Elements of Probability for Computer Scientists

Jean-Marc Vincent
LIG Laboratory, Mescal Projet,
51, avenue Jean Kuntzmann, F-38330 Montbonnot, France
Jean-Marc.Vincent@imag.fr

Abstract— **The aim of this short note is to provide a simple introduction of probabilistic modelling in computer science. Basic definition of probability language is given with its semantic. Then concepts of random variable and expectation are detailed for discrete spaces.**

CONTENTS

I Modeling Randomness 1

II Formal Probability Language 2
- II-A Reality and Events . . . 2
- II-B Probability Measure . . . 2
- II-C Conditional Probability . . . 2
- II-D Independence . . . 3

III Discrete Random Variables 3
- III-A Characteristics of a Probability Law . . 3
- III-B Mode . . . 3
- III-C Median and Quantiles . . . 3

IV The Expectation Operator 4
- IV-A Properties of $\mathbb{E}$ . . . 4
- IV-B Variance and Moments . . . 4
- IV-C Sums of Independent Random Variables 4

V Classical Probability Distributions 4
- V-A Bernoulli Law . . . 4
- V-B Vectors of bits . . . 4
- V-C Geometric law . . . 5
- V-D Poisson distribution . . . 5

VI Generating function 5

References 5

I. MODELING RANDOMNESS

In computer science, the aim is to build programs that run on computer and solve complex problems. These problems are instantiated at runtime so that the real input is not known when the program is designed and the programmer should deal with this uncertainty. Moreover, a program is usually written to solve repeatedly a problem with different instances and should be efficient on many inputs.

To take into account data variability, there are two kinds of approaches. The first one, worst case analysis, gives an upper bound on the complexity of algorithms. So the worst case guarantees that the program finishes before some time. But in many practical situations, this worst case does not happen frequently and the bound is rarely reached. The second approach, average case analysis suppose that the inputs of programs obey to some statistical regularity and as a consequence the execution time could be considered as a random variable with some probability distribution depending on the input distribution.

To model inputs variability and uncertainty a formal language is needed. This language is the probability language (section II) based on set axiomatic formalism enriched with specific axioms for probability. The formal probability theory build a framework of mathematical theorems, such as the *law of large numbers*, the *Central limit theorem*, etc. The main difficulty is not as deriving results but in the modelling process and the semantic associated to probabilities.

This difficulty comes from the fact that probabilities are commonly used in many real life situations : probability to have a sunny weather next day, that a baby is boy, to have a crash accident, to win in a lotterie, to have a full in a poker game, etc.

As an illustration example, try to solve these following questions

Example 1 : Boys and girls
Mr Smith has two children, one is a boy, what is the probability that the other is a girl ?

Example 2 : Pascal and Chevalier de Méré discussion
Consider the dice game with the following rules :

- bet 1
- throw two dices and sum the results
- if the result is 11 or 12 you earn 11 (including your bet)
- if not you loose your bet.

The Chevalier de Méré says 'playing this game a sufficiently long time and I'll get a fortune" and Pascal argues the contrary. Who is wrong and what were the two arguments ?

Example 3 : The Monty Hall problem
Consider the TV show game, there are 3 closed doors beside one there is a magnificent car, beside the two others nothing.

- TV host : Please choose one door. As example you choose door 2.
- TV host : I want to help you. I open one of the remaining door with nothing. For example he opens door 1.
- TV host : in fact you could modify your first choice, do you change your initial decision of choosing door 2.
- As example you decide to change and you open door 3.

You win if the car is beside.

What is a good strategy : change or not your initial decision ?

All the previous examples generate discussions among people, many websites give and comment solutions. The ambiguity is first in the modelling and next in the way the probability values are interpreted.

II. Formal Probability Language

The probability language was funded in 1938 by Andreï Nicolai Kolmogorov [1] in the famous monograph Grundbegriffe der Wahrscheinlichkeitsrechnung. His theoretical construction is based on the set theory (initiated by Cantor) coupled with the measure theory established by Lebesgue.

A. Reality and Events

The formalism and the composition rules are all based on set computation. Consider now a set Ω, a set $\mathcal{A}$ of parts of Ω is called a σ-**field** if it satisfies the following properties :

1) $\Omega \in \mathcal{A}$;
2) If $A \in \mathcal{A}$ then $\overline{A} \in \mathcal{A}$ (the complement of A in Ω is in $\mathcal{A}$);
3) Let $\{A_n\}_{n\in\mathbb{N}}$ a denumerable set of element of $\mathcal{A}$ then $\bigcup_{n\in\mathbb{N}} A_n \in \mathcal{A}$;
(σ-additivity property)

Interpretation The set Ω model the real world, which is impossible to capture with all of its complexity. Consequently we observe the reality with measurement tools and get partial information on it. An *event* is a fact we could observe on the real situation. It supposes the existence of an experience that produce the event which is observable.

