1

Selected Concepts from Probability

Hans Colonius

To appear in W. H. Batchelder, et al. (eds.), *New Handbook of Mathematical Psychology, Vol. 1.* Cambridge, U.K.: Cambridge University Press

Contents

1.1 Introduction

1.1.1 Goal of this chapter

Since the early beginnings of mathematical psychology, concepts from probability theory have always played a major role in developing and testing formal models of behavior and in providing tools for data analytic methods. Moreover, fundamental measurement theory, an area where such concepts have not been mainstream, has been diagnosed as wanting of a sound probabilistic base by founders of the field (see Luce, 1997). This chapter is neither a treatise on the role of probability in mathematical psychology nor does it give an overview of its most successful applications. The goal is to present, in a coherent fashion, a number of probabilistic concepts that, in my view, have not always found appropriate consideration in mathematical psychology. Most of these concepts have been around in mathematics for several decades, like coupling, order statistics, records, and copulas; some of them, like the latter, have seen a surge of interest in recent years, with copula theory providing a new means of modeling dependence in high-dimensional data (see Joe, 2015). A brief description of the different concepts and their interrelations follows in the second part of this introduction.

The following three examples illustrate the type of concepts addressed in this chapter. It is no coincidence that they all relate, in different ways, to the measurement of reaction time (RT), which may be considered a prototypical example of a random variable in the field. Since the time of Dutch physiologist Franciscus C. Donders (Donders, 1868/1969), mathematical psychologists have developed increasingly sophisticated models and methods for the analysis of RTs.[1] Nevertheless, the probabilistic concepts selected for this chapter are, in principle, applicable in any context where some form of randomness has been defined.

Example 1.1 (Random variables vs. distribution functions) Assume that the time to respond to a stimulus depends on the attentional state of the individual; the response may be the realization of a random variable with distribution function F_H in the high-attention state and F_L in the low-attention state. The distribution of observed RTs could then be modeled as a mixture distribution,

$$F(t) = pF_H(t) + (1-p)F_L(t),$$

for all $t \geq 0$ with $0 \leq p \leq 1$ the probability of responding in a state of high attention.

[1] For monographs, see Townsend and Ashby (1983), Luce (1986), Schweickert et al. (2012).

Alternatively, models of RT are often defined directly in terms of operations on random variables. Consider, for example, Donders' *method of subtraction* in the detection task; if two experimental conditions differ by an additional decision stage, D, total response time may be conceived of as the sum of two random variables, $D + R$, where R is the time for responding to a high intensity stimulus.
In the case of a mixture distribution, one may wonder whether or not it might also be possible to represent the observed RTs as the sum of two random variables H and L, say, or, more generally, if the observed RTs follow the distribution function of some $Z(H, L)$, where Z is a measurable two-place function of H and L. In fact, the answer is negative and follows as a classic result from the *theory of copulas* (Nelsen, 2006), to be treated later in this chapter.

Example 1.2 (Coupling for audiovisual interaction) In a classic study of intersensory facilitation, Hershenson (1962) compared reaction time to a moderately intense visual or acoustic stimulus to the RT when both stimuli were presented more or less simultaneously. Mean RT of a well-practiced subject to the sound (RT_A, say) was approximately 120 ms, mean RT to the light (RT_V) about 160 ms. When both stimuli were presented synchronously, mean RT was still about 120 ms. Hershenson reasoned that intersensory facilitation could only occur if the "neural events" triggered by the visual and acoustic stimuli occurred simultaneously somewhere in the processing. That is, "physiological synchrony", rather than "physical (stimulus) synchrony" was required. Thus, he presented bimodal stimuli with light leading sound giving the slower system a kind of "head start". In the absence of interaction, reaction time to the bimodal stimulus with presentation of the acoustic delayed by τ ms, denoted as $RT_{V\tau A}$, is expected to increase linearly until the sound is delivered 40 ms after the light (the upper graph in Figure 1.1). Actual results, however, looked more like the lower graph in Figure 1.1 where maximal facilitation occurs at about physiological synchrony. Raab (1962) suggested an explanation in terms of a probability summation (or, *race*) mechanism: response time to the bimodal stimulus, $RT_{V\tau A}$, is considered to be the winner of a race between the processing times for the unimodal stimuli, i.e. $RT_{V\tau A} \equiv \min\{RT_V, RT_A + \tau\}$. It then follows for the expected values (mean RTs):

$$\mathrm{E}[RT_{V\tau A}] = \mathrm{E}[\min\{RT_V, RT_A + \tau\}] \leq \min\{\mathrm{E}[RT_V], \mathrm{E}[RT_A + \tau]\},$$

a prediction that is consistent with the observed facilitation. It has later been shown that this prediction is not sufficient for explaining the observed

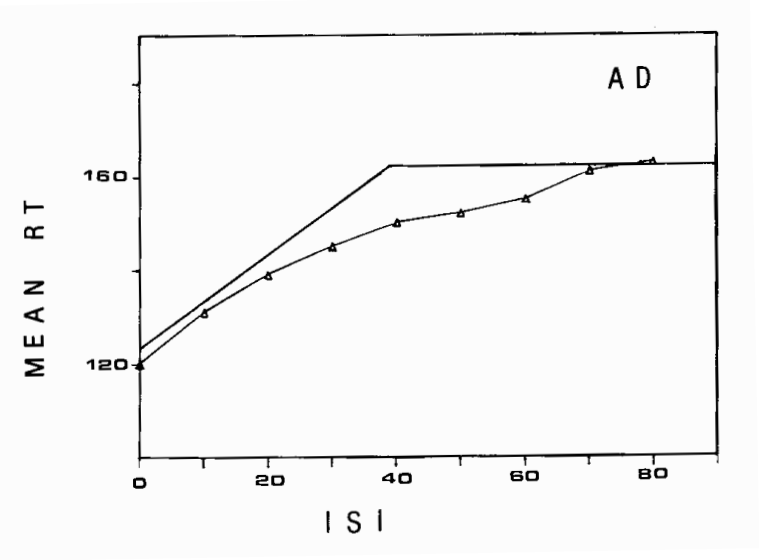

Figure 1.1 Bimodal (mean) reaction time to light and sound with interstimulus interval (ISI) and sound following light, $RT_V = 160$ ms, $RT_A = 120$ ms. Upper graph: prediction in absence of interaction, lower graph: observed mean RTs; data from Diederich and Colonius (1987).

amount of facilitation, and the discussion of how the effect should be modeled is still going on, attracting a lot of attention in both psychology and neuroscience.

However, as already observed by Luce (1986, p. 130), the above inequality only makes sense if one adds the assumption that the three random variables $RT_{V\tau A}$, RT_V, and RT_A are jointly distributed. Existence of a joint distribution is not automatic because each variable relates to a different underlying probability space defined by the experimental condition: visual, auditory, or bimodal stimulus presentation. From the *theory of coupling* (Thorisson, 2000), constructing such a joint distribution is always possible by assuming stochastic independence of the random variables. However –and this is the main point of this example– independence is not the only coupling possibility, and alternative assumptions yielding distributions with certain dependency properties may be more appropriate to describe empirical data.

Example 1.3 (Characterizing RT distributions: hazard function) Sometimes, a stochastic model can be shown to predict a specific parametric distribution, e.g. drawing on some asymptotic limit argument (central limit theorem or convergence to extreme-value distributions). It is often notori-

ously difficult to tell apart two densities when only a histogram estimate from a finite sample is available. Figure 1.2 provides an example of two theoretically important distributions, the gamma and the inverse gaussian densities with identical means and standard deviations, where the rather similar shapes make it difficult to distinguish them on the basis of a histogram.

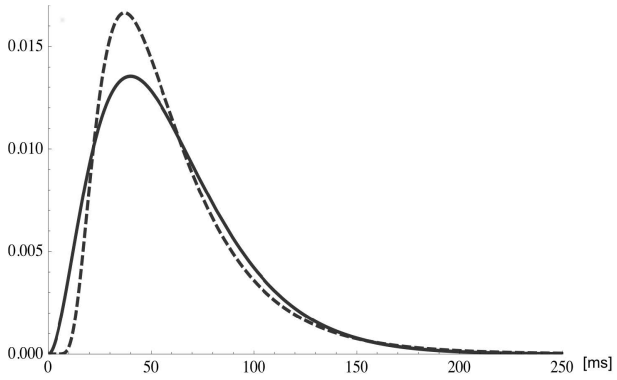

Figure 1.2 Inverse gaussian (dashed line) and gamma densities with identical mean (60 ms) and standard deviation (35 ms).

An alternative, but equivalent, representation of these distributions is terms of their *hazard functions* (see Section 1.10). The hazard function h_X of random variable X with distribution function $F_X(x)$ and density $f_X(x)$ is defined as

$$h_X(x) = \frac{f_X(x)}{1 - F_X(x)}.$$

As Figure 1.3 illustrates, the gamma hazard function is increasing with decreasing slope, whereas the inverse gaussian is first increasing and then decreasing. Although estimating hazard functions also has its intricacies (Kalbfleisch and Prentice, 2002), especially at the right tail, there is a better chance to tell the distributions apart based on estimates of the hazard function than on the density or distribution function. Still other methods to distinguish classes of distribution functions are based on the concept of *quantile function* (see Section 1.3.2), among them the method of *delta plots*, which has recently drawn the attention of researchers in RT modeling (Schwarz and Miller, 2012). Moreover, an underlying theme of this chapter is to provide tools for a model builder that do not depend on committing oneself to a particular parametric distribution assumption.

We hope to convey in this chapter that even seemingly simple situations,

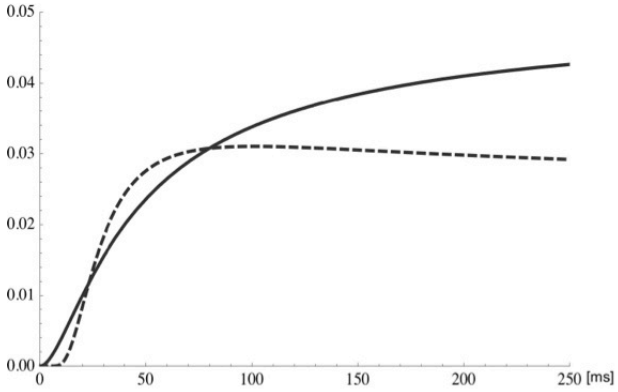

Figure 1.3 Hazard functions of the inverse gaussian (dashed line) and gamma distributions corresponding to the densities of Figure 1.2.

like the one described in Example 1.2, may require some careful consideration of the underlying probabilistic concepts.

1.1.2 Overview

In trying to keep the chapter somewhat self-contained, the first part presents basic concepts of probability and stochastic processes, including some elementary notions of measure theory. Because of space limitations, some relevant topics had to be omitted (e.g., random walks, Markov chains) or are only mentioned in passing (e.g., martingale theory). For the same reason, statistical aspects are considered only when suggested by the context[2]. Choosing what material to cover was guided by the specific requirements of the topics in the second, main part of the chapter.

The second part begins with a brief introduction to the notion of *exchangeability* (with a reference to an application in vision) and its role in the celebrated "theorem of de Finetti". An up-to-date presentation of quantile (density) functions follows, a notion that emerges in many areas including survival analysis. The latter topic, while central to RT analysis, has also found applications in diverse areas, like decision making and memory, and is treated next at some length, covering an important non-identifiability result. Next follow three related topics: order statistics, extreme values, and the theory of records. Whereas the first and, to a lesser degree, the second of these topics have become frequent tools in modeling psychological processes, the third one has not yet found the role that it arguably deserves.

[2] For statistical issues of reaction time analysis, see the competent treatments by Van Zandt (2000, 2002); and Ulrich and Miller (1994), for discussing effects of truncation.

The method of coupling, briefly mentioned in introductory Example 1.2, is a classic tool of probability theory concerned with the construction of a joint probability space for previously unrelated random variables (or, more general random entities). Although it is being used in many parts of probability, e.g., Poisson approximation, and in simulation, there are not many systematic treatises of coupling and it is not even mentioned in many standard monographs of probability theory. We can only present the theory at a very introductory level here, but the expectation is that coupling will have to play an important conceptual role in psychological theorizing. For example, its relevance in defining 'selective influence/contextuality' has been demonstrated in the work by Dzhafarov and colleagues (see also the chapter by Dzhafarov and Kujala in this volume).

While coupling strives to construct a joint probability space, existence of a multivariate distribution is presumed in the next two sections. *Fréchet classes* are multivariate distributions that have certain of their marginal distributions fixed. The issues are (i) to characterize upper and lower bounds for all elements of a given class and (ii) to determine conditions under which (bivariate or higher) margins with overlapping indices are compatible. Copula theory allows one to separate the dependency structure of a multivariate distribution from the specific univariate margins. This topic is pursued in the subsequent section presenting a brief overview of different types of multivariate dependence. Comparing uni- and multivariate distribution functions with respect to location and/or variability is the topic of the final section, stochastic orders.

A few examples of applications of these concepts to issues in mathematical psychology are interspersed in the main text. Moreover, the comments and reference section at the end gives a number of references to further pertinent applications.

1.2 Basics

Readers familiar with basic concepts of probability and stochastic processes, including some measure-theoretic terminology, may skip this first section of the chapter.

1.2.1 σ-algebra, probability space, independence, random variable, and distribution function

A fundamental assumption of practically all models and methods of response time analysis is that the response latency measured in a given trial of a

reaction time task is the realization of a random variable. In order to discuss the consequences of treating response time as a random variable or, more generally, as a function of several random variables, some standard concepts of probability theory will first be introduced.[3]

Let Ω be an arbitrary set, often referred to as *sample space* or set of *elementary outcomes* of a random experiment, and $\mathcal{F}$ a system of subsets of Ω endowed with the properties of a *σ-algebra (of events)*, i.e.,

(i) $\emptyset \in \mathcal{F}$ ("impossible" event $\emptyset$).
(ii) If $A \in \mathcal{F}$ then also its complement: $A^c \in \mathcal{F}$.
(iii) For a sequence of events $\{A_n \in \mathcal{F}\}_{n\geq 1}$, then also $\bigcup_{n=1}^{\infty} A_n \in \mathcal{F}$

The pair $(\Omega, \mathcal{F})$ is called *measurable space*. Let $\mathcal{A}$ be any collection of subsets of Ω. Since the power set, $\mathfrak{P}(\Omega)$, is a σ-algebra, it follows that there exists at least one σ-algebra containing $\mathcal{A}$. Moreover, the intersection of any number of σ-algebras is again a σ-algebra. Thus, there exists a unique *smallest* σ-algebra containing $\mathcal{A}$, defined as the intersection of all σ-algebras containing $\mathcal{A}$, called the σ-algebra *generated by* $\mathcal{A}$ and denoted as $\mathcal{S}(\mathcal{A})$.

Definition 1.4 (Probability space) The triple $(\Omega, \mathcal{F}, P)$ is a *probability space* if Ω is a sample space with σ-algebra $\mathcal{F}$ such that P satisfies the following (*Kolmogorov*) axioms:

(1) For any $A \in \mathcal{F}$, there exists a number $P(A) \geq 0$; the probability of A.
(2) $P(\Omega) = 1$.
(3) For any sequence of mutually disjoint events $\{A_n, n \geq 1\}$,

$$P(\bigcup_{n=1}^{\infty} A_n) = \sum_{n=1}^{\infty} P(A_n).$$

Then P is called *probability measure*, the elements of $\mathcal{F}$ are the *measurable* subsets of Ω, and the probability space $(\Omega, \mathcal{F}, P)$ is an example of *measure spaces* which may have measures other than P. Some easy to show consequences of the three axioms are, for measurable sets A, A_1, A_2,

1. $P(A^c) = 1 - P(A)$;
2. $P(\emptyset) = 0$;
3. $P(A_1 \cup A_2) \leq P(A_1) + P(A_2)$;
4. $A_1 \subset A_2 \rightarrow P(A_1) \leq P(A_2)$.

[3] Limits of space do not permit a completely systematic development here, so only a few of the most relevant topics will be covered in detail. For a more comprehensive treatment see the references in the final section (and Chapter 4 for a more general approach).

A set $A \subset \Omega$ is called a *null set* if there exists $B \in \mathcal{F}$, such that $B \supset A$ with $P(B) = 0$. In general, null sets need not be measurable. If they are, the probability space $(\Omega, \mathcal{F}, P)$ is called *complete*[4]. A property that holds everywhere except for those ω in a null set is said to hold (P-)*almost everywhere* (a.e.).

Definition 1.5 (Independence) The events $\{A_k, 1 \leq k \leq n\}$ are *independent* if, and only if,

$$P\left(\bigcap A_{i_k}\right) = \prod P(A_{i_k}),$$

where intersections and products, respectively, are to be taken over all subsets of $\{1, 2, \ldots, n\}$. The events $\{A_n, n \geq 1\}$ are independent if $\{A_k, 1 \leq k \leq n\}$ are independent for all n.

Definition 1.6 (Conditional probability) Let A and B be two events and suppose that $P(A) > 0$. The *conditional probability of B given A* is defined as

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

Remark 1.7 If A and B are independent, then $P(B|A) = P(B)$. Moreover, $P(\cdot|A)$ with $P(A) > 0$ is a probability measure. Because $0 \leq P(A \cap B) \leq P(A) = 0$, null sets are independent of "everything".

The following statements about any subsets (events) of Ω, $\{A_k, 1 \leq k \leq n\}$, turn out to be very useful in many applications in response time analysis and are listed here for later reference.

Remark 1.8 (Inclusion-exclusion formula)

$$\begin{aligned} P\left(\bigcup_{k=1}^{n} A_k\right) = \sum_{k=1}^{n} P(A_k) &- \sum_{1 \leq i \leq j \leq n} P(A_i \cap A_j) \\ &+ \sum_{1 \leq i < j < k \leq n} P(A_i \cap A_j \cap A_k) \\ &+ \cdots - (-1)^n P(A_1 \cap A_2 \cap \cdots \cap A_n) \end{aligned}$$

If the events are independent, this reduces to

$$P\left(\bigcup_{k=1}^{n} A_k\right) = 1 - \prod_{k=1}^{n} (1 - P(A_k)).$$

[4] Any given σ-algebra can be enlarged and the probability measure can be uniquely extended to yield a complete probability space, so it will be assumed in the following without further explicit mentioning that a given probability space is complete.

Definition 1.9 (Measurable function) Let $(\Omega, \mathcal{F})$ and $(\Omega', \mathcal{F}')$ be measure spaces and $T : \Omega \to \Omega'$ a mapping from Ω to Ω'. T is called *$\mathcal{F}$-$\mathcal{F}'$-measurable* if

$$T^{-1}(A') \in \mathcal{F} \text{ for all } A' \in \mathcal{F}',$$

where

$$T^{-1}(A') = \{\, \omega \in \Omega \mid T(\omega) \in A' \}$$

is called the *inverse image* of A'.

For the introduction of (real-valued) random variables, we need a special σ-algebra. Let $\Omega = \Re$, the set of real numbers. The σ-algebra of *Borel sets*, denoted as $\mathcal{B}(\Re) \equiv \mathcal{B}$, is the σ-algebra generated by the set of open intervals[5] of $\Re$. Importantly, two probability measures P and Q on $(\Re, \mathcal{B})$ that agree on all open intervals are identical, $P = Q$.

Definition 1.10 (Random variable) Let $(\Omega, \mathcal{F}, P)$ be a probability space. A (real valued) *random variable* X is a *$\mathcal{F}$-$\mathcal{B}$-measurable function* from the sample space Ω to $\Re$; that is, the inverse image of any Borel set A is $\mathcal{F}$-measurable:

$$X^{-1}(A) = \{\, \omega \mid X(\omega) \in A \} \in \mathcal{F}, \text{ for all } A \in \mathcal{B}.$$

If $X : \Omega \to [-\infty, , +\infty]$, we call X an *extended random variable*.

Random variables that differ only on a null set are called *equivalent* and for two random variables X and Y from the same equivalence class we write $X \sim Y$.

To each random variable X we associate an *induced probability measure*, Pr, through the relation

$$\Pr(A) = P(X^{-1}(A)) = P(\{\, \omega \mid X(\omega) \in A \}), \text{ for all } A \in \mathcal{B}.$$

The induced space $(\Re, \mathcal{B}, \Pr)$ can be shown to be a probability space by simply checking the above (Kolmogorov) axioms. Pr is also called the *distribution* of X.

Remark 1.11 Most often one is only interested in the random variables and "forgets" the exact probability space behind them. Then no distinction is made between the probability measures P and Pr, one omits the brackets $\{$ and $\}$ emphasizing that $\{X \in A\}$ actually is a set, and simply writes $P(X \in A)$, or $P_X(A)$, instead of Pr(A).

[5] It can be shown that $\mathcal{B}$ can equivalently be generated by the sets of closed or half-open intervals of real numbers. In the latter case, this involves extending $\mathcal{B}$ to a σ-algebra generated by the extended real line, $\Re \cup \{+\infty\} \cup \{-\infty\}$.

However, sometimes it is important to realize that two random variable are actually defined with respect to two different probability spaces. A case in point is our introductory Example 1.2 where the random variable representing reaction time to a visual stimulus and the one representing reaction time to the acoustic stimulus are not a-priori defined with respect to the same probability space. In such a case, e.g., it is meaningless to ask whether or not two events are independent (see Section 1.3.5).

"Equality" of random variables can be interpreted in different ways.

Remark 1.12 Random variables X and Y are *equal in distribution* iff they are governed by the same probability measure:

$$X =_d Y \iff P(X \in A) = P(Y \in A) \text{ , for all } A \in \mathcal{B}.$$

X and Y are *point-wise equal* iff they agree for almost all elementary events[6]:

$$X \overset{a.s.}{=} Y \iff P(\{\, \omega \mid X(\omega) = Y(\omega)\}) = 1,$$

i.e., iff X and Y are equivalent random variables, $X \sim Y$.

The following examples illustrates that two random variables may be equal in distribution, and at the same time there is no elementary event where they agree.

Example 1.13 (Gut, 2013, p. 27) Toss a fair coin once and set

$$X = \begin{cases} 1, & \text{if the outcome is heads,} \\ 0, & \text{if the outcome is tails,} \end{cases} \text{ and } Y = \begin{cases} 1, & \text{if the outcome is tails,} \\ 0, & \text{if the outcome is heads.} \end{cases}$$

Clearly, $P(X = 1) = P(X = 0) = P(Y = 1) = P(Y = 0) = 1/2$, in particular, $X =_d Y$. But $X(\omega)$ and $Y(\omega)$ differ for every ω.

Remark 1.14 A function $X : \Omega \to \Re$ is a random variable iff $\{\omega \mid X(\omega) \leq x\} \in \mathcal{F}$, for all $x \in \Re$. This important observation follows because the Borel σ-algebra can be generated from the set of half-open intervals of the form $(-\infty, x]$ (see footnote [5]).

Remark 1.15 Let X_1, X_2 be random variables. Then $\max\{X_1, X_2\}$ and $\min\{X_1, X_2\}$ are random variables as well. In general, it can be shown that a Borel measurable function of a random variable is a random variable.

Definition 1.16 (Distribution function) Let X be a real valued random variable. The *distribution function* of X is

$$F_X(x) = P(X \leq x), \;\; x \in \Re.$$

[6] Here, a.s. is for "almost sure".

From Remark 1.14 it is clear that the distribution function of X, $F_X(.)$, provides a complete description of the distribution $P(.)$ of X via the relation

$$F_X(b) - F_X(a) = P(X \in (a, b]), \text{ for all } a, b, \ -\infty < a \leq b < \infty.$$

Thus, every probability measure on $(\Re, \mathcal{B})$ corresponds uniquely to the distribution function of some random variable(s) but, as shown in Example 1.13, different random variables may have the same distribution function.

Next we list without proof the most important properties of any distribution function (Gut, 2013, p. 30). Let F be a distribution function. Then:

1. F is non-decreasing: $x_1 < x_2 \implies F(x_1) \leq F(x_2)$;
2. F is right-continuous: $\lim_{h \to 0^+} F(x + h) = F(x)$;
3. F has left-hand limits: $\lim_{h \to 0^+} F(x - h) = F(x-) = P(X < x)$;
4. $\lim_{x \to -\infty} F(x) = 0$ and $\lim_{x \to +\infty} F(x) = 1$;
5. F has at most a countable number of discontinuities.

Thus, the jump in F at x is

$$F(x) - F(x-) = P(X = x).$$

Conversely, some of the above properties can be shown to characterize the distribution function:

Proposition 1.17 *If a function $F : \Re \to [0, 1]$ is non-decreasing, right-continuous, $\lim_{x \to -\infty} F(x) = 0$, and $\lim_{x \to \infty} F(x) = 1$, then F is the distribution function of some random variable.*

Proof (see, e.g., Severini, 2005, p. 11) □

Example 1.18 (Weibull/exponential distribution) Define the *Weibull* distribution function for random variable T by

$$F(t) = \begin{cases} 0, & \text{if } t < 0, \\ 1 - \exp[-\lambda t^{\alpha}], & \text{if t} \geq 0, \end{cases}$$

with parameters λ and α, $\lambda > 0, \alpha > 0$. For $\alpha = 1$, this is the *exponential distribution*, Exp(λ), for short. While the exponential is not descriptive for empirically observed reaction times, it plays a fundamental role in response time modeling as a building block for more realistic distributions. Its usefulness derives from the fact that exponentially distributed random variables *have no memory*, or are *memoryless*, in the following sense:

$$P(X > s + t \,|\, X > t) = P(X > s), \text{ for } s, t \geq 0.$$

In the last example, the distribution function is a continuous function. In order to more fully describe the different types of distribution functions that exist, the concept of *Lebesgue integral* is required[7].

Definition 1.19 (Distribution function types) A distribution function F is called
(i) *discrete* iff for some countable set of numbers $\{x_j\}$ and point masses $\{p_j\}$,

$$F(x) = \sum_{x_j \leq x} p_j, \text{ for all } x \in \Re.$$

The function p is called *probability function*;
(ii) *continuous* iff it is continuous for all x;
(iii) *absolutely continuous* iff there exists a non-negative, Lebesgue integrable function f, such that

$$F(b) - F(a) = \int_a^b f(x)\, dx, \text{ for all } a < x < b.$$

The function f is called *density* of f;
(iv) *singular* iff $F \neq 0, F'$ exists and equals 0 a.e. (almost everywhere).

