Random variables

Last modified: 2026-11-20 00:10:36 (PST)

Definition 0.1 (Random variable)**.** A **random variable** is a function from a sample space Ω to the real numbers $\mathbb{R}$ [1]: it assigns a number to each possible outcome of a random experiment.

Remark

Remark. We use uppercase letters (X, Y, Z, …) for random variables, and lowercase letters (x, y, z, …) for their realized (observed) values. We write $\{X = x\}$ for the event $\{\omega \in \Omega : X(\omega) = x\}$, and similarly for $\{X \leq x\}$ and other conditions on X. A fully rigorous definition also requires the function to be *measurable*, so that every such set is an event with a probability; that requirement never binds in these notes.

Remark (A random variable as an uncertain quantity)

Remark 0.1 (A random variable as an uncertain quantity). In practice, a random variable is a quantity whose value we cannot deduce precisely from the information available to us at a given time. The number of heads in two coin flips is a random variable before the flips, because nothing we know fixes it. Once we see the flips, we know its value, and for us it is no longer uncertain.

Example 0.1 (Number of heads in two coin flips)**.**
Flip a fair coin twice. The sample space is $\Omega = \{HH, HT, TH, TT\}$, each outcome with probability $1/4$. Let X_1 equal 1 if the first flip lands heads and 0 otherwise, and let X_2 be the same for the second flip. Let $S = X_1 + X_2$ be the number of heads:

$$S(HH) = 2,$$
$$S(HT) = 1,$$
$$S(TH) = 1,$$
$$S(TT) = 0$$

Then $\{S = 1\} = \{HT, TH\}$, so

$$\begin{aligned}\Pr(S = 1) &= 1/4 + 1/4 \\ &= 1/2.\end{aligned}$$

Definition 0.2 (Range of a random variable)**.** The **range** of a random variable X, denoted $\mathcal{R}(X)$, is the set of values X can take:

$$\mathcal{R}(X) \stackrel{\text{def}}{=} \{X(\omega) : \omega \in \Omega\}$$

Remark

Remark. The range of X is the image of X as a function on Ω. For a general function, these notes say "image", because some sources use "range" for the codomain instead; for a random variable, "range" is the standard term.

Example 0.2 (Range of the number of heads)**.** In Example 0.1, $\mathcal{R}(X) = \{0, 1, 2\}$.

Definition 0.3 (Discrete random variable)**.** A random variable X is **discrete** if its range $\mathcal{R}(X)$ is countable.

Example 0.3 (Discrete random variables)**.** The number of heads in Example 0.1 is discrete, with a finite range. A count with no upper bound, such as the number of emergency-department visits in a day, is discrete with a countably infinite range, $\{0, 1, 2, \dots\}$.

Definition 0.4 (Continuous random variable)**.** A random variable X is **continuous** if $\Pr(X = x) = 0$ for every $x \in \mathbb{R}$.

Remark

Remark. Every continuous random variable in these notes also has a probability density function. Continuous random variables without a density exist, but they do not arise in

applications.

Example 0.4 (A continuous random variable)**.** Let X satisfy $\Pr(a \leq X \leq b) = b - a$ for all $0 \leq a \leq b \leq 1$. Taking

$$\begin{aligned} a &= b \\ &= x \end{aligned}$$

gives

$$\begin{aligned} \Pr(X = x) &= x - x \\ &= 0 \end{aligned}$$

for every $x \in [0, 1]$. Taking $a = 0$ and $b = 1$ gives $\Pr(0 \leq X \leq 1) = 1$, so by the complement rule X falls outside $[0, 1]$ with probability 0, and $\Pr(X = x) = 0$ for every x outside $[0, 1]$ as well. So X is continuous. Its range, $[0, 1]$, is uncountable.

Example 0.5 (A random variable that is neither discrete nor continuous)**.** Some random variables are neither discrete nor continuous: a time to event that equals exactly 0 with positive probability, and otherwise spreads over $(0, \infty)$, is one. For instance, let U be the continuous random variable of Example 0.4, and let $T = 1/(1-U) - 2$ when $U \in (1/2, 1)$, and $T = 0$ otherwise. On $\{U \in (1/2, 1)\}$, T is increasing in U and takes values in $(0, \infty)$, and solving $t = 1/(1-u) - 2$ for u gives $u = 1 - 1/(t+2)$. So $\{T > 0\} = \{U \in (1/2, 1)\}$, and $\{T = t\} = \{U = 1 - 1/(t+2)\}$ for each $t > 0$. Since U is continuous (Example 0.4), single points have probability zero:

$$\begin{aligned} \Pr(U = 1/2) &= \Pr(U = 1) \\ &= 0 \end{aligned}$$

Using these values:

$$\begin{aligned} \Pr(T > 0) &= \Pr(U \in (1/2, 1)) && (\{T > 0\} = \{U \in (1/2, 1)\}) \\ &= \Pr(1/2 \leq U \leq 1) - \Pr(U = 1/2) - \Pr(U = 1) && \text{(additivity)} \\ &= \tfrac{1}{2} - 0 - 0 && (\Pr(a \leq U \leq b) = b - a) \\ &= \tfrac{1}{2} && \text{(simplify)} \end{aligned}$$

T is not continuous, because

$$\begin{aligned} \Pr(T = 0) &= 1 - \Pr(T > 0) \\ &= 1/2 \end{aligned}$$

by the complement rule.

T is not discrete either. If it were, its range would be countable, so $\{T > 0\}$ would be the disjoint union of the countably many events $\{T = t\}$ with $t \in \mathcal{R}(T)$ and $t > 0$, and:

$$\begin{aligned}
\Pr(T > 0) &= \sum_{t \in \mathcal{R}(T),\, t>0} \Pr(T = t) && \text{(countable additivity)} \\
&= \sum_{t \in \mathcal{R}(T),\, t>0} \Pr\big(U = 1 - \tfrac{1}{t+2}\big) && (\{T = t\} = \{U = 1 - 1/(t+2)\}) \\
&= 0 && (U \text{ is continuous})
\end{aligned}$$

which contradicts $\Pr(T > 0) = 1/2$.

Definition 0.5 (Realization (observed value))**.** A **realization** of a random variable X is the value $x = X(\omega)$ it takes at the point ω of the sample space that the random experiment produced.

Remark

Remark. The random variable X is the rule, before the experiment is run; the realization x is one number, after it is run. Probabilities such as $\Pr(X = x)$ describe the rule, and a realization is a fixed number.

Example 0.6 (Realizations of the number of heads)**.** In Example 0.1, let X be the number of heads in two fair coin flips. Run the experiment once and suppose it produces $\omega = HT$. Then the realization of X is $x = X(HT) = 1$. Running the experiment again might produce TT, with realization $x = 0$. The variable X is the same both times; only its realization changes.

Characteristics of probability distributions

Definition 0.1 (Probability mass function (PMF))**.** If X is a discrete random variable, the **probability mass function** of X at value x, denoted $f(x)$, $f_X(x)$, $\mathbb{P}(x)$, $\mathbb{P}_X(x)$, or $\mathbb{P}(X = x)$, is the probability that X takes exactly the value x [2]:

$$\mathbb{P}(X = x) \overset{\text{def}}{=} \Pr(\{X = x\})$$

Example 0.1 (PMF of the number of heads)**.** For the number of heads X in two fair coin flips (Example 0.1):

$$\mathbb{P}(X=0)=\tfrac{1}{4},\quad \mathbb{P}(X=1)=\tfrac{1}{2},\quad \mathbb{P}(X=2)=\tfrac{1}{4}$$

Definition 0.2 (Bernoulli distribution)**.**
A discrete random variable X has the **Bernoulli distribution** with parameter $\pi \in [0,1]$, written $X \sim \text{Ber}(\pi)$ [3], if:

$$\begin{aligned}\Pr(X=x) &\stackrel{\text{def}}{=} 1_{x\in\{0,1\}}\pi^x(1-\pi)^{1-x}\\ &= \begin{cases}\pi, & x=1\\ 1-\pi, & x=0\end{cases}\end{aligned}$$

Example 0.2 (The first of two coin flips)**.** In Example 0.1, let X_1 indicate heads on the first flip, so

$$\begin{aligned}X_1(HH) &= X_1(HT)\\ &= 1\end{aligned}$$

and

$$\begin{aligned}X_1(TH) &= X_1(TT)\\ &= 0.\end{aligned}$$

Then

$$\begin{aligned}\mathbb{P}(X_1=1) &= \Pr(\{HH,HT\})\\ &= 1/2\end{aligned}$$

and

$$\begin{aligned}\mathbb{P}(X_1=0) &= \Pr(\{TH,TT\})\\ &= 1/2,\end{aligned}$$

so $X_1 \sim \text{Ber}(1/2)$.

Definition 0.3 (Probability density function (PDF))**.** If X is a continuous random variable, a **probability density function** of X, denoted $f(x)$, $f_X(x)$, $\mathbb{p}(x)$, $\mathbb{p}_X(x)$, or $\mathbb{p}(X=x)$, is a function f that satisfies [4]:

- $f(x) \geq 0$ for every x.

- The integral of f over any interval is the probability that X falls in that interval:

$$\Pr(a \le X \le b) = \int_a^b f(x)\,dx \quad \text{for all } a \le b$$

Definition 0.4 (Uniform distribution on a set)**.** Let S be a set with a measure μ such that $\mu(S) \in (0, \infty)$. The **uniform distribution** on S is the probability measure

$$\Pr(A) \stackrel{\text{def}}{=} \frac{\mu(A)}{\mu(S)}, \quad A \subseteq S.$$

Definition 0.5 (Uniform distribution on an interval)**.** The **uniform distribution** on an interval $[\alpha, \beta]$ (Definition 0.4 with μ the length), with $\alpha < \beta$, is the distribution of a random variable X, written $X \sim \text{Uniform}(\alpha, \beta)$, that is continuous with density [5]:

$$f(x) \stackrel{\text{def}}{=} \begin{cases} \frac{1}{\beta - \alpha}, & \alpha \le x \le \beta \\ 0, & \text{otherwise} \end{cases}$$

Example 0.3 (Subinterval probabilities of a uniform random variable)**.** For $X \sim \text{Uniform}(\alpha, \beta)$ (Definition 0.5) and any subinterval $[a, b] \subseteq [\alpha, \beta]$ (so $a, b \in [\alpha, \beta]$ and $a \le b$):

$$\begin{aligned} \Pr(a \le X \le b) &= \int_a^b f(x)\,dx && \text{(definition of a probability density function)} \\ &= \int_a^b \frac{1}{\beta - \alpha}\,dx && (f(x) = \tfrac{1}{\beta - \alpha} \text{ on } [\alpha, \beta]) \\ &= \frac{b - a}{\beta - \alpha} && \text{(integrate)} \end{aligned}$$

In particular, the probability that X falls in any subinterval is proportional to that subinterval's length. For the standard uniform distribution on $[0, 1]$, $f(x) = 1$ for $x \in [0, 1]$ and $f(x) = 0$ otherwise, giving $\Pr(a \le X \le b) = b - a$ for all $a, b \in [0, 1]$ with $a \le b$. An interval reaching outside $[\alpha, \beta]$ adds nothing to the probability, because $f = 0$ there.

