Tutorial notes on probability

by Ricardo Alonso
Summer School on Multiscale Modeling and Analysis
University of Texas at Austin
August, 2008

1 Basic definitions

Fix a set Ω which we refer as the *sample space* or the set of *outcomes*. We define for the sample space the following concepts:

Definition. The power set of Ω is defined as the set containing all the subsets of Ω

$$\mathcal{P}(\Omega) = \{A : A \subseteq \Omega\}.$$

Let $\mathcal{F} \subseteq \mathcal{P}(\Omega)$ with the following properties:

(i) $\phi \in \mathcal{F}$.

(ii) If $A \in \mathcal{F} \rightarrow A^c \in \mathcal{F}$.

(iii) Let $\{A_i\}$ a countable family of elements of $\mathcal{F}$, then

$$\bigcup_i A_i \in \mathcal{F}.$$

The family $\mathcal{F}$ is called σ*-field* and its elements are called *events*.

Definition. A probability function is a set function $P : \mathcal{F} \rightarrow [0,1]$ with the properties:

(i) $P(\Omega) = 1$.

(ii) For a mutually disjoint and countable family of events $\{A_i\}$

$$P\left(\bigcup_i A_i\right) = \sum_i P(A_i),$$

Since $P(A_i) \geq 0$ this sum always exists.

A good way to think in a probability function is that it is a function that measures the "size" of every event in $\mathcal{F}$.

Definition. The triplet $(\Omega, \mathcal{F}, P)$ is called *probability space*.

Examples.

(1) Fix $x_0 \in \mathbb{R}^n = \Omega$. Consider the set function defined as

$$P(A) = \begin{cases} 1 & x_0 \in A \\ 0 & \text{otherwise.} \end{cases}$$

The function P is a probability function in Ω.

(2) Take $\Omega = [0, \infty)$ and let $f(x) = \exp(-x)$. Then, the set function defined by

$$P(A) = \int_A f(s)ds$$

is a probability function in Ω.

1.1 Basic properties

The follow properties can be deduced from the properties *(i)* and *(ii)* of the probability function. Let $A, B \in \mathcal{F}$, then

(i) $0 \leq P(A) \leq 1$.

(ii) $P(\phi) = 0$.

(iii) $P(A^c) = 1 - P(A)$.

(iv) $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

1.2 Conditional probability

Let $A, B \in \mathcal{F}$, the probability of the event B given that the event A has occurred is defined by the ratio

$$P(B|A) := \frac{P(B \cap A)}{P(A)}.$$

The idea behind the definition of conditional probability is that the knowledge that the event A has occurred converts this event into the new sample space. Thus, the probability of any event B is referred to A using the intersection and then normalized to it using the quotient.

Definition. The event B is independent of A if

$$P(B|A) = P(B).$$

It turns out that if B is independent of A, then A is independent of B because

$$P(A|B) = P(B|A)\frac{P(A)}{P(B)} = P(A).$$

Therefore, we can simply say that A and B are independent. Clearly, in this case one has

$$P(B \cap A) = P(B)P(A).$$

2 Random variables

Let $(\Omega, \mathcal{F}, P)$ be a probability space, and assume that we can build a function $X : \Omega \to \mathbb{R}^n$ with the property that for any $A \in \mathcal{F}_{\mathbb{R}^n}$

$$\{\omega : X(\omega) \in A\} \in \mathcal{F}.$$

Here $\mathcal{F}_{\mathbb{R}^n}$ is a predetermined and sufficiently large σ-field of $\mathbb{R}^n$ (for example all the measurable sets of $\mathbb{R}^n$). Such a function is called *continuous random variable.* If the range of X is contained in $\mathbb{Z}^n$, we called it *discrete random variable.* Thus, a discrete random variable is a particular case of a continuous random variable.

Example. Flip two coins. The sample space of this experiment is $\Omega = \{ht, th, tt, hh\}$. All the following are different discrete random variables.

(1) $X : \Omega \to \mathbb{Z}$ such that $X(th) = X(ht) = 1$, $X(tt) = 2$, $X(hh) = 3$.

(2) $X : \Omega \to \mathbb{Z}$ such that $X(th) = 0$, $X(ht) = 1$, $X(tt) = 2$, $X(hh) = 3$.

(3) $X : \Omega \to \mathbb{Z}^2$ such that $X(th) = (0, 1)$, $X(ht) = (1, 0)$, $X(tt) = (0, 0)$, $X(hh) = (1, 1)$.

Random variables allow us to make computations of the probability and statistic of a particular experiment in the well-known spaces $\mathbb{R}^n$. Indeed, for any $A \in \mathcal{F}_{\mathbb{R}^n}$ we define the probability of A as

$$P(A) := P(\{\omega : X(w) \in A\}).$$

2.1 Probability distribution and density

Let X be a random variable $X : \Omega \to \mathbb{R}$.

