Probability Density Functions

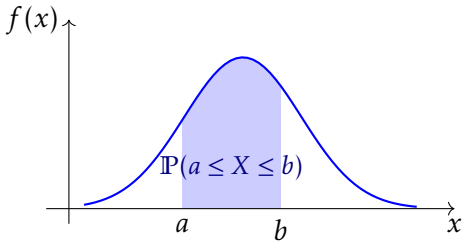
Last year we met random variables which take values in a *continuous* range – heights, masses, waiting times – rather than a discrete set of values. For such a variable there is no sensible way to list $\mathbb{P}(X = x)$ for every x: instead we describe how the probability is *spread out* along the number line.

Definition. A function f is a probability density function (pdf) for the continuous random variable X if

1. $f(x) \geq 0$ for all $x \in \mathbb{R}$,
2. $\displaystyle\int_{-\infty}^{\infty} f(x)\,\mathrm{d}x = 1$.

Probabilities are then given by *areas* under the graph of f:

$$\mathbb{P}(a \leq X \leq b) = \int_a^b f(x)\,\mathrm{d}x$$



Remark. For a continuous random variable, $\mathbb{P}(X = x) = 0$ for any individual value x (the area of a line segment of zero width is zero). Consequently

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

so we need not fuss over strict versus non-strict inequalities. This is emphatically *not* true for discrete variables!

Example

The continuous random variable X has pdf

$$f(x) = \begin{cases} kx(2-x) & 0 \leq x \leq 2 \\ 0 & \text{otherwise} \end{cases}$$

(a) Find the value of k.

(b) Find $\mathbb{P}(X > 1.5)$.

(a) The total area must be 1:

$$\int_0^2 kx(2-x)\,\mathrm{d}x = k\left[x^2 - \frac{x^3}{3}\right]_0^2 = k\left(4 - \frac{8}{3}\right) = \frac{4k}{3} = 1 \implies k = \frac{3}{4}$$

(b)

$$\mathbb{P}(X > 1.5) = \int_{1.5}^2 \frac{3}{4}x(2-x)\,\mathrm{d}x = \frac{3}{4}\left[x^2 - \frac{x^3}{3}\right]_{1.5}^2 = \frac{3}{4}\left(\frac{4}{3} - \frac{9}{8}\right) = \frac{5}{32}$$

Example (Piecewise pdf)

The continuous random variable X has pdf

$$f(x) = \begin{cases} \dfrac{x}{4} & 0 \le x \le 2 \\ \dfrac{4-x}{4} & 2 < x \le 4 \\ 0 & \text{otherwise} \end{cases}$$

(a) Verify that f is a valid pdf, and sketch it.

(b) Find $\mathbb{P}(1 \le X \le 3)$.

(a) Both pieces are non-negative on their domains. The graph is a triangle with base 4 and height $f(2) = \frac{1}{2}$, so the total area is $\frac{1}{2} \times 4 \times \frac{1}{2} = 1$. (Or integrate: $\int_0^2 \frac{x}{4}\,\mathrm{d}x + \int_2^4 \frac{4-x}{4}\,\mathrm{d}x = \frac{1}{2} + \frac{1}{2} = 1$.)

(b) The integral must be split at $x = 2$:

$$\mathbb{P}(1 \le X \le 3) = \int_1^2 \frac{x}{4}\,\mathrm{d}x + \int_2^3 \frac{4-x}{4}\,\mathrm{d}x = \left[\frac{x^2}{8}\right]_1^2 + \left[x - \frac{x^2}{8}\right]_2^3 = \frac{3}{8} + \frac{3}{8} = \frac{3}{4}$$

Mean, Variance and Percentiles

The formulae mirror the discrete case, with sums replaced by integrals.

Fact — For a continuous random variable X with pdf f:

$$\mu = \mathbb{E}[X] = \int_{-\infty}^{\infty} x f(x)\,dx$$

$$\sigma^2 = \mathrm{Var}[X] = \mathbb{E}[(X-\mu)^2] = \int_{-\infty}^{\infty} x^2 f(x)\,dx - \mu^2$$

$$\mathbb{E}[g(X)] = \int_{-\infty}^{\infty} g(x) f(x)\,dx \quad \text{for a function } g$$

In practice we only integrate over the interval where f is non-zero.

