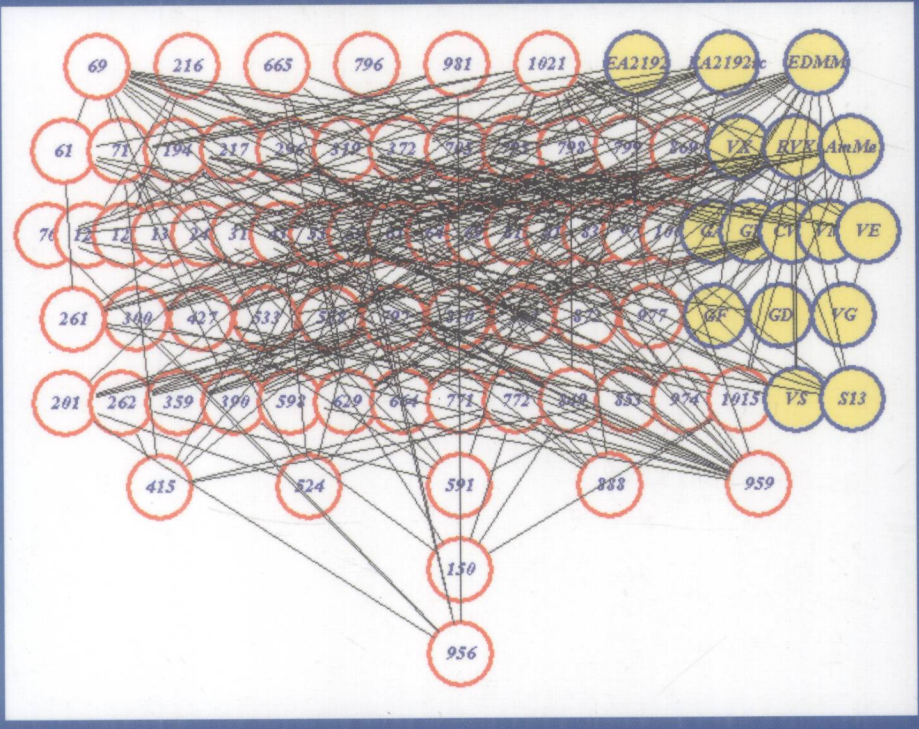
DATA HANDLING IN SCIENCE AND TECHNOLOGY

27

Scientific Data Ranking Methods: Theory and Applications

MANUELA PAVAN
ROBERTO TODESCHINI



Data Handling in

SCIENCE AND TECHNOLOGY

SCIENTIFIC DATA RANKING METHODS: THEORY AND APPLICATIONS

VOLUME **27**

Edited by

MANUELA PAVAN

Institute for Health and Consumer Protection
Joint Research Centre
European Commission
Ispra, Italy

ROBERTO TODESCHINI

Department of Environmental Sciences
University of Milano-Bicocca
Milano, Italy

ELSEVIER

Amsterdam • Boston • Heidelberg • London • New York • Oxford
Paris • San Diego • San Francisco • Singapore • Sydney • Tokyo

Elsevier
Linacre House, Jordan Hill, Oxford OX2 8DP, UK
Radarweg 29, PO Box 211, 1000 AE Amsterdam, The Netherlands

First edition 2008

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

Library of Congress Cataloging-in-Publication Data
A catalog record for this book is available from the Library of Congress

ISBN: 978-0-444-53020-2
ISSN: 0922-3487

Printed and bound in Hungary
08 09 10 11 12 10 9 8 7 6 5 4 3 2 1

Data Handling in

SCIENCE AND TECHNOLOGY

SCIENTIFIC DATA RANKING METHODS: THEORY AND APPLICATIONS

VOLUME **27**

DATA HANDLING IN SCIENCE AND TECHNOLOGY
Advisory Editors: S. Rutan and B. Walczak

Other volumes in this series:

Volume 1	Microprocessor Programming and Applications for Scientists and Engineers, by R.R. Smardzewski
Volume 2	Chemometrics: A Textbook, by D.L. Massart, B.G.M. Vandeginste, S.N. Deming, Y. Michotte and L. Kaufman
Volume 3	Experimental Design: A Chemometric Approach, by S.N. Deming and S.L. Morgan
Volume 4	Advanced Scientific Computing in BASIC with Applications in Chemistry, Biology and Pharmacology, by P. Valkó and S. Vajda
Volume 5	PCs for Chemists, edited by J. Zupan
Volume 6	Scientific Computing and Automation (Europe) 1990, *Preceedings of the Scientific Computing and Automation (Europe) Conference, 12–15 June, 1990, Maastricht, The Netherlands*, edited by E.J. Karjalainen
Volume 7	Receptor Modeling for Air Quality Management, edited by P.K. Hopke
Volume 8	Design and Optimization in Organic Synthesis, by R. Carlson
Volume 9	Multivariate Pattern Recognition in Chemometrics, illustrated by case studies, edited by R.G. Brereton
Volume 10	Sampling of Heterogeneous and Dynamic Material Systems: Theories of Heterogeneity, Sampling and Homogenizing, by P.M. Gy
Volume 11	Experimental Design: A Chemometric Approach (Second, Revised and Expanded Edition) by S.N. Deming and S.L. Morgan
Volume 12	Methods for Experimental Design: Principles and Applications for Physicists and Chemists, by J.L Goupy
Volume 13	Intelligent Software for Chemical Analysis, edited by L.M.C. Buydens and P.J. Schoenmakers
Volume 14	The Data Analysis Handbook, by I.E. Frank and R. Todeschini
Volume 15	Adaption of Simulated Annealing to Chemical Optimization Problems, edited by J. Kalivas
Volume 16	Multivariate Analysis of Data in Sensory Science, edited by T. Næs and E. Risvik
Volume 17	Data Analysis for Hyphenated Techniques, by E.J. Karjalainen and U.P. Karjalainen
Volume 18	Signal Treatment and Signal Analysis in NMR, edited by D.N. Rutledge
Volume 19	Robustness of Analytical Chemical Methods and Pharmaceutical Technological Products, edited by M.W.B. Hendriks, J.H. de Boer, and A.K. Smilde
Volume 20A	Handbook of Chemometrics and Qualimetrics: Part A, by D.L. Massart, B.G.M. Vandeginste, L.M.C. Buydens, S. de Jong, P.J. Lewi, and J. Smeyers-Verbeke
Volume 20B	Handbook of Chemometrics and Qualimetrics: Part B, by B.G.M. Vandeginste, D.L. Massart, L.M.C. Buydens, S. de Jong, P.J. Lewi, and J. Smeyers-Verbeke
Volume 21	Data Analysis and Signal Processing in Chromatography, by A. Felinger
Volume 22	Wavelets in Chemistry, edited by B. Walczak
Volume 23	Nature-inspired Methods in Chemometrics: Genetic Algorithms and Artificial Neural Networks, edited by R. Leardi
Volume 24	Handbook of Chemometrics and Qualimetrics, by D.L. Massart, B.M.G. Vandeginste, L.M.C. Buydens, S. de Jong, P.J. Lewi, and J. Smeyers-Verbeke
Volume 25	Statistical Design — Chemometrics, by R.E. Bruns, I.S. Scarminio and B. de Barros Neto
Volume 26	Practical Data Analysis in Chemistry, by Marcel Maeder, and Yorck-Michael Neuhold

CONTRIBUTORS

Numbers in parenthesis indicate the pages on which the authors' contributions begins.

