, if $\mathbf{E}[f(X)] = \mathbf{E}[f(Y)]$ for all Borel-measurable $f : \mathbf{R} \to \mathbf{R}$, then setting $f = \mathbf{1}_B$ shows that $\mathbf{P}[X \in B] = \mathbf{P}[Y \in B]$ for all Borel $B \subseteq \mathbf{R}$, i.e. that $\mathcal{L}(X) = \mathcal{L}(Y)$. ∎

Corollary 6.1.4. *If X and Y are random variables with $\mathbf{P}(X = Y) = 1$, then $\mathbf{E}[f(X)] = \mathbf{E}[f(Y)]$ for all Borel-measurable $f : \mathbf{R} \to \mathbf{R}$ for which either expectation is well-defined.*

Proof. It follows directly that $\mathcal{L}(X) = \mathcal{L}(Y)$. Then, letting $\mu = \mathcal{L}(X) = \mathcal{L}(Y)$, we have from Theorem 6.1.1 that $\mathbf{E}[f(X)] = \mathbf{E}[f(Y)] = \int_{\mathbf{R}} f\,d\mu$. ∎

6.2. Examples of distributions.

For a first example of a distribution of a random variable, suppose that $\mathbf{P}(X = c) = 1$, i.e. that X is always (or, at least, with probability 1) equal to some constant real number c. Then the distribution of X is the *point mass* δ_c, defined by $\delta_c(B) = \mathbf{1}_B(c)$, i.e. $\delta_c(B)$ equals 1 if $c \in B$ and equals 0 otherwise. In this case we write $X \sim \delta_c$, or $\mathcal{L}(X) = \delta_c$. From Corollary 6.1.4, since $\mathbf{P}(X = c) = 1$, we have $\mathbf{E}(X) = \mathbf{E}(c) = c$, and $\mathbf{E}(X^3+2) = \mathbf{E}(c^3+2) = c^3+2$, and more generally $\mathbf{E}[f(X)] = f(c)$ for any function f. In symbols, $\int_\Omega f(X(\omega))\,\mathbf{P}(d\omega) \equiv \int_{\mathbf{R}} f(t)\delta_c(dt) = f(c)$. That is, the mapping $f \mapsto \mathbf{E}[f(X)]$ is an *evaluation map*.

For a second example, suppose X has the **Poisson**(5) distribution considered earlier. Then $\mathbf{P}(X \in A) = \sum_{j\in A} e^{-5}5^j/j!$, which implies that $\mathcal{L}(X) = \sum_{j=0}^\infty (e^{-5}5^j/j!)\delta_j$, a convex combination of point masses. The following proposition shows that we then have $\mathbf{E}(f(X)) = \sum_{j=0}^\infty f(j)\,e^{-5}5^j/j!$ for any function $f : \mathbf{R} \to \mathbf{R}$.

Proposition 6.2.1. *Suppose $\mu = \sum_i \beta_i \mu_I$, where $\{\mu_i\}$ are probability distributions, and $\{\beta_i\}$ are non-negative constants (summing to 1, if we want μ to also be a probability distribution). Then for Borel-measurable functions $f : \mathbf{R} \to \mathbf{R}$,*

$$\int f d\mu \;=\; \sum_i \beta_i \int f d\mu_i\,,$$

provided either side is well-defined.

Proof. As in the proof of Theorem 6.1.1, it suffices (by linearity and the monotone convergence theorem) to check the equation when $f = \mathbf{1}_B$ is an indicator function of a Borel set B. But in this case the result follows immediately since $\mu(B) = \sum_i \beta_i \mu_i(B)$. ∎

Clearly, any other discrete random variable can be handled similarly to the **Poisson**(5) example. Thus, discrete random variables do not present any substantial new technical issues.

For a third example of a distribution of a random variable, suppose X has the **Normal**$(0,1)$ distribution considered earlier (henceforth denoted $N(0,1)$). We can define its law μ_N by

$$\mu_N(B) \;=\; \int_{-\infty}^{\infty} \phi(t)\mathbf{1}_B(t)\lambda(dt)\,, \qquad B \text{ Borel}\,, \tag{6.2.2}$$

where λ is Lebesgue measure on $\mathbf{R}$ (cf. (4.4.2)) and where $\phi(t) = \frac{1}{\sqrt{2\pi}}e^{-t^2/2}$. We note that for a mathematically complete definition, it is necessary to use the Lebesgue integral rather than the Riemann integral. Indeed, the Riemann integral is undefined unless B is a rather simple set (e.g. a finite union of intervals, or more generally a set whose boundary has measure 0), while we need $\mu_N(B)$ to be defined for all Borel sets B. Furthermore, since ϕ is continuous it is Borel-measurable (Proposition 3.1.8), so Lebesgue integrals such as (6.2.2) make sense.