Definition. The median m of X is the value such that $\mathbb{P}(X \le m) = \frac{1}{2}$, i.e.

$$\int_{-\infty}^{m} f(x)\,dx = \frac{1}{2}$$

The lower quartile Q_1 and upper quartile Q_3 satisfy $\mathbb{P}(X \le Q_1) = \frac{1}{4}$ and $\mathbb{P}(X \le Q_3) = \frac{3}{4}$; the nth percentile satisfies $\mathbb{P}(X \le x) = \frac{n}{100}$.

Example

The continuous random variable X has pdf

$$f(x) = \begin{cases} \dfrac{3x^2}{8} & 0 \le x \le 2 \\ 0 & \text{otherwise} \end{cases}$$

Find:

(a) $\mathbb{E}[X]$ and $\mathrm{Var}[X]$,

(b) $\mathbb{E}\left[\dfrac{1}{X}\right]$,

(c) the median of X.

(a)

$$\mathbb{E}[X] = \int_0^2 x \cdot \frac{3x^2}{8}\,dx = \frac{3}{8}\left[\frac{x^4}{4}\right]_0^2 = \frac{3}{2}$$

$$\mathbb{E}[X^2] = \int_0^2 x^2 \cdot \frac{3x^2}{8}\,dx = \frac{3}{8}\left[\frac{x^5}{5}\right]_0^2 = \frac{12}{5}$$

$$Var[X] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 = \frac{12}{5} - \frac{9}{4} = \frac{3}{20}$$

(b)

$$\mathbb{E}\left[\frac{1}{X}\right] = \int_0^2 \frac{1}{x} \cdot \frac{3x^2}{8}\,dx = \int_0^2 \frac{3x}{8}\,dx = \frac{3}{4}$$

Note that $\mathbb{E}[1/X] \neq 1/\mathbb{E}[X]$ *– expectation only passes through* linear *functions.*

(c) We need $\int_0^m \frac{3x^2}{8}\,\mathrm{d}x = \frac{1}{2}$, *i.e.* $\frac{m^3}{8} = \frac{1}{2}$, *so* $m = \sqrt[3]{4} \approx 1.59$.

Tip

Symmetry saves work: if the pdf is symmetric about $x = c$ (and the mean exists), then $\mathbb{E}[X] = c$ and the median is also c.

Example (OCR S2, June 2014)

A continuous random variable X has probability density function

$$f(x) = \begin{cases} \frac{1}{2}\pi \sin(\pi x) & 0 \le x \le 1 \\ 0 & \text{otherwise} \end{cases}$$

(i) Show that this is a valid probability density function.

(ii) Sketch the curve $y = f(x)$ and write down the value of $\mathbb{E}[X]$.

(iii) Find the value q such that $\mathbb{P}(X > q) = 0.75$.

(iv) Write down an expression, including an integral, for $\mathrm{Var}[X]$. (Do not attempt to evaluate the integral.)

(v) A student states that "X is more likely to occur when x is close to $\mathbb{E}[X]$." Give an improved version of this statement.

(i) $\sin(\pi x) \ge 0$ *for* $0 \le x \le 1$, *so* $f \ge 0$ *everywhere; and*

$$\int_0^1 \frac{\pi}{2}\sin(\pi x)\,\mathrm{d}x = \left[-\frac{1}{2}\cos(\pi x)\right]_0^1 = \frac{1}{2} - \left(-\frac{1}{2}\right) = 1.$$

(ii) One arch of a sine curve: from $(0,0)$ *up to a maximum of* $\frac{\pi}{2}$ *at* $x = \frac{1}{2}$, *back down to* $(1,0)$, *and zero outside* $[0,1]$. *The pdf is symmetric about* $x = \frac{1}{2}$, *so* $\mathbb{E}[X] = \frac{1}{2}$.

(iii) $\mathbb{P}(X > q) = 0.75$ *means* $\int_0^q \frac{\pi}{2}\sin(\pi x)\,\mathrm{d}x = 0.25$, *i.e.* $\frac{1}{2}(1 - \cos(\pi q)) = \frac{1}{4}$. *So* $\cos(\pi q) = \frac{1}{2}$, *giving* $\pi q = \frac{\pi}{3}$, *i.e.* $q = \frac{1}{3}$.