Davide Ballabio, *Milano Chemometrics and QSAR Research Group, Department of Environmental Sciences, University of Milano-Bicocca, P.za della Scienza, 1, 20126 Milano, Italy* (111,169,193)

Rainer Brüggemann, *Leibniz Institute of Freshwater Ecology and Inland Fisheries, Mueggelseedamm 310, 12587 Berlin, Germany* (73)

Sergio Canobbio, *Department of Environmental Sciences, University of Milano-Bicocca, P.za della Scienza, 1, 20126 Milano, Italy* (169)

Lars Carlsen, *Awareness Center, Hyldeholm 4, Veddelev, DK-4000 Roskilde, Denmark* (97, 139)

Viviana Consonni, *Milano Chemometrics and QSAR Research Group, Department of Environmental Sciences, University of Milano-Bicocca, P.za della Scienza, 1, 20126 Milano, Italy* (193)

Alberto Manganaro, *Milano Chemometrics and QSAR Research Group, Department of Environmental Sciences, University of Milano-Bicocca, P.za della Scienza, 1, 20126 Milano, Italy* (193)

Andrea Mauri, *Milano Chemometrics and QSAR Research Group, Department of Environmental Sciences, University of Milano-Bicocca, P.za della Scienza, 1, 20126 Milano, Italy* (111, 193)

Valeria Mezzanotte, *Department of Environmental Sciences, University of Milano-Bicocca, P.za della Scienza, 1, 20126 Milano, Italy* (169)

Wayne L. Myers, *School of Forest Resources, The Pennsylvania State University, University Park, PA 16802, USA* (159)

M. Cruz Ortiz, *Department of Chemistry, Faculty of Science, University of Burgos, Spain* (1)

Ganapati P. Patil, *Department of Statistics, The Pennsylvania State University, University Park, PA 16802, USA* (159)

Manuela Pavan, *Consumer Products Safety & Quality, Institute for Health and Consumer Protection, Joint Research Centre, European Commission, Via E. Fermi 2749, 21027 Ispra (VA), Italy* (51, 169, 193)

Stefan Pudenz, *Westlakes Scientific Consulting Ltd., Moor Row, Cumbria CA24 3LN, United Kingdom* (73)

M. Sagrario Sánchez, *Department of Mathematics and Computation, Faculty of Science, University of Burgos, Spain* (1)

Luis A. Sarabia, *Department of Mathematics and Computation, Faculty of Science, University of Burgos, Spain* (1)

Roberto Todeschini, *Milano Chemometrics and QSAR Research Group, Department of Environmental Sciences, University of Milano-Bicocca, P.za della Scienza 1, 20126 Milan, Italy* (51, 193)

Kristina Voigt, *GSF-National Research Centre for Environment and Health, Ingostaedter Landstr. 1, 85764 Neuahrberg, Germany* (73)

PREFACE

The intrinsic complexity of the systems analysed in scientific research together with the significant increase of available data require the availability of suitable methodologies for multivariate statistics analysis and motivate the endless development of new methods. Moreover, the increase of problem complexity leads to the decision processes becoming more complex, requiring the support of new tools able to set priorities and define rank order of the available options. Ordering is one of the possible ways to analyse data and to get an overview over the elements of a system. The different kinds of order ranking methods available can be roughly classified as total (called even-scoring) and partial order ranking methods, according to the specific order they provide. These methods are the ones needed to support and solve decision problems, setting priorities. Besides sophisticated multivariate statistics, used mostly in pre-processing and modelling data, priority setting makes use of quite simple methodologies. Total and partial ordering methods are described in several mathematical books, requiring different degrees of mathematical skills.

Our intention in writing this book has been to provide a comprehensive and widely accessible overview of the basic mathematical aspects of the total and partial ordering methods by a didactical approach and to explain their use by examples of relevant applications in different scientific fields.

In fact, in recent years, ranking methods have been applied in several different fields such as decision support, toxicology, chemical prioritization, environmental hazard, proteomics and genomics, analytical chemistry, food chemistry and quantitative structure–activity relationship (QSAR). Moreover, new researches based on ranking methods are under investigation providing new perspectives for DNA sequence comparisons and for analytical data pattern recognition.

Being usually based on simple algorithms, ranking methods (both total and partial rankings) can be easily understood and successfully applied, resulting in a new appealing multivariate computation tool.

