A Tutorial on Probability Theory

Paola Sebastiani
Department of Mathematics and Statistics
University of Massachusetts at Amherst

Corresponding Author: Paola Sebastiani. Department of Mathematics and Statistics, University of Massachusetts, Amherst, MA 01003, USA. PHONE: (413)545 0622, FAX: (413) 545 1801, EMAIL: sebas@math.umass.edu, URL: http://www.math.umass.edu.

Contents

1. Probability and Uncertainty

Probability measures the amount of uncertainty of an event: a fact whose occurrence is uncertain. Consider, as an example, the event R "Tomorrow, January 16th, it will rain in Amherst". The occurrence of R is difficult to predict — we have all been victims of wrong forecasts made by the "weather channel" — and we quantify this uncertainty with a number $p(R)$, called the probability of R. It is common to assume that this number is non-negative and it cannot exceed 1. The two extremes are interpreted as the probability of the impossible event: $p(R) = 0$, and the probability of the sure event: $p(R) = 1$. Thus, $p(R) = 0$ asserts that the event R will not occur while, on the other hand, $p(R) = 1$ asserts that R will occur with certainty.

Suppose now that you are asked to quote the probability of R, and your answer is $p(R) = 0.7$. There are two main interpretations of this number. The ratio $0.7/03$ represent the odds in favor of R. This is the *subjective probability* that measures your personal belief in R. *Objective probability* is the interpretation of $p(R) = 0.7$ as a relative frequency. Suppose, for instance, that in the last ten years, it rained 7 times on the day 16th January. Then $0.7 = 7/10$ is the relative frequency of occurrences of R, also given by the ratio between the favorable cases (7) and all possible cases (10).

There are other interpretations of $p(R) = 0.7$ arising, for instance, from logic or psychology (see Good (1968) for an overview.) Here, we will simply focus attention to rules for computations with probability.

2. Basic Definitions

Definition 1 (Sample Space) *The set of all possible events is called the* sample space *and is denoted by $\mathcal{S}$.*

If we denote events by capital letters $A, B, \ldots$, we write $\mathcal{S} = \{A, B, \ldots\}$. The identification of the sample space depends on the problem at hand. For instance, in the exercise of forecasting tomorrow weather, the sample space consists of all meteorological situations: rain (R), sun (S), cloud (C), typhoon (T) etc.

The sample space is a set, on which we define some algebraic operations between events.

Definition 2 (Algebraic Operations) *Let A and B be two events of the sample space $\mathcal{S}$. We will denote*

"A does not occur" *by $\bar{A}$;*

"either A or B occur" *by $A \cup B$;*

"both A and B occur" *by A, B;*

"A occurs and B does not" *by $A \backslash B \equiv A, \bar{B}$.*

The events A and B are *exhaustive* if $A \cup B = \mathcal{S}$, in other words we are sure that either A or B will occur. Thus, in particular $A \cup \bar{A} = \mathcal{S}$. The events A and B are *exclusive* if $A, B = \emptyset$, where $\emptyset$ is the impossible event, that is the event whose occurrence is known to be impossible. In this case, we are sure that if A occurs then B cannot. Clearly, we have $A, \bar{A} = \emptyset$.

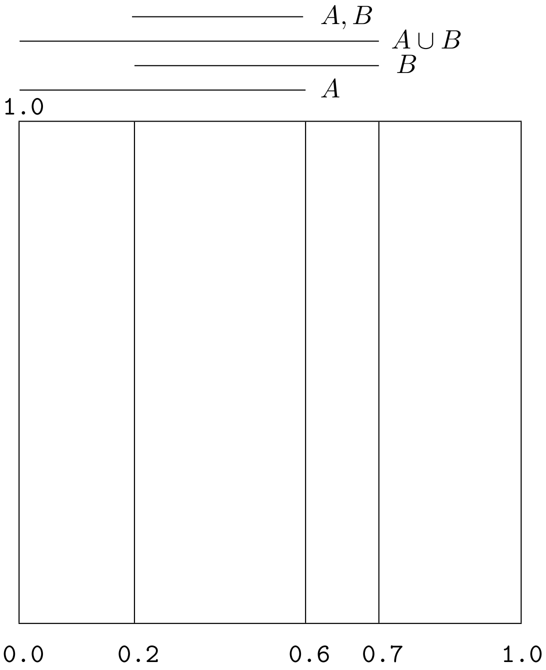

Figure 1: Graphical representation of operations with events.

These operations with events are easily represented via Venn's diagrams. Conventionally, we will represent events as rectangles, whose area is their probability. Because $\mathcal{S}$ is the union of all possible events, its probability is $p(\mathcal{S}) = 1$ and we represent $\mathcal{S}$ as a square of side 1. Under this conventional representation, we will call $\bar{A}$ the complementation of A, that is $\bar{A} = \mathcal{S} \backslash A$. Similarly, the events $A \cup B$, A, B and $A \backslash B$ will be called the union, the intersection and the difference of A and B.

Example 1 (Venn's Diagrams) *Figure 1 gives an example of a Venn's diagram. The rectangle of height 1.0 and basis $0.0 - 0.6$ — of area 0.6 — represents the event A, with $p(A) = 0.6$. The rectangle of height 1.0 and basis $0.2 - 0.7$ represents the event B, with $p(B) = 0.5$. The event $\bar{A}$ is the rectangle with basis $0.6 - 1.0$, and $\bar{B}$ is given by union of the rectangles with bases $0.0 - 0.2$ and $0.7 - 1.0$. The intersection of A and B is given by the rectangle with basis $0.2 - 0.6$, so that $p(A, B) = 0.4$. The union $A \cup B$ is given by the rectangle with basis $0.0 - 0.7$ with $p(A \cup B) = 0.7$. The event $A \backslash B$ is represented by the rectangle with basis $0.0 - 0.2$, and $B \backslash A$ by the rectangle with basis $0.6 - 0.7$. Thus, we have $p(A \backslash B) = 0.2$ and $p(B \backslash A) = 0.1$.*

3. Basic Axioms

In Example 1 the probability of an event is the area of the rectangle that represents the event, and the sample space is the union of all events. This representation can be generalized to more abstract spaces and leads to an axiomatic definition of probability (Kolmogorov, 1950) in terms of measure over a collection of subsets. This collection is assumed to contain the empty set, and to be closed under the complementation and countable union (i.e. $\cup_{i=1}^{\infty} A_i \in \mathcal{S}$.)

Theorem 1 *Let $\mathcal{S}$ denote the sample space. A set function $p(\cdot)$ defined in $\mathcal{S}$ is a probability function if:*

1. *For any event A in $\mathcal{S}$, then $p(A) \geq 0$;*
2. *$p(\mathcal{S}) = 1$;*
3. *If $A_1, A_2, \dots$ are exclusive events in $\mathcal{S}$ and hence $A_i, A_j = \emptyset$ for all i, j, then $p(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} p(A_i)$.*

From these axioms, the following elementary properties can be derived. See Karr (1992, page 25) for details.

Properties 1 *Let $p(\cdot)$ be a probability function defined over the sample space $\mathcal{S}$. Then $p(\cdot)$ satisfies the following properties:*

1. *$p(\emptyset) = 0$;*
2. *$p(\cdot)$ is* finitely additive*: if $A_1, A_2, \dots, A_n$ are events in $\mathcal{S}$, such that $A_i, A_j = \emptyset$ for all $i \neq j$, then*

$$p(\cup_{h=1}^{n} A_h) = \sum_{h=1}^{n} p(A_h). \tag{1}$$

   If these events form a partition of $\mathcal{S}$, i.e. they are such that $\cup_{h=1}^{n} A_h = \mathcal{S}$, then $p(\cup_{h=1}^{n} A_h) = 1$;

3. *$p(A \cup \bar{A}) = p(A) + p(\bar{A}) = 1$, so that $p(A) \leq 1$ for any A in $\mathcal{S}$;*
4. *if $A \subset B$ then $p(B \backslash A) = p(B) - p(A)$;*

Axiom *(iii)* is known as *countable additivity* and it is rejected by a school of probabilists who replace the countable additivity by finite additivity. Further details are in DeFinetti (1972).

Consider now the two events A and B in Figure 1. If we computed $p(A \cup B)$ as $p(A) + p(B)$ we would obtain $p(A \cup B) = 1.1$ that exceeds 1. The error here is that, in computing $p(A \cup B)$ as $p(A) + p(B)$, the event A, B is counted twice. Indeed, we can decompose A into $(A \backslash B) \cup (A, B)$ and similarly B into $(A, B) \cup (B \backslash A)$. Since the intersection $(A \backslash B), (A, B) = \emptyset$, the events $(A \backslash B)$ and (A, B) are exclusive and there follows, from item 3 in Theorem 1, that $p(A) = p(A \backslash B) + p(A, B)$, and similarly $p(B) = p(A, B) + p(B \backslash A)$. The event $A \cup B$ is given by $(A \backslash B) \cup (A, B) \cup (B \backslash A)$, and the three events are exclusive. Thus, from property (1) we have $p(A \cup B) = p(A \backslash B) + p(A, B) + p(B \backslash A) = p(A \backslash B) + p(A, B) + p(B \backslash A) + p(A, B) - p(A, B) = p(A) + p(B) - p(A, B)$. The rule derived in this example holds in general:

$$p(A \cup B) = p(A) + p(B) - p(A, B). \tag{2}$$

In particular, (2) reduces to (1) when $A, B = \emptyset$ and hence $p(A, B) = 0$.