(iv) $\mathrm{Var}[X] = \int_0^1 \frac{\pi}{2}x^2 \sin(\pi x)\,\mathrm{d}x - \left(\frac{1}{2}\right)^2$.

(v) $\mathbb{P}(X = x) = 0$ *for every individual value, so "more likely to occur" is meaningless as stated. Better: values of* X *in an interval close to* $\mathbb{E}[X]$ *are more likely than values in an interval of the same width further away.*

Textbook Exercises: [CUP.S] Ch 7 §1–3

The Cumulative Distribution Function

Definition. The cumulative distribution function (CDF) of a random variable X is

$$F(x) = \mathbb{P}(X \le x) = \int_{-\infty}^{x} f(t)\,\mathrm{d}t$$

Note the dummy variable: we are integrating *up to* x, so x is a limit of the integral and cannot also be the variable of integration. Inside the integral we use a fresh letter, t. Writing $\int_{-\infty}^{x} f(x)\,\mathrm{d}x$ is meaningless and loses marks.

> **Tip**
> If dummy variables feel awkward, you can instead find a general antiderivative $\int f(x)\,\mathrm{d}x$, then choose the constant of integration so that F equals 0 at the bottom of the domain. Both routes give the same answer; the dummy-variable notation is the one the examiners use.

Since $f \ge 0$, F is non-decreasing, with $F(x) \to 0$ as $x \to -\infty$ and $F(x) \to 1$ as $x \to \infty$. **Always write a CDF in full piecewise form, starting at 0 and finishing at 1:**

$$F(x) = \begin{cases} 0 & x < \text{lower end} \\ \cdots & \text{middle} \\ 1 & x > \text{upper end} \end{cases}$$

Fact (Relationship between pdf and CDF) — By the Fundamental Theorem of Calculus,

$$f(x) = F'(x)$$

wherever F is differentiable. So: *integrate* the pdf to get the CDF; *differentiate* the CDF to get the pdf. Also

$$\mathbb{P}(a \le X \le b) = F(b) - F(a),$$

and the median satisfies $F(m) = \frac{1}{2}$, the quartiles $F(Q_1) = \frac{1}{4}$, $F(Q_3) = \frac{3}{4}$, etc.

Example
Find the CDF of the triangular distribution from earlier:

$$f(x) = \begin{cases} \dfrac{x}{4} & 0 \le x \le 2 \\ \dfrac{4-x}{4} & 2 < x \le 4 \\ 0 & \text{otherwise} \end{cases}$$

For $0 \le x \le 2$:

$$F(x) = \int_0^x \frac{t}{4}\,\mathrm{d}t = \frac{x^2}{8}$$

For $2 < x \le 4$ we must carry forward the probability accumulated so far, $F(2) = \frac{1}{2}$:

$$F(x) = F(2) + \int_2^x \frac{4-t}{4}\,\mathrm{d}t = \frac{1}{2} + \left[t - \frac{t^2}{8}\right]_2^x = x - \frac{x^2}{8} - 1$$

Therefore, in full:

$$F(x) = \begin{cases} 0 & x < 0 \\ \dfrac{x^2}{8} & 0 \le x \le 2 \\ x - \dfrac{x^2}{8} - 1 & 2 < x \le 4 \\ 1 & x > 4 \end{cases}$$

Sanity checks: $F(4) = 4 - 2 - 1 = 1$ ✓, *and the two middle pieces agree at* $x = 2$ ✓.

Tip

The most common error with piecewise CDFs is forgetting the accumulated probability from earlier pieces. Always check the pieces *join up continuously* and that the last piece reaches exactly 1.

Example

The continuous random variable X has CDF

$$F(x) = \begin{cases} 0 & x < 1 \\ \dfrac{(x-1)^2}{9} & 1 \le x \le 4 \\ 1 & x > 4 \end{cases}$$

(a) Find $\mathbb{P}(2 \le X \le 3)$.

(b) Find the pdf of X.

(c) Find the median and the upper quartile of X.