The following *decomposition theorem* (Gut, 2013, pp. 36-38) shows that any distribution can be described uniquely by three types:

Theorem 1.20 *Every distribution function can be decomposed uniquely into a convex combination of three pure types, a discrete one, an absolutely continuous one, and a continuous singular one. Thus, if F is a distribution function, then*

$$F = \alpha F_{ac} + \beta F_d + \gamma F_{cs},$$

where $\alpha, \beta, \gamma \geq 0$ and $\alpha + \beta + \gamma = 1$. This means that

- *$F_{ac}(x) = \int_{-\infty}^{x} f(y)\, dy$, where $f(x) = F'_{ac}(x)$ a.e.;*
- *F_d is a pure jump function with at most a countable number of jumps;*
- *F_{cs} is continuous and $F'_{cs} = 0$ a.e.*

The exponential distribution is an example of an absolutely continuous distribution function, with density $f(t) = \lambda \exp[-\lambda t]$, for $t \geq 0$, and $f(t) = 0$, for $t < 0$.

[7] This topic is treated in introductory analysis books (see references in final section) and some familiarity is assumed here.

1.2.2 Random vectors, marginal and conditional distribution

Definition 1.21 (Random vector) An n-dimensional *random vector* (or, *vector-valued random variable*) $\mathbf{X}$ on a probability space $(\Omega, \mathcal{F}, P)$ is a measurable function from the sample space Ω to $\Re^n$: Thus, this requires that the inverse image of any Borel set is $\mathcal{F}$-measurable:

$$\mathbf{X}^{-1}(A) = \{\, \omega \mid \mathbf{X}(\omega) \in A\} \in \mathcal{F}, \text{ for all } A \in \mathcal{B}^n,$$

where $\mathcal{B}^n$ denotes the σ-algebra of Borel sets of $\Re^n$ generated by the n-dimensional "rectangles" of $\Re^n$. Random vectors are column vectors:

$$\mathbf{X} = (X_1, X_2, \ldots, X_n)',$$

where $'$ denotes transpose.

The *joint*, or *multivariate, distribution function* of $\mathbf{X}$ is

$$F_{X_1,X_2,\ldots,X_n}(x_1,, x_2, \ldots, x_n) = P(X_1 \leq x_1, X_2 \leq x_2, \ldots, X_n \leq x_n),$$

for $x_k \in \Re$, $k = 1, 2, \ldots, n$. This is written more compactly as $F_{\mathbf{X}}(\mathbf{x}) = P(\mathbf{X} \leq \mathbf{x})$, where the event $\{\mathbf{X} \leq \mathbf{x}\}$ is to be interpreted component wise. For discrete distributions, the joint probability function is defined by

$$p_{\mathbf{X}}(\mathbf{x}) = P(\mathbf{X} = \mathbf{x}), \;\; \mathbf{x} \in \Re^n,$$

and in the absolutely continuous case we have a *joint density*

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{\partial^n F_{\mathbf{X}}(\mathbf{x})}{\partial x_1 \partial x_2 \cdots \partial x_n}, \;\; \mathbf{x} \in \Re^n.$$

Let (X, Y) be a 2-dimensional random vector with joint probability distribution $p_{X,Y}(x, y)$. Then the *marginal probability function* is defined as

$$p_X(x) = P(X = x) = \sum_y p_{X,Y}(x, y),$$

and $p_Y(y)$ is defined analogously. The *marginal distribution function* is obtained by

$$F_X(x) = \sum_{u \leq x} p_X(u) = P(X \leq x, Y < +\infty).$$

In the absolutely continuous case, we have

$$F_X(x) = P(X \leq x, Y < \infty) = \int_{-\infty}^{x} \int_{-\infty}^{\infty} f_{X,Y}(u, v)\, dudv,$$

and differentiate to obtain the *marginal density function*,

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)\, dy.$$

The concept of conditional probability (Definition 1.5) extends to distributions:

Definition 1.22 Let X and Y be discrete, jointly distributed random variables. For $P(X = x) > 0$, the *conditional probability function of* Y *given that* $X = x$ equals

$$p_{Y|X=x}(y) = P(Y = y|X = x) = \frac{p_{X,Y}(x, y)}{p_X(x)},$$

and the *conditional distribution function of* Y *given that* $X = x$ is

$$F_{Y|X=x}(y) = \sum_{z \leq y} p_{Y|X=x}(z).$$

Definition 1.23 Let X and Y have a joint absolutely continuous distribution. For $f_X(x) > 0$, the *conditional density function of* Y *given* $X = x$ equals

$$f_{Y|X=x}(y) = \frac{f_{X,Y}(x, y)}{f_X(x)},$$

and the *conditional distribution of* Y *given that* $X = x$ *is*

$$F_{Y|X=x}(y) = \int_{-\infty}^{y} f_{Y|X=x}(z)dz.$$

Analogous formulas hold in higher dimensions and for more general distributions.

Sometimes, it is necessary to characterize a multivariate distribution function by its properties. For the bivariate case, we have, in analogy to Proposition 1.17,

Proposition 1.24 *Necessary and sufficient conditions for a bounded, nondecreasing, and right-continuous function* F *on* $\Re^2$ *to be a bivariate distribution function are:*

1. $\lim_{x_1 \to -\infty} F(x_1, x_2) = 0$, *and* $\lim_{x_2 \to -\infty} F(x_1, x_2) = 0$;
2. $\lim_{(x_1,x_2) \to (+\infty,+\infty)} F(x_1, x_2) = 1$;
3. (rectangle inequality or 2-increasing) *for any* (a_1, a_2), (b_1, b_2) *with* $a_1 < b_1$, $a_2 < b_2$,

$$F(b_1, b_2) - F(a_1, b_2) - F(b_1, a_2) + F(a_1, a_2) \geq 0.$$

This result can be extended to the n-dimensional case. For a proof, see, e.g., Leadbetter et al. (2014, p. 197)

So far, independence has only been defined for events (Definition 1.5). It can be used to define independence of random variables:

Definition 1.25 (Independence of random variables) The random variables $X_1, X_2, \ldots, X_n$ are *independent* iff, for arbitrary Borel measurable sets $A_1, A_2, \ldots, A_n$,

$$P\left(\bigcap_{k=1}^{n}\{X_k \in A_k\}\right) = \prod_{k=1}^{n} P(X_k \in A_k).$$

This can be shown to be equivalent to defining independence of the components $X_1, X_2, \ldots, X_n$ of a random vector $\mathbf{X}$ by

$$F_{\mathbf{X}}(\mathbf{x}) = \prod_{k=1}^{n} F_{X_k}(x_k), \quad \text{for all } \mathbf{x} \in \mathbb{R}^n.$$

For discrete and absolutely continuous distributions, independence is equivalent to factorization of the joint probability function and density into the product of the marginals, respectively.

Example 1.26 (Independent exponentials) Let $X_1, X_2, \ldots, X_n$ be independent random variables following the exponential distribution with parameters $\lambda_i\,(\lambda_i > 0), i = 1 \ldots, n$. Then, it is routine to show that

1. $Z = \min\{X_1, \ldots, X_n\}$ is also exponentially distributed, with parameter $\lambda_Z = \sum_{i=1}^{n} \lambda_i$;
2. for any i,

$$P(Z = X_i) = \frac{\lambda_i}{\lambda_Z}.$$

Proposition 1.27 *Let $X_1, X_2, \ldots, X_n$ be random variables and $h_1, h_2, \ldots, h_n$ be measurable functions. If $X_1, X_2, \ldots, X_n$ are independent, then so are $h_1(X_1), h_2(X_2), \ldots, h_n(X_n)$.*

This proposition easily follows from Definition 1.25. The following bivariate exponential distribution, often denoted as BVE, has generated much interest and will be referred to again in this chapter:

Example 1.28 (Marshall-Olkin BVE) Let $X_1 = \min\{Z_1, Z_{12}\}$, and $X_2 = \min\{Z_2, Z_{12}\}$ where Z_1, Z_2, Z_{12} are independent exponential random variables with parameters $\lambda_1, \lambda_2, \lambda_{12}$, respectively. Marshall and Olkin (1967) define a bivariate distribution by

$$\bar{F}(x_1, x_2) \equiv P(X_1 > x_1, X_2 > x_2) = \exp[-\lambda_1 x_1 - \lambda_2 x_2 - \lambda_{12} \max(x_1, x_2)].$$

Then,

$$\begin{aligned} P(X_1 = X_2) &= P(Z_{12} = \min\{Z_1, Z_2, Z_{12}\}) \\ &= \frac{\lambda_{12}}{\lambda} \\ &> 0, \end{aligned}$$

from Example 1.26, with $\lambda = \lambda_1 + \lambda_2 + \lambda_{12}$. Thus, the BVE distribution $F(x_1, x_2)$ has a singular component. We list without proof (see, e.g. Barlow and Proschan, 1975) some further properties of BVE:

(a) BVE has exponential marginals, with

$$\begin{aligned} \bar{F}_1(x_1) = P(X_1 > x_1) &= \exp[-(\lambda_1 + \lambda_{12})\, x_1], \quad \text{for } t \geq 0, \\ \bar{F}_2(x_2) = P(X_2 > x_2) &= \exp[-(\lambda_2 + \lambda_{12})\, x_1], \quad \text{for } t \geq 0; \end{aligned}$$

(b) BVE has the lack-of-memory property:

$$P(X_1 > x_1 + t, X_2 > x_2 + t \,|\, X_1 > x_1, X_2 > x_2) = P(X_1 > x_1, X_2 > x_2),$$

for all $x_1 \geq 0$, $x_2 \geq 0$, $t \geq 0$, and BVE is the only bivariate distribution with exponential marginals possessing this property;

(c) BVE distribution is a mixture of an absolutely continuous distribution $F_{ac}(x_1, x_2)$ and a singular distribution $F_{cs}(x_1, x_2)$:

$$\bar{F}(x_1, x_2) = \frac{\lambda_1 + \lambda_2}{\lambda}\, \bar{F}_{ac}(x_1, x_2) + \frac{\lambda_{12}}{\lambda} \bar{F}_s(x_1, x_2),$$

where $\bar{F}_s(x_1, x_2) = \exp[-\lambda \max(x_1, x_2)]$ and,

$$\begin{aligned} \bar{F}_{ac}(x_1, x_2) = \frac{\lambda}{\lambda_1 + \lambda_2} &\exp[-\lambda_1 x_1 - \lambda_2 x_2 - \lambda_{12} \max(x_1, x_2)] \\ &- \frac{\lambda_{12}}{\lambda_1 + \lambda_2} \exp[-\lambda \max(x_1, x_2)]. \end{aligned}$$

1.2.3 Expectation, other moments, and tail probabilities

The consistency of the following definitions is based on the theory of Lebesgue integration and the Riemann-Stieltjes integral (some knowledge of which is presupposed here).

Definition 1.29 (Expected value) Let X be a real-valued random variable

on a probability space $(\Omega, \mathcal{F}, P)$. The *expected value of* X (or *mean* of X) is the integral of X with respect to measure P:

$$\mathrm{E}\, X = \int_\Omega X(\omega)\, dP(\omega) = \int X\, dP.$$

For $X \geq 0$, $\mathrm{E}\, X$ is always defined (it may be infinite); for the general X, $\mathrm{E}\, X$ is defined if at least one of $\mathrm{E}\, X^+$ or $\mathrm{E}\, X^-$ is finite[8], in which case

$$\mathrm{E}\, X = \mathrm{E}\, X^+ - \mathrm{E}\, X^-;$$

if both values are finite, that is, if $\mathrm{E}\, |X| < \infty$, we say that X is *integrable*.

Let X, Y be integrable random variables. The following are basic consequences of the (integral) definition of expected value:

1. If $X = 0$ a.s., then $\mathrm{E}\, X = 0$;
2. $|X| < \infty$ a.s., that is, $P(|X| < \infty) = 1$;
3. If $\mathrm{E}\, X > 0$, then $P(X > 0) > 0$;
4. Linearity: $\mathrm{E}\,(aX + bY) = a\mathrm{E}\, X + b\mathrm{E}\, Y$, for any $a, b \in \Re$;
5. $\mathrm{E}\, XI\{X \neq 0\} = \mathrm{E}\, X$;[9]
6. Equivalence: If $X = Y$ a.s., then $\mathrm{E}\, X = \mathrm{E}\, Y$;
7. Domination: If $Y \leq X$ a.s., then $\mathrm{E}\, X \leq \mathrm{E}\, Y$;
8. Domination: If $|Y| \leq X$ a.s., then $\mathrm{E}\, |Y| \leq \mathrm{E}\, |X|$.

Remark 1.30 Define an integral over measurable sets $A \in \mathcal{F}$:

$$\mu_X(A) = \mathrm{E}\, XI\{A\} = \int_A X\, dP = \int_\Omega XI\{A\}\, dP.$$

In other words, $\mu_X(\cdot)$ is an"ordinary" expectation applied to the random variable $XI\{\cdot\}$. Note that $\mu_{I\{B\}}(A) = P(B \cap A)$, for $B \in \mathcal{F}$.

Expected value of random variables has so far been defined in terms of integrals over the sample space of the underlying probability space. Just as the probability space behind the random variables is "hidden" once they have been defined by the induced measure (see Remark 1.11), one can compute an integral on the real line rather than over the probability space invoking the Riemann-Stieltjes integral: Suppose X is integrable, then

$$\mathrm{E}\, X = \int_\Omega X\, dP = \int_\Re x\, dF_X(x).$$

[8] $X^+ = \sup\{+X, 0\}$ and $X^- = \sup\{-X, 0\}$.
[9] $I\{A\}$ is the indicator function for a set A.

Example 1.31 (Example 1.18 cont'ed) Expected value (mean) of an exponentially distributed random variable T is

$$\mathrm{E}\,T = \int_{\Re} t\, dF_T(t) = \int_{\Re} t\, f_T(t)\, dt = \int_0^{\infty} t\, \lambda \exp[-\lambda t]\, dt = 1/\lambda,$$

using the fact that T is absolutely continuous, that is, a density exists.

Since measurable functions of random variables are new random variables (Remark 1.15), one can compute expected values of functions of random variables: let X be a random variable and suppose that g is a measurable function such that $g(X)$ is an integrable random variable. Then

$$\mathrm{E}\, g(X) = \int_{\Omega} g(X)\, dP = \int_{\Re} g(x)\, dF_X(x).$$

For proof of this and the previous observations, see (Gut, 2013, p. 60).

Example 1.32 (Expected value of $1/T$) Let T be a positive random variable with density $f_T(t)$. Then T^{-1} has density $f_T(1/t)/t^2$. This is useful in RT analysis if one wants to go from reaction time (t) to "speed" ($1/t$).

An alternative way of computing expected values, which is often useful in RT analyses, is based on the following *tail probability formula*:

Remark 1.33 For a real-valued random variable with distribution function F,

$$\mathrm{E}X = \int_0^{\infty} (1 - F(x))\, dx - \int_{-\infty}^{0} F(x)\, dx,$$

assuming $\mathrm{E}|X| < \infty$. For nonnegative random variables this reduces to

$$\mathrm{E}X = \int_0^{\infty} (1 - F(x))\, dx,$$

where both sides of the equation are either finite or infinite. If X is a nonnegative, integer-valued random variable, then

$$\mathrm{E}X = \sum_{n=1}^{+\infty} P(X \geq n).$$

The expected value measures the "center of gravity" of a distribution, it is a *location* measure. Other properties of a distribution, like dispersion (spread), skewness (asymmetry), and kurtosis (peakedness), are captured by different types of *moments*:

Definition 1.34 Let X be a random variable. The

1. *moments* are $\mathrm{E}X^n, n = 1, 2, \ldots$;
2. *central moments* are $\mathrm{E}(X - \mathrm{E}X)^n, n = 1, 2, \ldots$;
3. *absolute moments* are $\mathrm{E}|X|^n, n = 1, 2, \ldots$;
4. *absolute central moments* are $\mathrm{E}|X - \mathrm{E}|^n, n = 1, 2, \ldots$.

Clearly, the first moment is the mean, $\mathrm{E}X = \mu$. The second central moment is called *variance*:

$$\mathrm{Var}X = \mathrm{E}(X - \mu)^2 \ \ (= \mathrm{E}X^2 - (\mathrm{E}X)^2).$$

A (standardized) measure of skewness is obtained using the third central moment:

$$\gamma_1 = \frac{\mathrm{E}(X - \mu)^3}{[\mathrm{Var}X]^{3/2}}. \tag{1.1}$$

A classic, but not undisputed[10], measure of kurtosis is based on the fourth (standardized) moment:

$$\beta_2 = \frac{\mathrm{E}(X - \mu)^4}{(\mathrm{VarX})^2}. \tag{1.2}$$

Although moments often provide a convenient summary of the properties of a distribution function, they are not always easy to work with or may not be available in simple, explicit form. Alternative ways of comparing distributions with respect to location, dispersion, skewness, and kurtosis will be discussed in Section 1.3.9.

Joint moments of two or more random variables, introduced next, are also commonly used.

Definition 1.35 Let $(X, Y)'$ be a random vector. The *covariance* of X and Y is

$$\mathrm{Cov}(X, Y) = \mathrm{E}(X - \mathrm{E}X)(Y - \mathrm{E}Y) \ \ (= \mathrm{E}XY - \mathrm{E}X\mathrm{E}Y).$$

A bivariate version of the tail-probability formula in Remark 1.33, also known as *Hoeffding's identity* (Hoeffding, 1940), will be quite helpful later on in Section 1.3.7:

Proposition 1.36 *Let F_{XY}, F_X, and F_Y denote the joint and marginal distribution functions for random variables X and Y, and suppose that $\mathrm{E}|XY|$, $\mathrm{E}|X|$, and $\mathrm{E}|Y|$ are finite; then*

$$\mathrm{Cov}(X, Y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} [F_{XY}(x, y) - F_X(x)F_Y(y)] \, dx \, dy.$$

[10] see (Balanda and MacGillivray, 1990)

An instructive proof of this is found in Shea (1983). Finally, the (linear) *correlation coefficient* is defined as

$$\mathrm{Cor}(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}X\mathrm{Var}Y}}$$

1.2.4 Product spaces and convolution

Definition 1.37 (Product space) For any finite number of measurable spaces,

$$(\Omega_1,\mathcal{F}_1),(\Omega_2,\mathcal{F}_2),\ldots,(\Omega_n,\mathcal{F}_n),$$

one can construct a *product measurable space* in a natural way. The set of all ordered n-tuples $(\omega_1,\ldots,\omega_n)$, with $\omega_i \in \Omega_i$ for each i, is denoted by $\Omega_1 \times \ldots \times \Omega_n$ is called the *product* of the $\{\Omega_i\}$. A set of the form

$$A_1 \times \ldots \times A_n = \{(\omega_1,\ldots,\omega_n) \in \Omega_1 \times \ldots \times \Omega_n \,|\, \omega_i \in \Omega_i \text{ for each } i\},$$

with $A_i \in \mathcal{F}_i$ for each i, is called a *measurable rectangle*. The *product σ-algebra* $\mathcal{F}_1 \otimes \ldots \otimes \mathcal{F}_n$ is defined to be the σ-algebra generated by the set of all measurable rectangles. Within the measure-theoretic approach to probability (Pollard, 2002), one can then construct a unique product measure $\mathbb{P}$ on the product σ-algebra such that

$$\mathbb{P}(A_1 \times \ldots \times A_n) = \prod_{k=1}^{n} P_k(A_k) \text{ for all } A_k \in \mathcal{F}_k, 1 \leq k \leq n.$$

Note that the product probability measure has built-in independence.

Given that models of response time typically make the assumption that observable RT is composed of different stages (e.g., Donders' method of subtraction) being able to calculate the distribution function for the sum (or difference) of random variables is clearly essential. Consider the product space $(\Omega_1 \times \Omega_2, \mathcal{F}_1 \times \mathcal{F}_2, P_1 \times P_2)$, and suppose that (X_1, X_2) is a two-dimensional random variable whose marginal distribution functions are F_1 and F_2, respectively. The convolution formula provides the distribution of $X_1 + X_2$.

Proposition 1.38 *In the above setting, for independent X_1 and X_2,*

$$F_{X_1+X_2}(u) = \int_{-\infty}^{\infty} F_1(u-y)\, dF_2(y).$$

If, in addition, X_2 is absolutely continuous with density f_2, then

$$F_{X_1+X_2}(u) = \int_{-\infty}^{\infty} F_1(u-y) f_2(y), dy.$$

If X_1 is absolutely continuous with density f_1, the density of the sum equals

$$f_{X_1+X_2}(u) = \int_{-\infty}^{\infty} f_1(u-y)\, dF_2(y).$$

If both are absolutely continuous, then

$$f_{X_1+X_2}(u) = \int_{-\infty}^{\infty} f_1(u-y) f_2(y)\, dy.$$

Proof (see, e.g., Gut, 2013, p. 68) □

Example 1.39 (Schwarz, 2001) Let D be a random variable following an *inverse Gauss* (or *Wald*) distribution (see Example 1.3) with density[11]

$$f(w \,|\, \mu, \sigma, a) = \frac{a}{\sigma\sqrt{2\pi w^3}} \exp\left[-\frac{(a-\mu w)^2}{2\sigma^2 w}\right];$$

The distribution function is

$$F(w \,|\, \mu, \sigma, a) = \Phi\left(\frac{\mu w - a}{\sigma\sqrt{w}}\right) + \exp\left(\frac{2a\mu}{\sigma^2}\right)\Phi\left(-\frac{\mu w + a}{\sigma\sqrt{w}}\right),$$

with Φ the standard normal distribution function. Let M be an exponentially distributed random variable with density

$$g(w \,|\, \gamma) = \gamma \exp(-\gamma w).$$

If D and M are independent, their sum has density h

$$h(t \,|\, \mu, \sigma, \gamma) = \int_0^t f(w \,|\, \mu, \sigma, a) \times g(t-w \,|\, \gamma)\, dw.$$

Inserting for f and g, after some algebraic manipulations, this yields

$$h(t \,|\, \mu, \sigma, a, \gamma) = \gamma \exp\left[-\gamma t + \frac{a(\mu-k)}{\sigma^2}\right] \times F(t \,|\, k, \sigma, a),$$

with $k \equiv \sqrt{\mu^2 - 2\gamma\sigma^2} \geq 0$. The latter condition amounts to $\mathrm{E}M = 1/\gamma \geq 2\mathrm{Var}D/\mathrm{E}D$; if D is interpreted as time to detect and identify a signal and M the time to execute the appropriate response, this condition is typically satisfied. Otherwise, a more elaborate computation of the convolution is required (see Schwarz, 2002). In the above interpretation of D, often the Wald

[11] This parametrization expresses D as the first-passage time through a level $a > 0$ in a Wiener diffusion process starting at $x = 0$ with drift $\mu > 0$ and variance $\sigma^2 > 0$.

distribution is replaced by the normal distribution (Hohle, 1965; Palmer et al., 2011).

1.2.5 Stochastic processes

A *stochastic process* X is a collection of random variables $\{X_t : t \in T\}$, all defined on the same probability space, say $(\Omega, \mathcal{F}, P)$. The *index set* T can be an arbitrary set but, in most cases, T represents time. X can be a *discrete-parameter* process, $T = \{1, 2, \ldots\}$, or a *continuous-parameter* process, with T an interval of the real line. For a fixed $\omega \in \Omega$, the function $X_t(\omega)$ of t is called a *realization* or *sample path* of X at ω. For each n-tuple $\mathbf{t} = (t_1, \ldots, t_n)$ of distinct elements of T, the random vector $(X_{t_1}, \ldots, X_{t_n})$ has a joint distribution function

$$F_{\mathbf{t}}(\mathbf{x}) = P(X_{t_1} \leq x_1, \ldots, X_{t_n} \leq x_n).$$

As $\mathbf{t}$ ranges over all n-tuples of T, the collection $\{F_{\mathbf{t}}\}$ is the *system of finite-dimension distributions* of X from which the distributional properties of the stochastic process can be studied[12].

A stochastic process $\{N(t), t \geq 0\}$ is called a *counting process* if $N(t)$ represents the total number of "events" that have occurred up to time t. Hence, a counting process $N(t)$ satisfies:

1. $N(t) \geq 0$;
2. $N(t)$ is integer-valued;
3. If $s < t$, then $N(s) \leq N(t)$;
4. For $s < t$, $N(t) - N(s)$ equals the number of events that have occurred in the interval $(s, t]$.

A counting process has *independent increments* if the numbers of events that occur in disjoint time intervals are independent random variables. It has *stationary increments* if the distribution of the number of events in the interval $(t_1 + s, t_2 + s]$, that is, $N(t_2 + s) - N(t_1 + s)$, does not depend on s, for all $t_1 < t_2$ and $s > 0$.

Let X_1 be the time of the first event of a counting process, and X_n the time between the $(n-1)$th and the nth event. The sequence $\{X_n, n \geq 1\}$ is called the *sequence of interarrival times*. Moreover,

$$S_n = \sum_{i=1}^{n} X_i, \quad n \geq 1,$$

[12] For the general measure-theoretic approach to stochastic processes, including *Kolmogorov's existence theorem*, see e.g. Leadbetter et al. (2014).

is the *waiting time* until the nth event.

The Poisson process

The most simple counting process is the Poisson process, which is a basic building block for many RT models. It can be defined in several, equivalent ways (see, e.g., Gut, 2013). Here we introduce it as follows:

Definition 1.40 The counting process $\{N(t), t \geq 0\}$ is said to be a *Poisson process with rate* λ, $\lambda > 0$, if:

1. $N(0) = 0$;
2. $N(t)$ has independent increments;
3. the number of events in any interval of length t is Poisson distributed with mean λt; that is, for all $s, t \geq 0$,

$$P(N(t+s) - N(s) = n) = \exp[-\lambda t]\frac{(\lambda t)^n}{n!}, \quad n = 0, 1, \ldots$$

Obviously, a Poisson process has stationary increments. Moreover, $\mathrm{E}[N(t)] = \lambda t$. For the interarrival times, one has

Proposition 1.41 *The interarrival times* $\{X_n, n \geq 1\}$ *of a Poisson process with rate* λ *are independent identically distributed (i.i.d.) exponential random variables with mean* $1/\lambda$.

Proof Obviously,

$$P(X_1 > t) = P(N(t) = 0) = \exp[-\lambda t],$$

so X_1 has the exponential distribution. Then

$$\begin{aligned} P(X_2 > t \,|\, X_1 = s) =& P(\text{no events in } (s, s+t] \,|\, X_1 = s) \\ =& P(\text{no events in } (s, s+t]) \ \text{ (by independent increments)} \\ =& \exp[-\lambda t] \ \text{ (by stationary increments).} \end{aligned}$$

□

Given the memoryless property of the exponential, the last result shows that the Poisson process *probabilistically* restarts itself at any point in time. For the waiting time until the nth event of the Poisson process, the density is

$$f(t) = \lambda \exp[-\lambda t]\,\frac{(\lambda t)^{n-1}}{(n-1)!}, \quad t \geq 0, \tag{1.3}$$

i.e., the gamma density, Gamma(n, λ). This follows using convolution (or

moment generating functions), but it can also be derived by appealing to the following logical equivalence:

$$N(t) \geq n \Leftrightarrow S_n \leq t.$$

Hence,

$$P(S_n \leq t) = P(N(t) \geq n) \tag{1.4}$$

$$= \sum_{j=n}^{\infty} \exp[-\lambda t] \frac{(\lambda t)^j}{j!}, \tag{1.5}$$

which upon differentiation yields the gamma density.

