Engineering Mathematics 2B

Module 13: Introduction to Probability

Nick Polydorides

School of Engineering

THE UNIVERSITY *of* EDINBURGH

Module 13 contents

Motivation

Nearly all measurements and observations from engineering or physical systems can be considered to be random variables.

To make sense of these 'noisy' and 'approximate' measurements we need to understand the main properties of random variables.

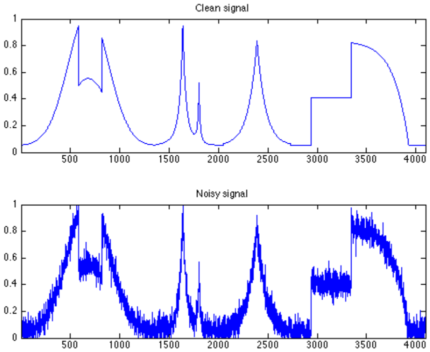
The RAC Foundation also received data from the bodies responsible for looking after Britain's strategic roads (motorways and some major A roads):

Authority	Number of bridges	Number of substandard bridges	Proportion of sub-standard bridges
Highways England	12,184	166	1%
Transport Scotland	2,440	33	1%
Welsh Assembly	1,263	48	4%

Source: www.racfoundation.org

Signal denoising - The making of an engineer

Information from digital systems (instruments) is captured with some form of noise. We measure the noisy (*data*) but we want the clean (*information*).



Introducing the random variable

A **random variable** $X : \Omega \to \mathbb{R}$ is a function that assigns a numerical value to each possible outcome of the trial/experiment. You may think of its outcomes $X = x$, as numerical random events.

Let x be a number, and X a random variable then

$$X = x, \quad \text{and} \quad X \leq x$$

are random events, to which we can assign probabilities.

$\mathbb{P}(X \leq x)$ is known as the cumulative density function of X.

$\mathbb{P}(X = x)$ is known as the probability mass function of X.

Introducing the random variable

Once a random variable X is introduced, the sample space (= the domain of the random variable) becomes irrelevant to everything we can possibly want to know about it.

Everything there's to know about it evolves around "the possible values of X and their corresponding probabilities".

It suffices to list the possible values of the variable and their corresponding probabilities. This information is contained within its cumulative probability density function.

A **function of a random variable** $f(X) := Y$ defines another random variable.

Random variables

A random variable X can be:

1. **discrete** if it takes values from a discrete countable set, or
2. **continuous** if it takes any real value on the x axis or an interval.

A **discrete** / **continuous** random variable X is completely characterised by its probability **mass** / **density** function $p_X(x)$, or its cumulative density function $F_X(x)$.

Knowing $p_X(x)$ we know the variable's mean and variance. Moreover and we can compute what X can be. We can also switch between $p_X(s)$ and $F_X(x)$.

Notation: Random variables are in capitals. Realisations of random variables are not random and are in small case.

Discrete random variables

Let X be a **discrete** random variable

The **Probability Mass Function** (PMF) $p_X(x)$ is a real discrete function that tells us 'what is the probability of a the event $X = x$ occurring'

$$p_X(x) = \mathbb{P}(X = x).$$

$\forall x \; p_X(x) \geq 0$ (non-negativity) and $\sum_i p_X(x_i) = 1$ (normalisation).

For example, in the fair dice roll $x_i \in \{1, 2, 3, 4, 5, 6\}$, and

dice outcome	probability
1	1/6
2	1/6
3	1/6
4	1/6
5	1/6
6	1/6

Table: For X discrete, PMF takes the form of a table.

Discrete random variables

Let X be a **discrete** random variable with **Probability Mass Function** PMF $p_X(x)$. PMF tells us 'what is the probability of a random variable taking a certain value'.

The **Cumulative Density Function** (CDF) $F_X(x)$ is a real piecewise constant function that tells us 'what is the probability that an outcome is no more than a value'

$$F_X(x) = \mathbb{P}(X \leq x) = \sum_{i \leq k} p_X(x_i),$$

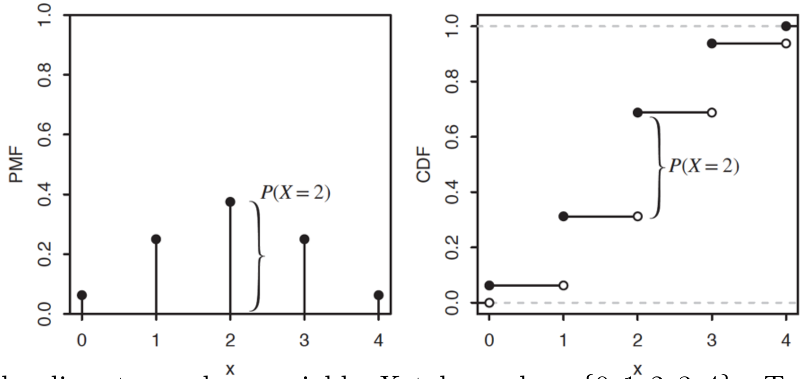
for an index k. (see example next)

PMF & CDF are related by

$$p_X(x_i) = F_X(x_i) - F_X(x_{i-1}), \quad i = 1, 2, \ldots$$

and so in principle we can calculate $F_X(x)$ directly from $p_X(x)$.