4. Conditional Probability

The probabilities considered so far are unconditional probabilities. In some situations, however, we may be interested in the probability of an event given the occurrence of some other event. For instance, the probability of R: "Tomorrow, January 16th, it will rain in Amherst" would change, if we happened to know that tomorrow is a cloudy day. Formally, if we denote by C the event "Tomorrow, 16th of January, will be cloudy", assuming the occurrence of C is equivalent to restricting our sample space, because other events as S (sunny day) are ruled out. We thus need to recompute the probability of R by taking into account this new piece of information. This is formally done by considering the *conditional probability* of R given that C occurs. This event is denoted by $R|C$.

Consider the events A and B in Figure 1. If we limit the scenario of possible events to A, the occurrence of B would be restricted to A, B. If we knew that A occurs, we would then deduce $p(B|A) = p(A, B)$. However, since $p(A) < 1$, we can only state that $p(B|A) = kp(A, B)$, where k is a proportionality constant that accounts for the uncertainty in the occurrence of A. Clearly, we have $p(A|A) = 1$ and also $p(A|A) = kp(A, A) = kp(A)$. From this, we deduce that $k = 1/p(A)$ and the conditional probability is thus defined as follows.

Definition 3 (Conditional Probability) *Let A and B events in $\mathcal{S}$, and suppose that $p(A) > 0$. The* conditional probability *of B given A is:*

P(A and B) / P(A)

$$p(B|A) = \frac{p(B, A)}{p(A)}. \tag{3}$$

To emphasize that $p(A)$ is unconditional, $p(A)$ is called *marginal probability*.

Example 2 (Conditional Probability) *Consider choosing a card from a well-shuffled standard deck of 52 playing cards. The probability that the first card extracted is an ace is clearly 4/52. Suppose that, after the first extraction, the card is not reinserted in the deck. What is the probability that the second card is an ace, given that the first card is an ace? Let A be the event that the first card is an ace, and let B be the event that the second card is an ace. The probability of A, B is*

$$\frac{12}{2652} = \frac{\textit{possible pairs of aces}}{\textit{all possible pairs of cards}}$$

and $p(A) = 4/52$. On using (3) we have

$$p(B|A) = \frac{\frac{12}{2652}}{\frac{4}{52}} = \frac{3}{51}.$$

Indeed, there are three aces left in a deck of 51 cards.

From the definition of conditional probability in (3), we derive the probability of the intersection of two events, called their *joint probability* in terms of conditional and marginal probabilities:

$$p(A, B) = p(B|A)p(A). \tag{4}$$

This rule can be applied to a larger number of events and produces the *multiplication rule* or *factorization rule*.

Definition 4 (Multiplication Rule) *The joint probability of a set of events $A_1, A_2, \ldots, A_n$ can be expressed as*

$$p(A_1, A_2, \ldots, A_n) = p(A_1)p(A_2|A_1)p(A_3|A_1, A_2)\ldots p(A_n|A_1, A_2, \ldots, A_n)$$

Consider again the events A and B represented in Figure 1. The events A and $\bar{A}$ form a partition of $\mathcal{S}$, so that we can decompose B into the union of the two exclusive events A, B and $\bar{A}, B$. Thus, if we use (4) and the third axiom of probability — item 3 in Theorem 1 — we have:

$$p(B) = p(A, B) + p(\bar{A}, B) = p(B|A)p(A) + p(B|\bar{A})p(\bar{A}). \tag{5}$$

Formula (5) is known as the *Total Probability Theorem* and expresses the marginal probability of B as a weighted average of the conditional probabilities $p(B|A)$ and $p(B|\bar{A})$ with weights given by $p(A)$ and $p(\bar{A})$. The importance of the Total Probability Theorem is that, sometimes, expressing conditional probabilities can be easier than expressing marginal probabilities, and (5) can be used to "break down" an event in more specific events, on which a more precise knowledge is available. Suppose, as an example, that B is the event that the result of a test to diagnose the presence of a disease A is positive. Quantifying the incidence of false positive $(B|\bar{A})$ and false negative $(\bar{B}|A)$ can be easier than quantifying the marginal probability of B. If, further, the incidence rate of A is known, then (5) can be used to derive $p(B)$.

The multiplication rule and the Total probability theorem can be extended to conditional probabilities. So,

$$p(A_1, A_2, \ldots, A_n|C) = p(A_1|C)p(A_2|A_1, C)p(A_3|A_1, A_2, C)\ldots p(A_n|A_1, A_2, \ldots, A_n, C)$$

and

$$p(B|C) = p(B|A, C)p(A|C) + p(B|\bar{A}, C)p(\bar{A}|C).$$

5. Bayes' Theorem

From $p(A|B) = p(B, A)/p(B)$ we can also write $p(A, B) = p(A|B)p(B)$. Putting these relations together, we derive the *Bayes' Theorem*:

$$p(A|B) = \frac{p(A, B)}{p(B)} = \frac{p(B|A)}{p(B)}p(A) = \frac{p(B|A)p(A)}{p(B|A)p(A) + p(B|\bar{A})p(\bar{A})}. \tag{6}$$

Bayes' Theorem can be regarded as a rule to update an initial probability $p(A)$, also called the *prior probability*, into an revised probability $p(A|B)$, called the *posterior probability*, that takes into account the updated knowledge currently available. In the example above Bayes' Theorem can be used to revise the probability that an individual has the disease, if the result of the test turns out to be positive. The next example is adapted from Casella and Berger (1990, page 21).

Example 3 (Bayes' Theorem) *Morse code uses "dots" and "dashes", which are known to occur in the proportion 3:4. Let D denote the event "a dot is sent", and let $\bar{D}$ denote the event "a dash is sent". Thus, we have $p(D) = 3/7$. When coded messages are sent, there can be errors in the* *transmission. Suppose that, with probability 1/8, a dot sent is erroneously received as a dash and vice-versa. If a dot is received, what is the probability that a dot was sent?*

Let R denote the event "a dot is received", we wish to calculate the posterior probability $p(D|R)$, which, by Bayes' Theorem, is:

$$p(D|R) = \frac{p(R|D)p(D)}{p(R)}.$$

The probabilities $p(R|D)$ and $p(D)$ are known. We only need to compute the marginal probability $p(R)$. We know that $p(\bar{R}|D) = 1/8$ as well as $p(R|\bar{D}) = 1/8$. Thus, we can compute $p(\bar{R}|\bar{D}) = 7/8$, and

$$p(R) = p(R|D)p(D) + p(R|\bar{D})p(\bar{D}) = \frac{7}{8}\frac{3}{7} + \frac{1}{8}\frac{4}{7} = \frac{25}{56}.$$

Thus, $p(D|R) = 21/25$.

6. Independence and Conditional Independence

When we apply Bayes' Theorem, we assume that the occurrence of an event B changes the probability of A. If, on the other hand, occurrence of B does not change the probability of A, then A and B are independent events. Thus, by definition we have

$$p(A) = p(A|B) = \frac{p(A,B)}{p(B)}$$

from which we derive that $p(A,B) = p(A)p(B)$.

Definition 5 (Independence) *Two events A and B, with $p(A) > 0$ and $p(B) > 0$ are independent if*

$$p(A,B) = p(A)p(B).$$

In words, the joint probability of two independent events factorizes into the product of the marginal probabilities. If A and B are independent, so are A and $\bar{B}$, $\bar{A}$ and B, as well as $\bar{A}$ and $\bar{B}$. For example, $p(\bar{A}, \bar{B}) = p(\mathcal{S} - (A \cup B))$, where $\mathcal{S}$ denotes the sample space, and $p(\mathcal{S} - (A \cup B)) = 1 - p(A \cup B) = (1 - p(A))(1 - p(B))$. The proof of the other properties is left as an exercise. Note also that exclusive events cannot be independent because $0 = p(A,B) = p(A)p(B)$ if and only if either $p(A) = 0$ or $p(B) = 0$.

Example 4 (Independence) *Consider an urn, containing r red balls and b black balls. A ball is drawn at random and then replaced, then a second ball is drawn at random. Let R_i be the event: "the i-th ball drawn is red", and let B_i be the event: "the i-th ball drawn is black". Since, at each drawning, the composition of the urn is unchanged, we have*

$$p(B_1) = \frac{b}{r+b} = p(B_2|B_1)$$

from which there follows that B_1 and B_2 are independent. For fixed i, the events R_i and B_i are exhaustive. We thus have independence of R_1 and B_2, R_1 and R_2, and B_1 and R_2.

The events $A_1, A_2, \ldots, A_n$ are called *pair-wise independent*, if $p(A_i, A_j) = p(A_i)p(A_j)$ for all possible pairs. When the factorization rule applies for all intersection of subsets of events, then $A_1, A_2, \ldots, A_n$ are called *mutually independent*. Pair-wise independence does not generally imply mutual independence. For example, suppose that A_1, A_2, A_3 are pair-wise independent, $p(A_i) > 0$ for all i and $A_1, A_2, A_3 = \emptyset$, then $p(A_1, A_2, A_3) = 0$ and the factorization rule does not apply.