Note that a function f is said to be $o(h)$ if

$$\lim_{h \to 0} \frac{f(h)}{h} = 0.$$

One alternative definition of the Poisson process is contained the following:

Proposition 1.42 *The counting process $\{N(t), t \geq 0\}$ is a Poisson process with rate $\lambda, \lambda > 0$, if and only if:*

1. $N(0) = 0$;
2. *the process has stationary and independent increments;*
3. $P(N(h) = 1) = \lambda h + o(h)$;
4. $P(N(h) \geq 2) = o(h)$.

Proof (see Ross, 1983, p. 32–34) □

The non-homogeneous Poisson process

Dropping the assumption of stationarity leads to a generalization of the Poisson process that also plays an important role in certain RT model classes, the *counter models*.

Definition 1.43 The counting process $\{N(t), t \geq 0\}$ is a *nonstationary*, or *non-homogeneous Poisson process with intensity function* $\lambda(t)$, $t \geq 0$, if:

1. $N(0) = 0$;
2. $\{N(t), t \geq 0\}$ has independent increments;
3. $P(N(t+h) - N(t) = 1) = \lambda(t)h + o(h)$;
4. $P(N(t+h) - N(t)) \geq 2) = o(h)$.

Letting

$$m(t) = \int_0^t \lambda(s)\, ds,$$

it can be shown that

$$P(N(t+s)-N(t)=n)= \exp[-(m(t+s)-m(t))][m(t+s)-m(t)]^n/n!,\ \ n=0,1,2\ldots, \quad (1.6)$$

that is, $N(t+s)-N(t)$ is Poisson distributed with mean $m(t+s)-m(t)$.

A non-homogeneous Poisson process allows for the possibility that events may be more likely to occur at certain times than at other times. Such a process can be interpreted as being a random sample from a homogeneous Poisson process, if function $\lambda(t)$ is bounded, in the following sense (p. 46 Ross, 1983):

Let $\lambda(t) \leq \lambda$, for all $t \geq 0$; suppose that an event of the process is counted with probability $\lambda(t)/\lambda$, then the process of counted events is a non-homogeneous Poisson process with intensity function $\lambda(t)$: Properties 1, 2, and 3 of Definition 1.43 follow since they also hold for the homogenous Poisson process. Property 4 follows since

$$\begin{aligned} P(\text{one counted event in } (t,t+h)) &= P(\text{one event in } (t,t+h))\frac{\lambda(t)}{\lambda}+o(h) \\ &= \lambda h\frac{\lambda(t)}{\lambda}+o(h) \\ &= \lambda(t)h+o(h). \end{aligned}$$

An example of this process in the context of *record values* is presented at the end of Section 1.3.4.

1.3 Specific topics

1.3.1 Exchangeability

When stochastic independence does not hold, the multivariate distribution of a random vector can be a very complicated object that is difficult to analyze. If the distribution is *exchangeable*, however, the associated distribution function F can often be simplifed and written in a quite compact form. Exchangeability intuitively means that the dependence structure between the components of a random vector is completely symmetric and does not depend on the ordering of the components (Mai and Scherer, 2012, p. 39 pp.).

Definition 1.44 (Exchangeability) Random variables $X_1,\ldots,X_n$ are *exchangeable* if, for the random vector $(X_1,\ldots,X_n)$,

$$(X_1,\ldots,X_n) =_d (X_{i_1},\ldots,X_{i_n})$$

for all permutations $(i_1, \ldots, i_n)$ of the subscripts $(1, \ldots, n)$. Furthermore, an infinite sequence of random variables $X_1, X_2, \ldots$ is *exchangeable* if any finite subsequence of $X_1, X_2, \ldots$ is exchangeable.

Notice that a subset of exchangeable random variables can be shown to be exchangeable. Moreover, any collection of independent identically distributed random variables is exchangeable. The distribution function of a random vector of exchangeable random variables is invariant with respect to permutations of its arguments. The following proposition shows that the correlation structure of a finite collection of exchangeable random variables is limited:

Proposition 1.45 *Let $X_1, \ldots, X_n$ be exchangeable with existing pairwise correlation coefficients $\rho(X_i, X_j) = \rho$, $i \neq j$. Then $\rho \geq -1/(n-1)$.*

Proof Without loss of generality, we scale the variables such that $\mathrm{E}Z_i = 0$ and $\mathrm{E}Z_i^2 = 1$. Then

$$0 \leq \mathrm{E}(\sum Z_i)^2 = \sum_i \mathrm{E}Z_i^2 + \sum_{i \neq j} \mathrm{E}Z_i Z_j = n + n(n-1)\rho,$$

from which the inequality follows immediately. □

Note that this also implies that $\rho \geq 0$, for an infinite sequence of exchangeable random variables. The assumption that a finite number of random variables are exchangeable is significantly different from the assumption that they are a finite segment of an infinite sequence of exchangeable random variables (for a simple example, see Galambos, 1978, pp. 127–8).

The most important result for exchangeable random variables is known as "de Finetti's theorem", which is often stated in words such as "An infinite exchangeable sequence is a mixture of i.i.d. sequences". The precise result, however, requires measure theoretic tools like *random measures*. First, let us consider the "Bayesian viewpoint" of sampling.

Let $X_1, \ldots, X_n$ be observations on a random variable with distribution function $F(x, \theta)$, where θ is a random variable whose probability function p is assumed to be independent of the X_j. For a given value of θ, the X_j are assumed to be i.i.d. Thus, the distribution of the sample is

$$P(X_1 \leq x_1, \ldots, X_n \leq x_n) = \int \prod_{j=1}^{n} F(x_j, \theta) p(\theta)\, d\theta = \mathrm{E}\left[\prod_{j=1}^{n} F(x_j, \theta)\right]. \tag{1.7}$$

Obviously, $X_1, \ldots, X_n$ are exchangeable random variables. The following theorem addresses the converse of this statement.

Theorem 1.46 ("de Finetti's theorem") *An infinite sequence $X_1, X_2, \ldots$ of random variables on a probability space $(\Omega, \mathcal{F}, P)$ is exchangeable if, and only if, there is a sub-σ-algebra $\mathcal{G} \subset \mathcal{F}$ such that, given $\mathcal{G}$, $X_1, X_2, \ldots$ are a.s. independent and identically distributed, that is,*

$$P(X_1 \leq x_1, \ldots, X_n \leq x_n \,|\, \mathcal{G}) = \prod_{j=1}^{n} P(X_1 \leq x_j \,|\, \mathcal{G}) \text{ a.s.},$$

for all $x_1, \ldots x_n \in \Re$.

Consequently, the n-dimensional distribution of the X_j can always be expressed as in Equation 1.7. A more complete statement of the theorem and a proof can be found, e.g., in Aldous (1985). The above theorem is considered a key result in the context of Bayesian (hierarchical) modeling, with the probability function of θ taking the role of a *prior distribution.*

1.3.2 Quantile functions

While specific distributions are typically described by certain moments (mean and variance) and transforms thereof, in particular skewness (γ_1 from Equation 1.1) and kurtosis (β_2 from Equation 1.2), this approach has certain shortcomings. For example, the variance may not be finite; there exist asymmetric distributions with $\gamma_1 = 0$; and β_2 does not reflect the sign of the difference between mean and median, which is a traditional basis for defining skewness. The concept of a *quantile function* is closely related to that of a distribution function and plays an increasingly important role in RT analysis.

The specification of a distribution through its quantile function takes away the need to describe a distribution through its moments. Alternative measures in terms of quantiles are available, e.g. the median as a measure of location, defined as $M = Q(0.5)$, or the interquartile range, as a measure of dispersion, defined as $IQR = Q(.75) - Q(.25)$. Another measure of dispersion in terms of quantiles, the *quantile spread*, will be discussed in section 1.3.9.

Definition 1.47 (Quantile function) Let X be a real-valued random variable with distribution function $F(x)$. Then the *quantile function* of X is defined as

$$Q(u) = F^{-1}(u) = \inf\{x \,|\, F(x) \geq u\}, \quad 0 \leq u \leq 1. \tag{1.8}$$

For every $-\infty < x < +\infty$ and $0 < u < 1$, we have

$$F(x) \geq u \quad \text{if, and only if, } Q(u) \leq x.$$

Thus, if there exists x with $F(x) = u$, then $F(Q(u)) = u$ and $Q(u)$ is the smallest value of x satisfying $F(x) = u$. If $F(x)$ is continuous and strictly increasing, $Q(u)$ is the unique value x such that $F(x) = u$.

Example 1.48 Let X be an $\text{Exp}(\lambda)$ random variable, $F(x) = 1-\exp[-\lambda x]$, for $x > 0$ and $\lambda > 0$. Then $Q(u) = \lambda^{-1}(-\log(1-u))$.

If $f(x)$ is the probability density function of X, then $f(Q(u))$ is called the *density quantile function*. The derivative of $Q(u)$, i.e.,

$$q(u) = Q'(u),$$

is known as the *quantile density function* of X. Differentiating $F(Q(u)) = u$, we find

$$q(u)f(Q(u)) = 1. \tag{1.9}$$

We collect a number of properties of quantile functions:

1. $Q(u)$ is non-decreasing on $(0, 1)$ with $Q(F(x)) \leq x$, for all $-\infty < x < +\infty$ for which $0 < F(x) < 1$;
2. $F(Q(u)) \geq u$, for any $0 < u < 1$;
3. $Q(u)$ is continuous from the left, or $Q(u-) = Q(u)$;
4. $Q(u+) = \inf\{x : F(x) > u\}$ so that $Q(u)$ has limits from above;
5. any jumps of $F(x)$ are flat points of $Q(u)$ and flat points of $F(x)$ are jumps of $Q(u)$;
6. a non-decreasing function of a quantile function is a quantile function.
7. the sum of two quantile functions is again a quantile function;
8. two quantile density functions, when added, produce another quantile density function;
9. if X has quantile function $Q(u)$, then $1/X$ has quantile function $1/Q(1-u)$.

The following simple but fundamental fact shows that any distribution function can be conceived as arising from the uniform distribution transformed by $Q(u)$:

Proposition 1.49 *If U is a uniform random variable over $[0, 1]$, then $X = Q(U)$ has its distribution function as $F(x)$.*

Proof

$$P(Q(U) \leq x) = P(U \leq F(x)) = F(x).$$

□

1.3.3 Survival analysis

Survival function and hazard function

Given that time is a central notion in survival analysis, actuarial, and reliability theory it is not surprising that many concepts from these fields have an important role to play in the analysis of psychological processes as well. In those contexts, realization of a random variable is interpreted as a "critical event", e.g., death of an individual or breakdown of a system of components.

Let $F_X(x)$ be the distribution function of random variable X; then $S_X(x) = 1 - F_X(x) = P(X > x)$, for $x \in \Re$, is called the *survival function* (or, *tail function*) of X. From the properties of distribution functions, it follows that the survival function is non-increasing, right-continuous and such that

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

While a more general framework for the following concepts exists, e.g., including discrete distribution functions, in the remainder of this section we assume all random variables or vectors to be non-negative with absolutely continuous distribution functions.

Definition 1.50 The hazard rate h_X of X with distribution function $F_X(x)$ and density $f_X(x)$ is defined as

$$h_X(x) = \frac{f_X(x)}{S_X(x)} = -\frac{d \log S_X(x)}{dx}. \tag{1.10}$$

In reliability theory, the hazard rate is often called *failure rate*. The following, equivalent definition helps in understanding its meaning.

Proposition 1.51

$$h_X(x) = \lim_{\Delta x \downarrow 0} \frac{P(x < X \leq x + \Delta x \,|\, X > x)}{\Delta x} \tag{1.11}$$

Proof For $P(X > x) > 0$,

$$\begin{aligned} P(x < X \leq x + \Delta x \,|\, X > x) &= \frac{P(x < X \leq x + \Delta x)}{1 - F_X(x)} \\ &= \frac{F_X(x + \Delta x) - F_X(x)}{1 - F_X(x)}, \end{aligned}$$

whence it follows that

$$\lim_{\Delta x \downarrow 0} \frac{P(x < X \leq x + \Delta x \,|\, X > x)}{\Delta x} = \frac{1}{1 - F_X(x)} \lim_{\Delta x \downarrow 0} \frac{F_X(x + \Delta x) - F_X(x)}{\Delta x}$$
$$= \frac{1}{S_X(x)} f_X(x)$$

□

Thus, for "sufficiently" small Δx, the product $h_X(x) \cdot \Delta x$ can be interpreted as the probability of the "critical event" occurring in the next instant of time given it has not yet occurred. Or, written with the $o()$ function:

$$P(x < X \leq x + \Delta x \,|\, X > x) = h_X(x)\Delta x + o(\Delta x).$$

The following conditions have been shown to be necessary and sufficient for a function $h(x)$ to be the hazard function of some distribution (see Marshall and Olkin, 2007).

1. $h(x) \geq 0$;
2. $\int_0^x h(t)\, dt < +\infty$, for some $x > 0$;
3. $\int_0^{+\infty} h(t)\, dt = +\infty$;
4. $\int_0^x h(t)\, dt = +\infty$ implies $h(y) = +\infty$, for every $y > x$.

It is easy to see that the exponential distribution with parameter $\lambda > 0$ corresponds to a constant hazard function, $h_X(x) = \lambda$ expressing the "lack of memory" property of that distribution.

A related quantity is the *cumulative* (or, *integrated*) *hazard function* $H(x)$ defined as

$$H(x) = \int_0^x h(t)\, dt = -\log S(x).$$

Thus,

$$S(x) = \exp\left[-\int_0^x h(t)\, dt\right],$$

showing that the distribution function can be regained from the hazard function.

Example 1.52 (Hazard function of the minimum) Let $X_1, X_2, \ldots, X_n$ be independent random variables, defined on a common probability space, and $Z = \min(X_1, X_2, \ldots, X_n)$. Then

$$\begin{aligned} S_Z(x) &= P(Z > z) \\ &= P(X_1 > x, X_2 > x, \ldots, X_n > x) \\ &= S_{X_1}(x) \ldots S_{X_n}(x). \end{aligned}$$

Logarithmic differentiation leads to

$$h_Z(x) = h_{X_1}(x) + \ldots + h_{X_n}(x).$$

In reliability theory, this model constitutes a *series system* with n independent components, each having possibly different life distributions: the system breaks down as soon as any of the components does, i.e. the "hazards" cumulate with the number of components involved. For example, the Marshall-Olkin BVE distribution (Example 1.28) can be interpreted to describe a series system exposed to three independent "fatal" break-downs with a hazard function $\lambda = \lambda_1 + \lambda_2 + \lambda_{12}$.

Hazard quantile function

For many distributions, like the normal, the corresponding quantile function cannot be obtained in closed form, and the solution of $F(x) = u$ must be obtained numerically. Similarly, there are quantile functions for which no closed-form expressions for $F(x)$ exists. In that case, analyzing the distribution via its hazard function is of limited use, and a translation in terms of quantile functions is more promising (Nair et al., 2013).

We assume an absolutely continuous $F(x)$ so that all quantile related functions are well defined. Setting $x = Q(u)$ in Equation 1.10 and using the relationship $f(Q(u)) = [q(u)]^{-1}$, the *hazard quantile function* is defined as

$$H(u) = h(Q(u)) = [(1-u)q(u)]^{-1}. \tag{1.12}$$

From $q(u) = [(1-u)H(u)]^{-1}$, we have

$$Q(u) = \int_0^u \frac{dv}{(1-v)H(v)}.$$

The following example illustrates these notions.

Example 1.53 (Nair et al., 2013, p. 47) Taking

$$Q(u) = u^{\theta+1}(1 + \theta(1-u)), \quad \theta > 0,$$

we have

$$q(u) = u^{\theta}[1 + \theta(\theta+1)(1-u)],$$

and so

$$H(u) = [(1-u)u^{\theta}(1 + \theta(\theta+1)(1-u))]^{-1}.$$

Note that there is no analytic solution for $x = Q(u)$ that gives $F(x)$ in terms of x.

Competing risks models: hazard-based approach

An extension of the hazard function concept, relevant for response times involving, e.g., a choice between different alternatives, is to assume that not only the time of breakdown but also its *cause* C are observable. We define a joint *sub-distribution function* F for (X, C), with X the time of breakdown and C its cause, where $C = 1, \ldots, p$, a (small) number of labels for the different causes:

$$F(x, j) = P(X \leq x, C = j),$$

and a *sub-survival distribution* defined as

$$S(x, j) = P(X > x, C = j).$$

Then, $F(x, j) + S(x, j) = p_j = P(C = j)$. Thus, $F(x, j)$ is not a proper distribution function because it only reaches the values p_j instead of 1 when $x \to +\infty$. It is assumed implicitly that $p_j > 0$ and $\sum_1^p p_j = 1$.

Note that $S(x, j)$ is not, in general, equal to the probability that $X > x$ for breakdowns of type j; rather, that is the conditional probability

$$P(X > x \,|\, C = j) = S(x, j)/p_j.$$

For continuous X, the *sub-density function* is $f(x, j) = -dS(x, j)/dx$. The *marginal survival function* and *marginal density* of X are calculated from

$$S(x) = \sum_{j=1}^{p} S(x, j) \quad \text{and} \quad f(x) = -dS(x)/dx = \sum_{j=1}^{p} f(x, j). \tag{1.13}$$

Next, one defines the hazard function for breakdown from cause j, in the presence of all risks, in similar fashion as the overall hazard function:

Definition 1.54 (Sub-hazard function) The sub-hazard function (for breakdown from cause j in the presence of all risks $1, \ldots, p$) is

$$\begin{aligned} h(x, j) &= \lim_{\Delta x \downarrow 0} \frac{P(x < X \leq x + \Delta x, C = j \,|\, X > x)}{\Delta x} \\ &= \lim_{\Delta x \downarrow 0} (\Delta x)^{-1} \, \frac{S(x, j) - S(x + \Delta x, j)}{S(x)} \\ &= \frac{f(x, j)}{S(x)}. \end{aligned} \tag{1.14}$$

Sometimes, $h(x, j)$ is called *cause-specific* or *crude hazard function*. The overall hazard function[13] is $h(x) = \sum_{j=1}^{p} h(x, j)$.

[13] Dropping subscript X here for simplicity.

Example 1.55 (Weibull competing risks model) One can specify a Weibull *competing risks model* by assuming sub-hazard functions of the form

$$h(x, j) = \alpha_j \beta_j^{-\alpha_j} x^{\alpha_j - 1},$$

with $\alpha_j, \beta_j > 0$. The overall hazard function then is

$$h(x) = \sum_{j=1}^{p} \alpha_j \beta_j^{-\alpha_j} x^{\alpha_j - 1},$$

and the marginal survival function follows as

$$S(x) = \exp\left\{ -\int_0^x h(y)\, dy \right\} = \exp\left\{ -\sum_{j=1}^{p} (x/\beta_j)^{\alpha_j} \right\}.$$

In view of Example 1.52, this shows that X has the distribution of $\min(X_1, \ldots, X_p)$, the X_j being independent Weibull variates with parameters (α_j, β_j). The sub-densities $f(x, j)$ can be obtained as $h(x, j)S(x)$, but integration of this in order to calculate the sub-survival distributions is generally intractable.

A special type of competing risks model emerges when *proportionality of hazards* is assumed, that is, when $h(x, j)/h(x)$ does not dependent on x for each j. It means that, over time, the instantaneous risk of a "critical event" occurring may increase or decrease, but the relative risks of the various causes remain invariant. The meaning of the assumption is captured more precisely by the following lemma (for proof see, e.g., Kochar and Proschan, 1991).

Lemma 1.56 (Independence of time and cause) *The following conditions are equivalent:*

1. *$h(x, j)/h(x)$ does not dependent on x for all j (proportionality of hazards);*
2. *the time and cause of breakdown are independent;*
3. *$h(x, j)/h(x, k)$ is independent of x, for all j and k.*

In this case, $h(x, j) = p_j\, h(x)$ or, equivalently, $f(x, j) = p_j\, f(x)$ or $F(x, j) = p_j\, F(x)$. The second condition in the lemma means that breakdown during some particular period does not make it any more or less likely to be from cause j than breakdown in some other period.

Example 1.57 (Weibull sub-hazards) With $h(x, j) = \alpha_j \beta_j^{-\alpha_j} x^{\alpha_j - 1}$ (see

Example 1.55) proportional hazards only holds if the α_j are all equal, say to α. Let $\beta^+ = \sum_{j=1}^{p} \beta_j^{-\alpha}$; in this case,

$$S(x) = \exp[-\beta^+ x^\alpha] \text{ and } f(x,j) = h(x,j)S(x) = \alpha\beta_j^{-\alpha} x^{\alpha-1} \exp[-\beta^+ x^\alpha].$$

With $\pi_j = \beta_j^{-\alpha}/\beta^+$, the sub-survival distribution is

$$S(x,j) = \pi_j \exp[-\beta^+ x^\alpha],$$

showing that this is a mixture model in which $p_j = P(C = j) = \pi_j$ and

$$P(X > x \,|\, C = j) = P(X > x) = \exp[-\beta^+ x^\alpha],$$

i.e., time and cause of breakdown are stochastically independent.

Competing risks models: latent-failure times approach

While the sub-hazard function introduced above has become the core concept of competing risks modeling, the more traditional approach was based on assuming existence of a *latent failure time* associated with each of the p potential causes of breakdown. If X_j represents the time to system failure from cause j ($j = 1, \ldots, p$), the smallest X_j determines the time to overall system failure, and its index is the cause of failure C. This is the prototype of the concept of *parallel processing* and, as such, central to reaction time modeling as well.

For the latent failure times we assume a vector $\mathbf{X} = (X_1, \ldots, X_p)$ with absolutely continuous survival function $G(\mathbf{x}) = P(\mathbf{X} > \mathbf{x})$ and marginal survival distributions $G_j(x) = P(X_j > x)$. The latent failure times are of course not observable, nor are the marginal hazard functions

$$h_j(x) = -d \log G_j(x)/dx.$$

However, we clearly must have an identity between $G(\mathbf{x})$ and the marginal survival function $S(x)$ introduced in Equation 1.13,

$$G(x\mathbf{1_p}) = S(x), \tag{1.15}$$

where $\mathbf{1_p}' = (1, \ldots, 1)$ of length p. It is then easy to show (see, e.g., Crowder, 2012, p. 236) that the sub-densities can be calculated from the joint survival distribution of the latent failure times as

$$f(x,j) = [-\partial G(\mathbf{x})/\partial x_j]_{x\mathbf{1_p}},$$

and the sub-hazards follow as

$$h(x,j) = [-\partial \log G(\mathbf{x})/\partial x_j]_{x\mathbf{1_p}},$$

where the notation $[\ldots]_{x\mathbf{1_p}}$ indicates that the enclosed function is to be evaluated at $\mathbf{x} = (x, \ldots, x)$.

Independent risks Historically, the main aim of competing risks analysis was to estimate the marginal survival distributions $G_j(x)$ from the sub-survival distributions $S(x, j)$. While this is not possible in general, except by specifying a fully parametric model for $G(\mathbf{x})$, it can also be done by assuming that the risks act independently. In particular, it can be shown (see Crowder, 2012, p. 245) that the set of sub-hazard functions $h(x, j)$ determines the set of marginals $G_j(x)$ via

$$G_j(x) = \exp\left\{-\int_0^x h(y, j)\, dy\right\}.$$

Without making such strong assumptions, it is still possible to derive some bounds for $G_j(x)$ in terms of $S(x, j)$. Obviously,

$$S(x, j) = P(X_j > x, C = j) \leq P(X_j > x) = G_j(x).$$

More refined bounds are derived in Peterson (1976):

Let $y = \max\{x_1, \ldots, x_p\}$. Then

$$\sum_{j=1}^{p} S(y, j) \leq G(\mathbf{x}) \leq \sum_{j=1}^{p} S(x_j, j). \tag{1.16}$$

Setting $x_k = x$ and all other $x_j = 0$ yields, for $x > 0$ and each k,

$$\sum_{j=1}^{p} S(x, j) \leq G_k(x) \leq S(x, k) + (1 - p_k). \tag{1.17}$$

These bounds cannot be improved because they are attainable by particular distributions, and they may be not very restrictive; that is, even knowing the $S(x, j)$ quite well may not suffice to say much about $G(\mathbf{x})$ or $G_j(x)$. This finding is echoed in some general non-identifiability results considered next.

Some non-identifiability results

A basic result, much discussed in the competing risks literature, is (for proof, see Tsiatis, 1975):

Proposition 1.58 (Tsiatis' Theorem) *For any arbitrary, absolutely continuous survival function $G(\mathbf{x})$ with dependent risks and sub-survival distributions $S(x, j)$, there exist a unique "proxy model" with independent risks*

yielding identical $S(x, j)$. It is defined by

$$G^*(\mathbf{x}) = \prod_{j=1}^{p} G_j^*(x_j), \text{ where } G_j^*(x_j) = \exp\{-\int_0^x h(y, j)dy\} \tag{1.18}$$

and the sub-hazard function $h(x, j)$ derives from the given $S(x, j)$.

The following example illustrates why this non-identifiability is seen as a problem.

Example 1.59 (Gumbel's bivariate exponential distribution) The joint survival distribution is

$$G(\mathbf{x}) = \exp[-\lambda_1 x_1 - \lambda_2 x_2 - \nu x_1 x_2],$$

with $\lambda_1 > 0$ and $\lambda_2 > 0$ and $0 \leq \nu \leq \lambda_1\lambda_2$. Parameter ν controls dependence: $\nu = 0$ implies stochastic independence between X_1 and X_2. The univariate marginals are $P(X_j > x_j) = G_j(x_j) = \exp[-\lambda_j x_j]$. Then the marginal survival distribution is $S(x) = \exp[-(\lambda_1 + \lambda_2)x - \nu x^2]$ and the sub-hazard functions are $h(x, j) = \lambda_j + \nu x$, $j = 1, 2$. From the Tsiatis theorem, the proxy model has

$$G_j^*(x) = \exp\left\{-\int_0^x (\lambda_j + \nu y)dy\right\} = \exp[-(\lambda_j x + \nu x^2/2)]$$
$$G^*(\mathbf{x}) = \exp[-(\lambda_1 x_1 + \nu x_1^2/2) - (\lambda_2 x_2 + \nu x_2^2/2)].$$

Thus, to make predictions about X_j one can either use $G_j(x)$ or $G_j^*(x)$, and the results will not be the same (except if $\nu = 0$). One cannot tell which one is correct from just (X, C) data.