The integration of the theory and application of ranking methods has been of central concern to us in writing this book, based on the idea that the constant development of the ranking field strongly depends on this synergy.

Thus, the first chapter provides an extensive overview of the basic theory of ranking methods in statistics; it gives clear definition of order relations and it covers different aspects of the "order" concept in statistics, including random variable, order statistics, non-parametric methods and rank-based methods. The direct link between order and graphs and the one between order and optimization problems is also presented. The following two chapters are intended to

illustrate the fundamental basis of the mostly known total and partial ordering methods.

A number of different and up-to-date interest applications of order ranking methods are described in the following chapters. Examples on the use of the Hasse diagrams partial ranking method for the evaluation of chemicals and of databases are provided in more detail together with its use for prioritizing polluted sites. A new similarity/diversity measure is also described as a new approach for the analysis of sequential data, where useful information is obtained by the ordering relationships between the sequence elements. The advantageous interplay between partial order ranking and QSAR is illustrated by selected examples from recent studies, including risk assessment, selecting safer alternatives and as a tool in the process of suggesting new substances with specific characteristics.

Furthermore, a case study on the application of total order ranking methods to river functionality assessment is illustrated with the purpose to generate information and to provide further understanding as a basis for of river restoration strategies.

Finally, a software tool called DART (Decision Analysis by Ranking Techniques) that implements most of the different ranking methods described above is presented.

We believe that the book can be of value to various researchers in several scientific fields such as

- Research centres and universities
- Governmental organizations
- Pharmaceutical companies
- Health control organizations
- Environmental control organizations

We hope this book will expand the knowledge and application of order ranking methods.

Manuela Pavan and Roberto Todeschini
April 2008

CONTENTS

CHAPTER 1

Introduction to Ranking Methods

L.A. Sarabia, M.S. Sánchez, and M.C. Ortiz

Contents

1. DEFINITION OF ORDER RELATIONS

From the mathematical point of view, an ordering (order relation) R in a set E is a binary relation among the elements in E that verifies:

(i) $u\ R\ u$, for each u in E (reflexive);
(ii) if $u\ R\ v$ and $v\ R\ u$, then $u = v$ (antisymmetric);
(iii) if $u\ R\ v$ and $v\ R\ w$, then $u\ R\ w$ (transitive).

A *total ordering* (or linear ordering or simple ordering) is, in addition, connected (complete), that is, every two members of the set are comparable (either $u\ R\ v$ or $v\ R\ u$, for all u, v in E), and thus enables every member to be ordered relative to every other, and that generates a unique "linear" chain.

If this is not so, R is called a *partial ordering* (that is, it is transitive and antisymmetric but not necessarily connected) and thus generates possibly different chains of comparable elements; members of distinct chains may be incomparable.

Typical examples are the relation "less than or equal to" in the real numbers which is a total order whereas the set inclusion is a partial order. Hence, the real

Data handling in Science and Technology, Vol. 27
ISSN 0922-3487, DOI: 10.1016/S0922-3487(08)00001-4

numbers form a unique chain of comparable elements (which is usually represented by the real line), whereas, for instance, the set of even natural numbers and the set of odd natural numbers are two incomparable sets; neither the set of odd numbers is included in the set of even numbers nor vice versa, so that they will be in different chains of comparable elements.

Let us consider an order and denote it as $\leq$, for simplicity. There are some special elements within such an order. One of the most important is the *least element* of a (sub)set S, which is an element u such that $u \leq v$, for all elements $v \in S$. A value that is less than or equal to all elements of a set of given values is called a *lower bound*. The infimum or greatest lower bound is the unique largest member of the set of lower bounds for some given set, and it is equal to its *minimum* if the given set has a least element.

Analogously, the *greatest element* of a subset S of a partially ordered set (poset) is an element of S, which is greater than or equal to any other element of S, that is, u, such that $v \leq u$, for all elements v of S; an *upper bound* is a value greater than or equal to all of a set of given values. The unique smallest member of the set of upper bounds for a given set is the supremum or least upper bound, and it is equal to its *maximum* if the given set has a greatest member.