Definition 6 (Conditional Independence) *Two events A and B, with $p(A) > 0$ and $p(B) > 0$ are* independent*, given C, if*

$$p(A, B|C) = p(A|C)p(B|C).$$

Independence does not imply conditional independence, so that it may happen that two events A and B are independent given C, but they are not independent when considered alone.

Example 5 (Conditional Independence and Marginal Dependence) *Let A, B conditional independent events given C, and suppose that $p(C) = 1/2$, $p(A|C) = 1/3$, $p(A|\bar{C}) = 1/2$, and $p(B|C) = 1/2$, $p(B|\bar{C}) = 1/3$. By the total probability theorem, we have $p(A) = p(A|C)p(C) + p(A|\bar{C})p(\bar{C}) = 5/12$ and, similarly, $p(B) = 5/12$. Now, we compute the joint probability $p(A, B)$ as*

$$p(A, B) = p(A, B|C)p(C) + p(A, B|\bar{C})p(\bar{C})$$

and by the conditional independence of A and B given C,

$$p(A, B|C)p(C) + p(A, B|\bar{C})p(\bar{C}) = p(A|C)p(B|C)p(C) + p(A|\bar{C})p(B|\bar{C})p(\bar{C}) = 1/6.$$

There follows that $p(A, B) \neq p(A)p(B)$ and hence the two events are not independent.

It may also happen that two events A and B are independent, but they are not independent given C.

Example 6 (Marginal Independence and Conditional Dependence) *Let A, B independent events, with $p(A) = 1/2$ and $p(B) = 2/3$. By independence, there follows that $p(A, B) = 1/3$. Let now C be an event with $p(C) = 1/4$, and suppose that $p(A|C) = p(B|C) = 1/3$, and $p(A|\bar{C}) = 5/9$ and $p(B|\bar{C}) = 7/9$. It is easy to check that this conditional probabilities are such that $p(A) = p(A|C)p(C) + p(A|\bar{C})p(\bar{C})$ and $p(B) = p(B|C)p(C) + p(B|\bar{C})p(\bar{C})$. Suppose now that $p(A, B|C) = 1/6$ and that $p(A, B|\bar{C}) = 7/18$. By the total probability theorem, we have $p(A, B|C)p(C) + p(A, B|\bar{C})p(\bar{C}) = 1/24 + 7/24 = 1/3 = p(A, B)$, so that the conditional probabilities are consistently defined. However, $p(A, B|C) = 1/6 \neq 1/9 = p(A|C)p(B|C)$. There follows that A and B are independent by definition but we have found an event C such that $A|C$ and $B|C$ are not independent.*

7. Discrete Random Variables

Suppose the events $A_1, A_2, \ldots, A_n$ form a partition of $\mathcal{S}$, hence $p(A_i) \geq 0$ and $\sum_{i=1}^{n} p(A_i) = 1$. We can associate each event A_i with the value of a discrete variable X, for instance, by setting $X = i$ if A_i occurs. In this way, we construct a variable which maps the sample space $\mathcal{S}$ into the integer

x_i	$p(X = x_i)$
0	$1 - p$
1	p

Table 1: *Probability mass function of the binary variable X*

numbers $1, 2, \ldots, n$. The uncertain occurrence of the events $A_1, A_2, \ldots, A_n$ is then inherited by the variable X, and we can write $p(X = i) = p(A_i)$, or $p(X = x_i) = p(A_i)$. Because this variable X takes on values with some probability, it is called a *random variable.* A random variable is defined by the values x_i — they are also called categories, or states — and the probabilities $p(X = x_i)$. The set of pairs $x_i, p(X = x_i)$ defines the probability mass function of X (p.m.f.). Note that, since $A_1, A_2, \ldots, A_n$ form a partition of $\mathcal{S}$, then $p(A_1 \cup A_2 \cup \ldots \cup A_n) = 1$. We then have that $\sum_{i=1}^{n} p(X = i) = 1$ and the p.m.f. is *normalized.* Random variables that take a finite (or countable) number of states are referred to as *discrete random variables.*

Example 7 (Bernoulli Trials) *During an opinion poll, British people are asked to whether they favor a common European market or not. Let E denote the event that the answer of a person is yes, and let $p(E) = p$. By coding the two types of answers into 1 (yes) and 0 (no), we define the binary variable X. The* p.m.f. *is reported in Table 1. Experiments as the one described in this example, in which the outcome is binary, are called* Bernoulli trials.

Associating events with values of a random variable enables us to compute summaries — description — of the phenomenon under investigation. For example, in the opinion poll described in Example 7, a sample of n people may be interviewed, and the interest of the interviewer may be in the number of people who answer yes. In this case, we have $X_1, X_2, \ldots, X_n$ random variables, where X_i codes the answer of the ith interviewed, and $X = \sum_{i=1}^{n} X_i$ is the frequency of positive answers.

Example 8 (Binomial Distribution) *Suppose that a sample of 3 British people are asked to whether they favor a common European market or not and let X_1, X_2, X_3 be the variables coding the 3 answers. Let $X = \sum_{i=1}^{3} X_i$ be the frequency of 1s in the sample, that is the frequency of positive answers. The first column in Table 2 reports the eight possible ordered combinations c_i of answers in the sample. The second column reports the value of X. Suppose also that the answers given by different individuals are independent, and that every individual gives a positive answer with the same probability p. By independence, we can then compute the probability of each value of X. These probabilities are in the last column of Table 2.*

The range of values of X is given by the integers $x = 0, 1, 2, 3$, and the probability that $X = x$ is independent of the order of the answers, e.g. $p(X = 1) = p(1 - p)^2$ for the three possible orders of answers $E, \bar{E}, \bar{E}$, $\bar{E}, E, \bar{E}$ and $\bar{E}, \bar{E}, E$. By grouping combinations of answers that have the same probability of occurrence, we have

$$p(X = x) = \frac{3!}{x!(3 - x)!} p^x (1 - p)^{3-x} \qquad x = 0, 1, 2, 3$$

where $3!/[x!(3 - x)!]$ is the number of different sequences of 3 elements containing E x times. This random variable is said to have a Binomial distribution.

c_i	x	$p(X = x)$
$\bar{E}, \bar{E}, \bar{E}$	0	$(1-p)^3$
$E, \bar{E}, \bar{E}$	1	$p(1-p)^2$
$\bar{E}, E, \bar{E}$	1	$p(1-p)^2$
$\bar{E}, \bar{E}, E$	1	$p(1-p)^2$
$E, E, \bar{E}$	2	$p^2(1-p)$
$E, \bar{E}, E$	2	$p^2(1-p)$
$\bar{E}, E, E$	2	$p^2(1-p)$
E, E, E	3	p^3

Table 2: *Probability mass function of the variable* $X = \sum_{i=1}^{3} X_i$

Definition 7 (Binomial Distribution) *Let* X_i $(i = 1, ..., n)$ *be* n *binary random variables associated to* n *independent Bernoulli trials, with a common* $p(X_i = 1) = p$ *for all* i*. Then, the variable* $X = \sum_{i=1}^{n} X_i$ *has a Binomial distribution whose* p.m.f. *is*

$$p(X = x) = \frac{n!}{n!(n-x)!} p^x (1-p)^{n-x} \qquad x = 0, ..., n$$

where $n!/[x!(n-x)!]$ *is the* Binomial coefficient*, which gives the number of ways* n *objects can be grouped into two classes having* x *and* $n - x$ *objects respectively, and* $p^x(1-p)^{n-x}$ *is the probability of observing one such ordered combination. The distribution of this variable will be denoted as*[1] $X \sim Bin(n, p)$*. The quantities* n *and* p *are* parameters *that identify a specific Binomial distribution.*

In general, we will denote the range of values taken by the variable as $\{x_1 < x_2 \ldots < x_n\}$, and the p.m.f. as $p(X = x_i) = p(x_i)$. The p.m.f. describes the probability of individual events. In many applications, we may be interested in evaluating the probability of the union of events, as $p(X \leq x_j)$ or $p(x_i < X \leq x_j)$ for some given x_i, x_j. Note first the difference between the events $X < x_j$ and $X \leq x_j$. The former is the event $X = x_1 \cup X = x_2 \cup ... \cup X = x_{j-1}$ — as $X < x_j$ denotes the event that X takes a value in the set $\{x_1 < x_2 \ldots < x_{j-1}\}$ — while the latter is the event $X = x_1 \cup X = x_2 \cup ... \cup X = x_j$, since the value x_j can be taken, as well as any of the values in $\{x_1 < x_2 \ldots < x_{j-1}\}$. The states of the variable X are associated with exclusive events and hence the probability of the event $X = x_1 \cup X = x_2 \cup ... \cup X = x_j$ is $\sum_{i=1}^{j} p(x_i)$. The value of $p(x_i < X \leq x_j)$ is the probability of the event $X = x_{i+1} \cup X = x_{i+2} \cup ... \cup X = x_j$ and it can be written as $p(X \leq x_j) - p(X \leq x_i)$. there follows that the evaluation of complex events can be done by evaluating first the cumulative probabilities, say $p(X \leq x_j)$ and $p(X \leq x_i)$, and by then manipulating them. When x is a variable, the function $p(X \leq x)$ is called the cumulative distribution function (c.d.f). The c.d.f fully characterizes a random variable (Casella and Berger, 1990).