There are many refinements of Tsiatis' theorem and related results (see e.g., Crowder, 2012) and, ultimately, the latent failure times approach in survival analysis has been superseded by the concept of sub-hazard function within the more general context of counting processes (Aalen et al., 2008).

1.3.4 Order statistics, extreme values, and records

The concept of order statistics plays a major role in RT models with a parallel processing architecture. Moreover, several theories of learning, memory, and automaticity have drawn upon concepts from the asymptotic theory of specific order statistics, the extreme values. Finally, we briefly treat theory of record values, a well developed statistical area which, strangely enough, has not yet played a significant role in RT modeling.

Order statistics

Suppose that $(X_1, X_2 \ldots, X_n)$ are n jointly distributed random variables. The corresponding *order statistics* are the X_i's arranged in ascending order of magnitude, denoted by $X_{1:n} \leq X_{2:n} \leq \ldots \leq X_{n:n}$. In general, the X_i are assumed to be i.i.d. random variables with $F(x) = P(X_i \leq x)$, $i = 1, \ldots n$, so that $(X_1, X_2 \ldots, X_n)$ can be considered as random sample from a population with distribution function $F(x)$. The distribution function of $X_{i:n}$ is obtained realizing that

$$\begin{aligned} F_{i:n}(x) &= P(X_{i:n} \leq x) \\ &= P(\text{at least } i \text{ of } X_1, X_2, \ldots, X_n \text{ are less than or equal to } x) \\ &= \sum_{r=i}^{n} P(\text{exactly } r \text{ of } X_1, X_2, \ldots, X_n \text{ are less than or equal to } x\ ) \\ &= \sum_{r=i}^{n} \binom{n}{r} [F(x)]^r [1 - F(x)]^{n-r}, \quad -\infty < x < +\infty. \end{aligned} \tag{1.19}$$

For $i = n$, this gives

$$F_{n:n}(x) = P(\max(X_1, \ldots, X_n) \leq x) = [F(x)]^n,$$

and for $i = 1$,

$$F_{1:n}(x) = P(\min(X_1, \ldots, X_n) \leq x) = 1 - [1 - F(x)]^n.$$

Assuming existence of the probability density for $(X_1, X_2 \ldots, X_n)$, the joint density for all n order statistics – which, of course, are no longer stochastically independent – is derived as follows (Arnold et al., 1992, p. 10). Given the realizations of the n order statistics to be $x_{1:n} < x_{2:n} < \cdots < x_{n:n}$, the original variables X_i are restrained to take on the values $x_{i:n}$ $(i = 1, 2, \ldots, n)$, which by symmetry assigns equal probability for each of the $n!$ permutations of $(1, 2, \ldots, n)$. Hence, the joint density function of all n order statistics is

$$f_{1,2,\ldots,n:n}(x_1, x_2, \ldots, x_n) = n! \prod_{r=1}^{n} f(x_r), \quad -\infty < x_1 < x_2 < \cdots < x_n < +\infty.$$

The marginal density of $X_{i:n}$ is obtained by integrating out the remaining $n - 1$ variables, yielding

$$f_{i:n}(x) = \frac{n!}{(i-1)!(n-i)!} [F(x)]^{i-1} [1 - F(x)]^{n-i} f(x), \quad -\infty < x < +\infty. \tag{1.20}$$

It is relatively straightforward to compute the joint density and the distribution function of any two order statistics (see Arnold et al., 1992).

Let $U_1, U_2, \ldots, U_n$ be a random sample from the standard uniform distribution (i.e., over $[0,1]$) and $X_1, X_2, \ldots, X_n$ be a random sample from a population with distribution function $F(x)$, and $U_{1:n} \leq U_{2:n} \leq \cdots \leq U_{n:n}$ and $X_{1:n} \leq X_{2:n} \leq \cdots \leq X_{n:n}$ the corresponding order statistics. For a continuous distribution function $F(x)$, the probability integral transformation $U = F(X)$ produces a standard uniform distribution. Thus, for continuous $F(x)$,

$$F(X_{i:n}) =_d U_{i:n}, \quad i = 1, 2, \ldots n.$$

Moreover, it is easy to verify that, for quantile function $Q(u) \equiv F^{-1}(u)$, one has

$$Q(U_i) =_d X_i, \quad i = 1, 2, \ldots n.$$

Since Q is non-decreasing, it immediately follows that

$$Q(U_{i:n}) =_d X_{i:n}, \quad i = 1, 2, \ldots n. \tag{1.21}$$

This implies simple forms for the quantile functions of the extreme order statistics, Q_n and Q_1,

$$Q_n(u_n) = Q(u_n^{1/n}) \text{ and } Q_1(u_1) = Q[1 - (1 - u_1)^{1/n}], \tag{1.22}$$

for the maximum and the minimum statistic, respectively.

Moments of order statistics can often be obtained from the marginal density (Equation 1.20). Writing $\mu_{i:n}^{(m)}$ for the m-th moment of $X_{i:n}$,

$$\mu_{i:n}^{(m)} = \frac{n!}{(i-1)!(n-i)!} \int_{-\infty}^{+\infty} x^m [F(x)]^{i-1} [1 - F(x)]^{n-i} f(x)\, dx, \tag{1.23}$$
$$1 \leq i \leq n, \quad m = 1, 2, \ldots .$$

Using relation (1.21), this can be written more compactly as

$$\mu_{i:n}^{(m)} = \frac{n!}{(i-1)!(n-i)!} \int_0^1 [F^{-1}(u)]^m u^{i-1} (1-u)^{n-i}\, du, \tag{1.24}$$
$$1 \leq i \leq n, \quad m = 1, 2, \ldots$$

The computation of various moments of order statistics can be reduced drawing upon various identities and recurrence relations. Consider the obvious identity

$$\sum_{i=1}^{n} X_{i:n}^m = \sum_{i=1}^{n} X_i^m, \quad m \geq 1.$$

Taking expectation on both sides yields

$$\sum_{i=1}^{n} \mu_{i:n}^{(m)} = n\mathrm{E}X^m = n\mu_{1:1}^{(m)}, \quad m \geq 1. \tag{1.25}$$

An interesting recurrence relation, termed triangle rule, is easily derived using Equation 1.24 (see Arnold et al., 1992, p. 112):

For $1 \leq i \leq n-1$ and $m = 1, 2 \ldots$,

$$i\mu_{i+1:n}^{(m)} + (n-i)\mu_{i:n}^{(m)} = n\mu_{i:n-1}^{(m)}. \tag{1.26}$$

This relation shows that it is enough to evaluate the mth moment of a single order statistic in a sample of size n if these moments in samples of size less than n are already available.

Many other recurrence relations, also including product moments and covariances of order statistics, are known and can be used not only to reduce the amount of computation (if these quantities have to be determined by numerical procedures) but also to provide simple checks of the accuracy of these computations (Arnold et al., 1992).

Characterization of distributions by order statistics. Given that a distribution F is known, it is clear that the distributions $F_{i:n}$ are completely determined, for every i and n. The interesting question is to what extent does knowledge of some properties of the distribution of an order statistic determine the parent distribution?

Clearly, the distribution of $X_{n:n}$ determines F completely since $F_{n:n}(x) = [F(x)]^n$ for every x, so that $F(x) = [F_{n:n}(x)]^{1/n}$. It turns out that, assuming the moments are finite, equality of the first moments of maxima,

$$\mathrm{E}X_{n:n} = \mathrm{E}X'_{n:n}, \quad n = 1, 2 \ldots,$$

is sufficient to infer equality of the underlying parent distributions $F(x)$ and $F'(x)$ (see Arnold et al., 1992, p. 145).

Moreover, distributional relations between order statistics can be drawn upon in order to characterize the underlying distribution. For example, straightforward computation yields that for a sample of size n from an exponentially distributed population,

$$nX_{1:n} =_d X_{1:1}, \quad \text{for } n = 1, 2, \ldots . \tag{1.27}$$

Notably, it has been shown (Desu, 1971) that the exponential is the *only* nondegenerate distribution satisfying the above relation. From this, a characterization of the Weibull is obtained as well:

Example 1.60 ("Power Law of Practice") Let, for $\alpha > 0$,

$$F(x) = 1 - \exp[-x^{\alpha}];$$

then, $W = X^{\alpha}$ is a unit exponential variable (i.e., $\lambda = 1$) and Equation 1.27 implies the following characterization of the Weibull:

$$n^{1/\alpha} X_{1:n} =_d X_{1:1}, \quad \text{for } n = 1, 2, \ldots . \tag{1.28}$$

At the level of expectations,

$$\mathrm{E}X_{1:n} = n^{-1/\alpha}\mathrm{E}X.$$

This has been termed the "power law of practice" (Newell and Rosenbloom, 1981): the time taken to perform a certain task decreases as a power function of practice (indexed by n). This is the basis of G. Logan's "theory of automaticity" (Logan, 1988, 1992, 1995) postulating Weibull-shaped RT distributions (see Colonius, 1995, for a comment on the theoretical foundations).

Extreme value statistics

The behavior of extreme order statistics, that is, of the maximum and minimum of a sample when sample size n goes to infinity, has attracted the attention of RT model builders, in particular in the context of investigating parallel processing architectures. Only a few results of this active area of statistical research will be considered here.

For the underlying distribution function F, its *right, or upper, endpoint* is defined as

$$\omega(F) = \sup\{x | F(x) < 1\}.$$

Note that $\omega(F)$ is either $+\infty$ or finite. Then the distribution of the maximum,

$$F_{n:n}(x) = P(X_{n:n} \leq x) = [F(x)]^n,$$

clearly converges to zero, for $x < \omega(F)$, and to 1, for $x \geq \omega(F)$. Hence, in order to obtain a nondegenerate limit distribution, a normalization is necessary. Suppose there exists sequences of constants $b_n > 0$ and a_n real $(n = 1, 2, \ldots)$ such that

$$\frac{X_{n:n} - a_n}{b_n}$$

has a nondegenerate *limit distribution*, or *extreme value distribution*, as $n \to +\infty$, i.e.,

$$\lim_{n \to +\infty} F^n(a_n + b_n x) = G(x), \tag{1.29}$$

for every point x where G is continuous, and G a nondegenerate distribution function[14]. A distribution function F satisfying Equation 1.29 is said to be in the *domain of maximal attraction* of a nondegenerate distribution function G, and we will write $F \in \mathcal{D}(G)$. Two distribution functions F_1 and F_2 are defined to be of the *same type* if there exist constants a_0 and $b_0 > 0$ such that $F_1(a_0 + b_0 x) = F_2(x)$.

It can be shown that (i) the choice of the norming constants is not unique and (ii) a distribution function cannot be in the domain of maximal attraction of more than one type of distribution. The main theme of extreme value statistics is to identify all extreme value distributions and their domains of attraction and to characterize necessary and/or sufficient conditions for convergence.

Before a general result is presented, let us consider the asymptotic behavior for a special case.

Example 1.61 Let $X_{n:n}$ be the maximum of a sample from the exponential distribution with rate equal to λ. Then,

$$\begin{aligned} P(X_{n:n} - \lambda^{-1}\log n \le x) &= \begin{cases} (1 - \exp[-(\lambda x + \log n)])^n, & \text{if } x > -\lambda^{-1}\log n \\ 0, & \text{if } x \le -\lambda^{-1}\log n \end{cases} \\ &= (1 - \exp[-\lambda x]/n)^n, \qquad \text{for } x > -\lambda^{-1}\log n \\ &\to \exp[-\exp[-\lambda x]], \quad -\infty < x < \infty, \end{aligned}$$

as $n \to +\infty$.

Possible limiting distributions for the sample maximum. The classic result on the limiting distributions, already described in Frechét (1927) and Fisher and Tippett (1928), is the following:

Theorem 1.62 *If Equation 1.29 holds, the limiting distribution function G of an appropriately normalized sample maximum is one of the following types:*

$$G_1(x;\alpha) = \begin{cases} 0, & x \le 0 \\ \exp[-x^{-\alpha}], & x > 0; \alpha > 0 \end{cases} \tag{1.30}$$

$$G_2(x;\alpha) = \begin{cases} \exp[-(-x)^{\alpha}], & x < 0; \alpha > 0 \\ 1, & x \ge 0 \end{cases} \tag{1.31}$$

$$G_3(x;\alpha) = \exp[-\exp(-x)], \quad -\infty < x < +\infty \tag{1.32}$$

[14] Note that convergence here is convergence "in probability", or "weak" convergence.

Distribution G_1 is referred to as the *Frechét type*, G_2 as the *Weibull type*, and G_3 as the *extreme value* or *Gumbel distribution*. Parameter α occurring in G_1 and G_2 is related to the tail behavior of the parent distribution F. It follows from the theorem that if F is in the maximal domain of attraction of some G, $F \in \mathcal{D}(G)$, then G is either G_1, G_2, or G_3.

Necessary and sufficient conditions on F to be in the domain of attraction of G_1, G_2, or G_3, respectively, have been given in Gnedenko (1943). Since these are often difficult to verify Arnold et al. (1992, pp. 210–212), here we list sufficient conditions for weak convergence due to von Mises (1936):

Proposition 1.63 *Let F be an absolutely continuous distribution function and let $h(x)$ be the corresponding hazard function.*

1. *If $h(x) > 0$ for large x and for some $\alpha > 0$,*

$$\lim_{x \to +\infty} x\, h(x) = \alpha,$$

   then $F \in \mathcal{D}(G_1(x;\alpha))$. Possible choices for the norming constants are $a_n = 0$ and $b_n = F^{-1}(1 - n^{-1})$.
2. *If $F^{-1}(1) < \infty$ and for some $\alpha > 0$,*

$$\lim_{x \to F^{-1}(1)} (F^{-1}(1) - x)h(x) = \alpha,$$

   then $F \in \mathcal{D}(G_2(x;\alpha))$. Possible choices for the norming constants are $a_n = F^{-1}(1)$ and $b_n = F^{-1}(1) - F^{-1}(1 - n^{-1})$.
3. *Suppose $h(x)$ is nonzero and is differentiable for x close to $F^{-1}(1)$ or, for large x, if $F^{-1}(1) = +\infty$. Then, $F \in \mathcal{D}(G_3)$ if*

$$\lim_{x \to F^{-1}(1)} \frac{d}{dx}\left(\frac{1}{h(x)}\right) = 0.$$

   Possible choices for the norming constants are $a_n = F^{-1}(1 - n^{-1})$ and $b_n = [nf(a_n)]^{-1}$.

Limiting distributions for sample minima. This case can be derived technically from the analysis of sample maxima, since the sample minimum from distribution F has the same distribution as the negative of the sample maximum from distribution F^*, where $F^*(x) = 1 - F(-x)$. We sketch some of the results for completeness here.

For parent distribution F one defines the *left*, or *lower endpoint* as

$$\alpha(F) = \inf\{x : F(x) > 0\}.$$

Note that $\alpha(F)$ is either $-\infty$ or finite. Then the distribution of the minimum,

$$F_{1:n}(x) = P(X_{1:n} \leq x) = 1 - [1 - F(x)]^n,$$

clearly converges to one, for $x > \alpha(F)$, and to zero, for $x \leq \alpha(F)$. The theorem paralleling Theorem 1.62 then is

Theorem 1.64 *Suppose there exist constants a_n^* and $b_n^* > 0$ and a non-degenerate distribution function G^* such that*

$$\lim_{n \to +\infty} 1 - [1 - F(a_n^* + b_n^* x)]^n = G^*(x)$$

for every point x where G^* is continuous, then G^* must be one of the following types:

1. $G_1^*(x; \alpha) = 1 - G_1(-x; \alpha)$,
2. $G_2^*(x; \alpha) = 1 - G_2(-x; \alpha)$,
3. $G_3^*(x) = 1 - G_3(-x)$,

where G_1, G_2, and G_3 are the distributions given in Theorem 1.62.

Necessary and/or sufficient conditions on F to be in the domain of (minimal) attraction of G_1, G_2, or G_3, respectively, can be found e.g., in Arnold et al. (1992, pp. 213-214).

Let us state some important practical implications of these results on asymptotic behavior:

- Only three distributions (Fréchet, Weibull, and extreme value/Gumbel) can occur as limit distributions of independent sample maxima or minima.
- Although there exist parent distributions such that the limit does not exist, for most continuous distributions in practice one of them actually occurs.
- A parent distribution with non-finite endpoint in the tail of interest cannot lie in a Weibull type domain of attraction.
- A parent distribution with finite endpoint in the tail of interest cannot lie in a Fréchet type domain of attraction.

One important problem with practical implications is the speed of convergence to the limit distributions. It has been observed that, e.g. in the case of the maximum of standard normally distributed random variables, convergence to the extreme value distribution G_3 is extremely slow and that $F_n(a_n + b_n x)^n$ can be closer to a suitably chosen Weibull distribution G_2 even for very large n. In determining the sample size required for the application of the asymptotic theory, estimates of the error of approximation

do not depend on x since the convergence is uniform in x. They do depend, however, on the choice of the norming constants (for more details, see Galambos, 1978, pp. 111–116). Because of this so called *penultimate approximation*, diagnosing the type of limit distribution even from rather large empirical samples may be problematic (see also the discussion in Colonius, 1995; Logan, 1995).

Record values

The theory of records should not be seen as a special topic of order statistics, but their study parallels the study of order statistics in many ways. In particular, things that are relatively easy for order statistics are also feasible for records, and things that are difficult for order statistics are equally or more difficult for records. Compared to order statistics, there are not too many data sets for records available, especially for the asymptotic theory, the obvious reason being the low density of the nth record for large n. Nevertheless, the theory of records has become quite developed representing an active area of research. Our hope in treating it here, although limiting exposition to the very basics, is that the theory may find some useful applications in the cognitive modeling arena.

Instead of a finite sample $(X_1, X_2 \ldots, X_n)$ of i.i.d. random variables, record theory starts with an infinite sequence $X_1, X_2, \ldots$ of i.i.d. random variables (with common distribution F). For simplicity, F is assumed continuous so that ties are not possible. An observation X_j is called a *record* (more precisely, *upper record*) if it exceeds in value all preceding observations, that is, $X_j > X_i$ for all $i < j$. *Lower records* are defined analogously.

Following the notation in Arnold et al. (1992, 1998), the sequence of *record times* $\{T_n\}_{n=0}^{\infty}$ is defined by

$$T_0 = 1 \text{ with probability 1,}$$

and for $n \geq 1$,

$$T_n = \min\{j \mid j > T_{n-1},\ X_j > X_{T_{n-1}}\}.$$

The corresponding *record value sequence* $\{R_n\}_{n=0}^{\infty}$ is defined as

$$R_n = X_{T_n},\ \ n = 0, 1, 2, \ldots.$$

Interrecord times, Δ_n, are defined by

$$\Delta_n = T_n - T_{n-1},\ \ n = 1, 2, \ldots$$

Finally, the *record counting process* is $\{N_n\}_{n=1}^{+\infty}$, where

$$N_n = \{\text{number of records among } X_1, \ldots, X_n\}.$$

Interestingly, only the sequence of record values $\{R_n\}$ depends on the specific distribution; monotone transformations of the X_i's will not affect the values of $\{T_n\}$, $\{\Delta_n\}$, and $\{N_n\}$. Specifically, if $\{R_n\}$, $\{T_n\}$, $\{\Delta_n\}$, and $\{N_n\}$ are the record statistics associated with a sequence of i.i.d. X_i's and if $Y_i = \phi(X_i)$ (where ϕ increasing) has associated record statistics $\{R'_n\}$, $\{T'_n\}$, $\{\Delta'_n\}$, and $\{N'_n\}$, then

$$T'_n =_d T_n,\ \Delta'_n =_d \Delta_n,\ N'_n =_d N_n,\ \text{and } R'_n =_d \phi(R_n).$$

Thus, it is wise to select a convenient common distribution for the X_i's which will make the computations simple. Because of its lack of memory property, the standard exponential distribution (Exp(1)) turns out to be the most useful one.

Distribution of the nth upper record R_n. Let $\{X_i^*\}$ denote a sequence of i.i.d. Exp(1) random variables. Because of the lack of memory property, $\{R_n^* - R_{n-1}^*, n \geq 1\}$, the differences beween successive records will again be i.i.d. Exp(1) random variables. It follows that

$$R_n^* \text{ is distributed as Gamma}(n+1, 1),\quad n = 0, 1, 2, \ldots\ . \tag{1.33}$$

From this we obtain the distribution of R_n corresponding to a sequence of i.i.d X_i's with common *continuous* distribution function F as follows: Note that if X has distribution function F, then $-\log[1 - F(X)]$ is distributed as Exp(1) (see Proposition 1.49), so that

$$X =_d F^{-1}(1 - \exp^{-X_n^*}),$$

where X^* is Exp(1). Because X is a monotone function of X^* and, consequently, the nth record of the $\{X_n\}$ sequence, R_n, is related to the nth record R_n^* of the exponental sequence by

$$R_n =_d F^{-1}(1 - \exp^{-R_n^*}).$$

From Equation 1.5, the survival of a Gamma$(n+1, 1)$ random variable is

$$P(R_n^* > r) = e^{-r^*} \sum_{k=0}^{n} (r^*)^k / k!$$

implying, for the survival function of the nth record corresponding to an i.i.d. F sequence,

$$P(R_n > r) = [1 - F(r)] \sum_{k=0}^{n} [-\log(1 - F(r))]^k / k!.$$

If F is absolutely continuous with density f, differentiating and simplifying yields the density for R_n

$$f_{R_n}(r) = f(r)[-\log(1 - F(r))]^n/n!.$$

The asymptotic behavior of the distribution of R_n has also been investigated. For example, when the common distribution of the X_i's is Weibull, the distribution of R_n is asymptotically normal.

Distribution of the nth record time T_n. Since $T_0 = 1$, we consider first the distribution of T_1.

For any $n \geq 1$, clearly

$$P(T_1 > n) = P(X_1 \text{ is largest among } X_1, X_2, \ldots, X_n).$$

Without loss of generality, we assume that $X_1, X_2, \ldots$ are independent random variables with a uniform distribution on $(0, 1)$. Then, by the total probability rule, for $n \geq 2$

$$\begin{aligned} P(T_1 = n) &= \int_0^1 P(T_1 = n \,|X_1 = x)\, dx \\ &= \int_0^1 x^{n-2}(1 - x)\, dx \\ &= \frac{1}{n-1} - \frac{1}{n} = \frac{1}{n(n-1)}. \end{aligned} \tag{1.34}$$

Note that the median of T_1 equals 2, but the expected value

$$\mathrm{E}T_1 = \sum_{n=2}^{+\infty} n\, P(T_1 = n) = +\infty.$$

Thus, after a maximum has been recorded by the value of X_1, the value of X_2 that will be even larger has a probability of $1/2$, but the expected "waiting time" to a larger value is infinity.

Introducing the concept of record indicator random variables, $\{I_n\}_{n=1}^{\infty}$, defined by $I_1 = 1$ with probability 1 and, for $n > 1$

$$I_n = \begin{cases} 1, & \text{if } X_n > X_1, \ldots, X_{n-1}, \\ 0, & \text{otherwise,} \end{cases}$$

it can be shown that the I_n's are independent Bernoulli random variables with

$$P(I_n = 1) = 1/n$$

for $n = 1, 2, \ldots$, and the joint density of the record times becomes

$$\begin{aligned} &P(T_1 = n_1, T_2 = n_2, \ldots, T_k = n_k) \\ &\quad = P(I_2 = 0, \ldots, I_{n_1 - 1} = 0, I_{n_1} = 1, I_{n_1 + 1} = 0, \ldots, I_{n_k} = 1) \\ &\quad = \frac{1}{2} \cdot \frac{2}{3} \cdot \frac{3}{4} \cdots \frac{n_1 - 2}{n_1 - 1} \cdot \frac{1}{n_1} \frac{n_1}{n_1 + 1} \cdots \frac{1}{n_k} \\ &\quad = [(n_1 - 1)(n_2 - 1) \cdots (n_k - 1) n_k]^{-1}. \end{aligned}$$

The marginal distributions can calculated to be

$$P(T_k = n) = S_{n-1}^k / n!,$$

where S_{n-1}^k are the *Stirling numbers of the first kind*, i.e., the coefficients of s^k in

$$\prod_{j=1}^{n-1} (s + j - 1).$$

One can verify that T_k grows rapidly with k and that it has in fact an infinitely large expectation. Moreover, about half of the waiting time for the kth record is spent between the $k - 1$th and the kth record.

Record counting process N_n. Using the indicator variables I_n, the record counting process $\{N_n, n \geq 1\}$ is just

$$N_n = \sum_{j=1}^{n} I_j,$$

so that

$$\begin{aligned} \mathrm{E}(N_n) &= \sum_{j=1}^{n} \frac{1}{j} \\ &\approx \log n + \gamma, \end{aligned} \tag{1.35}$$

and

$$\begin{aligned} \mathrm{Var}(N_n) &= \sum_{j=1}^{n} \frac{1}{j} \left(1 - \frac{1}{j}\right) \\ &\approx \log n + \gamma - \frac{\pi^2}{6}, \end{aligned} \tag{1.36}$$

where γ is *Euler's constant* $0.5772\ldots$ This last result makes it clear that records are not very common: in a sequence of 1000 observations one can expect to observe only about 7 records![15].

When the parent F is Exp(1) distributed, the record values R_n^* were shown to be distributed as Gamma$(n+1,1)$ random variables (Equation 1.33). Thus, the associated record counting process is identical to a homogeneous Poisson process with unit intensity ($\lambda = 1$). When F is an arbitrary continuous distribution with $F^{-1}(0) = 0$ and hazard function $\lambda(t)$, then the associated record counts become a nonhomogeneous Poisson process with intensity function $\lambda(t)$: to see this, note that there will be a record value between t and $t+h$ if, and only if, the first X_i whose values is greater than t lies between t and $t+h$. But we have (conditioning on which X_i this is, say $i = n$), by definition of the hazard function,

$$P(X_n \in (t, t+h) | X_n > t) = \lambda(t)h + o(h),$$

which proves the claim.

1.3.5 Coupling

The concept of *coupling* involves the joint construction of two or more random variables (or stochastic processes) on a common probability space. The theory of probabilistic coupling has found applications throughout probability theory, in particular in characterizations, approximations, asymptotics, and simulation. Before we will give a formal definition of coupling, one may wonder why a modeler should be interested in this topic.