The properties of events have the following meaning :

1) $\Omega \in \mathcal{A}$ means that the could observe the real world.
2) If $A \in \mathcal{A}$ then $\overline{A} \in \mathcal{A}$. If we could observe a given fact, we could also observe that this fact does not occur.
3) If A and B are events, then $A \cup B$ is an event. If we could observe two facts then we can to observe them simultaneously.

Example 4 : Requests on a web server

Consider a web server that delivers web pages according requests from clients spread on the Internet. The set Ω is the set of all possible clients with their own capacities, their location, etc. In this case an event is what you could observe. From the server point of view, the observables are the request type, the Internet address of the client. So *the request ask for page foo* is an event, *the origin of the request is from the US domain* is another event,... The fact that the end user that sends the request is a boy teenager is not observable and then is not an event.

Deriving from the set theory, the following properties are easy to prove and the semantic is left to the reader.

Proposition 1 (Events manipulation): Let A, B, C be events of Ω :

- $\overline{\overline{A}} = A$
- $A \cup (B \cup C) = (A \cup B) \cup C = A \cup B \cup C$
- $A \cap (B \cap C) = (A \cap B) \cap C = A \cap B \cap C$
- $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$
- $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$
- $\overline{A \cup B} = \overline{A} \cap \overline{B}$
- $\overline{A \cap B} = \overline{A} \cup \overline{B}$

B. Probability Measure

The idea of probability is to put some real value on events, then the probability function is defined on the set of events and associate to each event a real in $[0, 1]$.

$$\begin{array}{rccc} \mathbb{P}: & \mathcal{A} & \longrightarrow & [0,1]; \\ & A & \longmapsto & \mathbb{P}(A). \end{array}$$

It verifies the following rules :

1) $\mathbb{P}(\Omega) = 1$;
2) If $\{A_n\}_{n\in\mathbb{N}}$ is a sequence of disjoint events (for all (i, j), $A_i \cap A_j = \emptyset$) then
$$\mathbb{P}\left(\bigcup_n A_n\right) = \sum_n \mathbb{P}(A_n);$$
σ-additivity property.
3) If $\{A_n\}_{n\in\mathbb{N}}$ is a non-increasing sequence of events, $A_1 \supseteq A_2 \supseteq A_n \supseteq \cdots$ converging to $\emptyset$ ($\bigcap_{i=1}^{+\infty} A_i = \emptyset$) then
$$\lim_{n\to+\infty} \mathbb{P}(A_n) = 0.$$
This continuity axiom is usefull for non finite sets of events.

From these axioms we deduce many theorems verified by the formal system. The proof of the following propositions are left to the reader.

Proposition 2 (Probability properties): Let A and B events of Ω :

1) $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$;
2) $\mathbb{P}(\overline{A}) = 1 - \mathbb{P}(A)$;
3) $\mathbb{P}(\emptyset) = 0$;
4) If $A \subset B$, then $\mathbb{P}(A) \leqslant \mathbb{P}(B)$ ($\mathbb{P}$ is a non-decreasing function).
5) If $A \subset B$, then $\mathbb{P}(B - A) = \mathbb{P}(B) - \mathbb{P}(A)$.

Interpretation The semantic of a probability measure is related to experimentation. Consequently it supposes that we can repeat infinitively experiments in the same conditions. Then the probability of an event (observable) A is the abstraction of the proportion that this event is realized in a large number of experiments. Consequently the probability is an ideal proportion, assuming that we could produce an infinite number of experiments and compute the asymptotic of frequencies.

C. Conditional Probability

Consider an event B such that $\mathbb{P}(B) > 0$. The conditional probability of an event A knowing B, denoted by $\mathbb{P}(A|B)$ is given by

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

It defines a new probability measure on the set of event $\mathcal{A}$ (check it as an exercise).

Consider a partition of Ω in a countable set of observable events $\{B_n\}$ ($\mathbb{P}(B_n) > 0$). The law of total probability states that for all $A \in \mathcal{A}$

$$\mathbb{P}(A) = \sum_n \mathbb{P}(A|B_n)\mathbb{P}(B_n).$$

The Bayes' theorem reverse this scheme by

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

Interpretation The meaning of conditional probability comes from the fact that we could observe reality through several measurement instruments. For example, consider a transmission protocol. We observe both the size of messages and the transfer time of these. Then we want to explain the transfer time by the message size. From many experiments we deduce the probability distribution of the transfer given a size of message. The conditional probability considers external information (event) which is given a-priori. The law of total probability explains that if we have a set of disjoint alternatives, we could compute the probability of an event by computing its probability knowing each alternative and then combine all of them with the weight (probability) of each alternative.