(a) $\mathbb{P}(2 \le X \le 3) = F(3) - F(2) = \dfrac{4}{9} - \dfrac{1}{9} = \dfrac{3}{9} = \dfrac{1}{3}$

(b) Differentiating:

$$f(x) = \begin{cases} \dfrac{2(x-1)}{9} & 1 \le x \le 4 \\ 0 & \textit{otherwise} \end{cases}$$

(c) Median: $F(m) = \frac{1}{2} \implies (m-1)^2 = \frac{9}{2} \implies m = 1 + \frac{3}{\sqrt{2}} \approx 3.12$ *(taking the root in* $[1, 4]$*).*

Upper quartile: $(Q_3 - 1)^2 = \frac{27}{4} \implies Q_3 = 1 + \frac{3\sqrt{3}}{2} \approx 3.60$.

Example (OCR S3, June 2015)

A continuous random variable X has probability density function

$$f(x) = \begin{cases} kx & 0 \le x < 2 \\ \dfrac{k(4-x)^2}{2} & 2 \le x \le 4 \\ 0 & \text{otherwise} \end{cases}$$

where k is a constant.

(i) Show that $k = \frac{3}{10}$.

(ii) Find $\mathbb{E}[X]$.

(iii) Find the cumulative distribution function of X.

(iv) Find the upper quartile of X, correct to 3 significant figures.

(i) The total area is

$$\int_0^2 kx\,\mathrm{d}x + \int_2^4 \frac{k(4-x)^2}{2}\,\mathrm{d}x = 2k + \frac{k}{2}\left[-\frac{(4-x)^3}{3}\right]_2^4 = 2k + \frac{4k}{3} = \frac{10k}{3} = 1,$$

so $k = \frac{3}{10}$.

(ii)

$$\mathbb{E}[X] = \int_0^2 \frac{3x^2}{10}\,\mathrm{d}x + \int_2^4 \frac{3x(4-x)^2}{20}\,\mathrm{d}x = \frac{4}{5} + \frac{3}{20}\int_2^4 \left(16x - 8x^2 + x^3\right)\mathrm{d}x$$
$$= \frac{4}{5} + \frac{3}{20}\left[8x^2 - \frac{8x^3}{3} + \frac{x^4}{4}\right]_2^4 = \frac{4}{5} + \frac{3}{20}\cdot\frac{20}{3} = \frac{4}{5} + 1 = \frac{9}{5}$$

(iii) For $0 \le x < 2$: $F(x) = \dfrac{3x^2}{20}$, *so* $F(2) = \frac{3}{5}$. *For* $2 \le x \le 4$, *carrying this forward:*

$$F(x) = \frac{3}{5} + \int_2^x \frac{3(4-t)^2}{20}\,\mathrm{d}t = \frac{3}{5} + \left[-\frac{(4-t)^3}{20}\right]_2^x = 1 - \frac{(4-x)^3}{20}$$

In full:

$$F(x) = \begin{cases} 0 & x < 0 \\ \dfrac{3x^2}{20} & 0 \le x < 2 \\ 1 - \dfrac{(4-x)^3}{20} & 2 \le x \le 4 \\ 1 & x > 4 \end{cases}$$

(iv) Since $F(2) = \frac{3}{5} < \frac{3}{4}$, *the upper quartile lies in the second piece:*

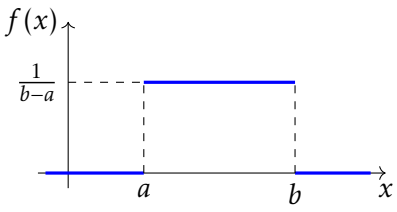
$$1 - \frac{(4-x)^3}{20} = \frac{3}{4} \implies (4-x)^3 = 5 \implies Q_3 = 4 - \sqrt[3]{5} = 2.29 \text{ (3 s.f.)}$$

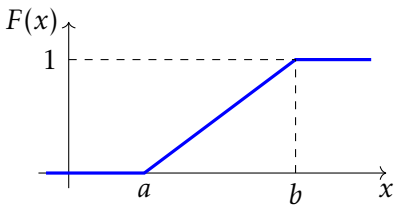
Textbook Exercises: [CUP.S] Ch 7 §4 [S2] Ch 1 [S3&4] S3 Ch 1

