Probability Density Functions

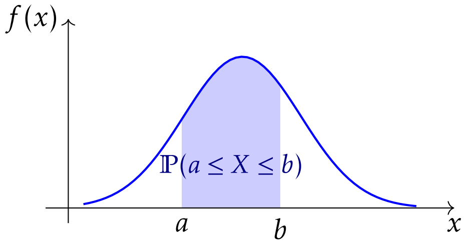
Last year we met random variables which take values in a *continuous* range – heights, masses, waiting times – rather than a discrete set of values. For such a variable there is no sensible way to list $\mathbb{P}(X = x)$ for every x: instead we describe how the probability is *spread out* along the 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 $\mathbb{P}(X > 1.5)$.

Example (Piecewise pdf)

The continuous random variable X has pdf

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

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

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

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 **$n$th 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) $\mathbb{E}\left[\frac{1}{X}\right]$,

(c) the median of X.

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 $\text{Var}[X]$. (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.

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:

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

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.

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.

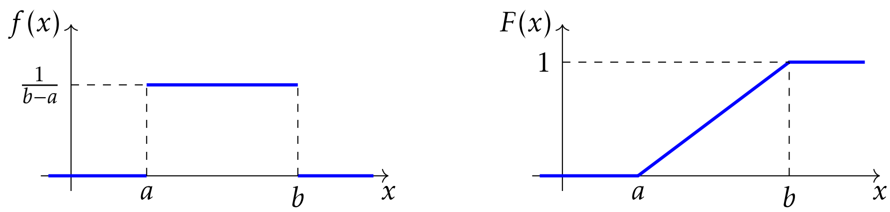
(iv) Find the upper quartile of X, correct to 3 significant figures.

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

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

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

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 $|E| > 0.3$.

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.

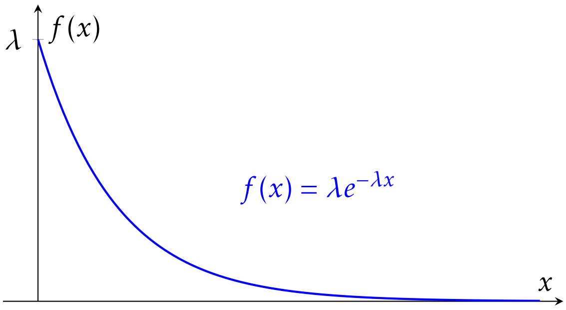
(ii) Write down the value of $\mathbb{E}[T]$ and find by integration the value of $\mathrm{Var}[T]$.

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 $x \geq 0$.

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.

The link with the Poisson distribution

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

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

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.

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 $F_Y(y) = \mathbb{P}(Y \leq y)$ and substitute: $\mathbb{P}(g(X) \leq y)$.
2. Rearrange the inequality $g(X) \leq y$ 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 $\mathbb{P}(Y > 1200)$.

Example

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

When the function is not one-to-one

Does $X^2 \leq y$ always mean $X \leq \sqrt{y}$? **No!** If X can be negative, then $X^2 \leq y$ means $-\sqrt{y} \leq X \leq \sqrt{y}$. 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.

> **Remark.** A decreasing function also needs care, because it *reverses* inequalities: if $Y = e^{-X}$ with $X \geq 0$, then
>
> $$\mathbb{P}(Y \leq y) = \mathbb{P}(e^{-X} \leq y) = \mathbb{P}(X \geq -\ln y) = 1 - F_X(-\ln y).$$

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?

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

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

(iv) Find the probability density function of A.

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