Definition 0.6 (Density with respect to a measure, Radon–Nikodym derivative)**.** Let μ be a measure on the range of a random variable X, written $\mathcal{R}(X)$. A **density with respect to a measure** μ, or **Radon–Nikodym derivative**, of the distribution of X is a non-negative measurable function f on $\mathcal{R}(X)$ whose Lebesgue integral against μ, not its Riemann integral, over each measurable set $A \subseteq \mathcal{R}(X)$ is the probability that

X falls in A [6, pp. 213, 423]:

$$\Pr(X \in A) = \int_A f(x)\, d\mu(x)$$

Example 0.4 (A fair die roll and counting measure)**.**
Let X be the roll of a fair six-sided die, so $\mathcal{R}(X) = D = \{1, 2, 3, 4, 5, 6\}$, and let μ be the counting measure on D. The constant function $f(x) = \frac{1}{6}$ on D is a density with respect to μ of the distribution of X. For each subset $A \subseteq D$:

$$\begin{aligned}
\int_A f(x)\, d\mu(x) &= \sum_{x \in A} \tfrac{1}{6} && \text{(an integral with respect to counting measure is a sum)} \\
&= \tfrac{|A|}{6} && (A \text{ has } |A| \text{ terms}) \\
&= \Pr(X \in A) && \text{(the six faces are equally likely)}
\end{aligned}$$

The first step is the integral with respect to counting measure. For example, $A = \{5, 6\}$ gives $\Pr(X \in A) = \frac{2}{6} = \frac{1}{3}$.

Proposition 0.1 (A PMF is a density with respect to counting measure)**.** *If X is a discrete random variable with PMF f_X, and μ is the counting measure on $\mathcal{R}(X)$, then f_X is a density with respect to μ of the distribution of X: for every subset $A \subseteq \mathcal{R}(X)$,*

$$\Pr(X \in A) = \int_A f_X(x)\, d\mu(x)$$

Proof

Proof. The counting measure is defined on every subset of $\mathcal{R}(X)$, so every function on $\mathcal{R}(X)$ is measurable, and $f_X \geq 0$, because each of its values is a probability. The range $\mathcal{R}(X)$ is countable, because X is discrete, so each $A \subseteq \mathcal{R}(X)$ is countable too, and the event $\{X \in A\}$ is the union of the disjoint events $\{X = x\}$ for $x \in A$:

$$\begin{aligned}
\Pr(X \in A) &= \sum_{x \in A} \Pr(X = x) && \text{(countable additivity)} \\
&= \sum_{x \in A} f_X(x) && \text{(definition of the PMF)} \\
&= \int_A f_X(x)\, d\mu(x) && \text{(an integral with respect to counting measure is a sum)}
\end{aligned}$$

The first step is countable additivity, and the last is the integral with respect to counting

measure. □

Example 0.5 (The number of heads, integrated against counting measure)**.**
For the number of heads X in two fair coin flips, with the PMF f_X of Example 0.1, let μ be the counting measure on $\mathcal{R}(X) = \{0, 1, 2\}$. For $A = \{1, 2\}$:

$$\begin{aligned}
\Pr(X \in \{1,2\}) &= \int_{\{1,2\}} f_X(x)\, d\mu(x) && \text{(a PMF is a density with respect to counting measure)} \\
&= f_X(1) + f_X(2) && \text{(an integral with respect to counting measure is a sum)} \\
&= \tfrac{1}{2} + \tfrac{1}{4} && \text{(the PMF of the number of heads)} \\
&= \tfrac{3}{4} && \text{(add)}
\end{aligned}$$

The first step is Proposition 0.1, and the third uses the PMF values in Example 0.1.

i Remark (PMFs and PDFs differ in the reference measure)

Remark 0.1 (PMFs and PDFs differ in the reference measure). A PDF plays the role of a PMF with Lebesgue measure in place of counting measure: Definition 0.3 integrates it with dx. Definition 0.6 uses the Lebesgue integral, because the Riemann integral has no version against counting measure. Against counting measure, the Lebesgue integral of a non-negative f over A adds up the values of f on A, which is how a PMF gives $\Pr(X \in A)$. Against Lebesgue measure, the Lebesgue integral of a non-negative f over $[a, b]$ agrees with its Riemann integral whenever f has one (mds), so Definition 0.3 gives the same values with either one. If the CDF F of a random variable is absolutely continuous, then F' is a Radon–Nikodym derivative of its distribution against Lebesgue measure [6, p. 423]. So a PMF and a PDF are both densities with respect to a measure, for different reference measures. Because the two share that role, these notes let f_X denote either one, and the joint-distribution form of Fubini–Tonelli covers PMFs, PDFs, and density-mass functions in one statement.

Theorem 0.1 (A density is not unique)**.** *If f is a density of a continuous random variable X, and $g \geq 0$ differs from f at only finitely many points, then g is also a density of X.*

i Proof

Proof. For $a \leq b$, the difference $g - f$ is 0 at all but finitely many points, so:

$$\begin{aligned}
\int_a^b g(x)\,dx &= \int_a^b f(x)\,dx + \int_a^b (g(x) - f(x))\,dx && \text{(linearity of the integral)} \\
&= \int_a^b f(x)\,dx + 0 && \text{(a function that is 0 at all but finitely many points integrat} \\
&= \Pr(a \le X \le b) && (f \text{ is a density of } X)
\end{aligned}$$

Together with $g \ge 0$, this is Definition 0.3 for g. □

Remark

Remark. So the value of a density at a single point carries no probability by itself. These notes use the version that is continuous wherever possible.

Theorem 0.2 (Some continuous random variables have no density)**.** *There exist continuous random variables that have no probability density function.*

Remark

Remark. The proof needs measure theory beyond these notes. The standard example is the Cantor distribution, which puts no probability on any single point, but puts all of its probability on a set of total length 0, so any candidate density would integrate to 0 over a set of probability 1.
Every continuous random variable in these notes also has a probability density function. Continuous random variables without a density do not arise in applications.

Definition 0.7 (Normal distribution (Gaussian distribution))**.** A random variable X has the **normal distribution** (or **Gaussian distribution**) with mean parameter $\mu \in \mathbb{R}$ and variance parameter $\sigma^2 > 0$, written $X \sim \mathrm{N}(\mu, \sigma^2)$ [7], if X is continuous with density:

$$\mathbb{p}(X = x) \stackrel{\text{def}}{=} \frac{1}{\sigma\sqrt{2\pi}} \mathrm{e}^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad x \in \mathbb{R}$$

Example 0.6 (Normal densities at their centers)**.** For $X \sim \mathrm{N}(0, 1)$, the density at $x = 0$ is:

$$
\begin{aligned}
\mathbb{p}(X=0) &= \frac{1}{1\cdot\sqrt{2\pi}}\mathrm{e}^{-\frac{(0-0)^2}{2\cdot 1}} && \text{(normal density with } \mu=0,\sigma^2=1\text{)}\\
&= \frac{1}{\sqrt{2\pi}} && (\mathrm{e}^0=1)\\
&\approx 0.399 && \text{(evaluate)}
\end{aligned}
$$

For $X\sim \mathrm{N}(0,0.1^2)$, the same steps give $\mathbb{p}(X=0)=1/(0.1\sqrt{2\pi})\approx 3.99$, a density greater than 1.

i Remark

Remark. A density is not a probability: it can exceed 1, as the second density in Example 0.6 does.

Theorem 0.3 (The normal density integrates to 1)**.** *If $X\sim \mathrm{N}(\mu,\sigma^2)$ (Definition 0.7), then:*

$$\int_{-\infty}^{\infty}\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}\,dx=1$$

Theorem 0.4 (The mean and variance of the normal distribution)**.** *If $X\sim \mathrm{N}(\mu,\sigma^2)$ (Definition 0.7), then X has mean $\mathbb{E}[X]=\mu$ and variance $\mathrm{Var}(X)=\sigma^2$.*

i Remark

Remark. The proofs of these two theorems evaluate three integrals by calculus: the total of the density, the mean integral and the variance integral. The total is the Gaussian integral. This page omits the proofs [8]. These facts are where the parameters' names come from.

Definition 0.8 (Standard normal distribution)**.** The **standard normal distribution** is the normal distribution with mean parameter 0 and variance parameter 1, $\mathrm{N}(0,1)$.

Example 0.7 (Half of a standard normal lies below 0)**.** If Z has the standard normal distribution, its density $\mathbb{p}(Z=z)=\frac{1}{\sqrt{2\pi}}\mathrm{e}^{-z^2/2}$ takes the same value at z and $-z$, so $\Pr(Z\le 0)=\Pr(Z\ge 0)$. Those two probabilities sum to 1, because $\Pr(Z=0)=0$ for a continuous Z, so each is $\frac{1}{2}$.

Definition 0.9 (Exponential distribution)**.** A random variable T has the **exponential distribution** with rate $\lambda > 0$ [9] if T is continuous with density:

$$f(t) \stackrel{\text{def}}{=} \begin{cases} \lambda \mathrm{e}^{-\lambda t}, & t \geq 0 \\ 0, & t < 0 \end{cases}$$

Example 0.8 (Probability of an event within one time unit)**.** Let T be exponential with rate λ. The probability that T falls in $[0, 1]$ is:

$$\begin{aligned} \Pr(0 \leq T \leq 1) &= \int_0^1 \lambda \mathrm{e}^{-\lambda t}\, dt && \text{(definition of a density)} \\ &= \left[-\mathrm{e}^{-\lambda t}\right]_{t=0}^{1} && \text{(antiderivative of } \lambda \mathrm{e}^{-\lambda t}\text{)} \\ &= 1 - \mathrm{e}^{-\lambda} && \text{(evaluate at the bounds)} \end{aligned}$$

With $\lambda = 0.5$, for example, $\Pr(0 \leq T \leq 1) = 1 - \mathrm{e}^{-0.5} \approx 0.393$.

Definition 0.10 (Cumulative distribution function (CDF))**.** For a random variable X (discrete or continuous), its **cumulative distribution function** [10] is:

$$\mathbb{F}(t) \stackrel{\text{def}}{=} \Pr(X \leq t), \quad t \in \mathbb{R}.$$

Remark

Remark. The CDF is always defined, whether X is discrete or continuous.