D. Independence

Two events A and B are independents if they satisfy

$$\mathbb{P}(A \cap B) = \mathbb{P}(A).\mathbb{P}(B).$$

This is rewritten, assuming $\mathbb{P}(B) > 0$

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

Interpretation Independence is related to the causality problem. If two events are not independent we could suspect a hidden relation between them, then an event could be the "cause" of the other. On the other side two events are independent if in the observed phenomenon there are no possible relations between the events. If we throw two dices, there is no reason that the result of the first dice depends on the result of the second dice. Moreover if the two observations are not independent (statistically on many experiments), we should search the physical reason (the cause) that couples the results of the dices.

III. Discrete Random Variables

In fact, the modelling process cannot capture the whole complexity of the physical reality. Then the Ω set is not accessible and the only way to get information is to observe the reality via filters (measurement instruments). Then we consider a random variable X as a function in a set E which is "well known" like (finite sets, integers, real numbers, real vectors,...) with an adequate system of events $\mathcal{B}$,

$$\begin{array}{rcl} X : \Omega & \longrightarrow & E \\ \omega & \longmapsto & X(\omega) \end{array}$$

such that event

$$\{X \in B\} \triangleq \{\omega \in \Omega \text{ such that } X(\omega) \in B\} \in \mathcal{A},$$

and then induces a probability measure on $\mathcal{B}$ denoted by $\mathbb{P}_X$ image of the probability measure $\mathbb{P}$ by X. The probability $\mathbb{P}_X$ is called the **law** of the random variable X.

Interpretation This is why the fundamental set is called Ω the ultimate limit which is practically unreachable. All things that could be observed are through random variables. Then in stochastic modelling, the Ω fundamental set is not used and the system is described by random variables with given laws. As an example, we model a dice trow by a random variable X with values in $E = \{1, 2, 3, 4, 5, 6\}$ and a uniform law of probability, $\mathbb{P}_X(i) = \mathbb{P}(X = i) = \frac{1}{|E|} = \frac{1}{6}$.

A. Characteristics of a Probability Law

When E is discrete, we usually choose the set $\mathcal{B}$ of events as the set of all parts of E. The law of X is consequently given by the probability of all elements (considered as elementary parts of E). We get the probability of any event B by

$$\mathbb{P}(X \in B) = \sum_{x \in B} \mathbb{P}(X = x).$$

The function f_X defined by

$$\begin{array}{rcl} f_X : E & \longrightarrow & [0, 1], \\ x & \longmapsto & f_X(x) = \mathbb{P}(X = x), \end{array}$$

is called the probability **density** function of X. We remark that f_X is non-negative and satisfies

$$\sum_{x \in E} f_X(x) = 1.$$

B. Mode

We define the mode of a discrete random variable distribution by

$$Mode(X) = \underset{x \in E}{\operatorname{argmin}}\, f_X(x),$$

it denotes a value that achieves the maximum of probability.

C. Median and Quantiles

When the set E is totally ordered, we define the **cumulative distribution function** F_X by

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

The function F_X is non-decreasing and if E is $\mathbb{Z}$ or isomorph to a part of $\mathbb{Z}$ we get

$$\lim_{n \to -\infty} F_X(n) = 0 \text{ and } \lim_{n \to +\infty} F_X(n) = 1.$$

Because E is ordered, we define the median of the probability law by the value $Median$ satisfying

$$Median = \underset{x \in E}{\operatorname{argmax}} \left\{ F_X(x) \leqslant \frac{1}{2} \right\}.$$

This is generalized by splitting the set E in parts with equal probability. For example, quartiles are defined by

$$q_1 = \operatorname*{argmax}_{x\in E}\left\{F_X(x) \leqslant \frac{1}{4}\right\},\ q_2 = Median,$$

$$q_3 = \operatorname*{argmax}_{x\in E}\left\{F_X(x) \leqslant \frac{3}{4}\right\}.$$

Deciles are given by

$$d_i = \operatorname*{argmax}_{x\in E}\left\{F_X(x) \leqslant \frac{i}{10}\right\},\ \text{for } 1 \leqslant i \leqslant 10.$$

IV. The Expectation Operator

When E is a richer structure (group, ring or field), we compute other parameters of the law and particularly moments. To simplify the text, we will consider only integer valued random variables.

