Lecture 1: Random variables, probability distributions and sampling

EMAT30007 Applied Statistics

Nikolai Bode & Filippo Simini

Department of Engineering Mathematics

Housekeeping arrangements

- **Lectures**
  - theory and examples
  - with Lecturers Filippo Simini in weeks 13-17 and Nikolai Bode in weeks 19-23
- **Labs**
  - hands-on exercises and applications using Matlab
  - with TAs Thomas, David and Elisa.
- **Assessments**
  - First coursework (**25%**)
    - Set in week 16 – due on 10 March (week 19)
  - Second coursework (**25%**)
    - Set in week 21 – due on 30 April (week 23)
  - Summer exam (**50%**)
    - Past papers on blackboard

Outline of the lecture

In this lecture you will learn:

- How to use Random Variables to model empirical data.
- How to plot empirical probability density (PDF) and cumulative distribution (CDF) functions.
- How to compute probability distributions of functions of Random Variables.
- How to generate synthetic data sampling from a Random Variable.

What is statistics?

Statistics aims to describe and interpret empirical data using the mathematical objects of probability theory.

A statistical model is a mathematical representation of a real world process that has some degree randomness.

Example: *Dice roll*

We roll a 6-sided dice and write down each outcome in a list:

$$\text{dice rolls} = \{⚂, ⚃, ⚄, ⚁, ⚅\} = [3, 4, 5, 2, 6]$$

An empirical data set is a collection of outcomes of a specific real-world process or experiment (a dice roll, in this case).
We imagine that each outcome has a given *probability* to be observed.

In statistics, we interpret an empirical data set as a collection of outcomes generated by a Random Variable.

Random Variables

A Random Variable (RV) is a function that maps any possible outcome of an experiment (sample space) to a number.
Example: if D is a RV for a dice roll, then $D(⚀) = 1, \ldots, D(⚅) = 6$.

A random variable has an associated probability distribution that provides the probability of occurrence of all possible outcomes.
Example: if D is a fair dice, then $P_D(1) = \cdots = P_D(6) = 1/6$.

Example: *Toss of a fair coin*

Possible outcomes (sample space): $\Omega = \{\text{'head'}, \text{'tail'}\}$.

Random Variable: $X(\omega) = \begin{cases} 1 & \text{if } \omega = \text{'head'} \\ 0 & \text{if } \omega = \text{'tail'} \end{cases}$

Probability distribution: $P_X(x) = p^x(1-p)^{1-x} = \begin{cases} p & \text{if } x = 1 \\ 1-p & \text{if } x = 0 \end{cases}$

where $p = 1/2$ is the probability of 'head' ('head' and 'tail' are equally likely).

Discrete Random Variables

A discrete RV maps outcomes to discrete (natural) numbers.

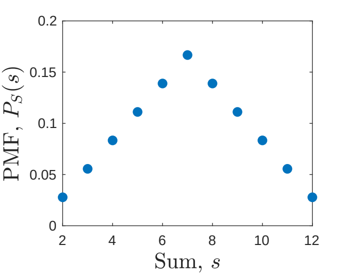
Example: *Sum of two dices*

Sample space (possible outcomes):

{(⚀, ⚀), (⚀, ⚁), . . . , (⚅, ⚄), (⚅, ⚅)}

$$\Omega = \{(d_1, d_2) \,|\, d_i \in [1, 2, 3, 4, 5, 6], i = [1, 2]\}$$

RV ‘Sum of two dices’: $S(d_1, d_2) = d_1 + d_2$



Probability Mass Function (PMF):

s	2	3	4	5	6	7	8	9	10	11	12
$P_S(s)$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

Continuous Random Variables

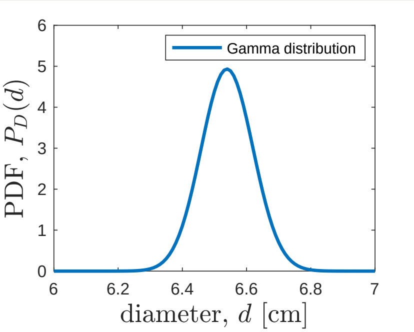
A continuous RV maps outcomes to continuous (real) numbers.

Example: *Diameter of a tennis ball*

Sample space (possible outcomes): any real number larger than zero. $\Omega = [0, +\infty] = \mathbb{R}^+$

RV: D is the measured diameter in centimetres.