In the introduction, we discussed an example of coupling in the context of audiovisual interaction (Example 1.2). Considering a more general context, data are typically collected under various experimental conditions, e.g., different stimulus properties, different number of targets in a search task, etc. Any particular condition generates a set of data considered to represent realizations of some random variable (e.g., RT), given a stochastic model has been specified. However, there is no principled way of relating, for example, an observed time to find a target in a condition with n targets, T_n, say, to that for a condition with $n+1$ targets, T_{n+1}, simply because T_n and T_{n+1} are not defined on a common probability space. Note that this would not prevent one from numerically comparing average data under both conditions, or even the entire distributions functions; any statistical hypothesis

[15] This number is clearly larger than typical empirical data sets and may be one reason why record theory has not seen much application in mathematical psychology (but see the remarks on time series analysis in Section 1.4).

about the two random variables, e.g., about the correlation or stochastic (in)dependence in general, would be void, however. Beyond this cautionary note, which is prompted by some apparent confusion about this issue in the RT modeling area, the method of coupling can also play an important role as modeling tool. For example, a coupling argument allows one to turn a statement about an ordering of two RT distributions, a-priori not defined on a common probability space, into a statement about a pointwise ordering of corresponding random variables on a common probability space. Moreover, coupling turns out to be a key concept in a general theory of "contextuality" being developed by E. N. Dzhafarov and colleagues (see Dzhafarov and Kujala, 2013, and also their chapter in this volume).

We begin with a definition of "coupling" for random variables without referring to the measure-theoretic details:

Definition 1.65 A *coupling* of a collection of random variables $\{X_i, i \in \mathrm{I}\}$, with I denoting some index set, is a family of jointly distributed random variables

$$(\hat{X}_i : i \in \mathrm{I}) \quad \text{such that} \quad \hat{X}_i =_d X_i, \;\; i \in \mathrm{I}.$$

Note that the joint distribution of the $\hat{X}_i$ need not be the same as that of X_i; in fact, the X_i may not even have a joint distribution because they need not be defined on a common probability space. However, the family $(\hat{X}_i : i \in \mathrm{I})$ has a joint distribution with the property that its marginals are equal to the distributions of the individual X_i variables. The individual $\hat{X}_i$ is also called a *copy of* X_i.

Before we mention a more general definition of coupling, let us consider a simple example.

Example 1.66 (Coupling two Bernoulli random variables) Let X_p, X_q be Bernoulli random variables, i.e.,

$$P(X_p = 1) = p \;\text{ and }\; P(X_p = 0) = 1 - p,$$

and X_q defined analogously. Assume $p < q$; we can couple X_p and X_q as follows:
Let U be a uniform random variable on $[0, 1]$, i.e., for $0 \le a < b \le 1$,

$$P(a < U \le b) = b - a.$$

Define

$$\hat{X}_p = \begin{cases} 1, & \text{if } 0 < U \le p; \\ 0, & \text{if } p < U \le 1 \end{cases}; \quad \hat{X}_q = \begin{cases} 1, & \text{if } 0 < U \le q; \\ 0, & \text{if } q < U \le 1. \end{cases}$$

Then U serves as a common source of randomness for both $\hat{X}_p$ and $\hat{X}_q$. Moreover, $\hat{X}_p =_d X_p$ and $\hat{X}_q =_d X_q$, and $\mathrm{Cov}(\hat{X}_p, \hat{X}_q) = p(1-q)$. The joint distribution of $(\hat{X}_p, \hat{X}_q)$ is presented in Table 1.1 and illustrated in Figure 1.4.

Table 1.1 *Joint distribution of two Bernoulli random variables.*

		$\hat{X}_q$		
		0	1	
$\hat{X}_p$	0	$1-q$	$q-p$	$1-p$
	1	0	p	p
		$1-q$	q	1

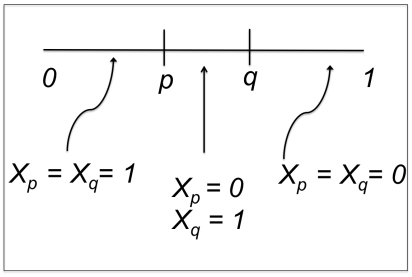

Figure 1.4 Joint distribution of two Bernoulli random variables

In this example, the two coupled random variables are obviously not independent. Note, however, that for any collection of random variables $\{X_i, i \in \mathrm{I}\}$ there always exists a *trivial* coupling, consisting of independent copies of the $\{X_i\}$. This follows from construction of the product probability measure (see Definition 1.37).

Before we consider further examples of coupling, coupling of probability measures is briefly introduced. The definition is formulated for the binary case only, for simplicity; let $(E, \mathcal{E})$ be a measurable space:

Definition 1.67 A *random element* in the space $(E, \mathcal{E})$ is the quadruple

$$(\Omega_1, \mathcal{F}_1, P_1, X_1),$$

where $(\Omega_1, \mathcal{F}_1, P_1)$ is the underlying probability space (the sample space) and X_1 is a $\mathcal{F}_1$-$\mathcal{E}$-measurable mapping from Ω_1 to E. A *coupling of the random elements* $(\Omega_i, \mathcal{F}_i, P_i, X_i)$, $i = 1, 2$, *in* $(E, \mathcal{E})$ is a random element $(\hat{\Omega}, \hat{\mathcal{F}}, \hat{P}, (\hat{X}_1, \hat{X}_2))$ in $(E^2, \mathcal{E} \otimes \mathcal{E})$ such that

$$X_1 =_d \hat{X}_1 \quad \text{and} \quad X_2 =_d \hat{X}_2,$$

with $\mathcal{E} \otimes \mathcal{E}$ the product σ-algebra on E^2.

This general definition is the starting point of a general theory of coupling for stochastic processes, but it is beyond the scope of this chapter (see Thorisson, 2000).

A *self-coupling* of a random variable X is a family $(\hat{X}_i : i \in \mathrm{I})$ where each $\hat{X}_i$ is a copy of X. A trivial, and not so useful, self-coupling is the one with all the $\hat{X}_i$ identical. Another example is the *i.i.d. coupling* consisting of independent copies of X. This can be used to prove

Lemma 1.68 *For every random variable X and nondecreasing bounded functions f and g, the random variables $f(X)$ and $g(X)$ are positively correlated, i.e.,*

$$\mathrm{Cov}[f(X), g(X)] \geq 0.$$

The lemma also follows directly from a general concept of dependence (association) to be treated in Section 1.3.8. A simple, though somewhat fundamental, coupling is the following:

Example 1.69 (Quantile coupling) Let X be a random variable with distribution function F, that is,

$$P(X \leq x) = F(x), \quad x \in \Re.$$

Let U be a uniform random variable on $[0, 1]$. Then, for random variable $\hat{X} = F^{-1}(U)$ (see Proposition 1.49),

$$P(\hat{X} \leq x) = P(F^{-1}(U) \leq x) = P(U \leq F(x)) = F(x), \quad x \in \Re,$$

that is, $\hat{X}$ is a copy of X, $\hat{X} =_d X$. Thus, letting F run over the class of all distribution functions (using the same U), yields a coupling of all differently distributed random variables, the *quantile coupling*. Since F^{-1} is nondecreasing, it follows from Lemma 1.68 that the quantile coupling consists of positively correlated random variables.

An important application of quantile coupling is in reducing a stochastic ordering between random variables to a pointwise (a.s.) ordering: Let X and X' be two random variables with distribution functions F and G, respectively. If there is a coupling $(\hat{X}, \hat{X}')$ of X and X' such that $\hat{X}$ is *pointwise dominated* by $\hat{X}'$, that is

$$\hat{X} \leq \hat{X}' \text{ (almost surely)},$$

then $\{\hat{X} \leq x\} \supseteq \{\hat{X}' \leq x\}$, which implies

$$P(\hat{X} \leq x) \geq P(\hat{X}' \leq x),$$

and thus

$$F(x) \geq G(x), \quad x \in \Re.$$

Then X is said to be *stochastically dominated* (or *dominated in distribution*) by X':

$$X \leq_d X'.$$

But the other direction also holds: *stochastic* domination implies *pointwise* domination: From $F(x) \geq G(x), \quad x \in \Re$, it follows that

$$G(x) \geq u \text{ implies } F(x) \geq u,$$

and thus,

$$\{x \in \Re \,|\, G(x) \geq u\} \subseteq \{x \in \Re \,|\, F(x) \geq u\}, \text{ thus,}$$

$$F^{-1}(u) \equiv \inf\{x|F(x) \geq u\} \leq \inf\{x|G(x) \geq u\} \equiv G^{-1}(u),$$

implying

$$F^{-1}(U) \leq G^{-1}(U), \text{ that is,}$$

$$\hat{X} \leq \hat{X}' \quad \text{pointwise (quantile coupling).}$$

Thus, we have proved a simple version of *Strassen's theorem*:

Theorem 1.70 (Strassen, 1965) *Let X and X' be random variables. Then*

$$X \leq_d X'$$

if, and only if, there is a coupling $(\hat{X}, \hat{X}')$ of X and X' such that a.s.

$$\hat{X} \leq \hat{X}'.$$

Note that it is often easier to carry out arguments using pointwise domination than stochastic domination. See the following example:

Example 1.71 (Additivity of stochastic domination) Let X_1, X_2, X_1', and X_2' be random variables such that X_1 and X_2 are independent, X_1' and X_2' are independent, and

$$X_1 \leq_d X_1' \quad \text{and} \quad X_2 \leq_d X_2'.$$

For $i = 1, 2$, let $(\hat{X}_i, \hat{X}_i')$ be a coupling of X_i and X_i' such that $\hat{X}_i \leq \hat{X}_i'$. Let $(\hat{X}_1, \hat{X}_1')$ and $(\hat{X}_2, \hat{X}_2')$ be independent. Then $(\hat{X}_1 + \hat{X}_2, \hat{X}_1' + \hat{X}_2')$ is a coupling of $X_1 + X_2$ and $X_1' + X_2'$, and

$$\hat{X}_1 + \hat{X}_2 \leq \hat{X}_1' + \hat{X}_2'$$

implying

$$X_1 + X_2 \leq_d X_1' + X_2'.$$

Coupling event inequality and maximal coupling

The following question is the starting point of many convergence and approximation results obtained from coupling arguments. Let X and X' be two random variables with non-identical distributions. How can one construct a coupling of X and X', $(\hat{X}, \hat{X}')$, such that $P(\hat{X} = \hat{X}')$ is maximal across all possible couplings? Here we follow the slightly more general formulation in Thorisson (2000) but limit presentation to the case of discrete random variables (the continuous case being completely analogous).

Definition 1.72 Suppose $(\hat{X}_i : i \in \mathrm{I})$ is a coupling of $X_i, i \in \mathrm{I}$, and let C be an event such that if it occurs, then all the $\hat{X}_i$ coincide, that is,

$$C \subseteq \{\hat{X}_i = \hat{X}_j, \text{ for all } i, j \in \mathrm{I}\}.$$

Such an event is called a *coupling event*.

Assume all the X_i take values in a finite or countable set E with $P(X_i = x) = p_i(x)$, for $x \in E$. For all $i, j \in \mathrm{I}$ and $x \in E$, we clearly have

$$P(\hat{X}_i = x, C) = P(\hat{X}_j = x, C) \leq p_j(x),$$

and thus for all $i \in \mathrm{I}$ and $x \in E$,

$$P(\hat{X}_i = x, C) \leq \inf_{j \in \mathrm{I}} p_j(x).$$

Summing over $x \in E$ yields the basic *coupling event inequality*:

$$P(C) \leq \sum_{x \in E} \inf_{j \in \mathrm{I}} p_i(x). \tag{1.37}$$

As an example, consider again the case of two discrete random variables X and X' with coupling $(\hat{X}, \hat{X}')$, and set $C = \{\hat{X} = \hat{X}'\}$. Then

$$P(\hat{X} = \hat{X}') \leq \sum_x \min\{P(X = x), P(X' = x)\}. \tag{1.38}$$

Interestingly, it turns out that, at least in principle, one can always construct a coupling such that the above coupling event inequality (1.37) holds with identity. Such a coupling is called *maximal* and C a *maximal coupling event*.

Proposition 1.73 *(Maximal coupling) Suppose $X_i, i \in \mathrm{I}$, are discrete random variables taking values in a finite or countable set E. Then there exists*

a maximal coupling, that is, a coupling with coupling event C such that

$$P(C) = \sum_{x \in E} \inf_{i \in \mathrm{I}} p_i(x).$$

Proof Put

$$c := \sum_{x \in E} \inf_{i \in \mathrm{I}} p_i(x) \quad \text{(the maximal coupling probability)} .$$

If $c = 0$, take the $\hat{X}_i$ independent and $C = \emptyset$. If $c = 1$, take the $\hat{X}_i$ identical and $C = \Omega$, the sample space. For $0 < c < 1$, these couplings are mixed as follows: Let J, V, and $W_i, i \in \mathrm{I}$, be independent random variables such that J is Bernoulli distributed with $P(J = 1) = c$ and, for $x \in E$,

$$\begin{aligned} P(V = x) =& \frac{\inf_{i \in \mathrm{I}} p_i(x)}{c} \\ P(W_i = x) =& \frac{p_i(x) - c\, P(V = x)}{1 - c}. \end{aligned}$$

Define, for each $i \in \mathrm{I}$,

$$\hat{X}_i = \begin{cases} V, & \text{if } J = 1, \\ W_i, & \text{if } J = 0. \end{cases} \tag{1.39}$$

Then

$$\begin{aligned} P(\hat{X}_i = x) &= P(V = x)P(J = 1) + P(W_i = x)P(J = 0) \\ &= P(X_i = x). \end{aligned}$$

Thus, the $\hat{X}_i$ are a coupling of the X_i, $C = \{J = 1\}$ is a coupling event, and it has the desired value c. The representation (1.39) of the X_i is known as *splitting representation*. □

Total variation distance and coupling. We conclude the treatment of coupling with a variation on the theme of maximal coupling: Given two random variables, what is a measure of closeness between them when an appropriate coupling is applied to make them as close to being identical as possible?

The *total variation distance* between two probability distributions μ and ν on Ω is defined as

$$\|\mu - \nu\|_{TV} := \max_{A \subseteq \Omega} |\mu(A) - \nu(A)|,$$

for all Borel sets A. Thus, the distance between μ and ν is the maximum

difference between the probabilities assigned to a single event by the two distributions. Using the coupling inequality, it can be shown that

$$\|\mu - \nu\|_{TV} = \inf\{P(X \neq Y) \mid (X, Y) \text{ is a coupling of } \mu \text{ and } \nu\}. \quad (1.40)$$

A splitting representation analogous to the one in the previous proof then assures that a coupling can be constructed so that the infimum is obtained (see Levin et al., 2008, pp.50–52).

1.3.6 Fréchet-Hoeffding bounds and Fréchet distribution classes

We consider general classes of multivariate distributions with given marginals. In contrast to the previous section on coupling, here existence of a multivariate distribution is taken for granted. Apart from studying the class of multivariate distributions $\mathcal{F}(F_1, \ldots, F_n)$ with fixed univariate marginals $F_1, \ldots, F_n$, classes of distributions with fixed bivariate or higher-order marginals have been studied as well. For example, the class $\mathcal{F}(F_{12}, F_{13})$ refers to trivariate distributions with fixed bivariate marginals F_{12} and F_{13}, where it is assumed that the first univariate margin F_1 is the same. General questions to ask are (cf. Joe, 1997):

1. What are the bounds for the multivariate distributions in a specified class $\mathcal{F}$?
2. Do the bounds correspond to proper multivariate distributions, and if so, is there a stochastic representation or interpretation of the bounds?
3. Are the bounds sharp if they do not correspond to proper multivariate distributions?
4. What are simple members of $\mathcal{F}$?
5. Can on construct parametric subfamilies in $\mathcal{F}$ with desirable properties?

The case of $n = 2$ is considered first.

Fréchet-Hoeffding bounds for $n = 2$

For two events A_1, A_2 in a probability space, the following sequence of inequalities is quite obvious:

$$\max\{0, P(A_1) + P(A_2) - 1\} \leq P(A_1) + P(A_2) - P(A_1 \cup A_2) = P(A_1 \cap A_2) \leq \min\{P(A_1), P(A_2)\}. \quad (1.41)$$

Defining $A_1 = \{X_1 \leq x_1\}$ and $A_2 = \{X_2 \leq x_2\}$ for random variables X_1, X_2 with joint distribution $F(x_1, x_2)$ and marginals $F_1(x_1)$ and $F_2(x_2)$ results in the *Fréchet-Hoeffding bounds* F^- and F^+:

Proposition 1.74 *Let $F(x_1, x_2)$ be a bivariate distribution function with marginals F_1 and F_2. Then for all x_1, x_2,*

$$F^-(x_1, x_2) = \max\{0, F_1(x_1) + F_2(x_2) - 1\} \leq F(x_1, x_2) \leq \min\{F_1(x_1), F_2(x_2)\} = F^+(x_1, x_2), \tag{1.42}$$

and both the lower bound F^- and the upper bound F^+ are distribution functions.

It is routine to check that the conditions of Proposition 1.24 are satified, i.e., that both F^- and F^+ are distribution functions as well. Moreover, it can be shown, for continuous marginals F_1, F_2, that F^+ and F^- represent perfect positive and negative dependence, respectively, in the following sense: X_1 is (almost surely) an increasing, respectively decreasing, function of X_2. This has also been referred to a *comonotonicity* and *countermonotonicity*, respectively[16].

Rearranging the inequalities in (1.41) yields

$$\max\{P(A_1), P(A_2)\} \leq P(A_1 \cup A_2) \leq \min\{P(A_1) + P(A_2), 1\},$$

and, for the distributions,

$$\max\{F_1(x_1), F_2(x_2)\} \leq F^*(x_1, x_2) \leq \min\{F_1(x_1) + F_2(x_2), 1\}, \tag{1.43}$$

with $F^*(x_1, x_2) = P(X_1 \leq x_1 \cup X_2 \leq x_2)$.

Example 1.75 (Race model inequality). This example relates to the audio-visual interaction context described in introductory Example 1.2, where visual (RT_V), auditory (RT_A), and visual-audio (RT_{VA}) reaction times are observed. According to the *race model*, RT_{VA} is considered a random variable identical in distribution to $\min\{RT_V, RT_A\}$, with RT_V and RT_A random variables coupled with respect to a common probability space.

Rewriting Equation 1.43 restricted to the diagonal, i.e., to all (x_1, x_2) with $x_1 = x_2 \equiv x$, one obtains for the corresponding distribution functions

$$\max\{F_V(x), F_A(x)\} \leq F_{VA}(x) \leq \min\{F_V(x) + F_A(x), 1\}, \quad x \geq 0. \tag{1.44}$$

The upper bound was suggested as a test of the race model in the reaction time literature (Miller, 1982); its testing has become routine procedure in empirical studies for discriminating "genuine" multisensory integration, that is, integration based the action of multisensory neurons, from mere probability summation (race model) (Colonius and Diederich, 2006).

[16] Note that perfect dependence does not mean that the correlation equals $+1$ or -1.

Fréchet-Hoeffding bounds for $n \geq 3$

Using[17] the simple generalization of the Inequalities (1.41) to $n \geq 3$,

$$\max\{0, \sum_j P(A_j) - (n-1)\} \leq P(A_1 \cap \cdots \cap A_n) \leq \min_j P(A_j), \qquad (1.45)$$

yields the general *Fréchet bounds* (short for Fréchet-Hoeffding b.)

$$\max\{0, F_1(x_1) + \cdots + F_n(x_n)\} \leq F(x_1, \ldots, x_n) \leq \min_j F_j(x_j), \qquad (1.46)$$

for all $x_j \in \Re$, $j = 1, \ldots, n$ and with $F \in \mathcal{F}(F_1, \ldots, F_n)$, the class of all n-variate distribution functions with fixed marginals $F_1, \ldots, F_n$. Moreover, it can be shown that the bounds are sharp, i.e., they cannot be improved in general.

The Fréchet upper bound, $\min_j F_j(x_j) \equiv F^+(\mathbf{x})$, $\mathbf{x} \in \Re^n$, is a distribution function. Indeed, if one of the univariate distributions is continuous, say F_1, then $F^+(\mathbf{x})$ is the joint distribution of $\mathbf{X}$ defined by quantile coupling via $U =_d F_1(X_1)$ (cf. Example 1.69):

$$\mathbf{X} =_d (U, F_2^{-1}(U), \ldots, F_n^{-1}(U)).$$

If all marginals are continuous, the random variables show perfect positive dependence (*comonotonicity*). For proof in the general case, see Joe (1997, pp. 58).

Note that the lower Fréchet bound is not in general a distribution function for $n \geq 3$. This is easy to see in the case of $n = 3$ with F_1, F_2, and F_3 non-degenerate. If the lower Fréchet bound $F^-(\mathbf{x})$ is a distribution for (X_1, X_2, X_3), then all three bivariate margins must also be (two-dimensional) lower Fréchet bounds. If one of the three distributions is continuous, say F_1, then, by using the probability transform and to satisfy countermonotonicity,

$$X_j = F_j^{-1}(1 - F_1(X_1)), \;\; j = 2, 3.$$

However, in this case X_2, X_3 are positively associated and this is a contradiction.

Nevertheless, a necessary and sufficient condition for the lower Fréchet

[17] All results in this and the subsequent section are from Joe (1997, 2015) where a wealth of results are gathered together.

bound of $\mathcal{F}(F_1, \ldots, F_n)$ to be a distribution can be stated for all $n \geq 3$:

$$\begin{aligned} &\text{either } \sum_j F_j(x_j) \leq 1, && \text{whenever } 0 < F_j(x_j) < 1,\ j = 1, \ldots, n; \\ &\text{or } \sum_j F_j(x_j) \geq n - 1, && \text{whenever } 0 < F_j(x_j) < 1,\ j = 1, \ldots, n. \end{aligned} \tag{1.47}$$

It turns out that these conditions imply that each of the F_j must have a discrete component.

Fréchet distribution classes with given higher-order marginals

We concentrate here on some examples for the case of $n = 3$.

(A) Class $\mathcal{F}(F_{12}, F_{13})$. We assume that the conditional distributions $F_{2|1}$ and $F_{3|1}$ are well defined. $\mathcal{F}(F_{12}, F_{13})$ is always non-empty since it contains the trivariate distribution with the second and third variables conditionally independent given the first, defined by

$$F_I(\mathbf{x}) = \int_{-\infty}^{x_1} F_{2|1}(x_2|y) F_{3|1}(x_3|y)\, dF_1(y). \tag{1.48}$$

Now let $F \in \mathcal{F}(F_{12}, F_{13})$ and write

$$F(\mathbf{x}) = \int_{-\infty}^{x_1} F_{23|1}(x_2, x_3|y)\, dF_1(y),$$

where $F_{23|1}$ is the bivariate conditional distribution of the second and third variable given the first. Applying the bivariate bounds from Proposition 1.74 yields the Fréchet upper bound

$$F^+(\mathbf{x}) = \int_{-\infty}^{x_1} \min\{F_{2|1}(x_2|y), F_{3|1}(x_3|y)\}\, dF_1(y)$$

and the Fréchet lower bound

$$F^-(\mathbf{x}) = \int_{-\infty}^{x_1} \max\{0, F_{2|1}(x_2|y) + F_{3|1}(x_3|y) - 1\}\, dF_1(y),$$

and both of these bounds are proper distributions.

One application of this result yields a generalization of the race model inequality (cf. Example 1.75) by conditioning on a third random variable, for example, an indicator variable.

(B) Class $\mathcal{F}(F_{12}, F_{13}, F_{23})$. This class is more difficult to analyze. There is clearly a need to check for compatibility of the three bivariate margins. For example, if F_{12}, F_{13} and F_{23} are bivariate standard normal distributions with corresponding correlations ρ_{12}, ρ_{13}, and ρ_{23}, then compatibility would mean that the correlation matrix with these values is non-negative definite.

It is straightforward, however, to obtain upper and lower bounds. For $F \in \mathcal{F}(F_{12}, F_{13}, F_{23})$, and $\mathbf{x} \in \Re^3$:

$$\max\{0, b_1(\mathbf{x}), b_2(\mathbf{x}), b_3(\mathbf{x})\} \leq F(\mathbf{x}) \leq \min\{a_{12}(\mathbf{x}), a_{13}(\mathbf{x}), a_{23}(\mathbf{x}), a_{123}(\mathbf{x})\}, \tag{1.49}$$

where

$$\begin{aligned}
a_{12}(\mathbf{x}) &= F_{12}(x_1, x_2),\ a_{13}(\mathbf{x}) = F_{13}(x_1, x_3), a_{23}(\mathbf{x}) = F_{23}(x_2, x_3),\\
a_{123}(\mathbf{x}) &= 1 - F_1(x_1) - F_2(x_2) - F_3(x_3)\\
&\quad + F_{12}(x_1, x_2) + F_{13}(x_1, x_3) + F_{23}(x_2, x_3),\\
b_1(\mathbf{x}) &= F_{12}(x_1, x_2) + F_{13}(x_1, x_3) - F_1(x_1),\\
b_2(\mathbf{x}) &= F_{12}(x_1, x_2) + F_{23}(x_2, x_3) - F_2(x_2),\\
b_3(\mathbf{x}) &= F_{13}(x_1, x_3) + F_{23}(x_2, x_3) - F_3(x_3).
\end{aligned}$$

For F_{12}, F_{13} and F_{23} to be compatible bivariate margins, the upper bound must be greater or equal to the lower bound in the above inequality for all $\mathbf{x}$. Some versions of these bounds play a role in describing properties of general parallel processing models.

Several methods for obtaining compatibility conditions are presented in Joe (1997, 2015). Some distribution-independent results exist, but no solution for the general case has been suggested. For continuous distributions, a result is based on considering *Kendall's tau* (τ), a measure of dependence defined as follows.

Definition 1.76 (Kendall's tau) Let F be a continuous bivariate distribution function and let $(X_1, X_2), (X_1', X_2')$ be independent pairs with distribution F. Then (the population version of) *Kendall's tau* equals the probability of concordant pairs minus the probability of discordant pairs, i.e.,

$$\begin{aligned}
\tau &= P[(X_1 - X_1')(X_2 - X_2') > 0] - P[(X_1 - X_1')(X_2 - X_2') < 0]\\
&= 2P[(X_1 - X_1')(X_2 - X_2') > 0] - 1\\
&= 4P[X_1 > X_1', X_2 > X_2'] - 1 = 4\int_{I^2} F\, dF - 1.
\end{aligned}$$

The following gives a necessary condition for compatibility:

Proposition 1.77 *(Joe, 1997, p. 76) Let* $F \in \mathcal{F}(F_{12}, F_{13}, F_{23})$ *and suppose*

F_{jk}, $j < k$, are continuous. Let $\tau_{jk} = \tau_{kj}$ be the value of Kendall's tau for F_{jk}, $j \neq k$. Then the inequality

$$-1 + |\tau_{ij} + \tau_{jk}| \leq \tau_{ik} \leq 1 - |\tau_{ij} - \tau_{jk}|$$

holds for all permutations (i, j, k) of $(1, 2, 3)$ and the bounds are sharp.