Definition 8 (Cumulative Distribution Function (c.d.f)) *Let* X *be a discrete random variable. The* cumulative distribution function *is defined as* $F_X(x) = p(X \leq x)$*, for all* $x \in (-\infty, +\infty)$*.*

For any $x_i \leq x < x_{i+1}$ the c.d.f is given by $F_X(x) = p(X \leq x) = \sum_{j=1}^{i} p(x_j)$. Clearly $F_X(x) = 0$ for all $x < x_1$, because X cannot take values smaller than x_1, and $F_X(x) = 1$ for all $x \geq x_n$, since, with

[1] the symbol $\sim$ is read as "distributed as"

x	$F_X(x)$
$(-\infty, 0)$	0
$[0, 1)$	$(1-p)^3$
$[1, 2)$	$(1-p)^3 + 3p(1-p)^2$
$[2, 3)$	$(1-p)^3 + 3p(1-p)^2 + 3p^2(1-p)$
$[3, \infty)$	$(1-p)^3 + 3p(1-p)^2 + 3p^2(1-p) + p^3 = (p+1-p)^3 = 1$

Table 3: *Cumulative distribution function of the variable* $X = \sum_{i=1}^{3} X_i$.

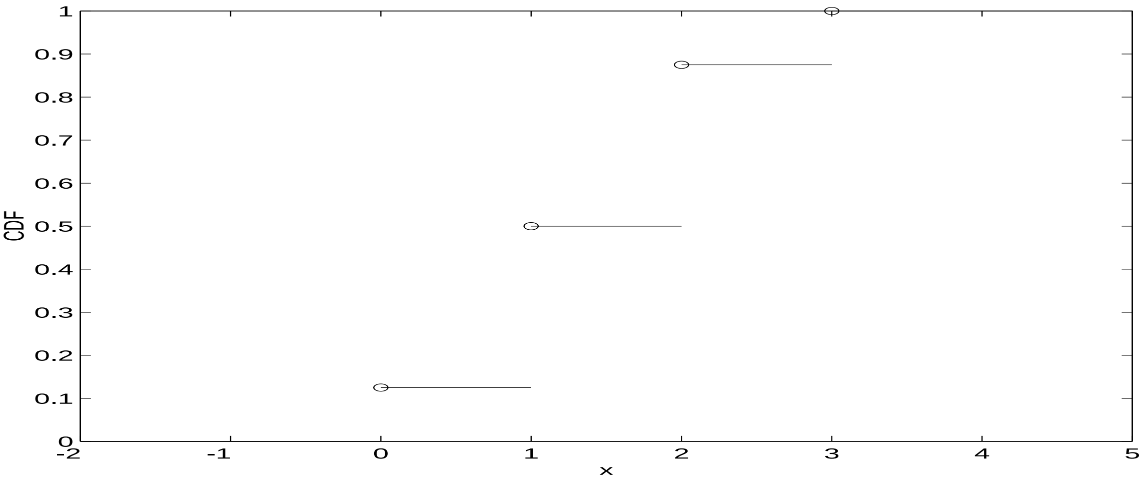

Figure 2: Graphical representation of the c.d.f in Table 7 for $p = 0.5$. Circles identify the value of the c.d.f at the jump points.

probability 1, X takes one of the values $x_1, \ldots, x_n$. The Table 7 reports the c.d.f of the variable X of Example 8. Figure 2 gives a graphical representation of the c.d.f in Table 7 for $p = 0.5$. The plot shows that the c.d.f is a step function with jumps that occur at the points x_i. At the jump point x_i, the c.d.f takes [2] the value $\sum_{j=1}^{i} p(x_j)$. The size of the jump is $p(x_i)$. Note that the c.d.f is a non-decreasing function.

The properties highlighted in this example hold in general and characterize a c.d.f as shown in the next Theorem.

Theorem 2 *A function $F(x)$ is a* c.d.f *of a variable X if and only if:*

1. $\lim_{x\to-\infty} F(x) = 0$ *and* $\lim_{x\to+\infty} F(x) = 1$;
2. *$F(x)$ is a non-decreasing function of x;*

[2] Technically, this property is called right-continuity (Casella and Berger, 1990)

3. $F(x)$ is right-continuous, that is the right-limit $\lim_{x \downarrow x_i} F(x) = F(x_i)$.

The examples described so far were focused on discrete random variables taking on a finite number of states. However, a discrete random variable can also take a countable number of states, e.g. the range of values can be the set of integer number $\mathcal{N} = \{0, 1, 2, ...\}$.

Example 9 (Poisson Distribution) *Let X be a discrete random variable taking values in the set $\mathcal{N}$ of integer numbers with probability*

$$p(X = x) = \frac{\lambda^x}{x!} e^{-\lambda} \quad x = 0, 1, 2, ...$$

It can be shown that $\sum_{x=0}^{\infty} p(X = x) = 1$. A random variable having this p.m.f. *is said to have a* Poisson distribution *and it is usually denoted by $P(\lambda)$. This distribution is usually associated with count data, and it can also be derived as limit of the Binomial distribution for $n \to \infty$. See Wilks (1963) for details.*

8. Continuous Random Variables

Theorem 2 does not exclude the possibility that the c.d.f of a random variable X be a continuous function. In this case, X is said to be *continuous* and it is assumed to be defined over the all real line.

Definition 9 (Density Function) *Let $F_X(x)$ be a continuous function satisfying the conditions stated in Theorem 2. The associated variable X is then called a continuous random variable. The function $f(\cdot)$ that verifies*

$$F(x) = \int_{-\infty}^{x} f(t)dt$$

is termed the density function *(*d.f.*) and has the properties of being non-negative ($f(x) \geq 0$ for all x in $(-\infty, \infty)$) and normalized ($\int_{-\infty}^{\infty} f(x)dx = 1$.)*

A continuous random variable usually represents events related to measurements. Examples are the variable representing the height of white males in United Kingdom the 31st of December 1998 or the variable representing the length of an object measured with an imprecise instrument. In both cases, the variables take real values and we need to be precise about the meaning of the event $X = x$. For example, the integer $x = 1$, or the rational $x = 1/2$ are examples of real numbers. But also $x = \sqrt{2} = 1.4142...$ or $x = \pi = 3.14...$ are examples of real numbers and the notation adopted is a conventional way to represent the fact that both $x = \sqrt{2} = 1.4142...$ or $x = \pi = 3.14...$ have an expression as infinite decimal numbers. What do we mean by the probability of the event "the height of a white male is 178.8976...cm" or "the length of the object is 3.9876...m"? Both events are so rare that their probability goes to zero. On the other hand, it is non-trivial to consider probability of events as "the height of a white male is between 178cm and 179cm" or "the length of the object is between 3.9m and 4.0m" and they both increase with the range of possible values. Hence, a non-zero probability is associated with intervals and not points and the limiting range of values we can consider is the infinitesimal interval dx. Thus, with continuous variables, we regard the event $X = x$ as actually the event that X takes values in the range $x, x + dx$ and $\int_{x}^{x+dx} f(x)dx$

is its probability. A consequence of this assumption is that the events $a \leq X < b$ and $a < X < b$ will have the same probability — while we have to treat them as different when X is discrete — since we are assuming that $X = a$ is essentially $a < X < a + dx$.

The probability of an event $a \leq X < b$ is computed by using the c.d.f. Hence, by writing $p(a \leq X < b) = p(X < b) - p(X \leq a)$ and using the definition of c.d.f, we have $p(a \leq X < b) = \int_a^b f(x)dx$. By plotting the d.f. against the x values, the probability of $a \leq X < b$ is represented by the area between the density function and the x-axis, with $a \leq x < b$.

Example 10 (Uniform Distribution) *The function*

$$f(x) = \begin{cases} 1 & \textit{if} \quad 0 \leq x \leq 1 \\ 0 & \quad\;\; \textit{otherwise} \end{cases}$$

defines a d.f.*, being non-negative and normalized as* $\int_{-\infty}^{\infty} f(x)dx = \int_0^1 dx = 1$*. The associated random variable* X *is said to have a* Uniform *distribution over the range 0–1, and sometimes it is denoted by* $X \sim U(0,1)$*. It is straightforward to show that this distribution assigns probability* $b - a$ *to an event* $a \leq X < b$*. The range of definition of a uniform distribution can be generalized to any finite interval* $[\alpha, \beta]$ *by defining the* d.f. $f(x) = 1/(\beta - \alpha)$*. In this case, the distribution will be denoted by* $U(\alpha, \beta)$ *and it assigns probability* $(b-a)/(\beta - \alpha)$ *to any event* $a \leq X \leq b$*.*

The c.d.f of a continuous random is in a one-to-one relationship with the d.f. and, often, continuous random variables are described in terms of their d.f.. We note that the properties of a d.f. of being non-negative and normalized are also sufficient to characterize a d.f.. A proof of this statement is given for instance by Karr (1992).