Example 0.9 (CDF of the number of heads)**.** Summing the PMF in Example 0.1 over values at or below t:

$$\mathbb{F}(t) = \begin{cases} 0, & t < 0 \\ \frac{1}{4}, & 0 \leq t < 1 \\ \frac{3}{4}, & 1 \leq t < 2 \\ 1, & t \geq 2 \end{cases}$$

Theorem 0.5 (Properties of a CDF)**.** *The CDF $\mathbb{F}$ of any random variable X is non-decreasing, and:*

$$\lim_{t \to -\infty} \mathbb{F}(t) = 0, \qquad \lim_{t \to \infty} \mathbb{F}(t) = 1$$

Proof

Proof. Non-decreasing. For $s < t$, $\{X \le t\}$ is the disjoint union of $\{X \le s\}$ and $\{s < X \le t\}$, so finite additivity gives:

$$\begin{aligned} \mathbb{F}(t) &= \Pr(X \le s) + \Pr(s < X \le t) && \text{(additivity, and the definition of the CDF)} \\ &\ge \Pr(X \le s) && \text{(probabilities are non-negative)} \\ &= \mathbb{F}(s) && \text{(definition of the CDF)} \end{aligned}$$

Limit at $-\infty$. The events $A_n = \{X \le -n\}$, $n = 1, 2, ...$, are decreasing, and their intersection is empty, because every outcome ω has $X(\omega) > -n$ once $n > -X(\omega)$. By continuity of probability, $\mathbb{F}(-n) = \Pr(A_n) \to 0$. For $t \le -n$, $0 \le \mathbb{F}(t) \le \mathbb{F}(-n)$, because $\mathbb{F}$ is non-decreasing, so $\mathbb{F}(t) \to 0$ as $t \to -\infty$.
Limit at ∞. The events $B_n = \{X > n\}$ are decreasing, and their intersection is empty, because every outcome ω has $X(\omega) \le n$ once $n \ge X(\omega)$. By continuity of probability, $\Pr(B_n) \to 0$, so, by the complement rule:

$$\begin{aligned} \mathbb{F}(n) &= 1 - \Pr(X > n) && \text{(complement rule)} \\ &\to 1 - 0 && (\Pr(B_n) \to 0) \end{aligned}$$

For $t \ge n$, $\mathbb{F}(n) \le \mathbb{F}(t) \le 1$, because $\mathbb{F}$ is non-decreasing, so $\mathbb{F}(t) \to 1$ as $t \to \infty$. □

Theorem 0.6 (The CDF determines the distribution)**.** *If random variables X and Y have the same CDF, then $\Pr(X \in B) = \Pr(Y \in B)$ for every Borel set $B \subseteq \mathbb{R}$, which includes every interval, and every set built from intervals by countably many unions, intersections, and complements.*

Remark

Remark. The proof needs measure theory beyond these notes [6], [10].

Definition 0.11 (Quantile function (population inverse CDF))**.** For a random variable X with cumulative distribution function (CDF) $\mathbb{F}$, its population **quantile function** (generalized inverse of $\mathbb{F}$) is [11]:

$$\mathbb{Q}(p) \stackrel{\text{def}}{=} \inf\{t : \mathbb{F}(t) \ge p\}, \quad p \in (0, 1).$$

Remark

Remark. The restriction to $p \in (0, 1)$ keeps $\mathbb{Q}(p)$ finite: for $p = 1$, the set $\{t : \mathbb{F}(t) \geq 1\}$ is empty whenever $\mathbb{F}(t) < 1$ for every t, as for a normal distribution, and the infimum of the empty set is $+\infty$.

Example 0.10 (Median of the number of heads)**.** From Example 0.9, $\mathbb{F}(t) \geq 0.5$ first holds at $t = 1$, so the median of the number of heads is $\mathbb{Q}(0.5) = 1$.

Remark

Remark. Figure 1 shows the mass function, CDF and quantile function of the number of heads in two fair coin flips (top row), beside the density, CDF and quantile function of a standard normal variable (bottom row).

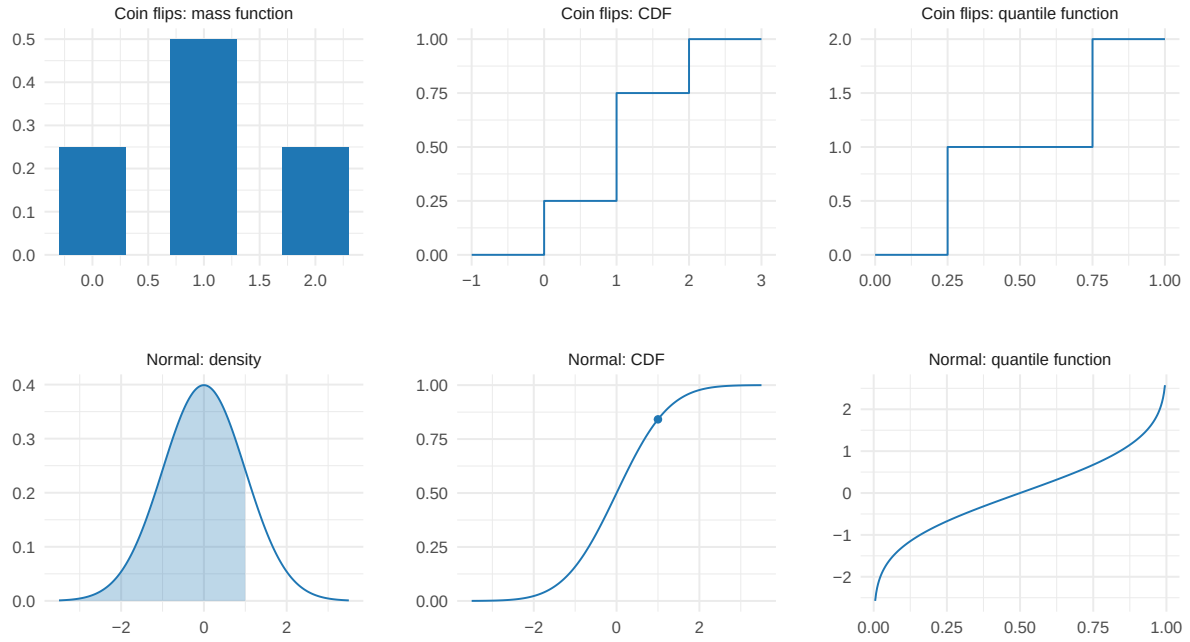


Figure 1: A probability mass function, cumulative distribution function and quantile function for a discrete random variable (top row: the number of heads in two fair coin flips) and a continuous one (bottom row: the standard normal distribution). The shaded area under the normal density up to a point equals the CDF at that point. The quantile function reverses the CDF: its axes are the CDF's, swapped.

Theorem 0.7 (A density integrates to the CDF)**.** *If X is a continuous random variable with density f and CDF $\mathbb{F}$, then for every t:*

$$\mathbb{F}(t) = \int_{-\infty}^{t} f(x)\,dx$$

Proof

Proof. For any $a < t$, $\{X \leq t\}$ is the disjoint union of $\{X \leq a\}$ and $\{a < X \leq t\}$, and $\Pr(X = a) = 0$ because X is continuous. So:

$$\begin{aligned}
\mathbb{F}(t) &= \Pr(X \leq a) + \Pr(a < X \leq t) && \text{(additivity, and the definition of the CDF)} \\
&= \mathbb{F}(a) + \Pr(a \leq X \leq t) && \text{(definition of the CDF; } \Pr(X = a) = 0\text{)} \\
&= \mathbb{F}(a) + \int_{a}^{t} f(x)\,dx && \text{(definition of the density)}
\end{aligned}$$

Letting $a \to -\infty$, $\mathbb{F}(a) \to 0$ (Theorem 0.5), which gives $\mathbb{F}(t) = \int_{-\infty}^{t} f(x)\,dx$. □

Theorem 0.8 (The derivative of the CDF is the density where the density is continuous)**.** *If X is a continuous random variable with density f and CDF $\mathbb{F}$, then at every t where f is continuous:*

$$f(t) = \frac{\partial}{\partial t}\mathbb{F}(t)$$

Proof

Proof. By Theorem 0.7, $\mathbb{F}(t) = \int_{-\infty}^{t} f(x)\,dx$. The fundamental theorem of calculus then gives $\frac{\partial}{\partial t}\int_{-\infty}^{t} f(x)\,dx = f(t)$ at every t where f is continuous. □

Corollary 0.1 (A CDF has its density as derivative at all but finitely many points)**.** *If X is a continuous random variable with CDF $\mathbb{F}$ and a density f that is continuous at all but finitely many points, then $\mathbb{F}$ is differentiable with $\mathbb{F}' = f$ at all but finitely many points.*

Proof

Proof. At every point t where f is continuous, Theorem 0.8 gives $\mathbb{F}'(t) = f(t)$, and by assumption those are all but finitely many points. □

Remark

Remark. Every density in these notes is continuous at all but finitely many points, so Corollary 0.1 applies to each of them.

Example 0.11 (CDF and density of a uniform random variable)**.** For $X \sim \text{Uniform}(0, 1)$ with the density f of Definition 0.5, and $t \in [0, 1]$:

$$\begin{aligned}
\mathbb{F}(t) &= \int_{-\infty}^{t} f(x)\,dx && \text{(the CDF is the integral of the density)} \\
&= \int_{-\infty}^{0} 0\,dx + \int_{0}^{t} 1\,dx && \text{(split at 0, and substitute } f\text{)} \\
&= t && \text{(integrate)}
\end{aligned}$$

For $t \in (0, 1)$, where f is continuous,

$$\begin{aligned}
\frac{\partial}{\partial t}\mathbb{F}(t) &= 1 \\
&= f(t).
\end{aligned}$$

At $t = 1$, f jumps from 1 to 0, and $\mathbb{F}$ has no derivative: its slope is 1 on the left and 0 on the right.

Theorem 0.9 (A piecewise-smooth CDF has its derivative as a density)**.** *If the CDF $\mathbb{F}$ of a random variable X is continuous everywhere, and has a continuous derivative at all but finitely many points, then X is continuous, and $\mathbb{F}'$ (given any values at the finitely many exceptional points) is a density of X.*

Proof

Proof. For events $A \subseteq B$, B is the disjoint union of A and $B \setminus A$, so additivity gives $\Pr(A) \leq \Pr(B)$. For any x and $\Delta > 0$, $\{X = x\} \subseteq \{x - \Delta < X \leq x\}$, so:

$$
\begin{aligned}
0 \le \Pr(X = x) & \qquad \text{(probabilities are non-negative)} \\
\le \Pr(x - \Delta < X \le x) & \qquad (\{X = x\} \subseteq \{x - \Delta < X \le x\}) \\
= \mathbb{F}(x) - \mathbb{F}(x - \Delta) & \qquad \text{(additivity, and the definition of the CDF)}
\end{aligned}
$$

and $\mathbb{F}(x) - \mathbb{F}(x - \Delta) \to 0$ as $\Delta \downarrow 0$, because $\mathbb{F}$ is continuous. So $\Pr(X = x) = 0$ for every x, and X is continuous. Then, for $a \le b$:

$$
\begin{aligned}
\Pr(a \le X \le b) &= \Pr(a < X \le b) && (\Pr(X = a) = 0) \\
&= \mathbb{F}(b) - \mathbb{F}(a) && \text{(additivity, and the definition of the CDF)} \\
&= \int_a^b \mathbb{F}'(x)\, dx && \text{(fundamental theorem of calculus)}
\end{aligned}
$$

The last step applies the fundamental theorem of calculus on each piece of $[a, b]$ between exceptional points, where $\mathbb{F}'$ is continuous, and joins the pieces using the continuity of $\mathbb{F}$ at the exceptional points. Finally, $\mathbb{F}' \ge 0$ wherever it exists, because $\mathbb{F}$ is non-decreasing (Theorem 0.5). □

Example 0.12 (A density from a CDF with a corner)**.** Let X have CDF $\mathbb{F}(t) = 0$ for $t < 0$, $\mathbb{F}(t) = t^2$ for $0 \le t \le 1$, and $\mathbb{F}(t) = 1$ for $t > 1$. This $\mathbb{F}$ is continuous everywhere, and has a continuous derivative everywhere except $t = 1$, where its slope drops from 2 to 0. By Theorem 0.9, X is continuous with density $f(t) = 2t$ for $0 \le t < 1$ and $f(t) = 0$ otherwise. As a check, $\int_0^1 2t\, dt = 1$.