Thus, if the above inequality does not hold for some (i, j, k), then the three bivariate margins are not compatible. Sharpness follows from the special trivariate normal case: Kendall's tau for the bivariate normal is $\tau = (2/\pi)\arcsin(\rho)$, so that the inequality becomes

$$-\cos(\tfrac{1}{2}\pi(\tau_{12} + \tau_{23})) \leq \sin(\tfrac{1}{2}\pi\tau_{13}) \leq \cos(\tfrac{1}{2}\pi(\tau_{12} - \tau_{23})),$$

with $(i, j, k) = (1, 2, 3)$.

Best bounds on the distribution of sums

Given that processing time in the psychological context is often assumed to be additively composed of certain sub-processing times, it is clearly of interest to establish bounds on the distribution of sums of random variables with arbitrary dependency structure. In keeping with the assumptions of the previous sections, existence of a multivariate distribution with given marginal distributions is presumed.

We start with the case of $n = 2$; consider the sum $S = X_1 + X_2$ with possibly dependent random variables X_1 and X_2. We are interested in finding random variables $S_{\min}$ and $S_{\max}$ such that

$$P(S_{\min} \leq s) \leq P(S \leq s) \leq P(S_{\max} \leq s), \quad s \in \Re,$$

and these bounds should be the best possible (in the stochastic order sense).

Let F_i, $i = 1, 2$, be the (marginal) distribution for X_i, and let $F_i^-(s) = P(X_i < s)$ be the left limit, defined for all $s \in \Re$, with $i = 1, 2$. We introduce

$$F_{\min}(s) = \sup_{x \in \Re} \max\left\{F_1^-(x) + F_2^-(s - x) - 1, 0\right\} \tag{1.50}$$

and

$$F_{\max}(s) = \inf_{x \in \Re} \min\left\{F_1(x) + F_2(s - x), 1\right\}. \tag{1.51}$$

It is routine to show that there exist random variables $S_{\min}$ and $S_{\max}$ such that $F_{\min}(s) = P(S_{\min} < s)$ and $F_{\max}(s) = P(S_{\max} \leq s)$ (see Denuit et al., 2005, p. 364).

Proposition 1.78 *If $S = X_1 + X_2$ and $F_{\min}(s)$ and $F_{\max}(s)$ are defined by Equation 1.50 and Equation 1.51, then*

$$F_{\min}(s) \leq F_S(s) \leq F_{\max}(s),$$

for all $s \in \Re$.

Proof (Denuit et al., 2005, p. 364) For arbitrary s and x in $\Re$, it is clear that $X_1 > x$ and $X_2 > s - x$ together imply $S > s$, so that

$$F_S(s) \leq P(X_1 \leq x \text{ or } X_2 \leq s - x) \leq F_1(x) + F_2(s - x),$$

obviously implying $F_S(s) \leq F_{\max}(s)$ for all s. Moreover, note that

$$P(X_1 < x) + P(X_2 < s - x) - P(X_1 < x, X_2 < s - x) \leq 1,$$

from which it follows that

$$\max\left\{F_1^-(x) + F_2^-(s - x) - 1, 0\right\} \leq P(X_1 < x, X_2 < s - x) \leq F_S(s),$$

which implies the desired result. □

Note that the distribution functions of $S_{\min}$ and $S_{\max}$ provide the best possible bounds on the distribution of S, but these random variables are no longer expressible as sums. Here is an example allowing an explicit computation of $F_{\min}$ and $F_{\max}$.

Example 1.79 Let $\text{Exp}(1/\lambda, \theta)$ denote a shifted exponential distribution $F(x) = 1 - \exp[-(x - \theta)/\lambda]$ for $\lambda > 0$ and $x \geq \theta$.

If X_i is distributed as $\text{Exp}(1/\lambda_i, \theta_i)$, $i = 1, 2$, then, using the method of Lagrange multipliers to determine the extrema of $F_S(s)$ under the constraint $x_1 + x_2 = s$, the distribution function of $S = X_1 + X_2$ is bounded below by

$$F_{\min}(s) = 1 - \exp[-(x - \theta^*)/(\lambda_1 + \lambda_2)],$$

with $\theta^* = \theta_1 + \theta_2 + (\lambda_1 + \lambda_2)\log(\lambda_1 + \lambda_2) - \lambda_1 \log(\lambda_1) - \lambda_2 \log(\lambda_2)$. It is also bounded above by

$$F_{\max}(s) = 1 - \exp[-(x - (\theta_1 + \theta_2))/\max(\lambda_1, \lambda_2)].$$

Best bounds can also be obtained for sums of $n \geq 3$ random variables and for non-decreasing continuous functions of the random variables. Moreover, assuming positive dependence among the random variables, the bounds in Proposition 1.78 can be sharpened.

1.3.7 Copula theory

There is a close connection between studying multivariate distributions with given marginals, i.e., the Fréchet distribution classes, and the theory of copulas to be introduced next.

Definition, Examples, and Sklar's theorem

Let (X, Y) be a pair of random variables with joint distribution function $F(x, y)$ and marginal distributions $F_X(x)$ and $F_Y(y)$. To each pair of real numbers (x, y), we can associate three numbers: $F_X(x)$, $F_Y(y)$, and $F(x, y)$. Note that each of these numbers lies in the interval $[0, 1]$. In other words, each pair (x, y) of real numbers leads to a point $(F_X(x), F_Y(y))$ in the unit square $[0, 1] \times [0, 1]$, and this ordered pair in turn corresponds to a number $F(x, y)$ in $[0, 1]$.

$$(x, y) \mapsto (F_X(x), F_Y(y)) \mapsto F(x, y) = C(F_X(x), F_Y(y)),$$
$$\Re \times \Re \longrightarrow [0, 1] \times [0, 1] \xrightarrow{C} [0, 1]$$

The mapping C is an example of a copula (it "couples" the bivariate distribution with its marginals). Using the probability integral transformation, a straightforward definition for any finite dimension n is the following:

Definition 1.80 A function $C : [0, 1]^n \longrightarrow [0, 1]$ is called n-dimensional *copula* if there is a probability space $(\Omega, \mathcal{F}, P)$ supporting a vector of standard uniform random variables $(U_1, \ldots, U_n)$ such that

$$C(u_1, \ldots, u_n) = P(U_1 \leq u_1, \ldots, U_n \leq u_n), \quad u_1, \ldots, u_n \in [0, 1].$$

The concept of copula has stirred a lot of interest in recent years in several areas of statistics, including finance, mainly for the following reasons (see e.g., Joe, 2015): it allows (i) to study the structure of stochastic dependency in a "scale-free" manner, i.e., independent of the specific marginal distributions, and (ii) to construct families of multivariate distributions with specified properties. The following important theorem laid the foundation of these studies (for a proof, see Nelsen, 2006, Theorem 2.10.9).

Theorem 1.81 (Sklar's Theorem, 1959) *Let $F(x_1, \ldots, x_n)$ be an n-variate distribution function with margins $F_1(x_1), \ldots, F_n(x_n)$; then there exists an n-copula $C : [0, 1]^n \longrightarrow [0, 1]$ that satisfies*

$$F(x_1, \ldots, x_n) = C(F_1(x_1), \ldots, F_n(x_n)), \quad (x_1, \ldots x_n) \in \Re^n.$$

If all univariate margins $F_1, \ldots, F_n$ are continuous, then the copula is unique. Otherwise, C is uniquely determined on $\mathrm{Ran}F_1 \times \mathrm{Ran}F_2 \times \ldots \mathrm{Ran}F_n$.

If $F_1^{-1}, \ldots, F_n^{-1}$ are the quantile functions of the margins, then for any $(u_1, \ldots, u_n) \in [0,1]^n$

$$C(u_1, \ldots, u_n) = F(F_1^{-1}(u_1), \ldots, F_n^{-1}(u_n)).$$

Sklar's theorem shows that copulas remain invariant under strictly increasing transformations of the underlying random variables. It is possible to construct a wide range of multivariate distributions by choosing the marginal distributions and a suitable copula.

Example 1.82 (Bivariate exponential) For $\delta > 0$, the distribution

$$F(x,y) = \exp\{-[e^{-x} + e^{-y} - (e^{\delta x} + e^{\delta y})^{-1/\delta}]\}, \quad -\infty < x, y, < +\infty,$$

with margins $F_1(x) = \exp\{-e^{-x}\}$ and $F_2(y) = \exp\{-e^{-y}\}$ corresponds to the copula

$$C(u,v) = uv \exp\{[(-\log u)^{-\delta} + (-\log v)^{-\delta}]^{-1/\delta}\},$$

an example of the class of *bivariate extreme value* copulas characterized by $C(u^t, v^t) = C^t(u,v)$, for all $t > 0$.

There is an alternative, analytical definition of copula based on the fact that distribution functions can be characterized as functions satisfying certain conditions, without reference to a probability space (cf. Proposition 1.24).

Definition 1.83 An n-dimensional copula C is a function on the unit n-cube $[0,1]^n$ that satisfies the following properties:

1. the range of C is the unit interval $[0,1]$;
2. $C(\mathbf{u})$ is zero for all $\mathbf{u}$ in $[0,1]^n$ for which at least one coordinate is zero (groundedness);
3. $C(\mathbf{u}) = u_k$ if all coordinates of $\mathbf{u}$ are 1 except the k-the one;
4. C is *n-increasing*, that is, for every $\mathbf{a} \leq \mathbf{b}$ in $[0,1]^n$ ($\leq$ defined componentwise) the volume assigned by C to the n-box $[\mathbf{a}, \mathbf{b}] = [a_1, b_1] \times \cdots \times [a_n, b_n]$ is nonnegative.

One can show that groundedness and the n-increasing property are sufficient to define a proper distribution function (for $n = 2$, see Proposition 1.24). Moreover, from the results in Section 1.3.6 it follows that every copula is bounded by the Fréchet-Hoeffding bounds,

$$\max(u_1, + \cdots + u_n - n + 1, 0) \leq C(u_1, \ldots, u_n) \leq \min(u_1, \ldots, u_n).$$

For $n = 2$, both of the bounds are copulas themselves, but for $n \geq 3$ the

lower bound is no longer n-increasing. Another well-known copula is the *independence* copula,

$$C(u_1, \ldots, u_n) = \prod_{i=1}^{n} u_i.$$

Moreover, copulas are uniformly continuous and all their partial derivatives exist almost everywhere, which is a useful property especially for computer simulations. Copulas with discrete margins have also been defined, but their treatment is less straightforward (for a review, see Pfeifer and Nešlehová, 2004).

Copula density and pair copula constructions (vines)

If the probability measure associated with a copula C is absolutely continuous (with respect to the Lebesgue measure on $[0,1]^n$), then there exists a *copula density* $c : [0,1]^n \longrightarrow [0,\infty]$ almost everywhere unique such that

$$C(u_1, \ldots, u_n) = \int_0^{u_1} \cdots \int_0^{u_n} c(v_1, \ldots, v_n)\, dv_n \ldots dv_1, \quad u_1, \ldots, u_n \in [0,1].$$

Such an absolutely continuous copula is n times differentiable and

$$c(u_1, \ldots, u_n) = \frac{\partial}{\partial u_1} \cdots \frac{\partial}{\partial u_n} C(u_1, \ldots, u_n), \quad u_1, \ldots, u_n \in [0,1].$$

For example, the independence copula is absolutely continuous with density equal to 1:

$$\Pi(u_1, \ldots, u_n) = \prod_{k=1}^{n} u_k = \int_0^{u_1} \cdots \int_0^{u_n} 1\, dv_n \ldots dv_1.$$

When the density of a distribution $F_{12}(x_1, x_2)$ exists, differentiating yields

$$f_{12}(x_1, x_2) = f_1(x_1) f_2(x_2)\, c_{12}(F_1(x_1), F_2(x_2)).$$

This equation shows how independence is "distorted" by copula density c whenever c is different from 1. Moreover, this yields an expression for the conditional density of X_1 given $X_2 = x_2$:

$$f_{1|2}(x_1|x_2) = c_{12}(F_1(x_1), F_2(x_2)) f_1(x_1) \tag{1.52}$$

This the starting point of a recent, important approach to constructing high-dimensional dependency structures from pairwise dependencies ("vine copulas"). Note that a multivariate density of dimension n can be decomposed

as follows, here taking the case for $n = 3$:

$$f(x_1, x_2, x_3) = f_{3|12}(x_3|x_1, x_2) f_{2|1}(x_2|x_1) f_1(x_1).$$

Applying the decomposition in Equation 1.52 to each of these terms yields,

$$\begin{aligned} f_{2|1}(x_2|x_1) &= c_{12}(F_1(x_1), F_2(x_2)) f_2(x_2) \\ f_{3|12}(x_3|x_1, x_2) &= c_{13|2}(F_{1|2}(x_1|x_2), F_{3|2}(x_3|x_2)) f_{3|2}(x_3|x_2) \\ f_{3|2}(x_3|x_2) &= c_{23}(F_2(x_2), F_3(x_3)) f_3(x_3), \end{aligned}$$

resulting in the "regular vine tree" representation

$$\begin{aligned} f(x_1, x_2, x_3) = f_3(x_3) f_2(x_2) f_1(x_1) \ \text{(marginals)} \\ \times c_{12}(F_1(x_1), F_2(x_2)) \cdot c_{23}(F_2(x_2), F_3(x_3)) \ \text{(unconditional pairs)} \\ \times c_{13|2}(F_{1|2}(x_1|x_2), F_{3|2}(x_3|x_2)) \ \text{(conditional pair)}. \end{aligned} \tag{1.53}$$

In order to visualize this structure, in particular for larger n, one defines a sequence of trees (acyclic undirected graphs), a simple version of it is depicted in Figure 1.5.

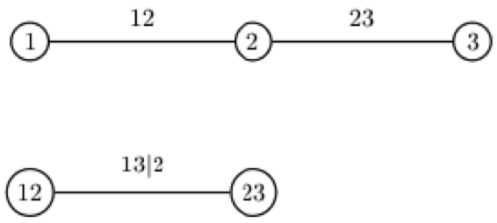

Figure 1.5 Graphical illustration of the decomposition in Equation 1.53. A line in the graph corresponds to the indices of a copula linking two distributions, unconditional in the upper graph, conditional in the lower graph.

Survival copula, dual and co-copula

Whenever it is more convenient to describe the multivariate distribution of a random vector $(X_1, \ldots, X_n)$ by means of its survival distribution, i.e.,

$$\hat{F}(x_1, \ldots, x_n) = P(X_1 > x_1, \ldots, X_n > x_n),$$

its *survival copula* can be introduced such that the analog of Sklar's theorem holds, with

$$\hat{F}(x_1, \ldots, x_n) = \hat{C}(\hat{F}_1(x_1), \ldots, \hat{F}_n(x_n)),$$

where the $\hat{F}_i$, $i = 1, \ldots, x_n$, are the marginal survival distributions. In the continuous case there is a one-to-one correspondence between the copula and its survival copula. For $n = 2$, this is

$$C(u_1, u_2) = \hat{C}(1 - u_1, 1 - u_2) + u_1 + u_2 - 1.$$

For the general case, we refer to (Mai and Scherer, 2012, pp. 20-21).

Two other functions closely related to copulas and survival copulas are useful in the response time modeling context. The *dual of a copula* C is the function $\tilde{C}$ defined by $\tilde{C}(u, v) = u + v - C(u, v)$ and the *co-copula* is the function C^* defined by $C^*(u, v) = 1 - C(1 - u, 1 - v)$. Neither of these is a copula, but when C is the (continuous) copula of a pair of random variables X and Y, the dual of the copula and the co-copula each express a probability of an event involving X and Y:

$$\tilde{C}(F(x), G(y)) = P(X \leq x \text{ or } Y \leq y)$$

and

$$C^*(1 - F(x), 1 - G(y)) = P(X > x \text{ or } Y > y).$$

Copulas with singular components

If the probability measure associated with a copula C has a singular component, then the copula also has a singular component which can often be detected by finding points $(u_1, \ldots, u_n) \in [0, 1]^n$, where some (existing) partial derivative of the copula has a point of discontinuity. A standard example is the upper *Fréchet bound copula*

$$M(u_1, \ldots, u_n) = \min(u_1, \ldots, u_n),$$

where the partial derivatives have a point of discontinuity;

$$\frac{\partial}{\partial u_k} M(u_1, \ldots, u_n) = \begin{cases} 1, u_k < \min(u_1, \ldots u_{k-1}, u_{k+1}, \ldots, u_n), \\ 0, u_k > \min(u_1, \ldots u_{k-1}, u_{k+1}, \ldots, u_n) \end{cases}.$$

The probability measure associated with $M(u_1, \ldots, u_n)$ assigns all mass to the diagonal of the unit n-cube $[0, 1]^n$ ("perfect positive dependence").

Archimedean copulas

The class of *Archimedean copulas* plays a major role in constructing multivariate dependency. We start by motivating this class via a mixture model interpretation, following the presentation in Mai and Scherer (2012). Consider a sequence of i.i.d. exponential random variables $E_1, \ldots, E_n$ defined on a probability space $(\Omega, \mathcal{F}, P)$ with mean $\mathrm{E}[E_1] = 1$. Independent of the E_i, let M be a positive random variable and define the vector of random variables $(X_1, \ldots, X_n)$ by

$$(X_1, \ldots, X_n) := \left(\frac{E_1}{M}, \ldots, \frac{E_n}{M} \right).$$

This can be interpreted as a two-step experiment. First, the intensity M is drawn. In a second step, an n-dimensional vector of i.i.d. $\mathrm{Exp}(M)$ variables is drawn. Thus, for each margin $k \in \{1, \ldots, n\}$, one has

$$\begin{aligned} \bar{F}_k(x) &= P(X_k > x) = P(E_k > xM) \\ &= \mathrm{E}\left[\mathrm{E}[\mathbb{1}_{\{E_k > xM\}} | M]\right] = \mathrm{E}\left[e^{-xM}\right] := \varphi(x), \quad x \geq 0, \end{aligned}$$

where φ is the *Laplace transform* of M. Thus, the margins are related to the mixing variable M via its Laplace transform. As long as M is not deterministic, the X_k's are dependent, since they are all affected by M: whenever the realization of M is large (small), all realizations of X_k are more likely to be small (large). The resulting copula is parametrized by the Laplace transform of M:

Proposition 1.84 *The survival copula of random vector* $(X_1, \ldots, X_n) \equiv \left(\frac{E_1}{M}, \ldots, \frac{E_n}{M} \right)$ *is given by*

$$\varphi\left(\varphi^{-1}(u_1) + \cdots + \varphi^{-1}(u_n)\right) := C_\varphi(u_1, \ldots, u_n), \tag{1.54}$$

where $u_1, \ldots, u_n \in [0, 1]$ *and* φ *is the Laplace transform of* M*.*

Copulas of this structure are called *Archimedean copulas*. Written differently,

$$P(X_1 > x_1, \ldots, X_n > x_n) = C_\varphi\left(\varphi(x_1), \ldots, \varphi(x_1)\right),$$

where $x_1, \ldots, x_n \geq 0$. From Equation 1.54 it clearly follows that the copula C_φ is symmetric, i.e., the vector of X_k's has an *exchangeable* distribution.

Proof of Proposition 1.84. The joint survival distribution of $(X_1, \ldots, X_n)$

is given by

$$\begin{aligned}\bar{F}(x_1,\dots,x_n) &= P(X_1 > x_1,\dots,X_n > x_n)\\ &= \mathrm{E}\left[\mathrm{E}[\mathbb{1}_{\{E_1>x_1M\}}\cdots\mathbb{1}_{\{E_n>x_nM\}}|M]\right]\\ &= \mathrm{E}[e^{-M\sum_{k=1}^n x_k}] = \varphi\left(\sum_{k=1}^n x_k\right),\end{aligned}$$

where $(x_1,\dots x_n) \in [0,\infty]^n$. The proof is established by transforming the marginals to $U[0,1]$ and by using the survival version of Sklar's theorem. □

Not all Archimedean copulas are obtained via the above probabilistic construction. The usual definition of Archimedean copulas is analytical: the function defined in Equation 1.54 is an *Archimedean copula generator* of the general form

$$\varphi\left(\varphi^{-1}(u_1)+\cdots+\varphi^{-1}(u_n)\right) = C_\varphi(u_1,\dots,u_n),$$

if φ satisfies the following properties:

(i) $\varphi : [0,\infty] \longrightarrow [0,1]$ is strictly increasing and continuously differentiable (of all orders);
(ii) $\varphi(0) = 1$ and $\varphi(\infty) = 0$;
(iii) $(-1)^k\varphi^{(k)} \geq 0, k = 1,\dots n$, i.e., the derivatives are alternating in sign up to order n.

If Equation 1.54 is a copula for all n, then φ is a *completely monotone function* and turns out to be the Laplace transform of a positive random variable M, as in the previous probabilistic construction (cf. Mai and Scherer, 2012, p. 64).

Example: Clayton copula and the copula of max and min order statistics

The generator of the Clayton family is given by

$$\varphi(x) = (1+x)^{-1/\vartheta}, \quad \varphi^{-1}(x) = x^{-\vartheta} - 1, \quad \vartheta \in (0,\infty).$$

It can be shown that this corresponds to a Gamma$(1/\vartheta, 1)$ distributed mixing variable M. The Clayton copula is hidden in the following application: Let $X_{(1)}$ and $X_{(n)}$ be the min and max order statistics of a sample $X_1,\dots,X_n$; one can compute the copula of $(-X_{(1)}, X_{(n)})$ via

$$P(-X_{(1)} \leq x, X_{(n)} \leq y) = P\left(\cap_{i=1}^n\{X_i \in [-x,y]\}\right) = (F(y) - F(-x))^n,$$

for $-x \leq y$ and zero otherwise. Marginal distributions of $-X_{(1)}$ and $X_{(n)}$ are, respectively,

$$F_{X_{(n)}}(y) = F(y)^n \quad \text{and} \quad F_{-X_{(1)}}(x) = (1-F(-x))^n.$$

Sklar's theorem yields the copula

$$C_{-X_{(1)}X_{(n)}}(u,v) = \max\{0, (u^{1/n} + v^{1/n} - 1)^n\},$$

which is the bivariate Clayton copula with parameter $\vartheta = 1/n$. Because $C_{X_{(1)}X_{(n)}}(u,v) = v - C_{-X_{(1)},X_{(n)}}(1-u,v)$, it can be shown with elementary computations (see Schmitz, 2004) that

$$\lim_{n\to+\infty} C_{X_{(1)}X_{(n)}}(u,v) = v - \lim_{n\to+\infty} C_{-X_{(1)}X_{(n)}}(1-u,v) = v - (1-u)v = uv,$$

for all $u, v \in (0,1)$, i.e., $X_{(1)}$ and $X_{(n)}$ are *stochastically independent* asymptotically.

Operations on distributions not derivable from operations on random variables

In the introductory section, Example 1.1 claimed that the mixture of two distributions F and G, say, $\phi(F,G) = pF + (1-p)G$, $p \in (0,1)$, was not derivable by an operation defined directly on the corresponding random variables, in contrast to, e.g., the convolution of two distributions which is derivable from addition of the random variables. Here we first explicate the definition of "derivability", give a proof of that claim, and conclude with a caveat.

Definition 1.85 A binary operation ϕ on the space of univariate distribution functions Δ is said to be *derivable* from a function on random variables if there exists a measurable two-place function V satisfying the following condition: for any distribution functions $F, G \in \Delta$ there exist random variables X for F and Y for G, defined on a common probability space, such that $\phi(F,G)$ is the distribution function of $V(X,Y)$.

Proposition 1.86 *The mixture $\phi(F,G) = pF + (1-p)G$, $p \in (0,1)$ is not derivable.*

Proof (Alsina and Schweizer, 1988) Assume ϕ is derivable, i.e., that a suitable function V exists. For any $a \in \Re$, define the unit step function ϵ_a in Δ by

$$\epsilon_a(t) = \begin{cases} 0, & \text{if } t \le a, \\ 1, & \text{if } t > a; \end{cases}$$

and, for any given x and y ($x \neq y$), let $F = \epsilon_x$ and $G = \epsilon_y$. Then F and G are, respectively, the distribution functions of X and Y, defined on a common probability space and respectively equal to x and y a.s. It follows that $V(X,Y)$ is a random variable defined on the same probability space

as X and Y which is equal to $V(x, y)$ a.s. The distribution of $V(X, Y)$ is $\epsilon_{V(x,y)}$. But since ϕ is derivable from V, the distribution of $V(X, Y)$ must be $\phi(\epsilon_x, \epsilon_y)$, which is $p\epsilon_x + (1-p)\epsilon_y$. Since $p\epsilon_x + (1-p)\epsilon_y \neq \epsilon_{V(x,y)}$, we have a contradiction. □

In Alsina and Schweizer (1988) it is shown that forming the geometric mean of two distributions is also not derivable. These examples demonstrate that the distinction between modeling with distribution functions, rather than with random variables, "...is intrinsic and not just a matter of taste" (Schweizer and Sklar, 1974) and that "the classical model for probability theory–which is based on random variables defined on a common probability space–has its limitations" (Alsina et al., 1993).

It should be noted, however, that by a slight extension of the copula concept, i.e., allowing for randomized operations, the mixture operation on Δ can be recovered by operations on random variables. Indeed, assume there is a Bernoulli random variable I independent of (X, Y) with $P(I = 1) = p$; then the random variable Z defined by $Z = IX + (1 - I)Y$ has mixture distribution $\phi(F, G)$. Such a representation of a mixture is often called a *latent variable representation* of ϕ and is easily generalized to a mixture of more than two components (see Faugeras, 2012, for an extensive discussion of this approach).

1.3.8 Concepts of dependence

It is obvious that copulas can represent any type of dependence among random variables. Moreover, we have seen that the upper and lower Fréchet-Hoeffding bounds correspond to maximal positive and negative dependence, respectively. There exist many, more or less strong, concepts of dependence that are often useful for model building. We can consider only a small subset of these.

