Modelling of Complex Systems

Randomness

Outline:

- Randomness and random numbers
- Probabilities and probability distributions
- Sampling random numbers

Randomness is the (actual or apparent) lack of predictability in events, sequences of symbols, etc. A random sequence has no order, and does not follow an understandable pattern.

Examples

Coin flipping: heads or tails?

Defining heads=1 and tails=0, we get sequences like 0010111010….

After N trials heads appears N_h times and tails appears N_t times, with the constraint $N_h + N_t = N$.

We can estimate the probabilities of finding heads and tails as:

$$P_h = \frac{N_h}{N} \quad \text{and} \quad P_t = \frac{N_t}{N}$$

The **normalization condition** imposes that

$$P_h + P_t = \frac{N_h}{N} + \frac{N_t}{N} = \frac{N_h + N_t}{N} = 1$$

It is important to understand that, rigorously speaking, the probabilities can be defined in this way only in the limit of $N \to \infty$, i.e.:

$$P_h = \lim_{N\to\infty} \frac{N_h}{N} \quad \text{and} \quad P_t = \lim_{N\to\infty} \frac{N_t}{N} \ .$$

For example, flipping the coin 2 times we can easily get 11, which would lead us to $P_h = 1$ and $P_t = 0$. But for a perfect coin we must have $P_h = P_t = 0.5$. Clearly, 2 flips are not enough to get accurate estimates for the probabilities, we must increase the number N.

Examples

Rolling dices:

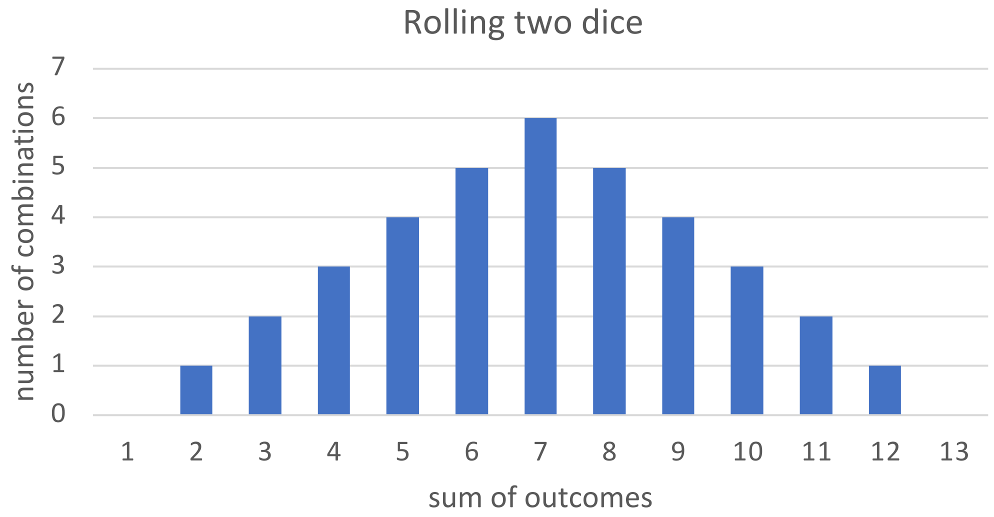
If the dice is fair, each number 1 to 6 has the same probability of 1/6.

We can effectively determine if a dice is fair by rolling it many times and analysing the sequence of events.

Examples

If we roll two dice, the sum of the outcomes is a number from 2 to 12.

In this case, the probability of each number is no longer uniform.



Examples

Buffon's Needle problem (1733):

What is the probability that a needle thrown to the floor will land across a line between two boards?

$$P = \frac{2}{\pi}\frac{l}{d}$$

This result can be used to approximate the number π in a Monte Carlo experiment.

Monte Carlo methods use randomness to solve numerical problems. Their precision increases with the number of trials.

Law of large numbers

The law of large numbers states that as the number of trial of a random process increases a sample average approaches its expected value.

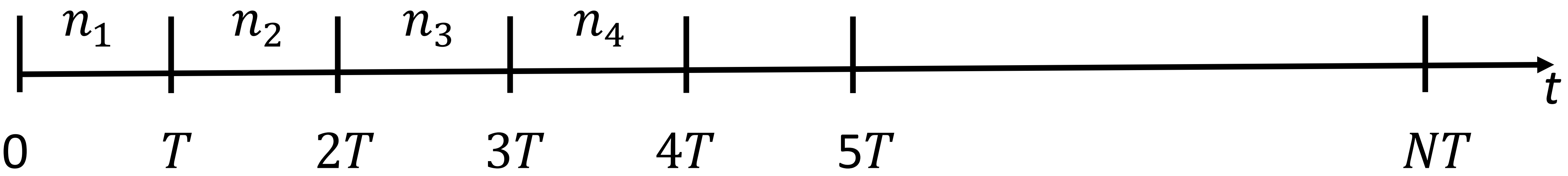
Consider a random variable X with k possible values, say $X_1, \ldots, X_k$, and corresponding probabilities $P_1, \ldots, P_k$. Then the expected value, or **expectation**, of X is $\langle X \rangle = \sum_{i=1}^{k} X_i P_i$.