Example 11 (Normal Distribution) *A continuous random variable* X *defined in* $(-\infty, \infty)$ *and with* d.f.*:*

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/(2\sigma^2)}$$

is said to have a Normal distribution *— or Gaussian distribution — with parameters* μ *and* σ^2*. The parameters* μ *and* σ^2 *identify, respectively, the point of symmetry of the distribution — the* d.f. *is symmetrical about* μ *— and the "spread" of the* d.f. *about* μ*. When* $\mu = 0$ *and* $\sigma^2 = 1$ *the distribution is called* Normal Standard*. Examples are in Figure 3 that plots three* d.f. *for* $\mu = 0$ *and increasing values of* σ^2 *(*$\sigma^2 = 0.25$ *dotted line;* $\sigma^2 = 1$ *continuous line;* $\sigma^2 = 2$ *dashed line.) Small values of* σ^2 *make the* d.f. *more concentrated about* μ*, while large values of* σ^2 *enlarge the spread of the* d.f.*. Consider the event* $X \leq -1$ *that, by symmetry, will have the same probability of the event* $X \geq 1$*. The area between the x-axis, the* d.f. *and the line* $x = -1$ *represents* $p(X \leq -1)$*, and this quantity will decrease as* σ^2 *decreases. On the other hand,* $p(-1 \leq X \leq 0)$ *— represented by the area between the*d.f.*, the x-axis and the lines* $x = -1$ *and* $x = 0$ *— will increase as* σ^2 *decreases. The symmetry of the* d.f. *ensures also that* $p(X \leq \mu) = p(X \geq \mu) = 0.5$*.*

The c.d.f *is the integral function*

$$\Phi(x) = \int_{-\infty}^{x} f(t)dt$$

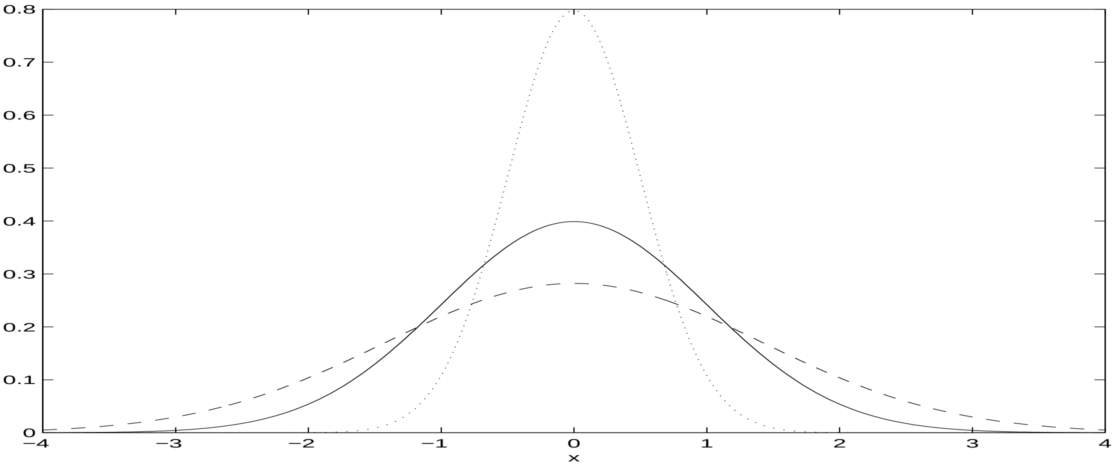

Figure 3: Density of Normal distributions with $\mu = 0$ and different σ^2: $\sigma^2 = 1$ continuous line; $\sigma^2 = 0.25$ dotted line; $\sigma^2 = 2$ dashed line.

that has no closed form solution. Hence, numerical integration *techniques are required to evaluate probability of events of interest.*

The Normal distribution can be also derived as limiting distribution from a Binomial distribution for $n \to \infty$ via the *Central Limit Theorem.* We state the Central Limit Theorem only to show that the limiting distribution of a Binomial random variable is a Normal distribution. More details are in Feller (1968).

Theorem 3 (Central Limit Theorem) *Let $X \sim (Bin(n,p)$. Then*

$$\lim_{n\to\infty} p(X \leq x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi[np(1-p)]^2}} e^{-(x-np)^2/(2[np(1-p)]^2)}$$

Hence, the limiting c.d.f *is that of a Normal distribution with parameters $\mu = np$ and $\sigma^2 = np(1-p)$.*

A family of p.m.f.s is called an *Exponential family* with parameter θ if the d.f. can be written:

$$f(x) = h(x)g(\theta)e^{\sum_{i=1}^{k} w_i(\theta)t_i(x)}$$

The exponential family contains many important discrete and continuous random variables. In the former case, the d.f. represents the p.m.f.. It is left as an exercise to show that the distributions in Table 8 belong to the exponential family.

A complete treatment of continuous random variables is given by Casella and Berger (1990).

Distribution	θ	$h(x)$	$g(\theta)$	$w_i(x)$	$t_i(\theta)$
$Bin(n,p)$	p	$\frac{n!}{x!(n-x)!}$	$(1-p)^n$	$w_1(\theta) = \log \frac{p}{(1-p)}$	$t_1(x) = x$
$P(\lambda)$	λ	$\frac{1}{x!}$	$e^{-\lambda}$	$w_1(\lambda) = \lambda$	$t_1(x) = 1$
$N(\mu, \sigma^2)$	(μ, σ^2)	1	$\frac{e^{-\mu^2/(2\sigma^2)}}{\sqrt{2\pi\sigma^2}}$	$w_1(\theta) = \frac{-1}{2\sigma^2}$	$t_1(x) = x^2$
				$w_2(\theta) = \frac{\mu}{\sigma^2}$	$t_2(x) = x$

Table 4: *Distributions that belong to the exponential family.*

9. Multivariate Distributions

The variables described so far are univariate variables that take on values in a subset of the real numbers. We can generalize the definition of random variables to vectors. A vector in R^c is a set of c univariate variables, and we will denote it by $\boldsymbol{X} = (X_1, ..., X_c)$. A vector $\boldsymbol{X}$ takes values $\boldsymbol{x} = (x_1, ..., x_c)$ in a subset of R^c. Note that we will use bold-face to denote a vector. The components of a vector will have an index identifying their position. When we refer to a particular value, as $\boldsymbol{X} = \boldsymbol{x}_j$, then we will use the notation $\boldsymbol{x}_j = (x_{1j}, ..., x_{cj})$.

Definition 10 (Random Vectors) *A* random vector $\boldsymbol{X} = (X_1, ..., X_c)$ *is a vector whose components X_i are univariate random variables. If X_i are all discrete, then $\boldsymbol{X}$ is a discrete random vector. If X_i are all continuous, $\boldsymbol{X}$ is called a continuous random vector.*

The distribution of a random vector is characterized by the joint c.d.f that is defined as

$$p(\boldsymbol{X} \leq \boldsymbol{x}) = F(\boldsymbol{x}) = p(X_1 \leq x_1, ..., X_c \leq x_c) \quad \mathbf{x} = (x_1, ..., x_c)$$

Note that $(X_1 \leq x_1, ..., X_c \leq x_c)$ is the intersection of events $X_i \leq x_i$, as it represents the event that the components X_i satisfy simultaneously the condition $X_i \leq x_i$, for all i. We can also describe the joint distribution of a random vector in terms of the joint p.m.f. or d.f., that are defined as

$$\begin{array}{ll} \text{discrete} & p(\boldsymbol{x}) = p(X_1 = x_1, ..., X_c = x_c) \\ \text{continuous} & f(\boldsymbol{x}) = f(x_1, ..., x_c) \end{array}$$

and, again, $X_1 = x_1, ..., X_c = x_c$ is the event that $X_i = x_i$ simultaneously, i.e. for all i. The p.m.f. obeys the rule that $\sum_{\boldsymbol{x}} p(\boldsymbol{x}) = 1$, while the d.f. is such that $\int_{-\infty}^{\infty} ... \int_{-\infty}^{\infty} f(x_1, ..., x_c) dx_1 ... dx_c = 1$. Hence, they are both normalized.