Similarly, given *any* Borel-measurable function (called a *density function*) f such that $f \geq 0$ and $\int_{-\infty}^{\infty} f(t)\lambda(dt) = 1$, we can define a law μ by

$$\mu(B) \;=\; \int_{-\infty}^{\infty} f(t)\mathbf{1}_B(t)\lambda(dt)\,, \qquad B \text{ Borel}\,.$$

We shall sometimes write this as $\mu(B) = \int_B f$ or $\mu(B) = \int_B f(t)\,\lambda(dt)$, or even as $\mu(dt) = f(t)\,\lambda(dt)$ (where such equalities of "differentials" have the interpretation that the two sides are equal when integrated over $t \in B$ for any Borel B, i.e. that $\int_B \mu(dt) = \int_B f(t)\lambda(dt)$ for all B). We shall also write this as $\frac{d\mu}{d\lambda} = f$, and shall say that μ *is absolutely continuous with respect to* λ, and that f *is the density for* μ *with respect to* λ. We then have the following.

Proposition 6.2.3. *Suppose μ has density f with respect to λ. Then for any Borel-measurable function $g : \mathbf{R} \to \mathbf{R}$,*

$$\mathbf{E}_\mu(g) \;\equiv\; \int_{-\infty}^{\infty} g(t)\mu(dt) \;=\; \int_{-\infty}^{\infty} g(t)f(t)\lambda(dt)\,,$$

provided either side is well-defined. In words, to compute the integral of a function with respect to μ, it suffices to compute the integral of the function times the density with respect to λ.

Proof. Once again, it suffices to check the equation when $g = \mathbf{1}_B$ is an indicator function of a Borel set B. But in that case, $\int g(t)\mu(dt) = \int \mathbf{1}_B(t)\mu(dt) = \mu(B)$, while $\int g(t)f(t)\lambda(dt) = \int \mathbf{1}_B(t)f(t)\lambda(dt) = \mu(B)$ by definition. The result follows. ∎

By combining Theorem 4.4.1 and Proposition 6.2.3, it is possible to do explicit computations with absolutely-continuous random variables. For example, if $X \sim N(0,1)$, then

$$\mathbf{E}(X) = \int t\,\mu_N(dt) = \int t\,\phi(t)\lambda(dt) = \int_{-\infty}^{\infty} t\,\phi(t)\,dt\,,$$

and more generally

$$\mathbf{E}(g(X)) = \int g(t)\,\mu_N(dt) = \int g(t)\,\phi(t)\,\lambda(dt) = \int_{-\infty}^{\infty} g(t)\,\phi(t)\,dt$$

for any Riemann-integrable function g; here the last expression is an ordinary, old-fashioned, calculus-style Riemann integral. It can be computed in this manner that $\mathbf{E}(X) = 0$, $\mathbf{E}(X^2) = 1$, $\mathbf{E}(X^4) = 3$, etc.

For an example combining Propositions 6.2.3 and 6.2.1, suppose that $\mathcal{L}(X) = \frac{1}{4}\delta_1 + \frac{1}{4}\delta_2 + \frac{1}{2}\mu_N$, where μ_N is again the $N(0,1)$ distribution. Then $\mathbf{E}(X) = \frac{1}{4}(1) + \frac{1}{4}(2) + \frac{1}{2}(0) = \frac{3}{4}$, $\mathbf{E}(X^2) = \frac{1}{4}(1) + \frac{1}{4}(4) + \frac{1}{2}(1) = \frac{7}{4}$, and so on. Note, however, that it is *not* the case that $\mathbf{Var}(X)$ equals the corresponding linear combination of variances (indeed, the variance of a point-mass is 0, so that the corresponding linear combinations of variances is $\frac{1}{4}(0) + \frac{1}{4}(0) + \frac{1}{2}(1) = \frac{1}{2}$); rather, the formula $\mathbf{Var}(X) = \mathbf{E}(X^2) - \mathbf{E}(X)^2 = \frac{7}{4} - (\frac{3}{4})^2 = \frac{19}{16}$ should be used.