Theorem 0.10 (A density as a limit of interval probabilities)**.** *If X is a continuous random variable with density f, and f is continuous at x, then $f(x)$ is the limit of the probability that X falls in an interval starting at x, divided by the width of that interval, as that width shrinks to 0:*

$$
f(x) = \lim_{\Delta \downarrow 0} \frac{\Pr(x \le X < x + \Delta)}{\Delta}
$$

Proof

Proof. For $\Delta > 0$:

$$
\begin{aligned}
\frac{\Pr(x \le X < x + \Delta)}{\Delta} &= \frac{\Pr(x < X \le x + \Delta)}{\Delta} && (\Pr(X = x) = 0 \text{ and } \Pr(X = x + \Delta) = 0) \\
&= \frac{\mathbb{F}(x + \Delta) - \mathbb{F}(x)}{\Delta} && \text{(additivity, and the definition of the CDF)}
\end{aligned}
$$

By Theorem 0.8, $\mathbb{F}$ is differentiable at x with $\mathbb{F}'(x) = f(x)$, so this difference quotient

converges to $f(x)$ as $\Delta \downarrow 0$. □

Remark

Remark. For small $\Delta > 0$, $f(x) \cdot \Delta \approx \Pr(x \le X < x + \Delta)$. Although a density is not unique (Theorem 0.1), this limit pins down its value at every point where it is continuous. See also Rothman et al. [12] (Chapter 22, p. 535).

Example 0.13 (The uniform density as a limit)**.** For $X \sim \text{Uniform}(0, 1)$ and $x \in [0, 1)$, $\Pr(x \le X < x + \Delta) = \Delta$ once $\Delta \le 1 - x$, so:

$$\begin{aligned} f(x) &= \lim_{\Delta \downarrow 0} \frac{\Delta}{\Delta} \\ &= 1 \end{aligned}$$

For $x < 0$ or $x > 1$, the interval eventually misses $[0, 1]$, so the limit is 0. Both agree with the density of Definition 0.5.

Proposition 0.2 (The density limit is the right derivative of the CDF)**.** *If X is a continuous random variable with CDF $\mathbb{F}$, then for every x:*

$$\lim_{\Delta \downarrow 0} \frac{\Pr(x \le X < x + \Delta)}{\Delta} = \lim_{\Delta \downarrow 0} \frac{\mathbb{F}(x + \Delta) - \mathbb{F}(x)}{\Delta}$$

whenever either limit exists; the right-hand side is the right derivative of $\mathbb{F}$ at x.

Proof

Proof. The first two lines of the proof of Theorem 0.10 use only that X is continuous, so for every $\Delta > 0$ the two quotients are equal. Equal functions of Δ have the same limit as $\Delta \downarrow 0$, or both have none. □

Example 0.14 (The density limit at a jump)**.** At a point where f is not continuous, the limit can differ from the chosen value $f(x)$. For $X \sim \text{Uniform}(0, 1)$ with the density f of Definition 0.5, which sets $f(1) = 1$ and $f(x) = 0$ for $x > 1$, and for $\Delta > 0$:

$$\begin{aligned} \Pr(1 \le X < 1 + \Delta) &= \Pr(1 \le X \le 1 + \Delta) && (\Pr(X = 1 + \Delta) = 0) \\ &= \int_1^{1+\Delta} f(x)\, dx && \text{(definition of the density)} \\ &= 0 && (f = 0 \text{ on } (1, 1 + \Delta]\text{; one point does not change an integral}) \end{aligned}$$

So the limit of $\Pr(1 \le X < 1+\Delta)/\Delta$ at $x = 1$ is 0, although $f(1) = 1$. By Proposition 0.2, 0 is also the right derivative of $\mathbb{F}$ at 1, matching the slope of 0 on the right in Example 0.11.

Theorem 0.11 (Density functions integrate to 1)**.** *If X is a continuous random variable with density f:*

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

Proof

Proof.

$$\begin{aligned}
\int_{-\infty}^{\infty} f(x)\,dx &= \lim_{b\to\infty} \int_{-\infty}^{b} f(x)\,dx && \text{(definition of an improper integral)} \\
&= \lim_{b\to\infty} \mathbb{F}(b) && \text{(the CDF is the integral of the density)} \\
&= 1 && \text{(limit of a CDF)}
\end{aligned}$$

The last step is Theorem 0.5. □

Example 0.15 (The uniform density integrates to 1)**.** For $X \sim \text{Uniform}(0, 1)$ (Definition 0.5),

$$\begin{aligned}
\int_{-\infty}^{\infty} f(x)\,dx &= \int_{0}^{1} 1\,dx \\
&= 1.
\end{aligned}$$

Definition 0.12 (Jointly distributed random variables)**.** Random variables $X_1, \ldots, X_n$ are **jointly distributed** when they are defined on the same sample space Ω, so that an event about several of them at once, such as $\{X_1 \le x_1\} \cap \{X_2 \le x_2\}$, has a probability.

Remark

Remark. We write $\Pr(X \in A, Y \in B)$ for $\Pr(\{X \in A\} \cap \{Y \in B\})$, and similarly for more variables.

Example 0.16 (The first flip and the total)**.** In Example 0.1, let X_1 indicate heads on the first flip (as in Example 0.2), and let X be the total number of heads. Both are

functions on the same sample space, $\{HH, HT, TH, TT\}$, so they are jointly distributed, and, for example:

$$\begin{aligned}\Pr(X_1 = 1, X = 1) &= \Pr(\{HH, HT\} \cap \{HT, TH\}) \\ &= \Pr(\{HT\}) \\ &= \tfrac{1}{4}\end{aligned}$$

Definition 0.13 (Joint distribution)**.** The **joint distribution** of jointly distributed random variables $X_1, \dots, X_n$ is the function that assigns to each set $A \subseteq \mathbb{R}^n$ for which $\{(X_1, \dots, X_n) \in A\}$ is an event the probability $\Pr((X_1, \dots, X_n) \in A)$ [13].

Example 0.17 (Joint distribution of the first flip and the total)**.** In Example 0.16, the joint distribution of X_1 and X assigns to the set $A = \{(x_1, x) : x_1 = 1,\, x \geq 1\}$ the probability

$$\begin{aligned}\Pr((X_1, X) \in A) &= \Pr(\{HH, HT\}) \\ &= \tfrac{1}{2}\end{aligned}$$

since the first flip is heads exactly on HH and HT, and each of those outcomes has at least one head.

Theorem 0.12 (A joint distribution is determined by its rectangle probabilities)**.** *If jointly distributed random variables $X_1, \dots, X_n$, and jointly distributed random variables $Y_1, \dots, Y_n$, satisfy*

$$\Pr(X_1 \in A_1, \dots, X_n \in A_n) = \Pr(Y_1 \in A_1, \dots, Y_n \in A_n)$$

for all sets $A_1, \dots, A_n$ of real numbers for which these are events, then $(X_1, \dots, X_n)$ and $(Y_1, \dots, Y_n)$ have the same joint distribution.

ℹ Remark

Remark. The proof needs measure theory beyond these notes: the sets $A_1 \times \cdots \times A_n$ form a π-system, and two probability measures that agree on a π-system agree on the σ-algebra it generates [6].

Definition 0.14 (Joint probability mass function (joint PMF))**.** For jointly distributed discrete random variables X and Y, the **joint probability mass function** (**joint PMF**) of X and Y is the probability that X takes the value x and Y takes the value y:

$$\mathbb{P}(X = x,\, Y = y) \overset{\text{def}}{=} \Pr(\{X = x\} \cap \{Y = y\})$$

Example 0.18 (Joint PMF of the first flip and the total)**.** Continuing Example 0.16, each outcome of the two flips fixes both X_1 and X, so the joint PMF puts probability $1/4$ on each outcome's pair of values:

	$X = 0$	$X = 1$	$X = 2$
$X_1 = 0$	$1/4$ (TT)	$1/4$ (TH)	0
$X_1 = 1$	0	$1/4$ (HT)	$1/4$ (HH)

Definition 0.15 (Marginal distribution)**.** When a random variable X is one of several jointly distributed random variables, the distribution of X by itself (its CDF, and its PMF or density) is called the **marginal distribution** of X [14].

i Remark

Remark. The word "marginal" only says that the other variables are being set aside; the marginal distribution of X is the same object as the distribution of X.

Example 0.19 (Marginal distribution of the total)**.** In Example 0.18, the marginal distribution of the total X is the PMF of Example 0.1: $\mathbb{P}(X = 0) = 1/4$, $\mathbb{P}(X = 1) = 1/2$, $\mathbb{P}(X = 2) = 1/4$.

Theorem 0.13 (Marginal PMF from a joint PMF)**.** *If X and Y are jointly distributed discrete random variables, then for every x:*

$$\mathbb{P}(X = x) = \sum_{y \in \mathcal{R}(Y)} \mathbb{P}(X = x,\, Y = y)$$

i Proof

Proof. The event $\{X = x\}$ is the disjoint union of the events $\{X = x\} \cap \{Y = y\}$ over the countably many values $y \in \mathcal{R}(Y)$, so:

$$\begin{aligned}
\mathbb{P}(X = x) &= \Pr\left(\bigcup_{y \in \mathcal{R}(Y)} (\{X = x\} \cap \{Y = y\})\right) && \text{(the events partition } \{X = x\}) \\
&= \sum_{y \in \mathcal{R}(Y)} \Pr(\{X = x\} \cap \{Y = y\}) && \text{(countable additivity)} \\
&= \sum_{y \in \mathcal{R}(Y)} \mathbb{P}(X = x,\, Y = y) && \text{(definition of the joint PMF)}
\end{aligned}$$

□

Example 0.20 (Summing a row of the joint PMF)**.** In Example 0.18, summing the $X_1 = 1$ row:

$$\begin{aligned}
\mathbb{P}(X_1 = 1) &= \mathbb{P}(X_1 = 1, X = 0) + \mathbb{P}(X_1 = 1, X = 1) + \mathbb{P}(X_1 = 1, X = 2) && \text{(marginal PMF from the joint PM} \\
&= 0 + \tfrac{1}{4} + \tfrac{1}{4} && \text{(read the table)} \\
&= \tfrac{1}{2} && \text{(add)}
\end{aligned}$$

which matches $X_1 \sim \text{Ber}(1/2)$ from Example 0.2.