Probability Density Function (PDF): $D \sim$ Gamma$(\alpha = 6540, \theta = 0.001)$



Gamma distribution:

$$P_D(d;\alpha,\theta) = \frac{1}{\Gamma(\alpha)\theta^\alpha} d^{\alpha-1} e^{-d/\theta}$$

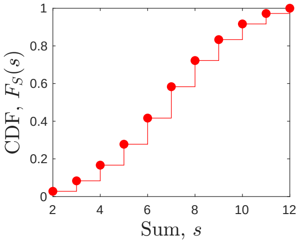
Cumulative Distribution Function (CDF)

The Cumulative Distribution Function (CDF) of a random variable X, $F_X(x)$, is the probability to get an outcome with value less or equal to x:

$$F_X(x) = Prob(X \leq x) = \sum_{y \leq x} P_X(y) \quad \text{for discrete RVs} \quad (1)$$

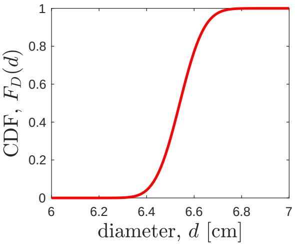
$$= \int_{-\infty}^{x} P_X(y)dy \quad \text{for continuous RVs} \quad (2)$$

Example: *Sum of two dices*



s	2	3	4	5	6	7	8	9	10	11	12
$F_S(s)$	$\frac{1}{36}$	$\frac{3}{36}$	$\frac{6}{36}$	$\frac{10}{36}$	$\frac{15}{36}$	$\frac{21}{36}$	$\frac{26}{36}$	$\frac{30}{36}$	$\frac{33}{36}$	$\frac{35}{36}$	$\frac{36}{36}$

Example: *Diameter of a tennis ball*



$$F_D(d; \alpha, \theta) = \int_0^d \frac{1}{\Gamma(\alpha)\theta^\alpha} y^{\alpha-1} e^{-y/\theta} dy$$

Properties of the CDF and PDF

- the CDF is a non decreasing function.
- the maximum value is 1 and the minimum is 0.
- $Prob(X > x) = 1 - F_X(x)$ is the Complementary CDF (CCDF).
- **From CDF to PDF**

  If the CDF of a continuous RV is differentiable, then the PDF is given by

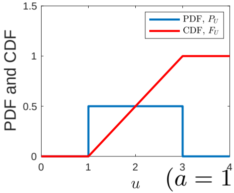
$$P_X(x) = \frac{dF_X(x)}{dx} = \lim_{dx \to 0} \frac{F_X(x + dx) - F_X(x)}{dx} \qquad (3)$$

- The PDF and PMF are normalised: $\int P_X(x)dx = 1$ and $\sum_x P_X(x) = 1$.

Example: *Uniform distribution*

The continuous RV U defined over the interval $u \in [a, b]$ has CDF $F_U(u) = (u - a)/(b - a)$ and its PDF is the Uniform distribution:

$P_U(u) = F'_U(u) = 1/(b - a)$.



$(a = 1, b = 3)$

How to compute the empirical CDF from data

Plot of the empirical CDF

1. Load data in a one-dim array, d.
2. Sort data in ascending order, x.
3. Create an array, y, of increasing integers from 1 to the length of x.
4. Divide each element of y by the length of x, so that the max of y is 1.
5. If there are (x_i, y_i) pairs with identical x-value, keep only the pair with the largest y-value. $y = F_X(x)$.

1. $d = [2, \quad 7, \quad 2, \quad 1, \quad 5]$
2. $x = [1, \quad 2, \quad 2, \quad 5, \quad 7]$
3. $$\begin{aligned} x &= [1, \quad 2, \quad 2, \quad 5, \quad 7] \\ y &= [1, \quad 2, \quad 3, \quad 4, \quad 5] \end{aligned}$$
4. $$\begin{aligned} x &= [1, \quad 2, \quad 2, \quad 5, \quad 7] \\ y &= [0.2, \quad 0.4, \quad 0.6, \quad 0.8, \quad 1] \end{aligned}$$
5. $$\begin{aligned} x &= [1, \quad ., \quad 2, \quad 5, \quad 7] \\ y &= [0.2, \quad ., \quad 0.6, \quad 0.8, \quad 1] \end{aligned}$$

In MATLAB, use `stairs(x, y)` or `cdfplot(d)` to plot the empirical CDF.