PMF & CDF Example



The discrete random variable X takes values $\{0, 1, 2, 3, 4\}$. To the left its PMF $p_X(x) = \{0.09, 0.22, 0.38, 0.22, 0.09\}$ and right its CDF $F_X(x)$.

Discrete random variables: Expectation & Variance

The **expectation (= mean)** of a discrete X represents the average of a large number of independent realisations of the random variable

$$\mathbb{E}[X] = \sum_i x_i \, p_X(x_i) = \mu_X$$

and similarly the expectation of a function $f(x)$ where x takes values from realisations of X is

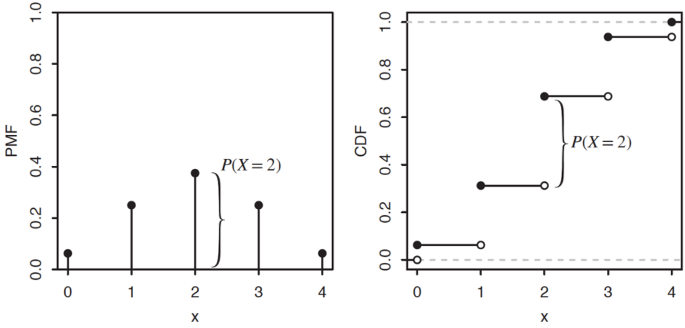
$$\mathbb{E}[f(x)] = \sum_i f(x_i) \, p_X(x_i).$$

The special case $f(x) = (x - \mu_X)^2$ defines the **variance** of X:

$$\mathrm{Var}(X) = \mathbb{E}[(X - \mu_X)^2] = \sigma_X^2.$$

Notation: The expectation and variance of X are conventionally denoted as μ_X and σ_X^2. σ_X is referred to as the **standard variation** of X.

Expectation & Variance Example



$$\mathbb{E}[X] = \sum_{i=0}^{4} x_i \, p_X(x_i) = 0 \cdot 0.09 + 1 \cdot 0.22 + 2 \cdot 0.38 + 3 \cdot 0.22 + 4 \cdot 0.09$$

yielding $\mu_X = 2$ and $\sigma_X^2 = 1.16$ as $\sigma_X^2 = \mathbb{E}[(x-2)^2]$ and

$$\sigma_X^2 = \sum_{i=0}^{4} (x_i - \mu_X)^2 p_X(x_i) = 4 \cdot 0.09 + 1 \cdot 0.22 + 0 \cdot 0.38 + 1 \cdot 0.22 + 4 \cdot 0.09.$$

Continuous random variables

Let X be a **continuous** random variable

Then X has infinite range of outcomes, each if which has zero probability!

The **Probability Density Function** (PDF)

$$\text{PDF} := \begin{cases} p_X(x) & a \leq x \leq b \\ 0 & \text{elsewhere} \end{cases}$$

is a real non-negative function, expressing the **relative likelihood** ($\neq$ **probability**) of an event $X = x$.

The limits satisfy $-\infty \leq a < b \leq +\infty$

Continuous random variables

Probability in continuous random variables is always defined in terms of integrals over a range of outcomes, thus

$$\mathbb{P}(X = x) = 0$$

If X is a **continuous** random variable with **Probability Density Function** PDF $p_X(x)$, the right question to ask is

"What is the probability that X is less or more than a value x?"

The density function satisfies the probability axioms,

- $\forall x \; p_X(x) \geq 0$ (non-negativity) and,
- $\int_{-\infty}^{+\infty} p_X(x) \mathrm{d}x = 1$ (normalisation).

Continuous random variables

The **Cumulative Density Function** (CFD) $F_X(x)$ is a real, continuous, non-decreasing function

$$F_X(x) = \mathbb{P}(X \leq x) = \int_{-\infty}^{x} p_X(t) \mathrm{d}t.$$

PDF & CDF are related by

$$p_X(x) = \frac{\mathrm{d}}{\mathrm{d}x} F_X(x).$$

Probability as an integral

The probability of a continuous random variable is always an **integral of the PDF over an interval** of outcomes.