Example 12 (A Bivariate Distribution) *Let X_1 and X_2 be two binary variables, taking values, say, 1 and 2. The joint* p.m.f. *of $\boldsymbol{X} = (X_1, X_2)$ can be represented by a 2×2 table, containing the probability $p(X_1 = x_i, X_2 = x_j)$, $(i, j = 1, 2)$, and $\sum_{i=1}^{2} \sum_{j=1}^{2} p(X_1 = x_i, X_2 = x_j) = 1$. An example is in Table 12.*

From the joint p.m.f., we can derive the *marginal distribution* of each component — or set of components — of the vector $\boldsymbol{X}$. Consider, for example, the p.m.f. in Table 12. The first row of the Table gives the probabilities $p(X_1 = 1, X_2 = 1)$ and $p(X_1 = 1, X_2 = 2)$. Summing up these two probabilities $p(X_1 = 1, X_2 = 1) + p(X_1 = 1, X_2 = 2)$ is equivalent to applying the Total Probability Theorem in (5) to the events $X_1 = 1, X_2 = 1$ and $X_1 = 1, X_2 =$, so that $p(X_1 = 1, X_2 = 1) + p(X_1 = 1, X_2 = 2)$

	X_2	
X_1	1	2
1	$p(X_1 = 1, X_2 = 1)$	$p(X_1 = 1, X_2 = 2)$
2	$p(X_1 = 2, X_2 = 1)$	$p(X_1 = 2, X_2 = 2)$

Table 5: *Joint* p.m.f. *of the random vector* $\mathbf{X} = (X_1, X_2)$

	X_2		Marginal
X_1	1	2	
1	$p(X_1 = 1, X_2 = 1)$	$p(X_1 = 1, X_2 = 2)$	$p(X_1 = 1)$
2	$p(X_1 = 2, X_2 = 1)$	$p(X_1 = 2, X_2 = 2)$	$p(X_1 = 2)$
Marginal	$p(X_2 = 1)$	$p(X_2 = 2)$	

Table 6: *Joint* p.m.f. *of the random vector* $\mathbf{X} = (X_1, X_2)$*, and marginal* p.m.f.*s of* X_1 *and* X_2*.*

is the marginal probability of $X_1 = 1$. Similarly, $p(X_1 = 2, X_2 = 1) + p(X_1 = 2, X_2 = 2) = p(X_1 = 2)$. Because the joint p.m.f. of X_1 and X_2 is normalized, the probabilities $p(X_1 = 1)$ and $p(X_1 = 2)$ sum up to one and define the p.m.f. of the univariate variable X_1. We can also derive the marginal distribution of X_2 as $p(X_2 = 1) = p(X_1 = 1, X_2 = 1) + p(X_1 = 2, X_2 = 1)$ and $p(X_2 = 2) = p(X_1 = 1, X_2 = 2) + p(X_1 = 2, X_2 = 2)$. The marginal p.m.f.s of both X_1 and X_2 can be displayed in Table 12 by adding a column and row as in Table 9.

When the random vector $\boldsymbol{X}$ has more than two components, we can also talk about the marginal distribution of a random vector having a subset of the components of $\boldsymbol{X}$. For instance, if $\boldsymbol{X} = (X_1, X_2, X_3)$ is a random vector in R^3, then (X_1, X_2) is a random vector in R^2. The p.m.f. of (X_1, X_2) is defined as $p(X_1 = x_{1i}, X_2 = x_{2j}) = \sum_k p(X_1 = x_{1i}, X_2 = x_{2j}, X_3 = x_{3k})$ for all i and j. We can now give a formal definition of marginal distribution for discrete random variables.

Definition 11 (Marginal Distribution: The Discrete Case) *Let* $\boldsymbol{X} = (X_1, ..., X_c)$ *be a discrete random vector, with* p.m.f. $p(\boldsymbol{X} = \boldsymbol{x})$*. The marginal distribution of any vector* $(X_1, ..., X_s)$ $(s < c)$ *has* p.m.f.

$$p((X_1, ..., X_s) = (x_1, ..., x_s)) = \sum_{\boldsymbol{x}_{s+1}, ..., \boldsymbol{x}_c} p(X_1 = x_1, ..., X_s = x_s, ..., X_c = x_c)$$

where $\boldsymbol{x}_{s+1}, ..., \boldsymbol{x}_c$ *denotes all values that* $X_{s+1}, ..., X_c$ *can take.*

Example 13 (Multinomial Distribution) *Let* X *be a discrete variable taking values* $1, ..., c$ *with probability* $p(X = i) = p_i$ *and* $\sum_{i=1}^{c} p_i = 1$*. Suppose we can observe the outcomes of* n *independent observations of the variable* X*, and that we are interested in the probability of the possible distributions of the* n *values into the* c *possible categories of* X*. We shall denote one such outcome as* $\boldsymbol{x} = (x_1, ..., x_c)$*,* $\sum_i x_i = n$*, and define by* $\boldsymbol{X}$ *the random vector that takes values* $\boldsymbol{x}$*. Then, it can be shown that the* p.m.f. *of* $\boldsymbol{X}$ *is*

$$p(\boldsymbol{X} = \boldsymbol{x}) = \frac{n!}{x_1!...x_c!} \prod_{i=1}^{c} p_i^{x_i} \tag{7}$$

where $n!/(x_1!...x_c!)$ is the multinomial coefficient *that represents the number of ways in which n objects can be grouped into c classes, each containing x_i objects, and $\prod_{i=1}^{c} p_i^{x_i}$ is the probability of one such ordered distribution of values. A random vector having* p.m.f. *as in (7) is said to have a* Multinomial distribution *with parameters n and $\boldsymbol{p} = (p_1, ..., p_c)$ and we will denote it as $\boldsymbol{X} \sim Mult(n, \boldsymbol{p})$. Note that if X is a bivariate distribution then, by setting $1 - p_1 = p_2$ and $n - x_1 = x_2$, we have*

$$p(\boldsymbol{X} = \boldsymbol{x}) = \frac{n!}{x_1!(n - x_1)!} p_1^{x_1} (1 - p)^{n - x_1}$$

that characterizes a Binomial distribution. A property of a multinomial distribution is that every individual component X_i has a Binomial distribution $Bin(n, p_i)$. A proof of this result can be found in Casella and Berger (1990), page 173.

From the joint and marginal p.m.f.s we can then derive the conditional probability $p(X_2 = 1|X_1 = 1) = p(X_1 = 1, X_2 = 1)/p(X_1 = 1)$ as well $p(X_2 = 2|X_1 = 1) = p(X_1 = 1, X_2 = 2)/p(X_1 = 1$. Note that $p(X_2 = 1|X_1 = 1) + p(X_2 = 2|X_1 = 1) = 1$ and hence the two probabilities $p(X_2 = 1|X_1 = 1), p(X_2 = 2|X_1 = 1)$ define a p.m.f.. Since $p(X_2 = 1|X_1 = 1), p(X_2 = 2|X_1 = 1)$ represent conditional probabilities when $X_1 = 1$, the distribution having p.m.f. defined by $p(X_2 = 1|X_1 = 1), p(X_2 = 2|X_1 = 1)$ is called the *conditional distribution* of X_2 given $X_1 = 1$. We will use the notation $X_2|X_1 = 1$ to denote the conditional variable. We can similarly define the conditional distribution of $X_2|X_1 = 2$ via the p.m.f. $p(X_2 = 1|X_1 = 2), p(X_2 = 2|X_1 = 2)$ as well as the two conditional distributions of X_2 given $X_1 = 1, 2$ whose p.m.f.s will respectively be $p(X_1 = 1|X_2 = 1), p(X_1 = 2|X_2 = 1)$ and $p(X_1 = 1|X_2 = 2), p(X_1 = 2|X_2 = 2)$. The definition can be extended to the case in which the condition event is the intersection of several events, and it is represented by $\boldsymbol{X} = \boldsymbol{x}$.

Definition 12 (Conditional Distribution: The Discrete Univariate Case) *Let X_1 be a univariate random variable, and let $\boldsymbol{X}$ be a random vector. The conditional distribution $X_1|\boldsymbol{X} = \boldsymbol{x}$ has* p.m.f.

$$p(X_1 = x_1|\boldsymbol{X} = \boldsymbol{x}) = \frac{p(X_1 = x_1, \boldsymbol{X} = \boldsymbol{x})}{p(\boldsymbol{X} = \boldsymbol{x})}. \tag{8}$$

Definition 12 defines a univariate conditional distribution. The extension to multivariate conditional distributions is obtained by considering

$$p(\boldsymbol{X}_1 = \boldsymbol{x}_1|\boldsymbol{X} = \boldsymbol{x}) = \frac{p(\boldsymbol{X}_1 = \boldsymbol{x}_1, \boldsymbol{X} = \boldsymbol{x})}{p(\boldsymbol{X} = \boldsymbol{x})} \tag{9}$$

where now $\boldsymbol{X}_1$ is a random vector.

Example 14 (Multinomial Case) *Let $\boldsymbol{X} = (X_1, ..., X_c)$ be a random vector having a Multinomial distribution $Mult(n, \boldsymbol{p})$. A property of a multinomial distribution is the closure under conditioning, i.e. $(X_1, ..., X_{c-1})|X_c = x_c$ has a multinomial distribution with parameters $n - x_c$ and $(p_1/p_c, ..., p_{c-1}/p_c)$ (?)page 173]Casella90.*

For continuous random vectors, the d.f. is a multivariate function, that is used to compute the joint probability of an event as the solution of a multiple integral:

$$p(X_1 \leq x_1, ..., X_c \leq x_c) = \int_{-\infty}^{x_1} ... \int_{-\infty}^{x_c} f(t_1, ..., t_c) dt_1 ... dt_c$$

Example 15 (A Bivariate Normal) *Let $\boldsymbol{X} = (X_1, X_2)$ be a bivariate continuous random vector. Define the column vectors*

$$\boldsymbol{x} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \quad \boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}$$

and the 2×2 symmetric matrix

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{12}^2 \\ \sigma_{12}^2 & \sigma_2^2 \end{pmatrix}$$

Now, we use these parameters to define the d.f.*:*

$$f(x_1, x_2) = \frac{1}{2\pi \det \Sigma} e^{-(\boldsymbol{x}-\boldsymbol{\mu})^T \Sigma^{-1} (\boldsymbol{x}-\boldsymbol{\mu})/2}$$

where $(\boldsymbol{x} - \boldsymbol{\mu})^T$ is a row vector. Then, the vector $\boldsymbol{X} = (X_1, X_2)$ is said to have a bivariate Normal distribution with multivariate parameters *μ and Σ. From this* d.f.*, we can evaluate the probability, say, $p(X_1 < 0, X_2 < 1)$ as*

$$p(X_1 < 0, X_2 < 1) = \int_{-\infty}^{0} \int_{-\infty}^{1} f(x_1, x_2) dx_1 dx_2$$

Note that the integral can only be solved using numerical integration techniques (Evans and Swartz, 1995).