The Continuous Uniform Distribution

Definition. The random variable X has the continuous uniform distribution (or *rectangular distribution*) on $[a,b]$, written $X \sim U[a,b]$, if

$$f(x) = \begin{cases} \dfrac{1}{b-a} & a \le x \le b \\ 0 & \text{otherwise} \end{cases}$$





Fact — The CDF, written in full, is

$$F(x) = \begin{cases} 0 & x < a \\ \dfrac{x-a}{b-a} & a \le x \le b \\ 1 & x > b \end{cases}$$

Theorem

If $X \sim U[a,b]$ then

$$\mathbb{E}[X] = \frac{a+b}{2} \qquad \text{and} \qquad \operatorname{Var}[X] = \frac{(b-a)^2}{12}$$

The mean comes free from symmetry; the variance is one careful integral followed by some tidy algebra.

Mean. *The pdf is symmetric about the midpoint of $[a,b]$, so $\mathbb{E}[X] = \dfrac{a+b}{2}$ immediately.*
Variance. *First*

$$\mathbb{E}[X^2] = \int_a^b \frac{x^2}{b-a}\,\mathrm{d}x = \frac{1}{b-a}\cdot\frac{b^3-a^3}{3} = \frac{a^2+ab+b^2}{3},$$

using the difference of cubes $b^3 - a^3 = (b-a)(b^2+ab+a^2)$. Then

$$\begin{aligned} Var[X] &= \frac{a^2+ab+b^2}{3} - \left(\frac{a+b}{2}\right)^2 \\ &= \frac{4(a^2+ab+b^2) - 3(a^2+2ab+b^2)}{12} \\ &= \frac{a^2-2ab+b^2}{12} = \frac{(b-a)^2}{12} \end{aligned}$$

Example

A number is recorded to the nearest integer, so that the rounding error E is modelled as $E \sim U[-0.5, 0.5]$. Find the mean and standard deviation of E, and the probability that $|E| > 0.3$.

By the formulae, $\mathbb{E}[E] = 0$ *and* $Var[E] = \frac{1^2}{12} = \frac{1}{12}$*, so* $\sigma = \frac{1}{\sqrt{12}} \approx 0.289$*.*

$$\mathbb{P}(|E| > 0.3) = \mathbb{P}(E < -0.3) + \mathbb{P}(E > 0.3) = 0.2 + 0.2 = 0.4$$

(each tail is an interval of width 0.2 *with density* 1*).*

Example (OCR S2, January 2010 (parts))
The continuous random variable T is equally likely to take any value from 5.0 to 11.0 inclusive.

(i) Sketch the graph of the probability density function of T.

(ii) Write down the value of $\mathbb{E}[T]$ and find by integration the value of $\mathrm{Var}[T]$.

(i) $T \sim U[5, 11]$*: a horizontal line at height* $\frac{1}{6}$ *from* $t = 5$ *to* $t = 11$*, and zero on either side.*

(ii) By symmetry, $\mathbb{E}[T] = 8$*. Then*

$$\mathbb{E}[T^2] = \int_5^{11} \frac{t^2}{6}\,\mathrm{d}t = \left[\frac{t^3}{18}\right]_5^{11} = \frac{1331 - 125}{18} = 67,$$

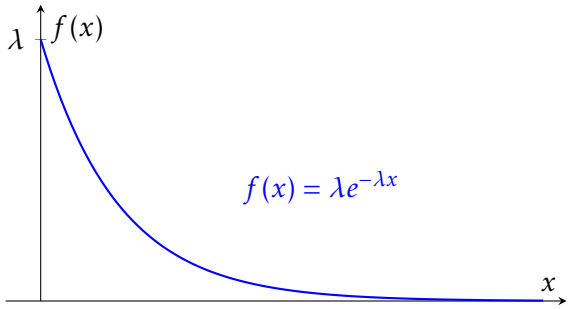
so $Var[T] = 67 - 8^2 = 3$*. (Check against the formula:* $\frac{(11-5)^2}{12} = 3$*.)* ✓

Textbook Exercises: [S2] Ch 1

The Exponential Distribution

Definition. The random variable X has the **exponential distribution** with rate $\lambda > 0$, written $X \sim \text{Exp}(\lambda)$, if

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \\ 0 & x < 0 \end{cases}$$



Example

Verify that f is a valid pdf, and show that the CDF is $F(x) = 1 - e^{-\lambda x}$ for $x \geq 0$.