How to compute the empirical PDF from data

Plot of the empirical PDF, using the definition in Eq. (3)

1. Divide the range of data d, $[\min(d), \max(d)]$, into bins, with bin edges b.
2. Compute the mid-point of each bin, $x_i = (b_{i+1} + b_i)/2$.
3. Count the number of data elements that fall in each bin, $c_i = |\{d_j : b_i \leq d_j < b_{i+1}\}|$.
4. Normalise. Rescale each element of c: $y_i = c_i/(|d|(b_{i+1} - b_i))$, so that the PDF integral is 1. $y = P_X(x)$.

1. $$d = [2, 7, 2, 1, 5]$$ $$b = [1, 3, 5, 7, 9]$$
2. $$x = [2, 4, 6, 8]$$
3. $$x = [2, \quad 4, \quad 6, \quad 8]$$ $$c = [3, \quad 0, \quad 1, \quad 1]$$
4. $$x = [2, \quad 4, \quad 6, \quad 8]$$ $$y = [0.3, \quad 0, \quad 0.1, \quad 0.1]$$

In MATLAB, use `histcounts` or `histogram(d, 'normalization', 'pdf')` to plot the empirical PDF of data in array `d` (use option `'BinEdges'` to specify bin edges).

Common Random Variables

- Uniform (continuous outcomes, all equally likely)
- Bernoulli (coin toss)
- Binomial (# of ‘heads’ in n tosses)
- Multinomial (# of ‘1’s, ‘2’s, ..., ‘6’s in n dice rolls)
- Poisson (# of independent arrivals/events)
- Exponential (time between consecutive independent arrivals/events)
- Normal (sum of many independent RVs)
- Lognormal (product of many independent RVs)
- Gamma (waiting times)
- Beta (distribution of probabilities)
- Pareto (distribution of wealth, city populations)
- . . . https://en.wikipedia.org/wiki/List_of_probability_distributions

Expectation of RVs

The expectation or expected value, $E(X)$, of Random Variable X with PDF $P_X(x)$ is defined as

$$E(X) \equiv \int P_X(x)\, x\, dx \qquad \text{or} \qquad E(X) \equiv \sum_x P_X(x) x \tag{4}$$

for continuous and discrete RVs. $E(X)$ is also called first moment or mean of X.

Properties of the expectation:

- Linearity of expectation: for any two RVs X and Y and any constant a $E(X+Y) = E(X) + E(Y)$ and $E(aX) = aE(X)$.
- Expectation of a product of **independent** RVs: if X and Y are independent RVs then $E(X \cdot Y) = E(X) \cdot E(Y)$. *This is not true in general.*
- Variance of a RV: $E[(X - E(X))^2] = E[X^2] - (E[X])^2$.
- Expectation of a function of a RV: $E(g(X)) = \int P_X(x)\, g(x)\, dx$.

Function of a Random Variable

Let X be a continuous RV with CDF $F_X(x)$ and PDF $P_X(x)$. If g is a strictly increasing or decreasing function with inverse $g^{-1} = h$, then the RV $Y = g(X)$ has the following CDF and PDF (if F_X and h differentiable):

$$F_Y(y) = \begin{cases} F_X(h(y)) & \text{if } h \text{ and } g \text{ increasing} \\ 1 - F_X(h(y)) & \text{if } h \text{ and } g \text{ decreasing} \end{cases} \tag{5}$$

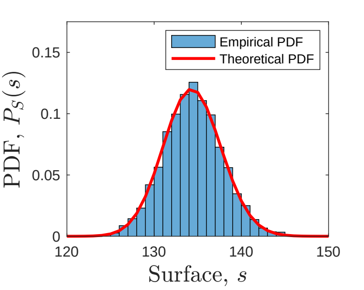
$$P_Y(y) = P_X(h(y)) \cdot \left| \frac{dh(y)}{dy} \right| \tag{6}$$

The domain of the new RV is transformed too: $Y \in [g(x_{min}), g(x_{max})]$ if g increasing or $Y \in [g(x_{max}), g(x_{min})]$ if g decreasing.

Example: *Surface area of a tennis ball*

Recall the RV D for the diameter of a tennis ball, which has a Gamma distribution: $P_D(d;\alpha,\theta) = \frac{1}{\Gamma(\alpha)\theta^\alpha} d^{\alpha-1} e^{-d/\theta}$. We can define the RV $S = \pi D^2$ for the surface of a tennis ball. What is the PDF of S?