When the components of $\boldsymbol{X}$ are continuous, the marginal distribution of a vector having, as elements, a subset of components of $\boldsymbol{X}$ is usually defined in terms of the d.f.

Definition 13 (Marginal Distribution: The Continuous Case) *Let $\boldsymbol{X} = (X_1, ..., X_c)$ be a continuous random vector, with* d.f. *$f(\boldsymbol{x})$. The marginal density of $(X_1, ..., X_s)$ is defined as*

$$f(x_1, ..., x_s) = \int_{-\infty}^{\infty} ... \int_{-\infty}^{\infty} f(\boldsymbol{x}) dx_{s+1} ... dx_c.$$

Example 16 (Closure of Multivariate Normal Distribution) *The multivariate Normal distribution has the same property of being closed under marginalization that we have seen for the Multinomial distribution. Consider, for instance, the bivariate normal distribution described in Example 15. Then, it can be shown that each individual component X_i has a univariate Normal distribution with parameters μ_i, σ_i. The more general case is considered in depth by (Whittaker, 1990).*

Suppose now X_1 and $\boldsymbol{X}$ are both continuous random variable and vector respectively. The distribution of $X_1|\boldsymbol{X} = \boldsymbol{x}$ is defined in terms of the *conditional density*

$$f(x_1|\boldsymbol{x}) = \frac{f(x_1, \boldsymbol{x})}{f(\boldsymbol{x})}$$

The definition of marginal and conditional distributions enable us to write the joint p.m.f. as a product of marginal and conditional p.m.f.s. For instance, from (8) we have

$$p(X_1 = x_1, \boldsymbol{X} = \boldsymbol{x}) = p(\boldsymbol{X} = \boldsymbol{x})p(X_1 = x_1|\boldsymbol{X} = \boldsymbol{x})$$

10. Summaries

It is common to give some summary measures of a random variable, usually the expected value and the mode, and some measure of the variability about these summaries. The expected value of a random variable is an average of the values that the variable can take, weighted by their probabilities. In summarizing a random variable via its expectation, we aim to give a reference value that is the quantity we expect the variable to take.

Definition 14 (Expected Value) *Let X be a discrete random variable taking values x_k with probability $p(x_k)$, $(k = 1, ..., n)$. The* expected value *of X is defined as*

$$E(X) = \sum_{i=k}^{n} x_k p(x_k)$$

If X is a continuous random variable with d.f. *$f(x)$ the expected value is defined as*

$$E(X) = \int_{-\infty}^{\infty} x f(x) dx$$

The expect value is a linear operator, that is, for any linear combination of m random variables $\sum_{j=1}^{m} a_j X_j$, then $E(\sum_{j=1}^{m} a_j X_j) = \sum_{j=1}^{m} a_j E(X_j)$.

Example 17 (Binomial Distribution: Expectation) *The linearity of the expected value let us compute easily the expected value of a variable having a Binomial distribution. By recalling that a Binomial random variable $X \sim Bin(n, p)$ is generated as the sum of n independent binary variables X_i with $p(X_i = 1) = p$, we can write*

$$E(X) = \sum_{i=1}^{n} E(X_i)$$

and we need to compute $E(X_i)$ that is given — simply apply Definition 14 — by $p \times 1 + (1-p) \times 0 = p$. Hence, we can conclude that $E(X) = np$.

Another summary measure of a random variable is the *mode*, that is defined as the value with the largest probability when X is a discrete random variable, and the value that maximizes the d.f. when X is a continuous random variable. If there is more than one value satisfying this requirement, the distribution is called multi-modal.

Example 18 (Discrete Uniform Distribution: Expectation) *Let X be a discrete random variable taking as values the integers k ($k = 1, ..., n$) with constant probability $p(x_k) = 1/n$. This distribution is called uniform, its expected value is easily found to be $E(X) = (n+1)/2$ and it is clearly multi-modal.*

When a probability distribution is summarized by the expected value, a measure of the variability around this quantity is the *variance*.

Definition 15 (Variance) *The variance of a discrete random variable X taking values x_k with probability $p(x_k)$ ($k = 1, ..., n$) is defined as*

$$V(X) = \sum_{i=1}^{n} p(x_k)(x_k - E(X))^2.$$

If X is a continuous random variable, then

$$V(X) = \int_{-\infty}^{\infty} (x - E(X))^2 f(x) dx$$

Since $(x_k - E(X))^2$ measure the distance of x_k from the expected value $E(X)$, the variance is the average of the distances between the values that the variable can actually take and the expectation, weighted by the probabilities. The *standard deviation* $\sqrt{V(X)}$ is then the Euclidean distance between the values of the variable X and the expectation. Note that the standard deviation is expressed in the same measurement scale of the variable X.

Example 19 (Binomial Distribution: Variance) *If $X \sim Bin(n,p)$ then $V(X) = np$. Note that, for given n, the variance is largest when $p = 0.5$ that represents the situation of maximum uncertainty, since both $X = 1$ and $X = 0$ are equally probable. The variance is symmetrical about $p = 0.5$ and decreases for p either approaching 1 or 0.*

If X has a Normal distribution with parameters μ and σ^2, then it can be shown that $E(X) = \mu$ and $V(X) = \sigma^2$. A property of the Normal distribution is that if $X \sim N(\mu, \sigma^2)$ and $Y = aX + b$ for real numbers a and b, then $Y \sim N(a\mu + b, a^2\sigma^2)$. This property allows to generate any Normal distribution from the Standard Normal distribution $X \sim N(0, 1)$. To generate a variable $Y \sim N(\mu, \sigma^2)$ we can simply define $Y = \mu X + \sigma$.

The variance is a measure of variability of a random variable about its expected value. A different measure of variability — actually a measure of the lack of information[3] conveyed by a variable — is the *entropy*. If X is a random variable with p.m.f. $p(x)$, the entropy of X is defined as

$$Ent(X) = -\sum_{x} p(x) \log p(x)$$

and becomes

$$Ent(X) = -\int_{-\infty}^{\infty} f(x) \log f(x) dx$$

[3] A derivation of the entropy as the lack of information from a sequence of binary digits that are sent with some probability of being inverted is given by (Cover and Thomas, 1991).

when X is a continuous random variable with d.f. $f(x)$. Note that the entropy is only a function of either the p.m.f. or d.f. of a distribution and does not take into account explicitly the values that the variable can take. If, for instance, X is a binary variable, taking values 1 with $p(X = 1) = p$ and 0 with probability $p(X = 0) = 1 - p$, then the entropy is $Ent(X) = -p \log p/(1 - p) - \log(1 - p)$. This is also the entropy of any other binary variable having p.m.f. $p, (1 - p)$, independently of the states in which the variable is coded. Note that when $p = 1/2$, $Ent(X) = \log 2$ that is the maximum achievable, and $Ent(X) \to 0$ as p approaches either 0 or 1. If $X \sim N(\mu, \sigma^2)$, then $Ent(X) = \log(2\pi\sigma^2)/2$ so that it is an increasing function of the variance. If $X \sim U(\alpha, \beta)$, then $Ent(X) = \log(\beta - \alpha)$ that is an increasing function of the range of values that X can take: the larger the range, the less informative the distribution. In all cases, the entropy measures the overall uncertainty of the distribution.

Mean and variance are simple summaries of univariate random variables. Suppose now $\boldsymbol{X}$ is a random vector. Every component is a univariate random variable, and hence it will have its own mean and variance. We can then define the expectation of a random vector $\boldsymbol{X} = (X_1, ..., X_c)$ as the vector $E(\boldsymbol{X}) = (E(X_1), ..., E(X_c))$ and

$$E(X_i) = \begin{cases} \sum_j p(X_i = x_{ij})p(x_{ij}) & \text{if } X_i \text{ is discrete} \\ \int_{-\infty}^{\infty} x_i f(x_i) dx_i & \text{if } X_i \text{ is continuous} \end{cases}$$

In both cases, the calculation are carried out using the marginal p.m.f. or d.f. of X_i.