Clearly $f \geq 0$. For $x \geq 0$,

$$F(x) = \int_0^x \lambda e^{-\lambda t}\,\mathrm{d}t = \left[-e^{-\lambda t}\right]_0^x = 1 - e^{-\lambda x}$$

As $x \to \infty$, $F(x) \to 1$, so the total area is indeed 1. In full:

$$F(x) = \begin{cases} 0 & x < 0 \\ 1 - e^{-\lambda x} & x \geq 0 \end{cases}$$

Theorem

If $X \sim \text{Exp}(\lambda)$ then

$$\mathbb{E}[X] = \frac{1}{\lambda} \qquad \text{and} \qquad \text{Var}[X] = \frac{1}{\lambda^2}$$

The derivation is integration by parts, used twice – and the second integral cleverly recycles the first.

$$\mathbb{E}[X] = \int_0^\infty x\lambda e^{-\lambda x}\,\mathrm{d}x = \left[-xe^{-\lambda x}\right]_0^\infty + \int_0^\infty e^{-\lambda x}\,\mathrm{d}x = 0 + \left[-\frac{1}{\lambda}e^{-\lambda x}\right]_0^\infty = \frac{1}{\lambda}$$

(the first bracket vanishes because $xe^{-\lambda x} \to 0$ as $x \to \infty$). Similarly, by parts with $u = x^2$:

$$\mathbb{E}[X^2] = \int_0^\infty x^2\lambda e^{-\lambda x}\,\mathrm{d}x = \left[-x^2e^{-\lambda x}\right]_0^\infty + \int_0^\infty 2xe^{-\lambda x}\,\mathrm{d}x = 0 + \frac{2}{\lambda}\int_0^\infty x\lambda e^{-\lambda x}\,\mathrm{d}x$$

Recognising the remaining integral as $\mathbb{E}[X]$, we get $\mathbb{E}[X^2] = \frac{2}{\lambda}\mathbb{E}[X] = \frac{2}{\lambda^2}$, so

$$Var[X] = \frac{2}{\lambda^2} - \frac{1}{\lambda^2} = \frac{1}{\lambda^2}$$

The link with the Poisson distribution

Fact — If occurrences follow a Poisson process with rate λ, the waiting time between successive occurrences has the Exp(λ) distribution. *Poisson counts occurrences; exponential measures gaps.*

The derivation is a single beautiful observation: the waiting time exceeds t exactly when *nothing happens* in $[0, t]$.

Suppose events occur singly, independently and at constant average rate λ per unit time, so the number *of occurrences in a time interval of length t is $N_t \sim Po(\lambda t)$. Let T be the waiting time until the first occurrence. Then*

$$\mathbb{P}(T > t) = \mathbb{P}(\textit{no occurrences in } [0,t]) = \mathbb{P}(N_t = 0) = e^{-\lambda t},$$

so

$$F_T(t) = \mathbb{P}(T \le t) = 1 - e^{-\lambda t},$$

which is exactly the CDF of Exp(λ).

Example

Calls arrive at a helpline at an average rate of 3 per minute, modelled by a Poisson process. Find the probability that the wait for the next call exceeds 30 seconds, and the median waiting time.

The waiting time (in minutes) is $T \sim Exp(3)$.

$$\mathbb{P}(T > 0.5) = e^{-3\times 0.5} = e^{-1.5} \approx 0.223$$

Median: $1 - e^{-3m} = \frac{1}{2} \implies e^{-3m} = \frac{1}{2} \implies m = \frac{\ln 2}{3} \approx 0.231$ *minutes, i.e. about* 13.9 *seconds. Note the median is less than the mean $\frac{1}{3}$ minute: the distribution has a long right tail.*

Textbook Exercises: [S3&4] S3 Ch 1

CDFs of Related Variables

Often we know the distribution of X but want the distribution of some function of it, $Y = g(X)$. This is widely regarded as one of the **hardest topics in FM Statistics**, but the method is mechanical if you follow it carefully.