Exercise 6.2.4. Why does Proposition 6.2.1 not imply that $\mathbf{Var}(X)$ equals the corresponding linear combination of variances?

6.3. Exercises.

Exercise 6.3.1. Let $(\Omega, \mathcal{F}, P)$ be Lebesgue measure on $[0,1]$, and set

$$X(\omega) = \begin{cases} 1\,, & 0 \le \omega < 1/4 \\ 2\omega^2, & 1/4 \le \omega < 3/4 \\ \omega^2, & 3/4 \le \omega \le 1\,. \end{cases}$$

Compute $P(X \in A)$ where
(a) $A = [0,1]$.
(b) $A = [\frac{1}{2}, 1]$.

Exercise 6.3.2. Suppose $P(Z = 0) = P(Z = 1) = \frac{1}{2}$, that $Y \sim N(0,1)$, and that Y and Z are independent. Set $X = YZ$. What is the law of X?

Exercise 6.3.3. Let $X \sim \mathbf{Poisson}(5)$.
(a) Compute $\mathbf{E}(X)$ and $\mathbf{Var}(X)$.
(b) Compute $\mathbf{E}(3^X)$.

Exercise 6.3.4. Compute $\mathbf{E}(X)$, $\mathbf{E}(X^2)$, and $\mathbf{Var}(X)$, where the law of X is given by
(a) $\mathcal{L}(X) = \frac{1}{2}\delta_1 + \frac{1}{2}\lambda$, where λ is Lebesgue measure on $[0,1]$.
(b) $\mathcal{L}(X) = \frac{1}{3}\delta_2 + \frac{2}{3}\mu_N$, where μ_N is the standard normal distribution $N(0,1)$.

Exercise 6.3.5. Let X and Z be independent, with $X \sim N(0,1)$, and with $\mathbf{P}(Z = 1) = \mathbf{P}(Z = -1) = 1/2$. Let $Y = XZ$ (i.e., Y is the product of X and Z).

(a) Prove that $Y \sim N(0,1)$.
(b) Prove that $\mathbf{P}(|X| = |Y|) = 1$.
(c) Prove that X and Y are *not* independent.
(d) Prove that $\mathbf{Cov}(X,Y) = 0$.
(e) It is sometimes claimed that if X and Y are normally distributed random variables with $\mathbf{Cov}(X,Y) = 0$, then X and Y must be independent. Is that claim correct?

Exercise 6.3.6. Let X and Y be random variables on some probability triple $(\Omega, \mathcal{F}, \mathbf{P})$. Suppose $\mathbf{E}(X^4) < \infty$, and that $\mathbf{P}[m \le X \le z] = \mathbf{P}[m \le Y \le z]$ for all integers m and all $z \in \mathbf{R}$. Prove or disprove that we necessarily have $\mathbf{E}(Y^4) = \mathbf{E}(X^4)$.

Exercise 6.3.7. Let X be a random variable, and let $F_X(x)$ be its cumulative distribution function. For fixed $x \in \mathbf{R}$, we know by right-continuity that $\lim_{y \searrow x} F_X(y) = F_X(x)$.
(a) Give a necessary and sufficient condition that $\lim_{y \nearrow x} F_X(y) = F_X(x)$.
(b) More generally, give a formula for $F_X(x) - (\lim_{y \nearrow x} F_X(y))$, in terms of a simple property of X.

Exercise 6.3.8. Consider the statement: $f(x) = \big(f(x)\big)^2$ for all $x \in \mathbf{R}$.
(a) Prove that the statement is true for all indicator functions $f = \mathbf{1}_B$.
(b) Prove that the statement is *not* true for the identity function $f(x) = x$.
(c) Why does this fact not contradict the method of proof of Theorem 6.1.1?

6.4. Section summary.

This section defined the distribution (or law), $\mathcal{L}(X)$, of a random variable X, to be a corresponding distribution on the real line. It proved that $\mathcal{L}(X)$ is completely determined by the cumulative distribution function, $F_X(x) = \mathbf{P}(X \le x)$, of X. It proved that expectation $\mathbf{E}(f(X))$ of any function of X can be computed (in principle) once $\mathcal{L}(X)$ or $F_X(x)$ is known. It then considered a number of examples of distributions of random variables, including discrete and continuous random variables and various combinations of them. It provided a number of results for computing expected values with respect to such distributions.