Example 20 (Expectation of a Multinomial Distribution) *Let $\boldsymbol{X} = (X_1, ..., X_c)$ have a multinomial distribution $Mult(n, \boldsymbol{p})$ with $\boldsymbol{p} = (p_1, ..., p_c)$. As shown in Example 13, every component X_i has a $Bin(n, p_i)$ distribution, so that $E(X_i) = np_i$ and $E(\boldsymbol{X}) = (np_1, ..., np_c)$.*

The variance of a random variable X is a measure of the variability of X about $E(X)$. Suppose now we have two discrete random variables X_1 and X_2. Then $V(X_1)$ and $V(X_2)$ are independent measures of the variability of X_1 and X_2. Consider the vector $\boldsymbol{X} = (X_1, X_2)$ with joint p.m.f. $p(x_{1i}, x_{2j})$. The joint p.m.f. of X_1 and X_2 gives a probability measure to the joint event $X_1 = x_{1i}, X_2 = x_{2j}$, and as the two variables are so related, we may wish to have a measure of their joint variability. One possible measure is the *covariance.*

Definition 16 (Covariance) *Let X_1 and X_2 be two random variables. The covariance of X_1 and X_2 is defined as:*

$$cov(X_1, X_2) = \begin{cases} \sum_{ij} (x_{1i} - E(X_1))(x_{2j} - E(X_2)) p(x_{1i}, x_{2j}) \\ \qquad \textit{if } X_i, X_j \textit{ are discrete} \\ \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x_1 - E(X_1))(x_2 - E(X_2)) f(x_1, x_2) dx_1 dx_2 \\ \qquad \textit{if } X_1, X_2 \textit{ are continuous} \end{cases}$$

Note that the computation of the covariance involves the joint distribution — in terms of either p.m.f. or d.f. — of X_1, X_2. Furthermore, if $X_1 \equiv X_2$, then $cov(X_1, X_2) = V(X_1)$ and also, clearly, $cov(X_1, X_2) = cov(X_2, X_1)$.

Let $\boldsymbol{X} = (X_1, ..., X_c)$ be a random vector. Every $c(c-1)$ pair of different components X_i, X_j will have a covariance that measures their pair-wise variability. We also have c variances $V(X_i)$. We can display these c^2 quantities into a $c \times c$ matrix $V(\boldsymbol{X})$ having (i, j)th element equal to $cov(X_i, X_j)$.

Hence the diagonal terms are the c variances $V(X_i)$, and the $c(c-1)$ off-diagonal terms are the covariances. Hence, this matrix will be called the *covariance matrix.* The symmetry of the covariance ensures that the covariance matrix is symmetric.

Example 21 (Expectation of a Bivariate Normal Distribution) *The Bivariate Normal distribution $\boldsymbol{X}$ introduced in Example 15 is characterized by the vector of parameters $\boldsymbol{\mu}$ and the matrix Σ. It can be shown that μ is the vector of expectations of $\boldsymbol{X}$, and hence $\mu_i = E(X_i)$ $(i = 1, 2)$. The matrix Σ is the covariance matrix. Hence $\sigma_i^2 = V(X_i)$ $(i = 1, 2)$ and $\sigma_{12}^2 = cov(X_1, X_2)$. Details of the calculation can be found, for instance, in Whittaker (1990).*

We conclude this chapter by describing in details a particular multivariate distribution — the *Dirichlet distribution.*

Definition 17 (Dirichlet Distribution) *Let $\boldsymbol{X} = (X_1, ..., X_c)$ be continuous random vector taking values in the c-dimension* simplex[4] *and with joint* d.f.

$$f(x_1, ..., x_c) = \frac{\Gamma(\alpha)}{\prod_{i=1}^{c} \Gamma(\alpha_i)} \prod_{i=1}^{c} x_i^{\alpha_i - 1} \quad \alpha_i > 0 \textit{ for all } i \tag{10}$$

where $\Gamma(x) = \int_0^\infty t^{x-1} e^{-t} dt$ is termed the Gamma function. *Then $\boldsymbol{X}$ is said to have a c-dimension Dirichlet distribution, with parameter vector $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_c)$ and $\alpha = \sum_{i=1}^{c} \alpha_i$. We will use the notation $\boldsymbol{X} \sim D(\alpha_1, \ldots, \alpha_c)$.*

A two-dimension Dirichlet distribution is also called a *Beta distribution.* The vector of expectations and the covariance matrix are function of the parameters α_i. We have:

$$E(X_i) = \frac{\alpha_i}{\alpha} \tag{11}$$

and the covariance matrix $V(\boldsymbol{X})$ has elements

$$cov(X_i, X_j) = -\frac{\alpha_i \alpha_j}{\alpha^2(\alpha + 1)} \quad i \neq j \tag{12}$$

$$V(X_i) = -\frac{\alpha_i(\alpha - \alpha_i)}{\alpha^2(\alpha + 1)} \tag{13}$$

Note that $V(X_i) = E(X_i)[1 - E(X_i)]/(\alpha + 1)$. An important property of the Dirichlet distribution is that it is closed under marginalization so that, if $\boldsymbol{X} \sim D(\alpha_1, \ldots, \alpha_c)$, then any random vector $\boldsymbol{X}_s = (X_1, ..., X_s)$ $(s < c)$ will have a Dirichlet distribution $D(\alpha_1, ..., \alpha_{s-1}, (\alpha - \sum_{i=1}^{s-1} \alpha_i))$. The proof of the next result can be found in Wilks (1963, page 179).

Property 1 *Let $\boldsymbol{X} \sim (D(\alpha_1, ..., \alpha_c)$ and define the random vector $\boldsymbol{Y} = (Y_1, ..., Y_{m-1})$ whose components are sum of different components of $\boldsymbol{X}$, that is $Y_1 = \sum_{j=1}^{c_1} X_i$, $Y_2 = \sum_{j=c_1+1}^{c_2} X_i$, ..., $Y_{m-1} = \sum_{j=c_{m-1}}^{c} X_i$. Then $\boldsymbol{Y} \sim D(\beta_1, ..., \beta_{m-1})$ where $\beta_i = \sum_{j=c_i}^{c_{i+1}} \alpha_j$.*

[4] The c-dimension simplex is such that if $\boldsymbol{X} = \boldsymbol{x} = (x_1, ..., x_c)$ then $\sum_{i=1}^{c} x_i = 1$

11. Special Distributions

We will add a summary of common random variables. For the time being, refer to Casella and Berger (1990).

12. Independence

We defined two events to be independent when their joint probability decomposes into the product of their marginal probabilities. In this section, we will extend the concept of independence to random variables that take values associated with the occurrence of an event. For example, let X and Y be binary variables and suppose that X takes values 1 if A occurs and 0 otherwise and that Y takes value 1 if the event B occurs, and 0 otherwise. If the event A and B be independent so are $\bar{A}$ and $\bar{B}$ as well as $A, \bar{B}$ and $\bar{A}, B$. The independence of A and B ensures that

$$p(X = 1, Y = 1) = p(A, B) = p(A)p(B) = p(X = 1)p(Y = 1)$$

and similarly, using the independence of $\bar{A}, B$, $A, \bar{B}$ and $\bar{A}, \bar{B}$ we can show $p(X = i, Y = j) = p(X = i)p(Y = j)$ all $i, j = 1, 2$. In other words, the joint p.m.f. of X, Y factorizes into the product of the marginal p.m.f.s and, in this case, we say the X and Y are *independent variables.*

Definition 18 (Independent Random Variables) *Let X and Y two discrete random variables taking values x_i and y_j. They are said to be independent if*

$$p(X = x_i, Y = y_j) = p(X = x_i)p(Y = y_j)$$

for all i and j.

The same concept can be extended to continuous random variables. In this case, the characterization of independence is given in terms of the joint d.f. and X and Y are independent if

$$f(x, y) = f(x)f(y)$$

Conditional independence is defined as for events. We say that X_1 and X_2 are independent given Y if

$$p(X_1 = x_{1i}, X_1 = x_{2i}|Y = y_j) = p(X_1 = x_{1i}|Y y_j)p(X_2 = x_{2i}|Y = y_j)$$

or

$$f(x_1, x_2|y) = f(x_1|y)f(x_2|y).$$

References

Casella, G., and Berger, R. L. (1990). *Statistical Inference.* Duxbury Press, Belmont, Ca.

Cover, T. M., and Thomas, M. (1991). *Elements of Information Theory.* Wiley, New York, NY.

DeFinetti, B. (1972). *Probability, Induction and Statistics.* Wiley, New York, NY.

Evans, M., and Swartz, T. (1995). Methods for approximating integrals in statistics with special emphasis on Bayesian integration problems. *Statistical Science*, *10*, 254–272.

Feller, W. (1968). *An Introduction to Probability Theory, Volume I.* Wiley, New York, NY.

Good, I. J. (1968). *The Estimation of Probability: An Essay on Modern Bayesian Methods.* MIT Press, Cambridge, MA.

Karr, A. F. (1992). *Probability.* Springer-Verlag, New York.

Kolmogorov, A. (1950). *Foundations of the Theory of Probability.* Chelsea Publishing Co., New York. English Translation of *Grundbegriffe der Wahrscheinlichkeitsrechnung*, Berlin, J. Springer 1933.

Whittaker, J. (1990). *Graphical Models in Applied Multivariate Statistics.* Wiley, New York, NY.

Wilks, S. S. (1963). *Mathematical Statistics.* Wiley, New York, NY.