Tip (The CDF method)

To find the distribution of $Y = g(X)$:

1. Write down $F_Y(y) = \mathbb{P}(Y \le y)$ and substitute: $\mathbb{P}(g(X) \le y)$.
2. Rearrange the inequality $g(X) \le y$ to isolate X – *this is the step that needs care*.
3. Express the result using the CDF of X: typically $F_Y(y) = F_X(g^{-1}(y))$.
4. State the new domain (apply g to the endpoints of the domain of X).
5. Differentiate to obtain the pdf: $f_Y(y) = F_Y'(y)$.

Never just substitute into the pdf – the pdf is a density, not a probability, and does not transform that way.

Example (Volume of a cube)

The edge length X cm of a cube is uniformly distributed on $[9, 11]$. Find the CDF and pdf of the volume $Y = X^3$, and find $\mathbb{P}(Y > 1200)$.

First, the CDF of X: for $9 \le x \le 11$, $F_X(x) = \dfrac{x-9}{2}$.
Since $y = x^3$ is increasing, $X^3 \le y \iff X \le y^{1/3}$. So for $9^3 = 729 \le y \le 11^3 = 1331$:

$$F_Y(y) = \mathbb{P}(X^3 \le y) = \mathbb{P}(X \le y^{1/3}) = F_X(y^{1/3}) = \frac{y^{1/3} - 9}{2}$$

In full:

$$F_Y(y) = \begin{cases} 0 & y < 729 \\ \dfrac{y^{1/3} - 9}{2} & 729 \le y \le 1331 \\ 1 & y > 1331 \end{cases}$$

Differentiating:

$$f_Y(y) = \begin{cases} \dfrac{1}{6} y^{-2/3} & 729 \le y \le 1331 \\ 0 & \textit{otherwise} \end{cases}$$

Note Y is not *uniform – the density is larger for smaller volumes. Finally*

$$\mathbb{P}(Y > 1200) = 1 - F_Y(1200) = 1 - \frac{1200^{1/3} - 9}{2} \approx 1 - 0.8157 = 0.184$$

Example

X has pdf $f_X(x) = 2x$ for $0 \le x \le 1$. Find the distribution of $Y = X^2$. Comment on your answer.

For $0 \le x \le 1$, $F_X(x) = x^2$. On this domain $x \mapsto x^2$ is increasing, so for $0 \le y \le 1$:

$$F_Y(y) = \mathbb{P}(X^2 \le y) = \mathbb{P}(X \le \sqrt{y}) = F_X(\sqrt{y}) = (\sqrt{y})^2 = y$$

So

$$F_Y(y) = \begin{cases} 0 & y < 0 \\ y & 0 \le y \le 1 \\ 1 & y > 1 \end{cases} \qquad f_Y(y) = 1 \text{ on } [0,1]$$

i.e. $Y \sim U[0,1]$: squaring this particular X produces a perfectly uniform variable.

When the function is not one-to-one

Does $X^2 \le y$ always mean $X \le \sqrt{y}$? **No!** If X can be negative, then $X^2 \le y$ means $-\sqrt{y} \le X \le \sqrt{y}$. Whenever g is not one-to-one on the domain of X, you must think about the inequality as a *region*, not just invert the function.

Example
$X \sim U[-1,1]$ and $Y = X^2$. Find the CDF and pdf of Y.

$F_X(x) = \dfrac{x+1}{2}$ for $-1 \le x \le 1$. For $0 \le y \le 1$:

$$\begin{aligned} F_Y(y) = \mathbb{P}(X^2 \le y) &= \mathbb{P}(-\sqrt{y} \le X \le \sqrt{y}) \\ &= F_X(\sqrt{y}) - F_X(-\sqrt{y}) \\ &= \frac{\sqrt{y}+1}{2} - \frac{-\sqrt{y}+1}{2} = \sqrt{y} \end{aligned}$$

So

$$F_Y(y) = \begin{cases} 0 & y < 0 \\ \sqrt{y} & 0 \le y \le 1 \\ 1 & y > 1 \end{cases} \qquad f_Y(y) = \frac{1}{2\sqrt{y}} \text{ for } 0 < y \le 1$$

Interesting features: the density is unbounded near $y = 0$ (yet the area is still 1), and had we carelessly written $\mathbb{P}(X \le \sqrt{y})$ we would have obtained $\frac{\sqrt{y}+1}{2}$, which is not even 0 at $y = 0$ – always sanity-check the endpoints.