The expectation operator associates to a probability law the quantity

$$\mathbb{E}X \triangleq \sum_{x\in E} x f_X(x) = \sum_{\omega\in\Omega} x\mathbb{P}(X(\omega) = x).$$

This is extended as an operator for any function h by

$$\mathbb{E}h(X) = \sum_{x\in E} h(x)\mathbb{P}(X = x).$$

A. Properties of $\mathbb{E}$

The operator $\mathbb{E}$ is linear; for X and Y random variables defined on the same probability space and λ, μ real coefficients

$$\mathbb{E}(\lambda X + \mu Y) = \lambda\mathbb{E}X + \mu\mathbb{E}Y.$$

The operator $\mathbb{E}$ preserve the natural order on $\mathbb{Z}$.That is

If $X \leqslant Y$ then $\mathbb{E}X \leqslant \mathbb{E}Y$.

When X and Y are independent variables then

$$\mathbb{E}(XY) = \mathbb{E}X.\mathbb{E}Y.$$

B. Variance and Moments

The order n moment of a probability law is defined by

$$M_n = \mathbb{E}X^n,$$

and the centralized moment of order 2 is called the **variance** of the law and is given by

$$\mathbb{V}arX = \mathbb{E}(X - \mathbb{E}X)^2 = \mathbb{E}X^2 - (\mathbb{E}X)^2.$$

Usually the variance is denoted by σ^2 and $\sigma = \sqrt{\mathbb{V}arX}$ is called the **standard deviation** of X. The variance of random variable satisfies the following properties :

$\mathbb{V}arX \geqslant 0$;

$\mathbb{V}arX = 0$ implies $\mathbb{P}(X = 0) = 1$;

$\mathbb{V}araX = a^2\mathbb{V}arX$;

If X and Y are independent then

$\mathbb{V}ar(X + Y) = \mathbb{V}arX + \mathbb{V}arY$.

Interpretation Parameters of probability laws try to synthesize the law in some value which could be compared with others. The concept of central tendency : most probable value (Mode), or value that splits the probability in two equal parts (Median), or arithmetic mean (Expectation) represent the law by one number that have its own semantic. Then one should indicates how far the law could be from the central tendency. The concept of variability around the central tendency gives information on the spreading of the law. For the mode an usual index of variability is the entropy (which is not in the topic of this note and indicates how far the probability is from the uniform distribution). For the median, quantiles shows the distance in of probability and the variance corresponds to the fluctuation around the mean value when experiments are repeated (error theory, central limit theorem).

C. Sums of Independent Random Variables

Consider X, Y two independent random variables on $\mathbb{N}$ with distribution p and q. The **convolution** product of the two distributions, denoted by $p \star q$, is given by

$$p \star q_k = \sum_i p_i.q_{k-i}.$$

The convolution product corresponds to the sum of independent variables, we have

$$\mathbb{P}(X + Y = k) = \sum_i \mathbb{P}(X = i, Y = k - i)$$

$$= \sum_i \mathbb{P}(X = i)\mathbb{P}(Y = k - i) = \sum_i p_i.q_{k-i}$$

$$= p \star q_k.$$

V. Classical Probability Distributions

A. Bernoulli Law

The simplest probability law is obtained from the coin game, a choice between 2 values, a random bit,... X follows the Bernoulli law with parameter $p \in [0, 1]$, denoted $\mathcal{B}(p)$, if

$$\mathbb{P}(X = 1) = p = 1 - \mathbb{P}(X = 0).$$

We easily get

$$\mathbb{E}X = p\ \ \mathbb{V}arX = p(1 - p).$$

This law represents the basic generator we could get for randomness in computer science. All other laws are deduced from infinite sequences of random bits.

B. Vectors of bits

We consider a random vector $X = [X_1, \cdots, X_n]$ composed of n independent random variables X_i identically distributed with law $\mathcal{B}(p)$. Then we obtain the law of the vector by

$$\mathbb{P}(X = x) = p^{H(x)}(1 - p)^{n-H(x)} \text{ where } x \in \{0, 1\}^n,$$

and $H(x)$ the Hamming weight of the bit vector x, if

$$x = [x_1, \cdots, x_n],\ \text{then } H(x) = \sum_{i=1}^{n} x_i.$$

From this formula many other laws could be computed

Binomial Law

The Binomial law $\mathcal{B}in(n,p)$ is the law of $H(X)$ where X is a random bit vector with size n and probability p for each bit. It takes values in $\{0,\cdots,n\}$ and the pdf is given by

$$\mathbb{P}(X=k)=\left(\begin{array}{c} n \\ k \end{array}\right)p^k(1-p)^{n-k}.$$

The mean $\mathbb{E}X=np$ and the variance $\mathbb{V}arX=np(1-p)$.