Aside asking $F_X(x) = \mathbb{P}(X \leq x)$: 'what is the probability that X is **less** than x?' we can also ask $\mathbb{P}(X \geq x)$: 'what is the probability that X is **more** than x?'

$$\mathbb{P}(X \geq x) = 1 - \mathbb{P}(X \leq x)$$

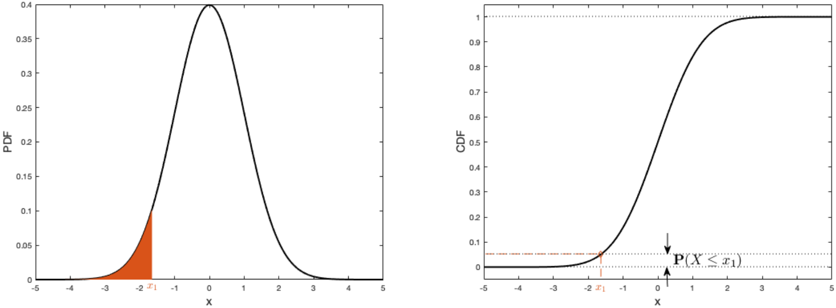
The probability of X taking values in the closed interval $x_1 \leq x \leq x_2$ is the **area** under the PDF above this interval

$$\mathbb{P}(x_1 \leq X \leq x_2) = \int_{x_1}^{x_2} p_X(x)\,\mathrm{d}x$$

If the interval shrinks to a **single point** x then the area vanishes and so does the probability

$$\mathbb{P}(X = x) = 0.$$

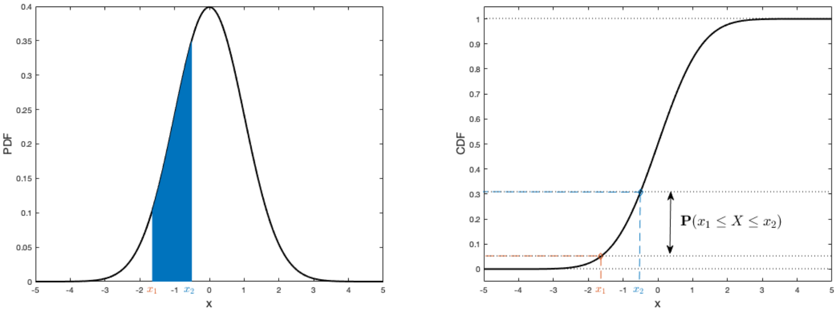
Probability of continuous random variables



For a continuous X with $p_X(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$ we can compute probabilities as

$$\mathbb{P}(X \leq x_1) = \int_{-\infty}^{x_1} p_X(x) \mathrm{d}x = F_X(x_1).$$

Probability of continuous random variables



For a continuous X with $p_X(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$ we can compute probabilities as

$$\mathbb{P}(x_1 \leq X \leq x_2) = \int_{x_1}^{x_2} p_X(x) \mathrm{d}x = F_X(x_2) - F_X(x_1).$$

Expectation

The expectation of a function $f(X)$ where X is a continuous variable with PDF $p_X(x)$ is given by

$$\mathbb{E}[f(X)] = \int_{-\infty}^{+\infty} f(x)\, p_X(x) \mathrm{d}x.$$

Setting $f(x) = x$ defines the **expectation** or average value of X as

$$\mathbb{E}[X] = \int_{-\infty}^{+\infty} x\, p_X(x) \mathrm{d}x = \mu_X$$

These default integral limits assume that $p_X(x)$ spans the whole x axis.

Variance

The **variance** of X, continuous rv with PDF $p_X(x)$ is given

$$\mathrm{Var}(X) = \int_{-\infty}^{+\infty} (x - \mu_X)^2 p_X(x) \mathrm{d}x = \sigma_X^2,$$

and this is simply the expectation integral for $f(x) = (x - \mu_X)^2$

$$\mathrm{Var}(X) = \mathbb{E}[(X - \mu_X)^2].$$

There is also a useful relation between variance and expectation, applicable to **both** discrete and continuous variables

$$\mathrm{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2.$$

Formulas

For X discrete:

- $\mathbb{P}(X = x_i) = p_X(x_i)$
- $p_X(x_i) = F_X(x_i) - F_X(x_{i-1})$
- $\mathbb{E}[X] = \sum_i x_i p_X(x_i)$
- $\mathrm{Var}(X) = \sum_i (x_i - \mathbb{E}[X])^2 p_X(x_i)$

For X continuous:

- $\mathbb{P}(X \leq x) = \int_{-\infty}^{x} p_X(x)\mathrm{d}x$
- $p_X(x) = \frac{\mathrm{d}}{\mathrm{d}x} F_X(x)$
- $\mathbb{E}[X] = \int_{-\infty}^{\infty} f(x) p_X(x)\mathrm{d}x$
- $\mathrm{Var}(X) = \int_{-\infty}^{\infty} (x - \mathbb{E}[X])^2 p_X(x)\mathrm{d}x$

For all X, $\mathrm{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2$

Main outcomes of module 13

You MUST know:

1. How to compute probabilities for continuous and discrete random random variables, given their PDF and PMF.
2. The meaning and formulas for the expectation and variance of a random variable.
3. The relation between PDF and CDF.

Good to know:

The R language. `https://www.r-project.org` Many employers looking to hire data scientists with engineering background will favour students with R or Python skills.