Remark. A decreasing function also needs care, because it *reverses* inequalities: if $Y = e^{-X}$ with $X \ge 0$, then

$$\mathbb{P}(Y \le y) = \mathbb{P}(e^{-X} \le y) = \mathbb{P}(X \ge -\ln y) = 1 - F_X(-\ln y).$$

Example (Class practice)
$X \sim \text{Exp}(\lambda)$.

(a) Find the CDF and pdf of $Y = X^2$.

(b) Find the CDF and pdf of $Z = e^{-X}$. What distribution is this?

(a) Since $X \geq 0$, here $X^2 \leq y \iff X \leq \sqrt{y}$ is safe. For $y \geq 0$:

$$F_Y(y) = F_X(\sqrt{y}) = 1 - e^{-\lambda\sqrt{y}}, \qquad f_Y(y) = \frac{\lambda}{2\sqrt{y}} e^{-\lambda\sqrt{y}} \quad (y > 0)$$

(b) Z takes values in $(0,1]$. The function e^{-x} is decreasing, so for $0 < z \leq 1$:

$$F_Z(z) = \mathbb{P}(e^{-X} \leq z) = \mathbb{P}(X \geq -\ln z) = e^{-\lambda(-\ln z)} = e^{\lambda \ln z} = z^\lambda$$

Hence $f_Z(z) = \lambda z^{\lambda-1}$ on $(0,1]$. (For $\lambda = 1$, $Z \sim U[0,1]$.)

Example (OCR S3, June 2014)

A rectangle of area A m^2 has a perimeter of 20 m, and each of the two shorter sides has length X m, where X is uniformly distributed between 0 and 2.

(i) Write down an expression for A in terms of X, and hence show that $A = 25 - (X-5)^2$.

(ii) Write down the probability density function of X.

(iii) Show that the cumulative distribution function of A is

$$F_A(a) = \begin{cases} 0 & a < 0 \\ \frac{1}{2}\left(5 - \sqrt{25-a}\right) & 0 \leq a \leq 16 \\ 1 & a > 16 \end{cases}$$

(iv) Find the probability density function of A.

(i) The sides are X and $10 - X$, so $A = X(10-X)$; completing the square, $A = 25 - (X-5)^2$.

(ii) $f_X(x) = \frac{1}{2}$ for $0 \leq x \leq 2$ (and 0 otherwise); hence $F_X(x) = \frac{x}{2}$ on $[0,2]$.

(iii) As x increases from 0 to 2, $A = x(10-x)$ increases from 0 to 16, so A takes values in $[0,16]$. For $0 \leq a \leq 16$:

$$F_A(a) = \mathbb{P}\left(25 - (X-5)^2 \leq a\right) = \mathbb{P}\left((X-5)^2 \geq 25 - a\right) = \mathbb{P}\left(5 - X \geq \sqrt{25-a}\right),$$

since $X \leq 2$ forces $X - 5 < 0$ (the other root, $X \geq 5 + \sqrt{25-a}$, is impossible). Hence

$$F_A(a) = \mathbb{P}\left(X \leq 5 - \sqrt{25-a}\right) = F_X\left(5 - \sqrt{25-a}\right) = \frac{1}{2}\left(5 - \sqrt{25-a}\right).$$

Sanity checks: $F_A(0) = \frac{1}{2}(5-5) = 0$ ✓ and $F_A(16) = \frac{1}{2}(5-3) = 1$ ✓.

(iv) Differentiating,

$$f_A(a) = \begin{cases} \dfrac{1}{4\sqrt{25-a}} & 0 \leq a \leq 16 \\ 0 & \text{otherwise} \end{cases}$$

Textbook Exercises: [CUP.S] Ch 7 §8 [S3&4] S3 Ch 1