Definition 0.16 (Joint probability density function (joint density))**.** For jointly distributed continuous random variables X and Y, a **joint probability density function** (**joint density**) of X and Y, denoted $f_{X,Y}(x, y)$ or $\mathbb{p}(X = x,\, Y = y)$, is a function $f_{X,Y}$ on $\mathbb{R}^2$ that satisfies:

- $f_{X,Y}(x, y) \geq 0$ for every (x, y).
- The integral of $f_{X,Y}$ over any region $A \subseteq \mathbb{R}^2$ is the probability that the pair (X, Y) falls in A:
$$\Pr((X, Y) \in A) = \iint_A f_{X,Y}(x, y)\, dx\, dy$$

i Remark

Remark. As with events, a fully rigorous version restricts A to a designated collection of regions; every region that arises in these notes is in it.

Example 0.21 (A joint density on a triangle)**.** Let $f_{X,Y}(x, y) = 2$ for $0 \leq x \leq y \leq 1$, and 0 otherwise. The triangle $\{(x, y) : 0 \leq x \leq y \leq 1\}$ has area $1/2$, so $f_{X,Y}$ gives the whole plane probability $2 \cdot \frac{1}{2} = 1$, and it is the joint density of a pair (X, Y) with $X \leq Y$

always. For example, the probability that both are at most $1/2$ is 2 times the area of the smaller triangle $\{0 \le x \le y \le 1/2\}$:

$$\begin{aligned}\Pr(X \le \tfrac{1}{2}, Y \le \tfrac{1}{2}) &= 2 \cdot \tfrac{1}{8} \\ &= \tfrac{1}{4}\end{aligned}$$

Theorem 0.14 (Not every pair of continuous random variables has a joint density)**.** *If X is a continuous random variable, then the pair (X, X) has no joint density.*

Proof

Proof. Suppose f were a joint density of (X, X), and let $L = \{(x, y) : y = x\}$ be the diagonal line. Every outcome ω has $X(\omega) = X(\omega)$, so $\{(X, X) \in L\} = \Omega$. For each x, the function $y \mapsto 1_{y=x} f(x, y)$ is 0 except at the single point $y = x$, so its integral over y is 0. Then:

$$\begin{aligned}
1 &= \Pr((X, X) \in L) && (\{(X, X) \in L\} = \Omega, \text{ and } \Pr(\Omega) = 1) \\
&= \iint_L f(x, y)\, dx\, dy && (\text{definition of a joint density}) \\
&= \int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} 1_{y=x} f(x, y)\, dy \right) dx && (\text{iterate the integral; Tonelli's theorem}) \\
&= \int_{-\infty}^{\infty} 0\, dx && (\text{the inner integrand is 0 except at } y = x) \\
&= 0 && (\text{integrate})
\end{aligned}$$

which is a contradiction. Tonelli's theorem allows the iterated integral because $f \ge 0$ (Fubini–Tonelli theorem; Billingsley [6], Theorem 18.3). □

Theorem 0.15 (Marginal density from a joint density)**.** *If X and Y have joint density $f_{X,Y}$, then X has the density:*

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

Proof

Proof. For $a \le b$, the event $\{a \le X \le b\}$ is the event that (X, Y) falls in the strip $[a, b] \times \mathbb{R}$, so:

$$
\begin{aligned}
\Pr(a \le X \le b) &= \Pr((X,Y) \in [a,b] \times \mathbb{R}) && \text{(same event)} \\
&= \iint_{[a,b]\times\mathbb{R}} f_{X,Y}(x,y)\,dx\,dy && \text{(definition of a joint density)} \\
&= \int_a^b \left(\int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dy \right) dx && \text{(iterate the integral; Tonelli's theorem)} \\
&= \int_a^b f_X(x)\,dx && \text{(definition of } f_X\text{)}
\end{aligned}
$$

Tonelli's theorem allows the iterated integral because $f_{X,Y} \ge 0$ (Fubini–Tonelli theorem; Billingsley [6], Theorem 18.3). So f_X satisfies Definition 0.3. □

Example 0.22 (Marginal densities on the triangle)**.** For the joint density of Example 0.21 and $x \in [0,1]$, $f_{X,Y}(x,y) = 2$ exactly when $x \le y \le 1$, so:

$$
\begin{aligned}
f_X(x) &= \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dy && \text{(marginal density from a joint density)} \\
&= \int_x^1 2\,dy && (f_{X,Y}(x,y) = 2 \text{ for } x \le y \le 1, \text{ else } 0) \\
&= 2(1-x) && \text{(integrate)}
\end{aligned}
$$

The same steps with the roles swapped give

$$
\begin{aligned}
f_Y(y) &= \int_0^y 2\,dx \\
&= 2y
\end{aligned}
$$

for $y \in [0,1]$.

Definition 0.17 (Joint density-mass function)**.** Let X be a discrete random variable and Y a continuous random variable, jointly distributed. A **joint density-mass function** of X and Y, denoted $\mathbb{p}(X = x, Y = y)$, is a function $\mathbb{p}(X = x, Y = y) \ge 0$ that is a probability mass in x and a probability density in y: for every x and every set $B \subseteq \mathbb{R}$,

$$
\Pr(X = x, Y \in B) = \int_B \mathbb{p}(X = x, Y = y)\,dy
$$

Remark

Remark. When X is continuous and Y is discrete, the roles swap: $\Pr(X \in B,\, Y = y) = \int_B \mathbb{p}(X = x,\, Y = y)\, dx$.

Corollary 0.2 (Marginal PMF from a joint density-mass function)**.** *If X and Y have joint density-mass function $\mathbb{p}(X = x,\, Y = y)$, then for every x:*

$$\mathbb{P}(X = x) = \int_{-\infty}^{\infty} \mathbb{p}(X = x,\, Y = y)\, dy$$

Proof

Proof. Every outcome has $Y(\omega) \in \mathbb{R}$, so $\{Y \in \mathbb{R}\} = \Omega$, and:

$$\begin{aligned}
\mathbb{P}(X = x) &= \Pr(\{X = x\} \cap \Omega) && (\{X = x\} \subseteq \Omega) \\
&= \Pr(X = x,\, Y \in \mathbb{R}) && (\{Y \in \mathbb{R}\} = \Omega) \\
&= \int_{-\infty}^{\infty} \mathbb{p}(X = x,\, Y = y)\, dy && \text{(definition of a joint density-mass function, with } B = \mathbb{R})
\end{aligned}$$

□

Example 0.23 (A coin flip and a waiting time)**.** Let $\mathbb{p}(X = 0,\, Y = y) = 1/2$ for $y \in [0, 1]$, $\mathbb{p}(X = 1,\, Y = y) = 1/4$ for $y \in [0, 2]$, and 0 otherwise. Integrating over y gives $\mathbb{P}(X = 0) = 1/2$ and $\mathbb{P}(X = 1) = 1/2$, which add to 1. For example, the probability that $X = 1$ and $Y \leq 1$ is:

$$\begin{aligned}
\Pr(X = 1,\, Y \leq 1) &= \int_0^1 \tfrac{1}{4}\, dy \\
&= \tfrac{1}{4}
\end{aligned}$$

Joint distributions and marginalization

For a discrete random variable X taking values in $\mathbb{X}$, a **probability mass function** (PMF) $\mathbb{P}(x) : \mathbb{X} \to [0, 1]$ gives the probability that X takes the value x:

$$\sum_{x \in \mathbb{X}} \mathbb{P}(x) = 1$$

The **support** of the distribution is the subset of values where the probability is strictly positive: $\{x \in \mathbb{X} : \mathbb{P}(x) > 0\}$. When $\mathbb{X}$ is finite, $\mathbb{P}(x)$ can be written as a probability vector.

For two discrete variables $X \in \mathbb{X}$ and $Y \in \mathbb{Y}$, their **joint probability distribution** $\mathbb{P}(x, y) : \mathbb{X} \times \mathbb{Y} \to [0, 1]$ satisfies:

$$\sum_{x \in \mathbb{X}} \sum_{y \in \mathbb{Y}} \mathbb{P}(x, y) = 1$$

For two variables, the joint probabilities can be laid out in a matrix or table. Summing across a row or column collapses that variable away, a process called **marginalization**:

$$\mathbb{P}(x) = \sum_{y \in \mathbb{Y}} \mathbb{P}(x, y), \qquad \mathbb{P}(y) = \sum_{x \in \mathbb{X}} \mathbb{P}(x, y)$$

The resulting distributions $\mathbb{P}(x)$ and $\mathbb{P}(y)$ are the **marginal distributions** of X and Y.

Example 0.24 (Joint probabilities and marginalization)**.** Consider an example adapted from Brian Hutchinson's notes [15]. Let $m \in \{0, 1\}$ indicate whether a meteorite hits your house on a given day ($m = 1$), and let $d \in \{0, 1\}$ indicate whether you have a good day ($d = 1$). In this distribution, 9 out of 10 days are good days ($\mathbb{P}(d = 1) = 0.9$), and the probability of a meteorite strike is 1 in a million ($\mathbb{P}(m = 1) = 10^{-6}$):

Table 1: Joint distribution of a meteorite strike (m) and day quality (d), with marginals.

	$d = 0$ (bad day)	$d = 1$ (good day)	Marginal $\mathbb{P}(m)$
$m = 0$ (no meteorite)	0.09999902	0.89999998	0.999999
$m = 1$ (meteorite hit)	0.00000098	0.00000002	0.000001
Marginal $\mathbb{P}(d)$	0.10000000	0.90000000	1.000000

Summing each row yields the marginal distribution of m; summing each column yields the marginal distribution of d. Conditioning inverts the perspective:

$$\begin{aligned}\mathbb{P}(d = 0 \mid m = 1) &= \frac{\mathbb{P}(m = 1, d = 0)}{\mathbb{P}(m = 1)} \\ &= \frac{0.00000098}{0.000001} \\ &= 0.98\end{aligned}$$

Given that a meteorite struck your house, you have a 98% chance of having a bad day. In reverse:

$$\begin{aligned}\mathbb{P}(m = 1 \mid d = 0) &= \frac{\mathbb{P}(m = 1, d = 0)}{\mathbb{P}(d = 0)} \\ &= \frac{0.00000098}{0.1} \\ &= 0.0000098\end{aligned}$$

Given that you are having a bad day, the probability of a meteorite hit is roughly 1 in 100,000. Bad days are common ($\mathbb{P}(d = 0) = 0.1$), so having one is very weak evidence that an astronomical rarity occurred.