Take a sample of N trials of the variable, denoted $x_1, \ldots, x_N$ (to avoid confusion). The **sample average** is $\bar{X} = \frac{1}{N} \sum_{j=1}^{N} x_j$.

The law of large numbers ensures that $\lim\limits_{N \to \infty} \bar{V} = \langle V \rangle$.

Another example

Let us count the number of cars that pass on a street under our window during an interval T



where N is the number of intervals. We get a sequence of random integer numbers: $n_1, n_2\,, n_3, n_4, \dots n_N$.

Every random number n_i takes a value 0,1, 2, 3, ….

Let us analyze this sequence:
We count the number of intervals where $n_i = n$.
We define this number as $N(n)$.

Another example

- Clearly $\sum_{n=0}^{\infty} N(n) = N$.
- **The probability to observe $\boldsymbol{n}$ cars** is $P(n) = \lim_{N\to\infty} \frac{N(n)}{N}$.
- Normalization condition: $\sum_{n=0}^{\infty} P(n) = \sum_{n=0}^{\infty} \frac{N(n)}{N} = 1.$
- **The mean value** of random numbers: $\langle n \rangle = \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} n_i$.
- We can rewrite this equation in another form,

$$\langle n \rangle = \frac{1}{N}\sum_{i=1}^{N} n_i = \frac{1}{N}\sum_{n=0}^{\infty} N(n)n = \sum_{n=0}^{\infty} \frac{N(n)}{N} n = \sum_{n=0}^{\infty} P(n)n$$

$$\langle n \rangle = \sum_{n=0}^{\infty} nP(n)$$

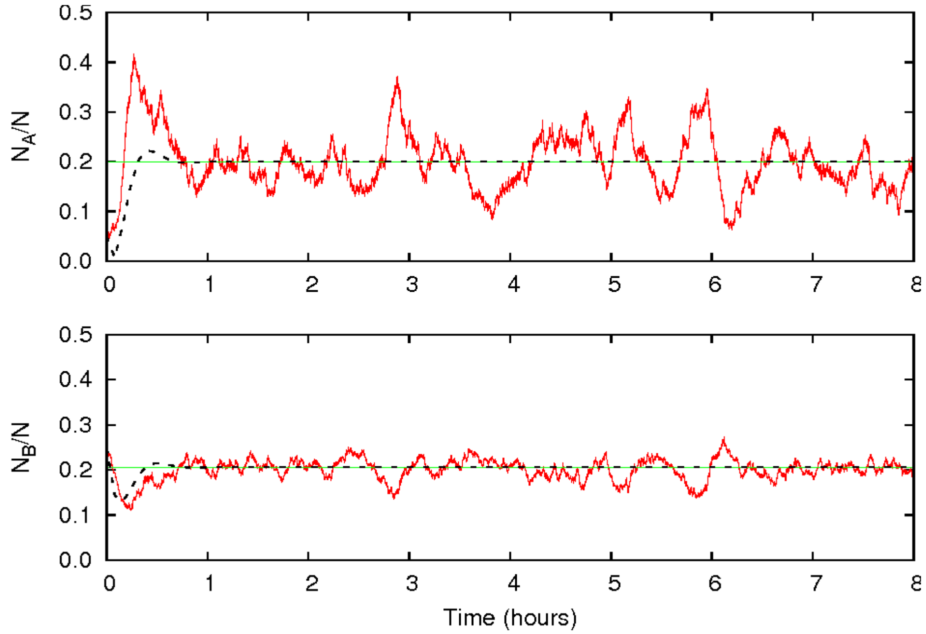
Another example

Fluctuations. We define a deviation of n_i from the mean value $\langle n \rangle$

$$\delta n_i = n_i \ - \ \langle n \rangle.$$

It is obvious that

$$\sum_{i=1}^{N} \delta n_i = \sum_{i=1}^{N} (n_i - \langle n \rangle) = \sum_{i=1}^{N} n_i - \sum_{i=1}^{N} \langle n \rangle$$
$$= N\langle n \rangle \ - N\langle n \rangle \ = 0$$



Another example

Variance. In order to measure how strong are fluctuations we introduce a so-called variance as follows:

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(\delta n_i)^2 = \frac{1}{N}\sum_{i=1}^{N}(n_i - \langle n\rangle)^2$$