In that respect, note the difference between minimum (resp. maximum) and minimal (resp. maximal) element. An element in an ordered set is *minimal* (resp. *maximal*) when there is no element smaller (resp. greater) than it, that is, it is the least (resp. greatest) element of a chain. If we do not need to specify, we use the generic term *optimal*.

Least and greatest elements may fail to exist but, if they exist, least and greatest elements are always unique. However, there can be many optimal elements in a set and some elements may be both maximal and minimal; thus a minimal (resp. maximal) element may not be the unique least (resp. greatest) element unless the order relation is a total order. Under total order relations, both terms coincide.

In that sense, an order in which every non-empty subset has at least one minimal element is an *inductive order*, whereas an ordering is a *well ordering* if every non-empty subset has a least member under the ordering, i.e. a unique minimal member that has the given relation to all members of the subset. A well order is an inductive order but not necessarily an inductive order is a well order.

There are some more properties that characterize different kinds of orders, see, for instance (Frank and Todeschini, 1994), and there are also some variations in the use of the terms. We have restricted ourselves to the most standard uses and define only the terms that we will use later on.

The fact is that it is not always possible to define a total order in a given set, and this fact affects in many scopes, because rankings (orderings) are used to compare nearly all variables that can be quantified in the interest of demonstrating differences. In general, many approaches have been made to jointly consider the information contained in the quantified variables and *summarize* it into one unique real number, giving rise to the so-called ranking methods. These can be a weighting of the characteristics, scaling and computing some statistics on them, etc. In (Allen and Sharpe, 2005) a case study is used to demonstrate the challenges

of creating a valid ranking structure, and references to different ranking methods are given.

2. ORDER IN STATISTICS

The statistical analysis based on the distribution of the ranks (order of the experimental values) has had an increasing development, passing from being a subject treated in the last chapter of books on applied statistics to being object of monographs. In 1956, Siegel published the first book (Siegel, 1956) dedicated to methods not based on normality, one of the most referenced books in statistics. From 1970, it is estimated that annually at least a book on non-parametric methods is published in which the methods based on ranks are a central subject. Some of these books are of interest for the users of the non-parametric statistics and have served as inspiration to write up this paragraph. Among the "classic ones", advisable books are the aforementioned by Siegel and the one by Kendall (1975), whose first edition is of 1948. Of practical character, usable by researchers with a basic formation in statistics, are the books by Sprent (1989) and by Conover (1999). The books by Lehmann (1975) and the one by Hettmansperger (1984) are centred exclusively in the methods based on ranks. A recent text that includes the last investigations in the subject but with an advanced level is the one by Govindarajulu (2007).

2.1 Random variables

Outcomes associated with an experiment may be numerical in nature, such as quantity in an analytical sample. The types of measurements are usually called *measurement scales* and are, from the weakest to the strongest, nominal, ordinal, interval and ratio scale.

The *nominal scale* of measurement uses numbers merely as a means of separating the properties or elements into different classes or categories, for example, the sites of a study about contamination.

The *ordinal scale* refers to measurements where only the comparisons "greater", "less" or "equal" are relevant, for example, the level of contamination: contamination in A is higher than in B, and contamination in B is higher than in C. If some of the values are equal to each other, we say ties exist.

The *interval scale* considers not only the relative order of the values but also the size of the interval between measurements as pertinent information. The *interval scale* involves the concept of a unit distance, and the distance between any two measurements may be expressed as a number of units, for example, the temperature. The actual value is merely a comparison with a reference value (the zero in scale) measured in units. A change in scale or location or both does not alter the principle of interval measurements.

The *ratio scale* is used when not only the order and interval size are important but also the ratio between two measurements has significance. It has sense to say that a measurement is twice or three times greater than another one. The ratio

scale is appropriate for measurements such as yields, quantities, weights and so on. The only distinction between the ratio scale and the interval scale is that the ratio scale has a natural measurement that is called zero, while the zero is arbitrarily defined in the interval scale.