Continuous distributions and densities

When a random variable X takes values in a continuous space $\mathcal{R}(X) \subseteq \mathbb{R}$, its behavior is described by a **probability density function** (PDF) $\mathbb{p}(x) : \mathcal{R}(X) \to \mathbb{R}_+$ satisfying:

$$\int_{\mathcal{R}(X)} \mathbb{p}(x)\,\mathrm{d}x = 1$$

Probabilities are assigned to subsets $A \subseteq \mathcal{R}(X)$ by integrating the density over that set:

$$\Pr(X \in A) = \int_A \mathbb{p}(x)\,\mathrm{d}x$$

Theorem 0.16 (Points have zero probability in continuous distributions)**.** *For any continuous random variable and any specific value $a \in \mathbb{R}$:*

$$\Pr(X = a) = 0$$

Why? As Brian Hutchinson notes [15], what is the probability that someone's height is *exactly* 6.0000000000 ... feet? Zero. A single real point has width zero, so the integral over a single point is zero.

Non-zero probabilities attach to intervals or regions of non-zero width:

$$\Pr(6-\varepsilon \le X \le 6+\varepsilon) = \int_{6-\varepsilon}^{6+\varepsilon} \mathbb{p}(x)\,\mathrm{d}x > 0 \qquad (\text{for } \varepsilon > 0)$$

Every rule developed for discrete variables carries over to continuous variables by replacing sums $\sum_{x\in\mathcal{R}(X)}$ with integrals $\int_{\mathcal{R}(X)} \mathrm{d}x$. For example, marginalizing out X from a joint density $\mathbb{p}(x,y)$ to find the marginal density $\mathbb{p}(y)$ becomes:

$$\mathbb{p}(y) = \int_{\mathcal{R}(X)} \mathbb{p}(x,y)\,\mathrm{d}x$$

Definition 0.18 (Kernel of a density)**.** A **kernel** of a density f is a function $g \ge 0$ such that $f(x) = c\,g(x)$ for every x, for some constant $c > 0$ that does not depend on x.

Definition 0.19 (Normalizing constant of a density)**.** Let g be a kernel of a density f. The **normalizing constant** of g is the constant $c > 0$ with

$$f(x) = c\,g(x) \quad \text{for every } x.$$

Theorem 0.17 (The normalizing constant makes the density integrate to 1)**.** *If g is a kernel of a density f, its normalizing constant is*

$$c = \frac{1}{\int_{-\infty}^{\infty} g(x)\,dx}.$$

Proof

Proof. A density integrates to 1 over the whole line, so

$$
\begin{aligned}
1 &= \int_{-\infty}^{\infty} f(x)\,dx && \text{(a density integrates to 1)} \\
&= \int_{-\infty}^{\infty} c\,g(x)\,dx && \text{(@def-normalizing-constant)} \\
&= c\int_{-\infty}^{\infty} g(x)\,dx. && \text{(constant factor out of the integral)}
\end{aligned}
$$

Dividing by $\int g(x)\,dx$ gives the formula. □

Definition 0.20 (Mode of a density or mass function)**.** Let X be a random variable with density or mass function f. The **modes** of f are the elements of the argmax of f over the range of X:

$$\operatorname{mode}(f) = \arg\max_{x\in\mathcal{R}(X)} f(x).$$

A **mode** of f is an element of $\operatorname{mode}(f)$.

ℹ Remark (A mode belongs to a density, not to a random variable)

Remark 0.2 (A mode belongs to a density, not to a random variable). A mode is a property of a density or mass function, so "the mode of X" is shorthand for the mode of a chosen density of X.
Two densities of the same random variable can differ on a set of probability zero, and then they can have different modes. A change of variables also moves the mode, because the density of a transformed variable includes a Jacobian factor.
The expectation of X is different: it depends only on the distribution of X, so every density of X gives the same value.

Theorem 0.18 (A density and its kernel have the same modes)**.** *If g is a kernel of the density f of a random variable X, then the modes of f are the elements of the argmax of g over the range of X:*

$$\operatorname{mode}(f) = \arg\max_{x\in\mathcal{R}(X)} g(x).$$

ℹ Proof

Proof. Write $f(x) = c\,g(x)$ with $c > 0$ (Definition 0.18). By the theorem that a positive multiple has the same argmax, $\arg\max_{x\in\mathcal{R}(X)} f(x) = \arg\max_{x\in\mathcal{R}(X)} g(x)$. The left side is $\operatorname{mode}(f)$ by Definition 0.20. □

Example 0.25 (The kernel, normalizing constant, and mode of an exponential density)**.** The exponential density with rate λ is $f(t) = \lambda e^{-\lambda t}$ for $t \geq 0$, and $f(t) = 0$ for $t < 0$. Its kernel is $g(t) = e^{-\lambda t}$ for $t \geq 0$, and $g(t) = 0$ for $t < 0$, and the normalizing constant is

$$\begin{aligned} c &= \frac{1}{\int_0^\infty e^{-\lambda t}\, dt} && \text{(@thm-normalizing-constant-formula)} \\ &= \frac{1}{[-e^{-\lambda t}/\lambda]_0^\infty} && \text{(antiderivative of } e^{-\lambda t}\text{)} \\ &= \frac{1}{1/\lambda} && \text{(evaluate at } \infty \text{ and } 0\text{)} \\ &= \lambda. && \text{(simplify)} \end{aligned}$$

The kernel is largest at $t = 0$, so by Theorem 0.18 the mode of f is 0.

Video lecture

Hutchinson's Probability Refresher (27 min) covers probability mass functions and probability density functions [16]. The login for the video site is posted on Canvas.

Watch

5MinuteAI explains discrete random variables and continuous random variables in five minutes each, following the University of Washington's CSE 312 [17].

Survival, hazard, and cumulative hazard functions

Definition 0.1 (Survival function (survivor function))**.** The **survival function** (or **survivor function**) of a random variable T, denoted $\mathbb{S}(t)$, is the probability that T exceeds t [18]:

$$\mathbb{S}(t) \stackrel{\text{def}}{=} \Pr(T > t)$$

Remark

Remark. The name comes from time-to-event analysis: if T is the time at which a participant dies, then $\mathbb{S}(t)$ is the probability that the participant is still alive at time t. The definition itself applies to any random variable.

Theorem 0.1 (Survival function and CDF)**.** *For any random variable T with CDF $F(t)$:*

$$\mathbb{S}(t) = 1 - F(t)$$

If T is continuous with density $f(t)$, then also:

$$\mathbb{S}(t) = \int_{u=t}^{\infty} f(u)\,du$$

Proof

Proof. The event $\{T > t\}$ is the complement of the event $\{T \leq t\}$, so the complement rule gives:

$$\begin{aligned}
\mathbb{S}(t) &\stackrel{\text{def}}{=} \Pr(T > t) && \text{(definition of the survival function)} \\
&= 1 - \Pr(T \leq t) && \text{(complement rule)} \\
&= 1 - F(t) && \text{(definition of the CDF)}
\end{aligned}$$

For continuous T, the density integrates to 1 over the real line (Theorem 0.11), and F is the integral of the density up to t (Theorem 0.7), so:

$$\begin{aligned}
1 - F(t) &= \int_{u=-\infty}^{\infty} f(u)\,du - \int_{u=-\infty}^{t} f(u)\,du && \text{(total integral is 1; } F \text{ is the integral of } f\text{)} \\
&= \int_{u=t}^{\infty} f(u)\,du && \text{(split the first integral at } t \text{ and cancel)}
\end{aligned}$$

□

Example 0.1 (Survival function of an exponential distribution)**.** Let T be exponential with rate $\lambda > 0$ (Definition 0.9), so T has density $f(t) = \lambda e^{-\lambda t}$ for $t \geq 0$. For $t \geq 0$, by Theorem 0.1:

$$\begin{aligned}
\mathbb{S}(t) &= \int_{u=t}^{\infty} \lambda e^{-\lambda u}\,du && \text{(integral form of the survival function)} \\
&= \left[-e^{-\lambda u}\right]_{u=t}^{\infty} && \text{(antiderivative of } \lambda e^{-\lambda u}\text{)} \\
&= 0 - \left(-e^{-\lambda t}\right) && \text{(evaluate at the bounds; } e^{-\lambda u} \to 0 \text{ as } u \to \infty\text{)} \\
&= e^{-\lambda t} && \text{(simplify)}
\end{aligned}$$

For $t < 0$,

$$\begin{aligned}\mathbb{S}(t) &= \Pr(T > t) \\ &= 1,\end{aligned}$$

since $T \geq 0$. With $\lambda = 0.5$, for example, $\mathbb{S}(2) = \mathrm{e}^{-1} \approx 0.368$.

Definition 0.2 (Hazard function (hazard rate, hazard rate function))**.** The **hazard function** (also called the **hazard rate** or **hazard rate function**) of a continuous random variable T at a value t with $\Pr(T \geq t) > 0$, typically denoted $\lambda(t)$ or $\mathbb{h}(t)$, is the limit of the conditional probability that T falls in an interval starting at t, given $T \geq t$, divided by the width of that interval, as that width shrinks to 0 [19]:

$$\lambda(t) \stackrel{\text{def}}{=} \lim_{\Delta \downarrow 0} \frac{\Pr(t \leq T < t + \Delta \mid T \geq t)}{\Delta}$$

Remark

Remark. Sources differ on the symbol: $\mathbb{h}(t)$ appears in Dobson and Barnett [20], Vittinghoff et al. [21], Klein and Moeschberger [22], and Kleinbaum and Klein [23], while $\lambda(t)$ appears in Rothman et al. [12] and Kalbfleisch and Prentice [24].
If T is the time at which an event occurs, then $\lambda(t)$ is a rate, not a probability: for a small interval width $\Delta > 0$, $\lambda(t) \cdot \Delta$ is approximately the probability that the event occurs in $[t, t + \Delta)$, given that it has not occurred before t. Many sources write the hazard as $\lambda(t) = \mathbb{p}(T = t \mid T \geq t)$, reading it as the density of T at t, conditional on the event $T \geq t$; that notation abbreviates the limit in Definition 0.2. For a discrete T, the analogous quantity is the discrete-time hazard.
The name "hazard" carries a connotation that the event is undesirable — death, relapse, equipment failure, and so on. When the event in question is neutral or desirable (recovery, conception, graduation, response to treatment), the same quantity $\lambda(t)$ is often called the **event incidence rate** instead. This terminology parallels the convention that conditional probabilities of undesirable events are called **risks**, while the same conditional probabilities for neutral or desirable events are simply called **probabilities**. The math is identical; only the name changes with the valence of the event.