Using the probability distribution function $P(n)$ we get

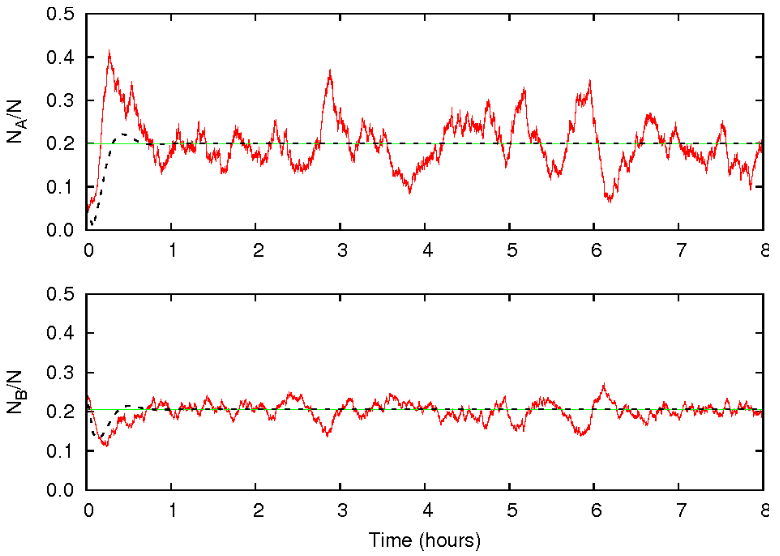
$$\sigma^2 = \sum_{n=0}^{\infty} P(n)(n-\langle n\rangle)^2 = \sum_{n=0}^{\infty} P(n)(n^2 - 2n\langle n\rangle + \langle n\rangle^2)$$

$$= \sum_{n=0}^{\infty} P(n)n^2 - 2\langle n\rangle \sum_{n=0}^{\infty} P(n)n + \langle n\rangle^2 \sum_{n=0}^{\infty} P(n)$$

Thus, $\sigma^2 = \langle n^2\rangle - \langle n\rangle^2$

Large σ^2

Small σ^2



Probability distribution

The statistics of a random variable X is described by its probability *mass* distribution function $P(X)$.

- $P(X)$ assigns a probability to each possible value of X.
- $P(X)$ must obey normalization $\sum_X P(X) = 1$.
- The mean value (expectation) of X is defined as $\langle X \rangle = \sum_X X\, P(X)$.
- The variance of X is: $\mathrm{var}(X) = \sum_X (X - \langle X \rangle)^2 P(X) = \langle X^2 \rangle - \langle X \rangle^2$
- More generally, the average of a function of X, say $f(X)$, is defined as $\langle f \rangle = \sum_X f(X)\, P(X)$.

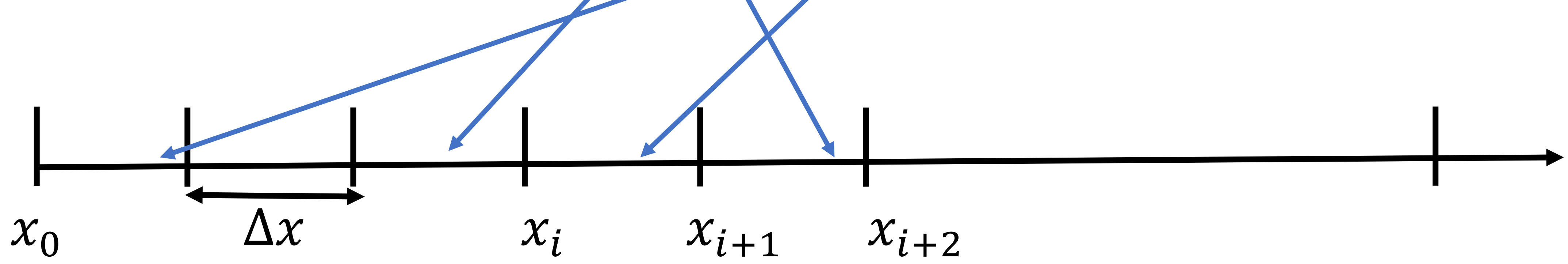
Continuous random variables

In many cases, random variables can take any value belonging to a continuous interval. Then, there infinitely many values allowed, each of them occurring with probability 0…

To deal with continuous random variables we introduce the concept of **probability density**, and probability density distribution function.

Continuous random variables

Lets consider, for example, velocities of cars v_i . We have a random sequence of real numbers $v_1, v_2, v_3, v_4, \ldots, v_N$. We divide the axis of real numbers into intervals (bins) of equal width Δx



We count the number values of v in the interval $(x_i, x_{i+1}]$, i.e.,

$x_i < v_n \le x_{i+1}$, and denote this number as $\Delta N(x_i, x_{i+1})$.