Most of the usual parametric statistical methods require an interval (or stronger) scale of measurement. Most non-parametric methods assume either the nominal or the ordinal scale to be appropriate. Of course, each scale has all of the properties of the weaker measurement scales, therefore statistical methods requiring only a weaker scale may be used with the stronger scales also.

A *random variable* is a function that assigns real numbers to the outcomes of an experiment or observation. We will usually denote random variables by capital letters, X, Y, T, with or without subscripts. The real numbers attained by the random variables will be denoted by lowercase letters. For example, if we have a sample of wastewater and we apply an analytical procedure to determine the content of triazines, the result is a random variable. If the procedure is applied to n aliquot samples, we obtain n outcomes, x_1, x_2,..., x_n that are not equal. The variability of the results caused by the analytical procedure is a characteristic of it and is modelled by means of a random variable.

A random variable is completely specified by its *cumulative distribution function* (cdf) $F_X(x)$, that is, the probability of the random variable being less than or equal to x for any value x. Symbolically, this is written as $F_X(x) = \text{pr}\{X \leq x\}$ for any real value x. In most of the applications, it is assumed that $F_X(x)$ is differentiable, which implies, among other things, that none of the possible outcomes has positive probability, that is, the probability of obtaining exactly a specific value is zero.

In the case of a differentiable distribution function, the derivative of $F_X(x)$ is the *probability density function* (pdf) $f_X(x)$. Any function $f(x)$ such that: (i) it is positive, $f(x) \geq 0$ and (ii) the area under the function is one, $\int_{\mathrm{R}} f(x)\mathrm{d}x = 1$, is the *probability density function* of a random variable.

The probability of the random variable X being in interval $[a, b]$ is the area under the pdf over the interval $[a, b]$, that is

$$\text{pr}\{X \in [a, b]\} = \int_a^b f(x)\mathrm{d}x \tag{1}$$

and the mean and the variance of X are written as

$$E(X) = \int_{\mathrm{R}} xf(x)\mathrm{d}x \tag{2}$$

$$V(X) = \int_{\mathrm{R}} (x - E(X))^2 f(x)\mathrm{d}x \tag{3}$$

These expressions are adapted in the obvious way for discrete random variables, that is, a random variable X that takes discrete values, x_i, $i \in I$. The set I can be a finite set, for example, $I = \{0, 1, 2, 3\}$, or an infinite one (a numerable set),

$I = \mathbb{N} = \{1, 2, 3, \ldots\}$, but always totally ordered. In both cases, we speak about the *probability function* instead of the probability *density* function. The probability function is greater than zero only for the values that the random variable takes, i.e. $f_X(x_i) = p_i = \text{pr}\{X = x_i\} > 0$ while $f_X(x) = 0$ if $x \neq x_i \; \forall x_i$. Also, $\Sigma_i p_i = 1$ must hold. The probability of a discrete variable X being in interval $[a, b]$ is thus the sum of the probabilities associated with the values x_i in $[a, b]$, that is

$$\text{pr}\{X \in [a, b]\} = \sum_i p_i \text{ for } x_i \in [a, b] \tag{4}$$

and the mean and the variance of X are

$$E(X) = \sum_i x_i p_i \tag{5}$$

$$V(X) = \sum_i (x_i - E(X))^2 p_i \tag{6}$$

Example 1

Consider the finite (discrete) random variable X that takes the values x_i with probabilities p_i written in Table 1. Its cumulative probability function is right continuous in x_i and constant in the intervals $[x_i, x_{i+1})$

$$F(x) = \begin{cases} 0, & \text{if } x < 1 \\ 0.20, & \text{if } 1 \leq x < 2 \\ 0.35, & \text{if } 2 \leq x < 4 \\ 0.60, & \text{if } 4 \leq x < 5 \\ 0.70, & \text{if } 5 \leq x < 6 \\ 1.00, & \text{if } 6 \leq x \end{cases} \tag{7}$$

This function is drawn in Figure 1A where the "jumps" corresponding to the values x_i have been joined with vertical lines. A simple calculation applying Eqs (5) and (6) to the data in Table 1 gives $E(X) = 3.8$, $V(X) = 3.66$ and, thus, the standard deviation is $\sqrt{3.66} = 1.91$.