Definition. The probability distribution of X is the function defined as

$$F(x) := P(\{\omega : X(w) \leq x\}) = P(X \leq x).$$

If F is differentiable, we can obtain the so called *density distribution* $f(x)$ of X from F using differentiation. Thus, we have the relation

$$F(x) = \int_{-\infty}^{x} f(s)ds.$$

In the case of a discrete random variable $X : \Omega \to \mathbb{Z}$, we adopt for convenience a slightly different definition for the density distribution

$$f(x) := P(X = x),$$

which leads to the relation

$$F(x) = P(X \leq x) = \sum_{u \leq x} f(u).$$

Examples.

(1) $X : \Omega \to \mathbb{R}$ is normally distributed with parameters (μ, σ) if

$$f(x) = (2\pi\sigma)^{-1/2} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right).$$

(2) $X : \Omega \to \{1, 2, \cdots, n\}$ is binomially distributed with parameter $0 \le p \le 1$ if

$$f(x) = \begin{pmatrix} n \\ x \end{pmatrix} p^x (1-p)^{n-x}.$$

We explain the binomial distribution in the following way. Assume we have an experiment that has two outcomes: $false = 0$ and $true = 1$. We run the experiment n times knowing that every run is independent of the previous ones. Thus, a possible outcome or *realization* of our experiment would be

$$\underbrace{000101100111\cdots0010110111110}_{n \text{ times}}$$

Assume that the probability of getting a false outcome is p, and thus, the probability of getting a true outcome is $1-p$. Since the runs are independent, the probability of one realization is $p^x(1-p)^{n-x}$, where x is the number of false outcomes in the realization. Now, the number of possible realizations having x false outcomes is $\begin{pmatrix} n \\ x \end{pmatrix}$, then, we deduce that the probability of having x false outcomes in n runs is precisely

$$\begin{pmatrix} n \\ x \end{pmatrix} p^x (1-p)^{n-x}.$$

If we define the random variable X as the number of false outcomes of this experiment after n runs, we conclude that X is binomially distributed.

2.2 Join distributions and independent random variables

Let X_i with $i = 1, 2, \ldots, n$ be random variables with $X_i : \Omega_i \to \mathbb{R}$. In order to fully describe the interaction of these random variables, we put them together in a single random vector $X : \times\Omega_i \to \mathbb{R}^n$ with a uniquely defined probability function

$$P : \times\Omega_i \to [0, 1].$$

In this setting, we define the join probability distribution of X by

$$F(x) = P\left(\bigcap_{i=1}^{n} \{X_i \le x_i\}\right),$$

where x_i is the i-entry of x. Similarly to the 1-dimensional case, we define the join density distribution as the function $f : \mathbb{R}^n \to [0, \infty)$ such that

$$F(x) = \int_{\cap\{s_i \le x_i\}} f(s)ds.$$

More generally, we have

$$P(A) = P(X \in A) = \int_A f(s)ds.$$

The functions F and f comprise all the statistics of the random variables X_i's (including their interactions). In fact, the individual statistics of the X_i's can be easily found from the join probability function by means of averaging. Thus, we have the following

Definition. The *marginal* distributions of X are the functions

$$f_{X_i}(x_i) = \int_{\mathbb{R}^{n-1}} f(x_i, s) ds.$$

The marginal distributions are nothing else than the density distributions of each particular X_i.

Definition. Let X and Y be random variables. These random variables are called independent if

$$f(x, y) = f_X(x) f_Y(y).$$

3 Expected value

Let X be a random variable. Then

Definition.

(i) The *expected* or *mean* value is defined as the average

$$E[X] := \int_{\mathbb{R}} s \ f(s) ds.$$

The notation $\mu_X = E[X]$ is commonly used.

(ii) The variance is defined as the average

$$Var(X) := E[(X - \mu_X)^2] = \int_{\mathbb{R}} (s - \mu_X)^2 \ f(s) ds.$$

The notation $\sigma_X^2 = E[(X - \mu_X)^2]$ is commonly used. The σ_X stands for the *standard deviation* of X.

The following are simple properties that hold for the expected value and variance

(1) $E[cX] = cE[X]$ for $c \in \mathbb{R}$.

(2) $E[X + Y] = E[X] + E[Y]$ for any two random variables X and Y.

(3) $\sigma_X^2 = E[X^2] - \mu_X^2$.

(4) $Var(cX) = c^2 Var(X)$ for $c \in \mathbb{R}$.

The following are properties that hold for any two **independent** random variables X and Y

(1) $E[XY] = E[X]E[Y]$.

(2) $Var(X + Y) = Var(X - Y) = Var(X) + Var(Y)$.

An important theorem that confirms that the outcome of a random variable is unlikely to be far from its mean value in terms of the variance scale is the

Theorem 3.1. *(Chebyshev's inequality) Let X be a random variable with mean value μ and variance σ^2, then*

$$P(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}.$$

Proof. Let f the probability density of X, then

$$\begin{aligned} P(|X - \mu| \geq k\sigma) &= \int_{\{|s-\mu| \geq k\sigma\}} f(s)ds \\ &\leq \int_{\{|s-\mu| \geq k\sigma\}} \left(\frac{|s - \mu|}{k\sigma}\right)^2 f(s)ds \\ &\leq \frac{1}{k^2\sigma^2} \int_{\mathbb{R}} |s - \mu|^2 f(s)ds = \frac{1}{k^2}. \end{aligned}$$

□