The total number of random numbers v_i in the sequence is

$$\sum_{i=0}^{\infty} \Delta N(x_i, x_{i+1}) = N$$

Probability density distribution

The probability density distribution function is defined as

$$P(x_i) = \lim_{N\to\infty,\Delta x\to 0} \frac{\Delta N(x_i, x_{i+1})}{N\Delta x}.$$

The probability density distribution can be understood as follows: $P(x)\Delta x$ gives the probability of the random variable taking a value between in the interval $(x, x + \Delta x]$ in the limit of $\Delta x \to 0$.

Probability density distribution

The normalization is $\sum_{i=0}^{\infty} P(x_i)\Delta x = \sum_{i=0}^{\infty} \frac{\Delta N(x_i, x_{i+1})}{N\Delta x} \Delta x = 1.$

In the limit $N \to \infty, \Delta x \to 0$, but still $\Delta N(x_i, x_{i+1})$>>1, we can use the integral representation

$$\sum_{i=0}^{\infty} P(x_i)\Delta x = \int_{-\infty}^{\infty} P(x)dx = 1$$

Mean value: $\langle x \rangle = \sum_{i=0}^{\infty} x_i P(x_i)\Delta x = \int_{-\infty}^{\infty} xP(x)dx$

Variance: $\sigma^2 = \sum_{i=0}^{\infty} (x_i - \langle x \rangle)^2 P(x_i)\Delta x = \int_{-\infty}^{\infty} (x - \langle x \rangle)^2 P(x)dx$

Probability density distribution

Suppose that we have a continuous random variable x with a known probability density distribution $P(x)$. Let us consider another variable that is a function of x, say $y = f(x)$.

What the probability density distribution of y say $Q(y)$?

Using a simple rational: the probability of x falling in an interval $(x, x + \Delta x]$ must be equal to the probability of y falling inside $(f(x), f(x + \Delta x)]$

$$P(x)\Delta x = Q(y)[f(x + \Delta x) - f(x)] = Q(y)\Delta y$$

$$\Leftrightarrow Q(y) = P(x)\frac{\Delta x}{\Delta y} = P(x)\left(\frac{\partial y}{\partial x}\right)^{-1}$$

Probability density distribution

$$Q(y) = \left(\frac{\partial y}{\partial x}\right)^{-1} P(x)$$

For example, a rescaling $y = cx$ gives $Q(y) = \frac{1}{c} P\left(\frac{y}{c}\right)$.

Or, a power-law $y = x^{\alpha}$ gives $Q(y) = \frac{y^{1/\alpha - 1}}{\alpha} P\left(y^{1/\alpha}\right)$.

Independence vs. correlations

When facing multiple random variables, it is essential to take into consideration their correlation or independence.

More concretely, let us consider two random variables, X and Y, that appear simultaneously, that is, trial i consists of a pair of random numbers (X_i, Y_i).

The joint statistics of X and Y is described by the **joint probability distribution function** $P(X, Y)$ that assigns a probability to each possible pair of values of X and Y.

In this case, the normalization condition is $\sum_X \sum_Y P(X, Y) = 1$, and an average is $\langle f \rangle = \sum_X \sum_Y f(X, Y) P(X, Y)$.

Independence vs. correlations

- Two events are **independent** if the occurrence of **one does not affect** the probability of occurrence of **the other**.

  For the joint probability distribution function independence means $P(X, Y) = P(X)P(Y)$.

- We say two events are **correlated** when $P(X, Y) \neq P(X)P(Y)$, which implies that **one affects the other**.

Independence vs. correlations

For **independent variables** we can write for the average of their product as $\langle X \cdot Y \rangle = \langle X \rangle \cdot \langle Y \rangle$:

$$\langle X\,Y \rangle = \sum_X \sum_Y X\,Y\,P(X,Y) = \sum_X \sum_Y XP(X)\,YP(Y)$$

$$= \left(\sum_X XP(X) \right) \left(\sum_Y YP(Y) \right) = \langle X \rangle \langle Y \rangle$$

More generally, for independent variables the average of a product of two functions is

$$\langle f(X) \cdot g(Y) \rangle = \langle f(X) \rangle \cdot \langle g(Y) \rangle$$

Independence vs. correlations

When two random variables are **correlated**, we can usually get a measure of their pair correlations from the so-called **covariance**

$$C_{XY} = \frac{1}{N}\sum_{i=1}^{N} \delta X_i \delta Y_i = \langle XY \rangle - \langle X \rangle \langle Y \rangle$$

Notice that when the variables are independent $C_{XY} = 0$.

Furthermore, the Pearson correlation coefficient is defined as

$$\rho_{XY} = \frac{C_{XY}}{\sigma_X \sigma_Y}$$

The same ideas are directly generalized to more than two variables.