Figure 1B shows the cdf of a normal distribution (continuous distribution) with the same mean and standard deviation as X, that is, a $N(3.8, 1.9)$. Its cdf is given by

$$F(x) = \int_{-\infty}^{x} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{u - \mu}{\sigma}\right)^2\right) \mathrm{d}u, \text{ where } \mu = 3.8 \text{ and } \sigma = 1.9 \tag{8}$$

When several random variables are defined jointly or when several experiments are considered as a combined experiment, each with its own one or more random

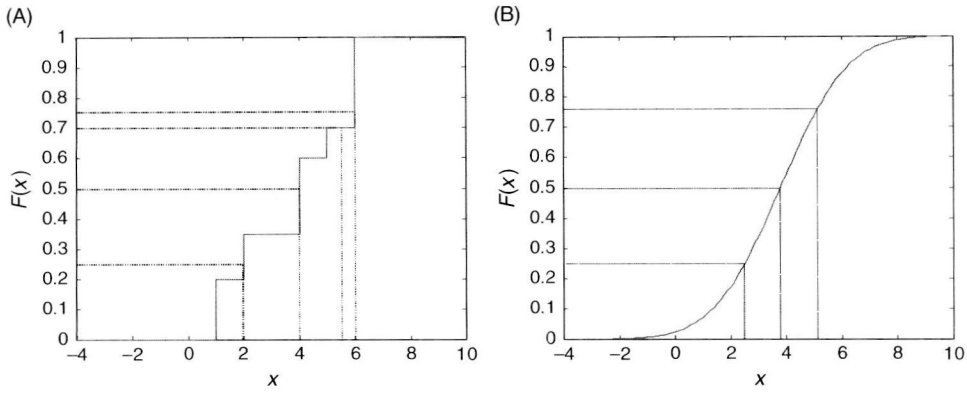


Figure 1 (A) Cumulative probability function of the discrete random variable of Table 1 (Eq. (7)). (B) Cumulative distribution function of a normal distribution with the same mean, 3.8, and standard deviation, 1.91.

Table 1 Values x_i of a discrete random variable and probabilities $p_i = \text{pr}\{X = x_i\}$

x_i	1	2	4	5	6
p_i	0.20	0.15	0.25	0.10	0.30

variables, it becomes useful to consider joint distributions, described by *joint probability functions* (discrete case), which are defined as follows:

$$f(x_1, x_2, \ldots, x_n) = \text{pr}(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) \tag{9}$$

The *joint distribution function* $F(x_1, x_2, \ldots, x_n)$ of the continuous random variables $X_1, \ldots, X_n$ is defined by means of the following equation:

$$F(x_1, x_2, \ldots, x_n) = \text{pr}(X_1 \leq x_1, X_2 \leq x_2, \ldots, X_n \leq x_n) \tag{10}$$

The random variables $X_1, \ldots, X_n$ are independent if

$$f(x_1, x_2, \ldots, x_n) = f_{x_1}(x_1) \times \cdots \times f_{x_n}(x_n) = \text{pr}(X_1 = x_1) \times \cdots \times \text{pr}(X_n = x_n) \tag{11}$$

or, for continuous random variables

$$\begin{aligned} F(x_1, x_2, \ldots, x_n) &= \text{pr}(X_1 \leq x_1, X_2 \leq x_2, \ldots, X_n \leq x_n) \\ &= \text{pr}\{X_1 \leq x_1\} \times \text{pr}\{X_2 \leq x_2\} \times \cdots \times \text{pr}\{X_n \leq x_n\} \\ &= F_{X_1}(x_1) \times F_{X_2}(x_2) \times \cdots \times F_{X_n}(x_n) \end{aligned} \tag{12}$$