C. Geometric law

For an infinite number of bits identically distributed the index X of the first occurrence of 0 follows the geometric distribution with parameter p denoted $\mathcal{G}eom(p)$. Its pdf is given by

$$\mathbb{P}(X=k)=(1-p)p^{k-1}.$$

The mean and variance distribution are

$$\mathbb{E}X=\frac{1}{1-p} \text{ and } \mathbb{V}ar\frac{p}{(1-p)^2}.$$

D. Poisson distribution

This law appears as an approximation of the Binomial distribution when n is large and $\lambda=np$ (the mean) is small with regards to n. In that case, the number X of bits 1 has the pdf

$$\mathbb{P}(X=k)=e^{-\lambda}\frac{\lambda^k}{k!},$$

and the mean $\mathbb{E}X=\lambda$ and the variance $\mathbb{V}arX=\lambda$.

VI. Generating function

Definition 1 (Convolution): Let X et Y independent integer valued random variables with respective density f_X et f_Y. The convolution of f_X and f_Y is the density of $X+Y$ denoted by $f_{X+Y}=f_X*f_Y$:

$$f_{X+Y}(k)=\sum_{j=0}^{k}f_X(j)f_Y(k-j).$$

Computing the convolution of probability densities, is usually difficult. The idea is to change the point of view and transform the density in another object (a function) such that operations of convolution are transformed in operations on functions (that should be simpler). The main mathematical tool that provides such principles is the Fourier transform that could be instantiated in the discrete case as generating functions.

Definition 2 (Generating function): Let X be an integer valued random variable with density f_X. The generating function associated to X is the power series

$$G_X(x)=\mathbb{E}(x^X)=\sum_k x^i\mathbb{P}(X=k).$$

This power series converges for at least $|x|\leqslant 1$ and the coefficients of the series could be obtained by successive derivation at 0,

$$\mathbb{P}(X=k)=\frac{G_X^{(k)}(0)}{k!},$$

where $G_X^{(k)}$ is the k^{th} derivative of G_X (this is just the application of the Taylor expansion of G_X in 0).

The integral of the density is 1 implies that $G_X(1)=1$. The bijection between positive formal series summing to 1 show that the generating function characterize completely the probability distribution of X.

Proposition 3 (Properties of the generating function): The generating function G_X satisfies the following properties:
- G_X is positive, non-decreasing and convex on $[0,1]$.
- Moments of X when finite are given by successive derivations of G_X in 1:

$$\mathbb{E}(X)=G'_X(1);$$
$$Var(X)=G''_X(1)+G'_X(1)-G'_X(1)^2;$$
and more generally
$$G_X^{(n)}(1)=\mathbb{E}\left[X(X-1)(X-2)\cdots(X-n+1)\right].$$

Proposition 4 (Generating function and convolution): Let X and Y independent integer valued random variables, then

$$G_{X+Y}=G_X.G_Y$$

References

[1] A. Kolmogorov. *Foundations of the theory of probability*. Chelsea Publishing Company, 1933. English translation of *Grundbegriffe der Wahrscheinlichkeitrechnung* published in Erg ebnisse Der Mathematic, 1950.

Laws and notations	$X(\Omega)$	$\mathbb{P}(X=k)$	$\mathbb{E}(X)$	$\mathbb{V}ar(X)$	$G_X(z)$
Uniform $\mathcal{U}(n)$	$[1,n]$	$\frac{1}{n}$	$\frac{n+1}{2}$	$\frac{(n^2-1}{12}$	$z\frac{1-z^n}{1-z}$
Bernoulli $\mathcal{B}(1,p)$	$\{0,1\}$	$\mathbb{P}(X=1)=p$ $\mathbb{P}(X=0)=1-p$	p	$p(1-p)$	$(1-p)+pz$
Binomial $\mathcal{B}(n,p)$	$[0,n]$	$C_n^k p^k(1-p)^{n-k}$	np	$np(1-p)$	$((1-p)+pz)^n$
Geometric $\mathcal{G}(p)$	$\mathbb{N}^*$	$(1-p)^{k-1}p$	$\frac{1}{p}$	$\frac{1-p}{p^2}$	$z\frac{1-p}{1-(1-p)z}$
Poisson $\mathcal{P}(\lambda)$	$\mathbb{N}$	$e^{-\lambda}\frac{\lambda^k}{k!}$	λ	λ	$e^{\lambda(z-1)}$

TABLE I

SUMMARY OF CLASSICAL DISCRETE PROBABILITY LAWS