Positive dependence

Here it is of particular interest to find conditions that guarantee multivariate positive dependence in the following sense: For a random vector $(X_1, \ldots, X_n)$

$$P(X_1 > x_1, \ldots, X_n > x_n) \geq \prod_{i=1}^{n} P(X_i > x_i), \tag{1.55}$$

for all real x, called *positive upper orthant dependence, (PUOD)*. Replacing "$>$" by "$\leq$" in the probabilities above yields the concept of *positive lower orthant dependence, (PLOD)*. An example for $PLOD$ are archimedean copulas

with completely monotone generators (see Proposition 1.84), this property following naturally from the construction where independent components are made positively dependent by the random variabel M:

$$(X_1, \ldots, X_n) = \left(\frac{E_1}{M}, \ldots, \frac{E_n}{M} \right).$$

Both *PUOD* and *PLOD* are implied by a stronger concept:

Definition 1.87 A random vector $\mathbf{X} = (X_1, \ldots, X_n)$ is said to be *associated* if

$$\mathrm{Cov}[f(\mathbf{X}), g(\mathbf{X})] \geq 0$$

for all non-decreasing real-valued functions for which the covariance exists.

Association implies the following properties:

1. Any subset of associated random variables is associated.
2. If two sets of associated random variables are independent of one another, then their union is a set of associated random variables.
3. The set consisting of a single random variable is associated.
4. Non-decreasing functions of associated random variables are associated.
5. If a sequence $\mathbf{X}^{(k)}$ of random vectors, associated for each k, converges to $\mathbf{X}$ in distribution, then $\mathbf{X}$ is also associated.

Examples for applying the concept include the following:

1. Let $X_1, \ldots X_n$ be independent. If $S_k = \sum_{i=1}^{k} X_i$, $k = 1, \ldots n$, then $(S_1, \ldots, S_n)$ is associated.
2. The order statistics $X_{(1)}, \ldots, X_{(n)}$ are associated.
3. Let Λ be a random variable and suppose that, conditional on $\Lambda = \lambda$, the random variables $X_1, \ldots, X_n$ are independent with common distribution F_λ. If $F_\lambda(x)$ is decreasing in λ for all x, then $X_1, \ldots, X_n$ are associated (without conditioning on λ). To see this let $U_1, \ldots, U_n$ be independent standard uniform random variables that are independent of Λ. Define $X_1, \ldots, X_n$ by $X_i = F_\Lambda^{-1}(U_i)), i = 1, \ldots, n$. As the X_i are increasing functions of the independent random variables $\Lambda, U_1, \ldots, U_n$ they are associated (Ross, 1996, p. 449).

An important property of association is that it allows inferring independence from uncorrelatedness. Although this holds for the general multivariate case, we limit presentation to the case of two random variables, X and Y, say.

Recall Hoeffding's lemma (Proposition 1.36),

$$\mathrm{Cov}(X,Y) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} [F_{XY}(x,y) - F_X(x)F_Y(y)]\, dx\, dy.$$

If X and Y are associated, then $\mathrm{Cov}(X,Y)$ must be greater than zero. Thus, uncorrelatedness implies $F_{XY}(x,y) - F_X(x)F_Y(y)$ is zero for all x, y, that is, X and Y are independent.

Whenever it is difficult to directly test for association, a number of stronger dependency concepts are available.

Definition 1.88 The bivariate density $f(x,y)$ is of *total positivity of order 2* (TP_2) if

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

for all $x < x'$ and $y < y'$ in the domain of $f(x,y)$.

This property has also been called "positive likelihood ratio dependence" and can be shown to imply association. A multivariate density

$$f(x_1,\ldots,x_i,\ldots,x_j,\ldots,x_n)$$

is called TP_2 *in pairs* if the density, considered as a function of x_i and x_j, with all other arguments fixed, is TP_2. In $\Re^n$, the concept of TP_2 *in pairs* is equivalent to the following:

Definition 1.89 A function $f : \Re^n \longrightarrow \Re$ is *multivariate total positive of order two,* MTP_2, if

$$f(\mathbf{x}\vee\mathbf{y})f(\mathbf{x}\wedge\mathbf{y}) \geq f(\mathbf{x})f(\mathbf{y}),$$

for all $\mathbf{x},\mathbf{y} \in \Re^n$, where $\mathbf{x}\vee\mathbf{y}$ denotes the vector of maxima, $\max(x_i,y_i)$, and $\mathbf{x}\wedge\mathbf{y}$ the vector of minima, $\min(x_i,y_i)$, $i = 1,\ldots,n$. If f is a density for random vector $\mathbf{X}$ satisfying MTP_2, then $\mathbf{x}$ is said to be MTP_2.

The class MTP_2 possesses a number of nice properties, among them:

1. If f, g are MTP_2, then fg is MTP_2;
2. A vector $\mathbf{X}$ of independent random variables is MTP_2;
3. the marginals of f which is MTP_2 are also MTP_2;
4. If $f(\mathbf{x},\mathbf{y})$ is MTP_2 on $\Re^n \times \Re^m$, $g(\mathbf{y},\mathbf{z})$ is MTP_2 on $\Re^m \times \Re^k$, then

$$h(\mathbf{x},\mathbf{z}) = \int f(\mathbf{x},\mathbf{y})g(\mathbf{y},\mathbf{z})\, d\mathbf{y}$$

is MTP_2 on $\Re^n \times \Re^k$;

5. If f is MTP_2, then $f(\phi_1(x_1), \ldots, \phi_n(x_n))$ is MTP_2, where ϕ_i are nondecreasing.

Moreover, it has been shown that MTP_2 implies association. Examples of MTP_2 random variables abound. For example, a multivariate normal density is MTP_2 if, and only if, the elements of the inverse of the covariance matrix, $B = \Sigma^{-1}$, $b_{ij} \leq 0$ for $i \neq j$. The vector of order statistics is MTP_2 as well, and so are negative multinomial variables. The following observation is of interest from a model-building point of view:

Let n-dimensional vectors $\mathbf{X}$ and $\mathbf{Y}$ be independent. If both $\mathbf{X}$ and $\mathbf{Y}$ are MTP_2, then the convolution $\mathbf{Z} = \mathbf{X} + \mathbf{Y}$ is no longer MTP_2, but $\mathbf{Z}$ is still associated.

Negative dependence

A concept of positive dependence for two random variables can easily be transformed into one of negative dependence. Specifically, if (X, Y) has positive dependence, then $(X, -Y)$ on $\Re^2$, or $(X, 1 - Y)$ on the unit square, have negative dependence. However, with more than two variables negative dependence is restricted. For example, if X_1 has perfect negative dependence with X_2 and with X_3, then X_2 and X_3 have perfect positive dependence, and the larger the number of variables, the weaker is negative dependence.

Reversing the Inequality (1.55) in the definition of $PUOD$ turns this multivariate positive dependence concept into *negative upper orthant dependence* ($NUOD$) and, similarly, $PLOD$ becomes $NLOD$ by reversing the inequality. Both $NUOD$ and $NLOD$ clearly capture the idea of multivariate negative dependence but the challenge has been to find conditions under which these inequalities are fulfilled. The most influential concept was introduced by Joag-Dev and Proschan (1983) (where all results of this section can be found), based on the simple idea that negative dependence means that splitting the set of random variables into two parts leads to a negative covariance between the two sets.

Definition 1.90 Random vector $\mathbf{X}$ is *negatively associated* (NA) if, for every subset $A \subseteq \{1, \ldots, n\}$,

$$\mathrm{Cov}(f(X_i, i \in A), g(X_j, j \in A^c)) \leq 0,$$

whenever f, g are nondecreasing.

NA may also refer to the set of random variables $\{X_1, \ldots, X_n\}$ or to the underlying distribution of $\mathbf{X}$. It has the following properties:

1. For a pair of random variables (X, Y), NA and *negative quadrant dependence*,
$$P(X \le x, Y \le y) \le P(X \le x)P(Y \le y)$$
are equivalent;
2. For disjoint subsets $A_1, \ldots, A_m$ of $\{1, \ldots, n\}$, and nondecreasing functions $f_1, \ldots, f_n$, $\mathbf{X}$ being NA implies
$$\mathrm{E}\prod_{i=1}^{m} f_i(\mathbf{X}_{A_i}) \le \prod_{i=1}^{m} \mathrm{E} f_i(\mathbf{X}_{A_i}),$$
where $\mathbf{X}_{A_i} = (X_j, j \in A_i)$.
3. If $\mathbf{X}$ is NA, then it is both $NUOD$ and $NLOD$.
4. A subset of NA random variables is NA.
5. If $\mathbf{X}$ has independent components, then it is NA.
6. Increasing functions defined on disjoint subsets of a set of NA variables are NA.
7. If $\mathbf{X}$ is NA and $\mathbf{Y}$ is NA and $\mathbf{X}$ is independent of $\mathbf{Y}$, then $(\mathbf{X}, \mathbf{Y})$ is NA.

Negative association is often created by conditioning. For example, let X_1, $\ldots, X_n$ be independent random variables. Then the joint conditional distribution of $X_1, \ldots, X_n$, given $\sum_{i=1}^{n} X_i = s$, is NA for almost all s.

A result on independence via uncorrelatedness, generalizing the result cited in the previous section on positive dependence, is the following

Proposition 1.91 *Suppose* $\mathbf{X}$ *is (positively) associated or negatively associated. Then*

(a) $\mathbf{X}_A$ *is independent of* $\mathbf{X}_B$ *if, and only if,* $\mathrm{Cov}(X_i, X_j) = 0$*, for* $i \in A, j \in B, A \cap B = \emptyset$*;*

(b) $X_1, \ldots, X_n$ *are mutually independent if, and only if,* $\mathrm{Cov}(X_i, X_j) = 0$ *for all* $i \ne j$*.*

For example, a multivariate normal density has negative (positive) association if and only if all its pairwise correlation coefficients are negative (positive).

Measuring dependence

The dependence properties discussed so far are characterizable as being *local*. For example, the inequality defining PQD holds for all pairs (x, y) of the random variables X, Y. From this point of view, covariance and correlation are *global* measures of dependency. Since most of the (local) dependence

concepts introduced above are invariant with respect to increasing transformations of the variables, two bivariate measures of monotone dependence for continuous variables are of special interest. The first, *Kendall's tau* was introduced in Definition 1.76. The other is the following.

Definition 1.92 (Spearman's rho) Let F be a continuous bivariate distribution function with univariate margins F_1, F_2 and let (X_1, X_2) have distribution F. Then the (population version) of *Spearman' rho*, ρ, is defined as the correlation of $F_1(X_1)$ and $F_2(X_2)$.

Since $F_1(X_1)$ and $F_2(X_2)$ are standard uniform random variables (U, V) with distribution function (copula) C, their expectations are $1/2$, their variances are $1/12$, and Spearman's rho becomes

$$\begin{aligned}\rho =& \frac{\mathrm{E}UV - \mathrm{E}U\mathrm{E}V}{\sqrt{\mathrm{Var}U\mathrm{Var}V}} = \frac{\mathrm{E}UV - 1/4}{1/12} = 12\,\mathrm{E}UV - 3 \\ =& 12\int_{I^2} uv\, dC(u,v) - 3. \end{aligned} \tag{1.56}$$

This parallels the copula version of Kendall's tau,

$$\tau = 4\int_{I^2} C\, dC - 1,$$

which can be interpreted as the expected value of the function $C(U, V)$ of standard uniform random variables, $4\,\mathrm{E}C(U, V) - 1$. Besides being invariant with respect to increasing transformations, both τ and ρ are equal to 1 for the bivariate Fréchet upper bound (one variable is an increasing function of the other) and to -1 for the Fréchet lower bound (one variable is a decreasing function of the other). There are numerous results for both measures in the literature, let us just mention an early one by Kruskal (1958) relating both measures: For the population versions of Kendall's tau and Spearman' s rho,

$$-1 \leq 3\tau - 2\rho \leq 1.$$

1.3.9 Stochastic orders

The simplest way of comparing two distribution functions is by comparison of the associated means (or medians). However, such a comparison is often not very informative, being based on only two single numbers[18]. In studies of reaction time, for example, analyses of empirical data have long been limited to simply comparing mean reaction times measured under different experimental conditions. However, it is now generally acknowledged that much

[18] Sometimes, means may not even exist for a distribution.

deeper information about the underlying mental processes can be uncovered by examining the entire distribution function (e.g., Palmer et al., 2011).

Several orders of distribution functions that compare the "location" or the "magnitude" of random variables will be considered here. Moreover, when one is interested in comparing two distributions that have the same mean, one usually considers the standard deviations or some other measure of variability of these distributions. But such a comparison is again based on only two single numbers. Several orders of distributions taking into account various forms of knowledge about the underlying distributions will be considered here. Orders associated with different notions of positive dependence as well multivariate orders are briefly mentioned. There is a wealth of results on the various order concepts (see Shaked and Shantikumar, 2007, for a comprehensive compilation) but, due to space limitations, our presentation must be very selective.

Univariate stochastic orders

Let X and Y be two random variables with distribution functions F and G, respectively, not necessarily defined on the same probability space. In Section 1.3.5, the usual stochastic domination order was defined in the context of quantile coupling:

Definition 1.93 X is said to be *stochastically dominated* (or *dominated in distribution*) by Y,

$$X \leq_d Y, \text{ if } F(x) \geq G(x), \quad x \in \Re.$$

By Strassen's theorem (Theorem 1.70), there is a coupling $(\hat{X}, \hat{Y})$ of X and Y such that $\hat{X} \leq \hat{Y}$ pointwise. This can be used to prove an equivalent condition for the above stochastic order which easily generalizes to multivariate orders (see below):

Proposition 1.94

$$X \leq_d Y \quad \textit{if, and only if,} \quad (*) \ \mathrm{E}[g(X)] \leq \mathrm{E}[g(Y)]$$

for all bounded nondecreasing functions g.

Proof Suppose $X \leq_d Y$ and obtain a coupling $(\hat{X}, \hat{Y})$ such that $\hat{X} \leq \hat{Y}$ pointwise. Then, for any bounded nondecreasing function g, $g(\hat{X}) \leq g(\hat{Y})$, implying $\mathrm{E}[g(X)] \leq \mathrm{E}[g(Y)$. Conversely, fix an $x \in \Re$ and take $g = 1_{(x,+\infty)}$ in $(*)$: then

$$P(X > x) \leq P(Y > x), \quad x \in \Re,$$

implying $X \leq_d Y$. □

Stochastic domination is closed under nondecreasing transformations, in the following sense: Let $X_1, \ldots, X_m$ be a set of independent random variables and let $Y_1, \ldots, Y_m$ be another set of independent random variables. If $X_i \leq_d Y_i$ for $i = 1, \ldots, m$, then, for any nondecreasing function $\psi : \Re^m \longrightarrow \Re$, one has

$$\psi(X_1, \ldots, X_m) \leq_d \psi(Y_1, \ldots, Y_m).$$

In particular,

$$\sum_{j=1}^{m} X_j \leq_d \sum_{j=1}^{m} Y_j,$$

that is, the order is closed under convolution.

Obviously, if both $X \leq_d Y$ and $Y \leq_d X$, then X and Y are equal in distribution, $X =_d Y$. However, sometimes a weaker condition already implies equality in distribution. Clearly, if $X \leq_d Y$ then $\mathrm{E}X \leq \mathrm{E}Y$. But we also have

Proposition 1.95 *If $X \leq_d Y$ and if $\mathrm{E}h(X) = \mathrm{E}h(Y)$ for some strictly increasing function h, then $X =_d Y$.*

Proof First, we assume $h(x) = x$; let $\hat{X}$ and $\hat{Y}$ be the variables of the corresponding coupling, thus $P(\hat{X} \leq \hat{Y}) = 1$. If $P(\hat{X} < \hat{Y}) > 0$, then $\mathrm{EX} = \mathrm{E}\hat{\mathrm{X}} < \mathrm{EY} = \mathrm{E}\hat{\mathrm{Y}}$, a contradiction to the assumption. Therefore, $X =_d \hat{X} = \hat{Y} =_d Y$. Now let h be some strictly increasing function. Observe that if $X \leq_d Y$, then $h(X) \leq_d h(Y)$ and, therefore, from the above result we have that $h(X) =_d h(Y)$. The strict monotonicity of h yields $X =_d Y$. □

The next order is based on the concept of hazard rate function (see Definition 1.50).

Definition 1.96 Let X and Y be two nonnegative random variables with absolutely continuous distribution functions F and G and with hazard rate functions r and q, respectively, such that

$$r(t) \geq q(t), \quad t \in \Re.$$

The X is said to be *smaller than Y in the hazard rate order*, denoted as $X \leq_{hr} Y$.

It turns out that $\leq_{hr}$ can be defined for more general random variables, not necessarily nonnegative and even without requiring absolute continuity. The condition is

$$\frac{\bar{G}(t)}{\bar{F}(t)} \text{ increases in } t \in (-\infty, \max(u_X, u_Y))$$

$(a/0)$ is taken to be equal to ∞ whenever $a > 0$.). Here u_X and u_Y denote the right endpoints of the supports of X and Y. The above can be written equivalently as

$$\bar{F}(x)\bar{G}(y) \geq \bar{F}(y)\bar{G}(x) \text{ for all } x \leq y, \tag{1.57}$$

which, in the absolutely continuous case, is equivalent to

$$\frac{f(x)}{\bar{F}(y)} \geq \frac{g(x)}{\bar{G}(y)} \quad \text{for all } x \leq y.$$

Setting $x = -\infty$ in (1.57), one can show that

$$X \leq_{hr} Y \implies X \leq_d Y.$$

Many more results on $\leq_{hr}$ are available, e.g., order statistics satisfy

$$X_{k:n} \leq_{hr} X_{k+1:n}$$

for $k = 1, 2, \ldots, n$. Moreover, if W is a random variable with (mixture) distribution function $pF + (1-p)G$ for some $p \in (0,1)$, then $X \leq_{hr} Y$ implies $X \leq_{hr} W \leq_{hr} Y$.

Definition 1.97 Let X and Y have densities f and g, respectively, such that

$$\frac{g(t)}{f(t)} \text{ increases in } t \text{ over the union of the supports of } X \text{ and } Y$$

(here $a/0$ is taken to be equal to ∞ whenever $a > 0$), or, equivalently,

$$f(x)g(y) \geq f(y)g(x) \text{ for all } x \leq y.$$

Then X is said to be *smaller than Y in the likelihood ratio order*, denoted by $X \leq_{lr} Y$.

An analogous definition can be given for discrete random variables, and even for mixed distributions. Moreover, it can be shown that, for distributions functions F and G for X and Y, respectively,

$$X \leq_{lr} Y \implies GF^{-1} \text{ is convex.}$$

The following (strict) hierarchy exists:

$$X \leq_{lr} Y \implies X \leq_{hr} Y \implies X \leq_d Y.$$

Results on order statistics and binary mixtures analogous to those for $\leq_{hr}$ cited above also hold for $\leq_{lr}$.

Univariate variability orders

The convex order. The first variability order considered here requires the definition of a *convex function* $\phi : \Re \longrightarrow \Re$, i.e.,

$$\phi(\alpha x + (1-\alpha)y) \leq \alpha\phi(x) + (1-\alpha)\phi(y),$$

for all $x, y \in \Re$ and $\alpha \in [0,1]$.

Definition 1.98 Let X and Y be two random variables such that

$$\mathrm{E}\phi(X) \leq \mathrm{E}\phi(Y) \text{ for all convex functions } \phi : \Re \longrightarrow \Re,$$

provided the expectations exist. Then X is said to be *smaller than Y in the convex order*, denoted as

$$X \leq_{cx} Y.$$

Roughly speaking, convex functions take on their (relatively) larger values over regions of the form $(-\infty, a) \cup (b, +\infty)$ for $a < b$. Therefore, when the condition in the above definition holds, Y is more likely to take on "extreme" values than X, that is, Y is "more variable" than X.[19]

The convex order has some strong implications. Using the convexity of functions $\phi(x) = x, \phi(x = -x)$ and $\phi(x) = x^2$, it is easily shown that

$$X \leq_{cx} Y \quad \text{implies} \quad \mathrm{E}X = \mathrm{E}Y \quad \text{and } \mathrm{Var}X = \mathrm{Var}Y.$$

Moreover, when $\mathrm{E}X = \mathrm{E}Y$, then a condition equivalent to $X \leq_{cx} Y$ is

$$\mathrm{E}[\max(X, a)] \leq \mathrm{E}[\max(Y, a)], \quad \text{for all } a \in \Re.$$

A deeper result, an extension of Strassen's theorem (Theorem 1.70) is the following:

Theorem 1.99 (Müller and Rüschendorf, 2001) *The random variables X and Y satisfy $X \leq_{cx} Y$ if, and only if, there exist two random variables $\hat{X}$ and $\hat{Y}$, defined on the same probability space, such that*

$$\hat{X} =_d X \quad \textit{and} \quad \hat{Y} =_d Y,$$

and $\{\hat{X}, \hat{Y}\}$ is a martingale, i.e., $\mathrm{E}[\hat{Y}|\hat{X}] = \hat{X}$ a.s. Furthermore, the random variables $\hat{X}$ and $\hat{Y}$ can be selected such that $[\hat{Y}|\hat{X} = x]$ is increasing in x in the stochastic order $\leq_d$.

[19] Clearly, it is sufficient to consider only functions ϕ that are convex on the union of the supports of X and Y rather than over the whole real line.

As noted above, the convex order only compares random variables that have the same means. One way to drop this requirement is to introduce a so-called *dilation order* by defining

$$X \leq_{dil} Y, \text{ if } [X - \mathrm{E}X] \leq_{cx} [Y - \mathrm{E}Y].$$

For nonnegative random variables X and Y with finite means, one can alternatively define the *Lorenz order* by

$$X \leq_{Lorenz} Y, \quad \text{if} \quad \frac{X}{\mathrm{E}X} \leq_{cx} \frac{Y}{\mathrm{E}Y},$$

which can be used to order random variables with respect to the *Lorenz curve* used, e.g., in economics to measure the inequality of incomes.

We conclude with a result that nicely connects the convex order with the multivariate dependence concepts from Section 1.3.8.

Proposition 1.100 *Let $X_1, \ldots, X_n$ be positively associated (cf. Definition 1.87) random variables, and let $Y_1, \ldots, Y_n$ be independent random variables such that $X_i =_d Y_i$, $i = 1, \ldots, n$. Then*

$$\sum_{i=1}^{n} X_i \geq_{cx} \sum_{i=1}^{n} Y_i.$$

If "positively" is replace by "negatively" above (cf. Definition 1.90), then the convex order sign is reversed.

This makes precise the intuition that, if random variables $X_1, \ldots, X_n$ are "more positively [negatively] associated" than random variables $Y_1, \ldots, Y_n$ in some sense, but otherwise have identical margins for each $i = 1, \ldots, n$, then the sum of the X_i should be larger [smaller] in the convex order than the Y_i.

The dispersive order. This order is based on comparing difference between quantiles of the distribution functions.

Definition 1.101 Let X and Y be random variables with quantile functions F^{-1} and G^{-1}, respectively. If

$$F^{-1}(\beta) - F^{-1}(\alpha) \leq G^{-1}(\beta) - G^{-1}(\alpha), \quad \text{whenever} \quad 0 < \alpha \leq \beta < 1,$$

then X is said to be *smaller than Y in the dispersive order*, denoted by $X \leq_{disp} Y$.

In contrast to the convex order, the dispersive order is clearly *location free*:

$$X \leq_{disp} Y \Leftrightarrow X + c \leq_{disp} Y, \quad \text{for any real } c.$$

The dispersive order is also *dilative*, i.e., $X \leq_{disp} aX$ whenever $a \geq 1$ and, moreover,

$$X \leq_{disp} Y \Leftrightarrow -X \leq_{disp} -Y.$$

However, it is not closed under convolutions. Its characterization requires two more definitions.

First, a random variable Z is called *dispersive* if $X + Z \leq_{disp} Y + Z$ whenever $X \leq_{disp} Y$ and Z is independent of X and Y. Second, a density (more general, any nonnegative function) g is called *logconcave* if $\log g$ is concave, or, equivalently,

$$g[\alpha x + (1-\alpha)y] \geq [g(x)]^{\alpha}[g(y)]^{1-\alpha}.$$

Proposition 1.102 *The random variable X is dispersive if, and only if, X has a logconcave density.*[20]

It can be shown that the dispersion order implies the dilation order, that is, for random variables with finite means,

$$X \leq_{disp} Y \Longrightarrow X \leq_{dil},$$

from which it immediately follows that also $\mathrm{Var}X \leq \mathrm{Var}Y$.

The quantile spread order. All variability orders considered so far are strong enough to imply a corresponding ordering of the variances. Moreover, distributions with the same finite support are not related in terms of the dispersive order unless they are identical. Hence a weaker concept is desirable. In fact, Townsend and Colonius (2005) developed a weaker concept of variability order in an attempt to describe the effect of sample size on the shape of the distribution of the order statistics $X_{1:n}$ and $X_{n:n}$. It is based on the notion of *quantile spread*:

Definition 1.103 The *quantile spread* of random variable X with distribution F is

$$QS_X(p) = F^{-1}(p) - F^{-1}(1-p),$$

for $0.5 < p < 1$.

With $S = 1 - F$ (the survival function), $F^{-1}(p) = S^{-1}(1-p)$ implies

$$QS_X(p) = S^{-1}(1-p) - S^{-1}(p).$$

[20] Note that a logconcave density follows from its distribution function G being *strongly unimodal*, i.e., if the convolution $G * F$ is unimodal for every unimodal F.

The quantile spread of a distribution[21] describes how probability mass is placed symmetrically about its median and hence can be used to formalize concepts such as peakedness and tailweight traditionally associated with kurtosis. This way it allows to separate concepts of kurtosis and peakedness for asymmetric distributions.

For the extreme order statistics (cf. Equation 1.22) one gets simple expressions for the quantile spread, making it easy to describe their behavior as a function of n:

$$QS_{min}(p) = S^{-1}[(1-p)^{1/n}] - S^{-1}(p^{1/n})$$

and

$$QS_{max}(p) = F^{-1}(p^{1/n}) - F^{-1}[(1-p)^{1/n}].$$

Definition 1.104 Let X and Y be random variables with quantile spreads QS_X and QS_Y, respectively. Then X is called *smaller than Y in quantile spread order*, denoted as $X \leq_{QS} Y$, if

$$QS_X(p) \leq QS_Y(p), \quad \text{for all } p \in (0.5, 1).$$

The following properties of the quantile spread order are easily collected:

1. The order $\leq_{QS}$ is *location-free*, i.e.,

$$X \leq_{QS} Y \text{ iff } X + c \leq_{QS} Y, \text{ for any real } c$$

2. The order $\leq_{QS}$ is *dilative*, i.e, $X \leq_{QS} aX$, whenever $a \geq 1$.
3.

$$X \leq_{QS} Y \text{ if, and only if } -X \leq_{QS} -Y.$$

4. Assume F_X and F_Y are symmetric, then

$$X \leq_{QS} Y \text{ if, and only if, } F_X^{-1}(p) \leq F_Y^{-1}(p) \text{ for } p \in (0.5, 1).$$

5. $\leq_{QS}$ implies ordering of the *mean absolute deviation around the median*, MAD,

$$\text{MAD}(X) = \text{E}[|X - \text{M}(X)|]$$

(M the median), i.e.,

$$X \leq_{QS} Y \quad \text{implies} \quad \text{MAD}(X) \leq \text{MAD}(Y).$$

[21] It turns out that the concept of *spread function* defined by Balanda and MacGillivray (1990) is equivalent to the quantile spread.