Example 0.2 (A hazard greater than 1)**.** A hazard can exceed 1, just as a density can (Example 0.6). Let T be exponential with rate $\lambda = 2$, so $\mathbb{S}(t) = \mathrm{e}^{-2t}$ for $t \geq 0$ (Example 0.1). Since T is continuous, $\Pr(T = s) = 0$ for every s, so

$$\begin{aligned}\Pr(T \geq s) &= \Pr(T > s) \\ &= \mathbb{S}(s).\end{aligned}$$

For $t \geq 0$ and $\Delta > 0$, $\{T \geq t\}$ is the disjoint union of $\{t \leq T < t + \Delta\}$ and $\{T \geq t + \Delta\}$,

so:

$$\begin{aligned}
\Pr(t \le T < t+\Delta \mid T \ge t) &= \frac{\Pr(\{t \le T < t+\Delta\} \cap \{T \ge t\})}{\Pr(T \ge t)} && \text{(definition of conditional probability)} \\
&= \frac{\Pr(t \le T < t+\Delta)}{\Pr(T \ge t)} && \text{(subset property)} \\
&= \frac{\Pr(T \ge t) - \Pr(T \ge t+\Delta)}{\Pr(T \ge t)} && \text{(additivity)} \\
&= \frac{\mathbb{S}(t) - \mathbb{S}(t+\Delta)}{\mathbb{S}(t)} && (\Pr(T \ge s) = \mathbb{S}(s)) \\
&= \frac{\mathrm{e}^{-2t} - \mathrm{e}^{-2(t+\Delta)}}{\mathrm{e}^{-2t}} && \text{(exponential survival function)} \\
&= 1 - \mathrm{e}^{-2\Delta} && \text{(divide by } \mathrm{e}^{-2t}\text{)}
\end{aligned}$$

The subset property applies because $\{t \le T < t+\Delta\} \subseteq \{T \ge t\}$. Then:

$$\begin{aligned}
\lambda(t) &= \lim_{\Delta \downarrow 0} \frac{1 - \mathrm{e}^{-2\Delta}}{\Delta} && \text{(definition of the hazard function)} \\
&= \frac{\partial}{\partial \Delta}\left(1 - \mathrm{e}^{-2\Delta}\right)\Big|_{\Delta=0} && \text{(definition of the derivative; } 1 - \mathrm{e}^{0} = 0\text{)} \\
&= 2\mathrm{e}^{0} && \text{(chain rule)} \\
&= 2 && (\mathrm{e}^{0} = 1)
\end{aligned}$$

So $\lambda(t) = 2 > 1$ at every $t \ge 0$: a hazard is a rate, not a probability.

Definition 0.3 (Discrete-time hazard)**.** The **discrete-time hazard** of a discrete random variable T at a value t with $\Pr(T \ge t) > 0$ is the conditional probability that T equals t, given $T \ge t$:

$$\Pr(T = t \mid T \ge t)$$

Example 0.3 (Rolling until the first six)**.** Roll a fair die repeatedly, and let T be the number of the roll that first shows a six. Then $T \ge t$ exactly when the first $t-1$ rolls are not sixes, which has probability $(5/6)^{t-1}$, and $T = t$ when, in addition, roll t is a six. So, for each $t = 1, 2, \ldots$:

$$\begin{aligned}
\Pr(T = t \mid T \geq t) &= \frac{\Pr(T = t,\, T \geq t)}{\Pr(T \geq t)} && \text{(definition of conditional probability)} \\
&= \frac{\Pr(T = t)}{\Pr(T \geq t)} && (T = t \text{ implies } T \geq t) \\
&= \frac{(5/6)^{t-1} \cdot (1/6)}{(5/6)^{t-1}} && \text{(rolls are independent)} \\
&= \frac{1}{6} && \text{(cancel)}
\end{aligned}$$

The discrete-time hazard is the same at every roll: having gone without a six so far does not change the chance of a six on the next roll.

Corollary 0.1 (A discrete-time hazard lies in $[0, 1]$)**.** *The discrete-time hazard of a discrete random variable T at a value t with $\Pr(T \geq t) > 0$ satisfies:*

$$0 \leq \Pr(T = t \mid T \geq t) \leq 1$$

Proof

Proof. $\{T \geq t\}$ is the disjoint union of $\{T = t\}$ and $\{T > t\}$, so, by additivity, $\Pr(T \geq t) = \Pr(T = t) + \Pr(T > t) \geq \Pr(T = t)$. Since $\{T = t\} \subseteq \{T \geq t\}$:

$$\begin{aligned}
\Pr(T = t \mid T \geq t) &= \frac{\Pr(\{T = t\} \cap \{T \geq t\})}{\Pr(T \geq t)} && \text{(definition of conditional probability)} \\
&= \frac{\Pr(T = t)}{\Pr(T \geq t)} && \text{(subset property)}
\end{aligned}$$

The numerator is non-negative and at most the positive denominator, so the ratio lies in $[0, 1]$. □

Remark

Remark. Unlike the continuous-time hazard function, which is a rate and can exceed 1 (Example 0.2), the discrete-time hazard is a probability.

Theorem 0.2 (A hazard is non-negative)**.** *If $\lambda(t)$ is the hazard function of a continuous random variable T at a value t with $\Pr(T \geq t) > 0$, then:*

$$\lambda(t) \geq 0$$

Proof

Proof. For each $\Delta > 0$, the quantity that defines the hazard is non-negative:

$$\frac{\Pr(t \le T < t + \Delta \mid T \ge t)}{\Delta} \ge 0 \quad \text{(a conditional probability is non-negative, and } \Delta > 0\text{)}$$

A limit of non-negative values is non-negative, so $\lambda(t) \ge 0$. □

Remark (The range of a hazard)

Remark 0.1 (The range of a hazard). A hazard can exceed 1: Example 0.2 has $\lambda(t) = 2$. Together with Theorem 0.2, the hazards of a continuous random variable take values in $[0, \infty)$. The discrete-time hazard differs: it lies in $[0, 1]$ by Corollary 0.1.

Theorem 0.3 (Hazard equals density over survival)**.** *If T is a continuous random variable with density $f(t)$ and survival function $\mathbb{S}(t)$, then for every t with $\mathbb{S}(t) > 0$ at which f is continuous:*

$$\lambda(t) = \frac{f(t)}{\mathbb{S}(t)}$$

Proof

Proof. The proof uses three facts. First, for $\Delta > 0$ the event $\{t \le T < t + \Delta\}$ is a subset of the event $\{T \ge t\}$, so intersecting them leaves $\{t \le T < t + \Delta\}$ (the subset property). Second, $\Pr(T = t) = 0$ for a continuous T, so

$$\begin{aligned} \Pr(T \ge t) &= \Pr(T > t) \\ &= \mathbb{S}(t), \end{aligned}$$

which is positive. Third, because f is continuous at t, Theorem 0.10 gives $f(t)$ as the limit of $\Pr(t \le T < t + \Delta)/\Delta$.

$$
\begin{aligned}
\lambda(t) &\stackrel{\text{def}}{=} \lim_{\Delta \downarrow 0} \frac{\Pr(t \le T < t+\Delta \mid T \ge t)}{\Delta} && \text{(definition of the hazard function)} \\
&= \lim_{\Delta \downarrow 0} \frac{1}{\Delta} \cdot \frac{\Pr(\{t \le T < t+\Delta\} \cap \{T \ge t\})}{\Pr(T \ge t)} && \text{(definition of conditional probability)} \\
&= \lim_{\Delta \downarrow 0} \frac{1}{\Delta} \cdot \frac{\Pr(t \le T < t+\Delta)}{\Pr(T \ge t)} && \text{(subset property)} \\
&= \frac{1}{\Pr(T \ge t)} \cdot \lim_{\Delta \downarrow 0} \frac{\Pr(t \le T < t+\Delta)}{\Delta} && (\Pr(T \ge t) \text{ does not depend on } \Delta) \\
&= \frac{f(t)}{\Pr(T \ge t)} && \text{(density as a limit; } f \text{ is continuous at } t) \\
&= \frac{f(t)}{\mathbb{S}(t)} && (\Pr(T = t) = 0 \text{ for continuous } T)
\end{aligned}
$$

□

Example 0.4 (Hazard function of an exponential distribution)**.** Continuing Example 0.1, for $t > 0$, where f is continuous:

$$
\begin{aligned}
\lambda(t) &= \frac{f(t)}{\mathbb{S}(t)} && \text{(hazard equals density over survival)} \\
&= \frac{\lambda \mathrm{e}^{-\lambda t}}{\mathrm{e}^{-\lambda t}} && \text{(substitute the exponential density and survival function)} \\
&= \lambda && \text{(cancel } \mathrm{e}^{-\lambda t}\text{)}
\end{aligned}
$$

At $t = 0$, where f jumps from 0 to λ, Definition 0.2 gives the same value directly: $\Pr(0 \le T < \Delta \mid T \ge 0)/\Delta = \left(1 - \mathrm{e}^{-\lambda\Delta}\right)/\Delta \to \lambda$ as $\Delta \downarrow 0$. So the exponential distribution has a constant hazard for $t \ge 0$, equal to its rate parameter; for $t < 0$, $f(t) = 0$, so $\lambda(t) = 0$.

Definition 0.4 (Cumulative hazard function)**.** The **cumulative hazard function** of a continuous random variable T, often denoted $\Lambda(t)$ or $\mathbb{H}(t)$, is the integral of its hazard function up to t:

$$
\Lambda(t) \stackrel{\text{def}}{=} \int_{u=-\infty}^{t} \lambda(u)\, du
$$

Example 0.5 (Cumulative hazard function of an exponential distribution)**.** Continuing Example 0.4, the hazard is $\lambda(u) = \lambda$ for $u \ge 0$ and 0 for $u < 0$, so for $t \ge 0$:

$$\begin{aligned}
\Lambda(t) &= \int_{u=-\infty}^{0} 0\,du + \int_{u=0}^{t} \lambda\,du && \text{(split the integral at 0)} \\
&= 0 + \lambda t && \text{(integrate each piece)} \\
&= \lambda t && \text{(simplify)}
\end{aligned}$$

and $\Lambda(t) = 0$ for $t < 0$.