We can apply the function $s = \pi d^2$ to our diameter data, obtaining the corresponding surface areas, and compute the empirical PDF (blue histogram).
We can also apply Eq. (6) and obtain $P_S(s) = P_D(\sqrt{s/\pi})|\frac{1}{2\pi}\sqrt{\pi/s}|$ (red curve).



Sampling from a Random Variable

Sampling from a RV X with PDF P_X means to generate numbers such that the probability to generate number x is $P_X(x)$.

Being able to sample from a RV allows us to simulate real-world processes.

Computers have (pseudo) random numbers generators that produce numbers that are uniformly distributed in $[0, 1)$.
How can we use a computer's uniform random number generator to sample from a RV with a given distribution?

Example: *Simulate a fair coin.*

A fair coin can be modelled as a RV X with PDF $P_X(0) = 1/2$ and $P_X(1) = 1/2$, so we want to generate 0 ('tail') with probability $1/2$ and 1 ('head') with probability $1/2$.

We can sample a random number u from $U(0, 1)$, the computer's uniform random number generator: u will be smaller than 0.5 with probability $1/2$ and larger than 0.5 with probability $1/2$. To simulate the RV X we can transform u into x such that:
$x = 0$ if $0 \leq u < 0.5$ and $x = 1$ if $0.5 \leq u < 1$.

Inverse Probability Integral Transform

The Inverse Probability Integral Transform (IPIT) is a method to transform uniformly distributed random numbers into numbers with a different distribution.

The IPIT is a sampling method based on Eq. (5) for the function of a RV: to sample numbers from RV X we only need to know the inverse of its CDF, F_X^{-1}:

Sampling using the Inverse Probability Integral Transform (IPIT)

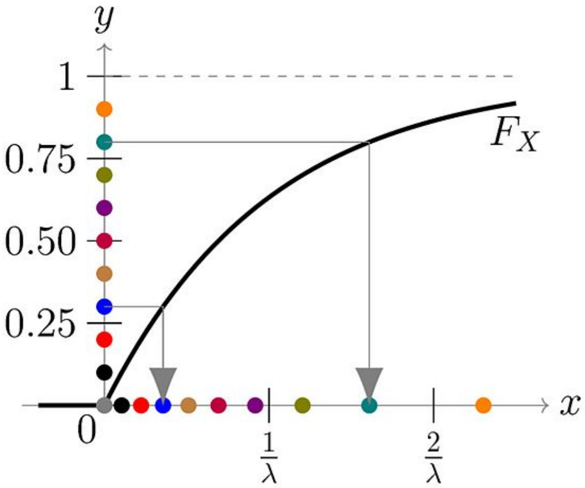
1. Generate a number u from the uniform distribution $U(0, 1)$.
2. Compute $x = F_X^{-1}(u)$

The CDF of X will be F_X because if $X = F_X^{-1}(U)$ then from Eq. (5)
$F_X(x) = F_U((F_X^{-1})^{-1}(x)) = F_U(F_X(x)) = F_X(x)$.
The inverse of the CDF, F_X^{-1}, is called the quantile function.

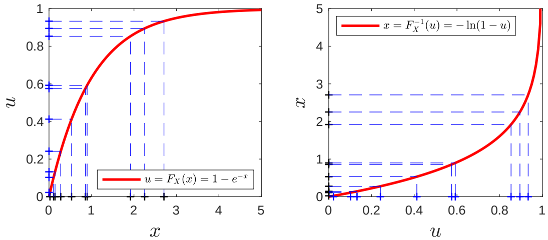
Example: *Exponential distribution*

To sample from $P_X(x) = ae^{-ax}$ with CDF $F_X(x) = 1 - e^{-ax}$, compute the inverse of $u = F_X(x) = 1 - e^{-ax}$, obtaining $x = F_X^{-1}(u) = -\ln(1 - u)/a$.

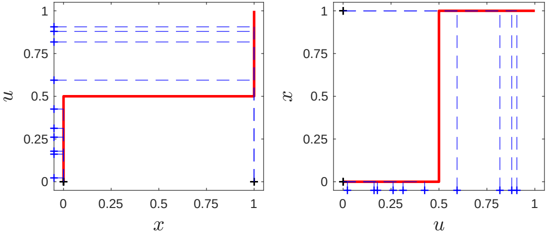
Visual intuition for the IPIT



https://en.wikipedia.org/wiki/Inverse_transform_sampling



IPIT for the Exponential distribution, $X \sim Exp(1)$.



IPIT for the fair coin example, $X \sim Bernoulli(0.5)$.