Corollary 0.2 (Cumulative hazard of a non-negative random variable)**.** *If T is a continuous random variable with $T \geq 0$, such as a time to event, then $\lambda(u) = 0$ for every $u < 0$, and for every $t \geq 0$:*

$$\Lambda(t) = \int_{u=0}^{t} \lambda(u)\,du$$

Proof

Proof. Let $u < 0$. Every outcome has $T(\omega) \geq 0 > u$, so $\{T \geq u\} = \Omega$, which has probability $1 > 0$, and $\lambda(u)$ is defined. For $0 < \Delta \leq -u$, the event $\{u \leq T < u + \Delta\}$ is empty, because $u + \Delta \leq 0$, so:

$$\begin{aligned}
\lambda(u) &= \lim_{\Delta \downarrow 0} \frac{\Pr(u \leq T < u + \Delta \mid T \geq u)}{\Delta} && \text{(definition of the hazard function)} \\
&= \lim_{\Delta \downarrow 0} \frac{\Pr(\{u \leq T < u + \Delta\} \cap \{T \geq u\})}{\Delta \cdot \Pr(T \geq u)} && \text{(definition of conditional probability)} \\
&= \lim_{\Delta \downarrow 0} \frac{\Pr(\emptyset \cap \Omega)}{\Delta \cdot \Pr(\Omega)} && \text{(the two events found above)} \\
&= \lim_{\Delta \downarrow 0} \frac{0}{\Delta \cdot 1} && (\emptyset \cap \Omega = \emptyset;\ \Pr(\emptyset) = 0;\ \Pr(\Omega) = 1) \\
&= 0 && \text{(simplify)}
\end{aligned}$$

Then, for $t \geq 0$:

$$
\begin{aligned}
\Lambda(t) &= \int_{u=-\infty}^{t} \lambda(u)\,du && \text{(definition of the cumulative hazard)} \\
&= \int_{u=-\infty}^{0} \lambda(u)\,du + \int_{u=0}^{t} \lambda(u)\,du && \text{(split the integral at 0)} \\
&= \int_{u=-\infty}^{0} 0\,du + \int_{u=0}^{t} \lambda(u)\,du && (\lambda(u) = 0 \text{ for } u < 0) \\
&= \int_{u=0}^{t} \lambda(u)\,du && \text{(integrate)}
\end{aligned}
$$

□

Remark

Remark. The form in Corollary 0.2, with lower limit 0, is the form most survival-analysis texts use.

Corollary 0.3 (Survival function from the cumulative hazard)**.** *If T is a continuous random variable whose density f is continuous at all but finitely many points, then for every t with $\mathbb{S}(t) > 0$:*

$$\Lambda(t) = -\log\{\mathbb{S}(t)\}$$

and therefore:

$$\mathbb{S}(t) = \exp\{-\Lambda(t)\} \tag{1}$$

Proof

Proof. Since $\mathbb{S}(t) = 1 - F(t)$ (Theorem 0.1) and F has derivative f wherever f is continuous (Theorem 0.8), $\mathbb{S}$ has derivative $-f$ at those points. At each such u with $\mathbb{S}(u) > 0$:

$$
\begin{aligned}
\frac{d}{du}(-\log\{\mathbb{S}(u)\}) &= -\frac{1}{\mathbb{S}(u)} \cdot \frac{d}{du}\mathbb{S}(u) && \text{(chain rule)} \\
&= -\frac{1}{\mathbb{S}(u)} \cdot (-f(u)) && (\mathbb{S}' = -f) \\
&= \frac{f(u)}{\mathbb{S}(u)} && \text{(simplify)} \\
&= \lambda(u) && \text{(hazard equals density over survival)}
\end{aligned}
$$

The function $-\log\{\mathbb{S}(u)\}$ is continuous, because $\mathbb{S}$ is continuous for a continuous T, and it has derivative $\lambda(u)$ at all but finitely many points, so it is an antiderivative of λ for

the fundamental theorem of calculus. As $u \to -\infty$, $\mathbb{S}(u) \to 1$, so $-\log\{\mathbb{S}(u)\} \to 0$. Therefore:

$$\begin{aligned}
\Lambda(t) &\stackrel{\text{def}}{=} \int_{u=-\infty}^{t} \lambda(u)\, du && \text{(definition of the cumulative hazard)} \\
&= [-\log\{\mathbb{S}(u)\}]_{u=-\infty}^{t} && (-\log \mathbb{S} \text{ is an antiderivative of } \lambda) \\
&= -\log\{\mathbb{S}(t)\} - 0 && \text{(evaluate at the bounds)} \\
&= -\log\{\mathbb{S}(t)\} && \text{(simplify)}
\end{aligned}$$

Exponentiating both sides of $-\Lambda(t) = \log\{\mathbb{S}(t)\}$ gives Equation 1. □

Example 0.6 (Recovering the exponential survival function from its cumulative hazard)**.** Continuing Example 0.5, for $t \geq 0$:

$$\begin{aligned}
\mathbb{S}(t) &= \exp\{-\Lambda(t)\} && \text{(survival function from the cumulative hazard)} \\
&= \exp\{-\lambda t\} && \text{(substitute } \Lambda(t) = \lambda t)
\end{aligned}$$

which matches the survival function computed directly in Example 0.1.

Table 2: Probability distribution functions of a continuous random variable T

Name	Symbols	Definition
Probability density function (PDF)	$f(t), \mathbb{p}(t)$	$f \geq 0$ with $\int_a^b f(u)\, du = \Pr(a \leq T \leq b)$
Cumulative distribution function (CDF)	$F(t)$	$\Pr(T \leq t)$
Survival function	$\mathbb{S}(t), \bar{F}(t)$	$\Pr(T > t)$
Hazard function	$\lambda(t), \mathbb{h}(t)$	$\lim_{\Delta \downarrow 0} \Pr(t \leq T < t + \Delta \mid T \geq t)/\Delta$
Cumulative hazard function	$\Lambda(t), \mathbb{H}(t)$	$\int_{u=-\infty}^{t} \lambda(u)\, du$
Log-hazard function	$\eta(t)$	$\log\{\lambda(t)\}$

Remark

Remark. For a continuous random variable $T \geq 0$ whose density is continuous at all but finitely many points, the following results connect the functions in Table 2:

- Theorem 0.1
- Theorem 0.3
- Corollary 0.2
- Corollary 0.3

Each arrow in the following diagram converts one function into the next:

$$f(t) \xleftarrow[\mathbb{S}(t)\lambda(t)]{-\mathbb{S}'(t)} \mathbb{S}(t) \xleftarrow{\exp\{-\Lambda(t)\}} \Lambda(t) \xleftarrow{\int_{u=0}^{t} \lambda(u)\,du} \lambda(t) \xleftarrow{\exp\{\eta(t)\}} \eta(t)$$

$$f(t) \xrightarrow[\int_{u=t}^{\infty} f(u)\,du]{f(t)/\lambda(t)} \mathbb{S}(t) \xrightarrow[-\log\{\mathbb{S}(t)\}]{} \Lambda(t) \xrightarrow[\Lambda'(t)]{} \lambda(t) \xrightarrow[\log\{\lambda(t)\}]{} \eta(t)$$

Further reading

Another explanation

Evans and Rosenthal's free textbook [25] covers random variables in Chapter 2.

References

[1] Wikipedia contributors, "Random variable — Wikipedia, the free encyclopedia." 2026. Available: https://en.wikipedia.org/wiki/Random_variable

[2] Wikipedia contributors, "Probability mass function — Wikipedia, the free encyclopedia." 2026. Available: https://en.wikipedia.org/wiki/Probability_mass_function

[3] Wikipedia contributors, "Bernoulli distribution — Wikipedia, the free encyclopedia." 2026. Available: https://en.wikipedia.org/wiki/Bernoulli_distribution

[4] Wikipedia contributors, "Probability density function — Wikipedia, the free encyclopedia." 2026. Available: https://en.wikipedia.org/wiki/Probability_density_function

[5] Wikipedia contributors, "Continuous uniform distribution — Wikipedia, the free encyclopedia." 2026. Available: https://en.wikipedia.org/wiki/Continuous_uniform_distribution

[6] P. Billingsley, *Probability and measure*, 3rd ed. in Wiley series in probability and mathematical statistics. New York: Wiley, 1995.

[7] Wikipedia contributors, "Normal distribution — Wikipedia, the free encyclopedia." 2026. Available: https://en.wikipedia.org/wiki/Normal_distribution

[8] G. Casella and R. Berger, *Statistical inference*, 2nd ed. Cengage Learning, 2002. Available: https://www.cengage.com/c/statistical-inference-2e-casella-berger/9780534243128/

[9] Wikipedia contributors, "Exponential distribution — Wikipedia, the free encyclopedia." 2026. Available: https://en.wikipedia.org/wiki/Exponential_distribution

[10] Wikipedia contributors, "Cumulative distribution function — Wikipedia, the free encyclopedia." 2026. Available: https://en.wikipedia.org/wiki/Cumulative_distribution_function

[11] Wikipedia contributors, "Quantile function — Wikipedia, the free encyclopedia." 2026. Available: https://en.wikipedia.org/wiki/Quantile_function

[12] K. J. Rothman, T. L. Lash, T. J. VanderWeele, and S. Haneuse, *Modern epidemiology*, Fourth edition. Philadelphia: Wolters Kluwer, 2021.

[13] Wikipedia contributors, "Joint probability distribution — Wikipedia, the free encyclopedia." 2026. Available: https://en.wikipedia.org/wiki/Joint_probability_distribution

[14] Wikipedia contributors, "Marginal distribution — Wikipedia, the free encyclopedia." 2026. Available: https://en.wikipedia.org/wiki/Marginal_distribution

[15] B. Hutchinson, "DATA 471/571: Machine learning." Western Washington University, 2024.

[16] B. Hutchinson, "DATA 471/571 (machine learning) and CSCI 481/581 (deep learning) video lectures." Western Washington University. Accessed: Sep. 28, 2026. [Online]. Available: https://facultyweb.cs.wwu.edu/~hutchib2/video_lectures/data371/

[17] 5MinuteAI, "Probability & statistics." Video playlist on YouTube. Accessed: Oct. 12, 2026. [Online]. Available: https://www.youtube.com/playlist?list=PLeB45KifGiuHesi4PALNZSYZFhViVGQJK

[18] Wikipedia contributors, "Survival function — Wikipedia, the free encyclopedia." 2026. Available: https://en.wikipedia.org/wiki/Survival_function

[19] Wikipedia contributors, "Failure rate — Wikipedia, the free encyclopedia." 2026. Available: https://en.wikipedia.org/wiki/Failure_rate

[20] A. J. Dobson and A. G. Barnett, *An introduction to generalized linear models*, 4th ed. CRC press, 2018. doi: 10.1201/9781315182780.

[21] E. Vittinghoff, D. V. Glidden, S. C. Shiboski, and C. E. McCulloch, *Regression methods in biostatistics: Linear, logistic, survival, and repeated measures models*, 2nd ed. Springer, 2012. doi: 10.1007/978-1-4614-1353-0.

[22] J. P. Klein and M. L. Moeschberger, *Survival analysis: Techniques for censored and truncated data*, 2nd ed. Springer, 2003. Available: https://link.springer.com/book/10.1007/b97377

[23] D. G. Kleinbaum and M. Klein, *Survival analysis: A self-learning text*, 3rd ed. Springer, 2012. Available: https://link.springer.com/book/10.1007/978-1-4419-6646-9

[24] J. D. Kalbfleisch and R. L. Prentice, *The statistical analysis of failure time data*. John Wiley & Sons, 2011.

[25] M. J. Evans and J. S. Rosenthal, *Probability and statistics: The science of uncertainty*, 2nd ed. Freeman, 2010. Available: https://utstat.utoronto.ca/mikevans/jeffrosenthal/