The last point follows from writing

$$\mathrm{MAD}(X) = \int_{0.5}^{1} [F^{-1}(p) - F^{-1}(1-p)]\, dp.$$

Example 1.105 (Cauchy distribution) The Cauchy distribution with density

$$f(x) = \frac{\gamma}{\pi(x^2+\gamma^2)}, \;\; \gamma > 0,$$

has no finite variance; the quantile function is

$$F^{-1}(p) = \gamma \tan[\pi(p-0.5)],$$

yielding quantile spread

$$QS(p) = \gamma\{\tan[\pi(p-0.5)] - \tan[\pi(0.5-p)]\}.$$

Thus, parameter γ clearly defines the QS order for the Cauchy distribution.

The next example demonstrates how QS order can be used to describe the effect of sample size on the minimum order statistic.

Example 1.106 The quantile spread for the Weibull minimum (cf. Example 1.18) is

$$QS_{min}(p;n) = (n\lambda)^{-1/\alpha}\left[\left(\ln\frac{1}{1-p}\right)^{1/\alpha} - \left(\ln\frac{1}{p}\right)^{1/\alpha}\right].$$

For $p \in (0.5, 1)$, this is easily seen to decrease in n. Thus,

$$X_{1:n} \leq_{QS} X_{1:n-1},$$

i.e., the quantile spread for the Weibull minimum decreases as a function of sample size n.

The quantile spread order has been used by Mitnik and Baek (2013) in ordering the *Kumaraswamy distribution*, an alternative to the Beta distribution, which possesses the advantage of having an invertible closed form (cumulative) distribution function.

Multivariate stochastic orders

Various extensions of both univariate stochastic orders and univariate variability orders to the multivariate case exist. We only sketch two multivariate versions of the usual stochastic order $\leq_d$ here.

For real vectors $\mathbf{x} = (x_1, \ldots, x_n)$, $\mathbf{y} = (y_1, \ldots, y_n)$ we define the coordinatewise ordering $\mathbf{x} \leq \mathbf{y}$ by $x_i \leq y_i$, $i = 1, \ldots, n$. A real-valued function $f : \Re^n \longrightarrow \Re$ is *nondecreasing* if $\mathbf{x} \leq \mathbf{y}$ implies $f(\mathbf{x}) \leq f(\mathbf{y})$ for all $\mathbf{x}, \mathbf{y} \in \Re^n$.

The following definition is a straightforward generalization of the condition stated in Proposition 1.94:

Definition 1.107 Random vectors $\mathbf{X} = (X_1, \ldots, X_n)$ and $\mathbf{Y} = (Y_1, \ldots, Y_n)$ are *strongly stochastically ordered*, denoted as $\mathbf{X} \leq_{st} \mathbf{Y}$, if

$$\mathrm{E}[g(\mathbf{X})] \leq \mathrm{E}[g(\mathbf{Y})]$$

for all nondecreasing functions g for which the expectation exists.[22]

A multivariate version of Strassen's theorem (Theorem 1.70) holds:

Theorem 1.108 *The random vectors* $\mathbf{X}$ *and* $\mathbf{Y}$ *satisfy* $\mathbf{X} \leq_{st} \mathbf{Y}$ *if, and only if, there exist two random vectors* $\hat{\mathbf{X}}$ *and* $\hat{\mathbf{Y}}$*, defined on the same probability space, such that*

$$\hat{\mathbf{X}} = \mathbf{X}, \;\; \hat{\mathbf{Y}} = \mathbf{Y}, \quad \textit{and} \quad P(\hat{\mathbf{X}} \leq \hat{\mathbf{Y}}) = 1.$$

The actual construction of $\mathbf{X}$ and $\mathbf{Y}$ is achieved by an iterative coupling argument (see e.g. Szekli, 1995, pp. 51-52). Moreover, there exists a straightforward generalization of Proposition 1.95 also allowing the implication $\mathbf{X} =_{st} \mathbf{Y}$.

Another, straightforward generalization of the univariate stochastic order $\leq_d$ is based on the multivariate distribution and survival functions. For a random vector $\mathbf{X} = (X_1, \ldots, X_n)$ with distribution function F, let $\bar{F}$ be the multivariate survival function of $\mathbf{X}$, that is,

$$\bar{F}(x_1, \ldots, x_n) \equiv P(X_1 > x_1, \ldots, X_n > x_n).$$

for all $\mathbf{x} = (x_1, \ldots x_n)$. Let $\mathbf{Y}$ be another n-dimensional random vector with distribution function G and survival function $\bar{G}$.

Definition 1.109 $\mathbf{X}$ is said to be *smaller than* $\mathbf{Y}$ *in the upper orthant order*, denoted as $\mathbf{X} \leq_{uo} \mathbf{Y}$, if

$$\bar{F}(x_1, \ldots, x_n) \leq \bar{G}(x_1, \ldots, x_n), \quad \textit{for all} \quad \mathbf{x};$$

moreover, $\mathbf{X}$ is said to be *smaller than* $\mathbf{Y}$ *in the lower orthant order*, denoted as $\mathbf{X} \leq_{lo} \mathbf{Y}$, if

$$F(x_1, \ldots, x_n) \leq G(x_1, \ldots, x_n) \quad \textit{for all} \quad \mathbf{x}.$$

The following characterization of $\leq_{uo}$ and $\leq_{lo}$ demonstrates that these orders are actually, in general, weaker than $\leq_{st}$:

[22] Note that, for $n = 1$, the orders $\leq_d$ and $\leq_{st}$ are the same.

Proposition 1.110 *Let* $\mathbf{X}$ *and* $\mathbf{Y}$ *be two* n*-dimensional random vectors. Then* $\mathbf{X} \leq_{uo} \mathbf{Y}$ *if, and only if,*

$$\mathrm{E}[\psi(\mathbf{X})] \leq \mathrm{E}[\psi(\mathbf{Y})] \quad \textit{for every distribution function } \psi;$$

moreover, $\mathbf{X} \leq_{lo} \mathbf{Y}$ *if, and only if,*

$$\mathrm{E}[\psi(\mathbf{X})] \leq \mathrm{E}[\psi(\mathbf{Y})] \quad \textit{for every survival function } \psi.$$

Clearly, we have the implication

$$\mathbf{X} \leq_{st} \mathbf{Y} \Longrightarrow (\mathbf{X} \leq_{uo} \mathbf{Y} \text{ and } \mathbf{X} \leq_{lo} \mathbf{Y}).$$

Finally, it should be mentioned that multivariate versions of the hazard rate order $\leq_{hr}$ and of the likelihood ratio order $\leq_{lr}$ have been developed, as well as generalizations of the variability orders $\leq_{cx}$ and $\leq_{disp}$.

Positive dependence orders

Whenever a concept of dependence has been defined, one can compare a given distribution with one obtained under stochastic independence. More generally, it may be interesting to compare two distributions with respect to their strength of dependence. However, comparing two distributions seems meaningful only if their marginals are the same. Therefore, comparisons will always be performed within a given *Fréchet* distribution class.

From Section 1.3.6, we know that, in the bivariate case, the upper and lower Fréchet bounds F^+ and F^- represent distributions with perfect positive and negative dependence, respectively, providing natural upper and lower bounds in the class $\mathcal{F}(F_1, F_1)$. The most basic dependence order is based on the concept of *positive quadrant dependence* (PQD),

$$P(X_1 \leq x_1, X_2 \leq x_2) \geq P(X_1 \leq x_1)P(X_2 \leq x_2),$$

for all x_1, x_2. PQD is the $n = 2$ version of the multivariate $PLOD$ order introduced in Section 1.3.8.

Definition 1.111 Let random vector (X_1, X_2) with distribution function F and random vector (Y_1, Y_2) with distribution function G be members of the Fréchet class $\mathcal{F}(F_1, F_1)$. Then, (X_1, X_2) is said to be *smaller than* (Y_1, Y_2) *in the PQD order*, denoted by $(X_1, X_2) \leq_{PQD} (Y_1, Y_2)$, if

$$F(x_1, x_2) \leq G(x_1, x_2), \quad \text{for all } x_1, x_2.$$

Sometimes it will be useful to write this as $F \leq_{PQD} G$. From Hoeffding's identity (Proposition 1.36), it follows readily that, if $(X_1, X_2) \leq_{PQD} (Y_1, Y_2)$, then

$$\mathrm{Cov}(X_1, X_2) \leq \mathrm{Cov}(Y_1, Y_2),$$

and, since $\mathrm{Var}X_i = \mathrm{Var}Y_i, i = 1, 2$, we have

$$\mathrm{Cor}(X_1, X_2) \leq \mathrm{Cor}(Y_1, Y_2).$$

It can also be shown that Kendall's tau and Spearman's rho are preserved under the PQD order. Moreover, for every distribution $F \in \mathcal{F}(F_1, F_1)$, we have

$$F^- \leq_{PQD} F \leq_{PQD} F^+.$$

For random vectors (X_1, X_2) and (Y_1, Y_2) with distribution functions in $\mathcal{F}(F_1, F_1)$, we have

$$(X_1, X_2) \leq_{PQD} (Y_1, Y_2) \Longleftrightarrow (X_1, X_2) \leq_{uo} (Y_1, Y_2)$$

and

$$(X_1, X_2) \leq_{PQD} (Y_1, Y_2) \Longleftrightarrow (X_1, X_2) \geq_{lo} (Y_1, Y_2);$$

however, for the right-hand order relations it is not required that (X_1, X_2) and (Y_1, Y_2) have the same marginals. Therefore, whereas the upper and lower orthant orders measure the size (or location) of the underlying random vectors, the PQD order measures the degree of positive dependence of the underlying random vectors.

We conclude with an example of PQD relating to Archimedean copulas (cf. Section 1.3.7).

Example 1.112 (Genest and MacKay, 1986) Let ϕ and ψ be two Laplace transforms of nonnegative random variables. Then F and G, defined by

$$F(x_1, x_2) = \phi(\phi^{-1}(x_1) + \phi^{-1}(x_2)), \quad (x_1, x_2) \in [0, 1]^2,$$

and

$$G(y_1, y_2) = \psi(\psi^{-1}(y_1) + \psi^{-1}(y_2)), \quad (y_1, y_2) \in [0, 1]^2,$$

are bivariate distribution functions with standard uniform marginals, i.e., F and G are Archimedean copulas. Then it can be shown that $(X_1, X_2) \leq_{PQD}$ (Y_1, Y_2) if, and only if $\psi^{-1}\phi$ is *superadditive*, i.e.,

$$\psi^{-1}\phi(x + y) \geq \psi^{-1}\phi(x) + \psi^{-1}\phi(y),$$

for all $x, y \geq 0$.

Several other positive dependence orders, derived from multivariate dependence concepts, like association, have been introduced in various contexts.

1.4 Bibliographic references

1.4.1 Monographs

The material in Section 1.2 is presented at the level of Durrett (2010) or Gut (2013); some of the measure-theoretic concepts are found in Leadbetter et al. (2014) and Pollard (2002); Severini (2005) develops distribution theory without measure theory. An extended discussion of the bivariate exponential distribution (Example 1.28) is found in Galambos and Kotz (1978). Aldous (1985) is a standard reference on exchangeability; see also the review of Kallenberg (2005) by Diaconis (2009); more recent presentations from a Bayesian viewpoint are Goldstein (2013) and Dawid (2013) (see also O'Hagen and Forster, 2004, pp. 108–114). A modern treatment of quantile functions is the quantile-based presentation of reliability analysis in Nair et al. (2013). A comprehensive discussion of multivariate survival analysis, including competing risks theory, is given in Crowder (2012). The order statistics section follows the presentation in Arnold et al. (1992) and, for the extreme value statistics and asymptotic results, Galambos (1978) was consulted. The introduction to the theory of records follows Arnold et al. (1998). The basics of coupling theory are found in the first chapter of Thorisson (2000); see also Lindvall (2002) and Levin et al. (2008). The results on Fréchet distribution classes are from Joe (1997) (see also Denuit et al., 2005). A good introduction to copula theory is Nelsen (2006), recent results are collected in Joe (2015); see Mai and Scherer (2012) for simulation issues and Rüschendorf (2013) for copulas and risk analysis. Concepts of dependence are treated in Mari and Kotz (2004), Joe (1997, 2015), and Barlow and Proschan (1975). The standard reference for stochastic orders is Shaked and Shantikumar (2007) and also Müller and Stoyan (2002) and Szekli (1995); for a compilation of papers on recent order results, see Li and Li (2013).

1.4.2 Selected applications in mathematical psychology

This final section presents a list, clearly not representative and far from being exhaustive, of mathematical psychology applications relating to the probabilistic concepts treated in this chapter.

An application of the concept of exchangeability is found in the study by Diaconis and Freedman (1981) disproving a conjecture of Bela Julesz concerning the discriminability of a class of random patterns. Quantile function, survival function, and hazard rate have become standard tools in RT analysis (e.g., Schwarz and Miller, 2012; Townsend and Eidels, 2011; Houpt and Townsend, 2012; Colonius, 1988). Chechile (2011) characterizes properties

of the "reverse hazard rate". The non-identifiability results in competing risks theory have found an interpretation in random utility theory (Marley and Colonius, 1992) and RT modeling (Townsend, 1976; Dzhafarov, 1993; Jones and Dzhafarov, 2014). Order statistics and (generalized) Poisson processes are essential parts of the RT models in Smith and Van Zandt (2000), Miller and Ulrich (2003), and Schwarz (2003). Issues of generalizing extreme-value statistics results have been addressed in Cousineau et al. (2002). The theory of coupling seems not to have played a role in mathematical psychology until the work by Dzhafarov and Kujala (Dzhafarov and Kujala, 2013), although the requirement of a coupling assumption had already been recognized explicitly in earlier work (cf. Ashby and Townsend, 1986; Luce, 1986; Colonius and Vorberg, 1994). Fréchet-Hoeffding bounds have been introduced in the context of audiovisual interaction models of RT (see Colonius, 1990; Diederich, 1992) and were further discussed in Townsend and Nozawa (1995), Townsend and Wenger (2004), and Colonius (2015). Copula theory has been applied in psychometric modeling (see Braeken et al., 2013, and further references therein). The multivariate dependence concept of association was applied in a random utility context by Colonius (1983). Several of the stochastic order concepts presented here have been discussed in Townsend (1990). Stochastic orders related to the *delta plot method* have been characterized in Speckman et al. (2008) (see also Schwarz and Miller, 2012).

1.5 Acknowledgment

I am grateful to Maike Tahden for pointing out numerous typos and imprecisions in the original draft of the chapter.

References

Aalen, O.O., Borgan, Ø., and Gjessing, H.K. 2008. *Survival and event history analysis*. New York, NY: Springer Verlag.

Aldous, D.J. 1985. *École d'Éte de Probabilités de Saint-Flour XIII-1983. Lecture Notes in Mathematics*. Vol. 1117. Berlin: Springer. Chap. Exchangeability and related topics, pages 1–198.

Alsina, C., and Schweizer, B. 1988. Mixtures are not derivable. *Foundations of Physics Letters*, **1**, 171–174.

Alsina, C., Nelson, R.B., and Schweizer, B. 1993. On the characterization of class of binary operations on distribution functions. *Statist. Probab. Lett.*, **17**, 85–89.

Arnold, B.C., Balakrishnan, N., and Nagaraja, H.N. 1992. *A first course in order statistics*. New York, NY: John Wiley & Sons.

Arnold, B.C., Balakrishnan, N., and Nagaraja, H.N. 1998. *Records*. New York, NY: John Wiley & Sons.

Ashby, F.G., and Townsend, J.T. 1986. Varities of perceptual independence. *Psychological Review*, **93**, 154–179.

Balanda, K.P., and MacGillivray, H.L. 1990. Kurtosis and spread. *The Canadian Journal of Statistics*, **18**(1), 17–30.

Barlow, R.E., and Proschan, F. 1975. *Statistical theory of reliability and life testing*. New York, NY: Holt, Rinehart and Winston.

Braeken, J., Kuppens, P., De Boeck, P., and Tuerlincks, F. 2013. Contextualized personality questionnaires: a case for copulas in structural equation models for categorical data. *Multivariate Behavioral Research*, **48**, 845–870.

Chechile, R.A. 2011. Properties of reverse hazard function. *Journal of Mathematical Psychology*, **55**, 203–222.

Colonius, H. 1983. A characterization of stochastic independence, with an application to random utility theory. *Journal of Mathematical Psychology*, **27**, 103–106.

Colonius, H. 1988. Modeling the redundant signals effect by specifying the hazard function. *Perception & Psychophysics*, **43**, 605–607.

Colonius, H. 1990. Possibly dependent probability summation of reaction time. *Journal of Mathematical Psychology*, **34**, 253–275.

Colonius, H. 1995. The instance theory of automaticity: Why the Weibull? *Psychological Review*, **102**(4), 744–750.

Colonius, H., and Diederich, A. 2006. The race model inequality: interpreting a

geometric measure of the amount of violation. *Psychological Review*, **113**(1), 148–154.

Colonius, H., and Vorberg, D. 1994. Distribution inequalities for parallel models with unlimited capacity. *Journal of Mathematical Psychology*, **38**, 35–58.

Colonius, Hans. 2015. Behavioral measures of multisensory integration: bounds on bimodal detection probability. *Brain Topography*, **28**(1), 1–4.

Cousineau, D., Goodman, V.W., and Shiffrin, R.M. 2002. Extending statistics of extremes to distributions varying on position and scale, and implications for race models. *Journal of Mathematical Psychology*, **46**(4), 431–454.

Crowder, M. 2012. *Multivariate survival analysis and competing risks*. Boca Raton, FL: CRC Press.

Dawid, A.P.. 2013. Exchangeabiltiy and its ramifications. Pages 19–29 of: Damien, P., Dellaportas, P., Polson, N.G., and Stephens, D.A. (eds), *Bayesian theory and applications*. Oxford, UK: Oxford University Press.

Denuit, M., Dhaene, J., Goovaerts, M., and Kaas, R. 2005. *Actuarial theory for dependent risks*. John Wiley and Sons, Ltd.

Desu, M.M. 1971. A characterization of the exponential distribution by order statistics. *Annals of Mathematical Statistics*, **42**, 837–838.

Diaconis, P. 2009. Review of "Probabilistic symmetries and invariance principles", by Olav Kallenberg. *Bulletin of the American Mathematical Society*, **46**(4), 691–696.

Diaconis, P., and Freedman, D. 1981. On the statistics of vision: the Julesz conjecture. *Journal of Mathematical Psychology*, **124**(2), 112–138.

Diederich, A. 1992. Probability inequalities for testing separate activation models of divided attention. *Perception & Psychophysics*, **52**(6), 714–716.

Diederich, A., and Colonius, H. 1987. Intersensory facilitation in the motor component? *Psychological Research*, **49**, 23–29.

Donders, F.C. 1868/1969. Over de snelheid van psychische processen [On the speed of mental processes]. *Onderzoekingen gedaan in het Physiologisch Laboratorium der Utrechtsche Hoogeschool, 1868-1869, Tweede reeks, II, 92-120. Transl. by Koster, W. G. (1969). In Attention and performance II, Koster, W. G., Ed., Acta Psychologica*, **30**, 412–431.

Durrett, R. 2010. *Probability: Theory and Examples*. 4rd edn. Cambridge, UK: Cambridge University Press.

Dzhafarov, E. N. 1993. Grice-representability of response time distribution families. *Psychometrika*, **58**(2), 281–314.

Dzhafarov, E. N., and Kujala, J. V. 2013. All-possible-couplings approach to measuring probabilistic context. *PLoS ONE*, **8**(5), e61712.

Faugeras, O.P. 2012. Probabilistic constructions of discrete copulas. *Unpublished manuscript*, hal–00751393.

Fisher, R. A., and Tippett, L. H. C. 1928. Limiting forms of the frequency distribution of the largest or smallest member of a sample. *Proc. Cambridge Philos. Soc.*, **24**, 180–190.

Frechét, M. 1927. Sur la loi de probabilité de l'écart maximum. *Ann. Soc. Polonaise Math.*, **6**, 92–116.

Galambos, J. 1978. *The asymptotic theory of extreme order statistics*. New York, NY: John Wiley & Sons.

Galambos, J., and Kotz, S. 1978. *Characterizations of probability distributions*. Lecture Notes in Mathematics No. 675. New York, NY: Springer.

Genest, C., and MacKay, J. 1986. Copules archimédiennes et familles de loi bidimensionnelles dont les marges sont données. *The Canadian Journal of Statistics*, **14**(2), 145–159.

Gnedenko, B. 1943. Sur la distribution limite du terme maximum d'une série aléatoire. *Ann. Math.*, **44**, 423–453.

Goldstein, M. 2013. Observables and models: exchangeability and the inductive argument. Pages 3–18 of: Damien, P., Dellaportas, P., Polson, N.G., and Stephens, D.A. (eds), *Bayesian theory and applications*. Oxford, UK: Oxford University Press.

Gut, A. 2013. *Probability: A graduate course*. Second edn. New York, NY: Springer.

Hershenson, M. 1962. Reaction time as a measure of intersensory facilitation. *Journal of Experimental Psychology*, **63**, 289–293.

Hoeffding, W. 1940. Maßstabinvariante Korrelationstheorie. *Schriften des Mathematischen Instituts und des Instituts für Angewandte Mathematik der Universität Berlin*, **5**, 181–233.

Hohle, R.H. 1965. Inferred components of reaction times as a function of foreperiod duration. *Journal of Experimental Psychology*, **69**, 382–386.

Houpt, J.W., and Townsend, J.T. 2012. Statistical Measures for Workload Capacity Analysis. *Journal of Mathematical Psychology*, **56**, 341–355.

Joag-Dev, K., and Proschan, F. 1983. Negative association of random variables with applications. *Annals of Statistics*, **11**, 286–295.

Joe, H. 1997. *Multivariate models and dependence concepts*. London, UK: Chapman & Hall.

Joe, H. 2015. *Dependence modeling with copulas*. Boca Raton, FL: CRC Press, Taylor & Francis.

Jones, M., and Dzhafarov, E. N. 2014. Unfalsifiability and mutual translatability of major modeling schemes for choice reaction time. *Psychological Review*, **121**, 1–32.

Kalbfleisch, J.D., and Prentice, R.L. 2002. *The statistical analysis of failure time data*. 2nd edition edn. Hoboken, NJ: Wiley-Interscience.

Kallenberg, O. 2005. *Probabilistic symmetries and invariance principles*. New York, NY: Springer-Verlag.

Kochar, S.C., and Proschan, F. 1991. Independence of time and cause of failure in the multiple dependent competing risks model. *Statistica Sinica*, **1**(1), 295–299.

Kruskal, W.H. 1958. Ordinal measures of association. *Journal of the American Statistical Association*, **53**, 814–861.

Leadbetter, R., Cambanis, S., and Pipiras, V. 2014. *A basic course in measure and probability*. New York, NY: Cambridge University Press.

Levin, D. A., Peres, Y., and Wilmer, E. L. 2008. *Markov chains and mixing times*. Providence, R. I.: American Mathematical Society.

Li, H., and Li, X. 2013. *Stochastic orders in reliability and risk*. Lecture Notes in Statistics, vol. 208. New York, NY: Springer-Verlag.

Lindvall, T. 2002. *Lectures on the coupling method*. Reprint of wiley 1992 edn. Mineola, NY: Dover.

Logan, G. D. 1988. Toward an instance theory of automatization. *Psychological Review*, **95**, 492–527.

Logan, G. D. 1992. Shapes of reaction-time distributions and shapes of learning curves: A test of the instance theory of automaticity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **18**, 883–914.

Logan, G. D. 1995. The Weibull distribution, the power law, and the instance theory of automaticity. *Psychological Review*, **102**(4), 751–756.

Luce, R.D. 1986. *Response times: Their role in inferring elementary mental organization*. New York, NY: Oxford University Press.

Luce, R.D. 1997. Several unresolved conceptual problems of mathematical psychology. *Journal of Mathematical Psychology*, **41**(1), 79–87.

Mai, J.F., and Scherer, M. 2012. *Simulating copulas: Stochastic Models, Sampling Algorithms and Applications*. Series in Quantitative Finance, no. Vol. 4. Imperial College Press.

Mari, D.D., and Kotz, S. 2004. *Correlation and dependence*. Reprint of 2001 edn. London, UK: Imperial College Press.

Marley, A.A.J., and Colonius, H. 1992. The "horse race" random utility model for choice probabilities and reaction times, and its competing risks interpretation. *Journal of Mathematical Psychology*, **36**, 1–20.

Marshall, A.W., and Olkin, I. 1967. A generalized bivariate exponential destribution. *Journal of Applied Probability*, **4**, 291–302.

Marshall, A.W., and Olkin, I. 2007. *Life distributions*. New York, NY: Springer-Verlag.

Miller, J., and Ulrich, R. 2003. Simple reaction time and statistical facilitation: a parallel grains model. *Cognitive Psychology*, **46**, 101–151.

Miller, J. O. 1982. Divided attention: Evidence for coactivation with redundant signals. *Cognitive Psychology*, **14**, 247—279.

Mitnik, P.A., and Baek, S. 2013. The Kumaraswamy distribution: median-dispersion re-parameterizations for regression modeling and simulation-based estimation. *Statistical Papers*, **54**, 177–192.

Müller, A., and Rüschendorf, L. 2001. On the optimal stopping values induced by general dependence structures. *Journal of Applied Probability*, **38**(3), 672—684.

Müller, A., and Stoyan, D. 2002. *Comparison methods for stochastic models and risks*. Vol. 389. Wiley.

Nair, N.U., Sankaran, P.G., and Balakrishnan, N. 2013. *Quantile-based reliability analysis*. New York, NY: Springer Verlag.

Nelsen, R. B. 2006. *An introduction to copulas*. 2nd edn. Lecture Notes in Statistics, vol. 139. New York, NY: Springer-Verlag.

Newell, A., and Rosenbloom, P. S. 1981. Mechanisms of skill acquisition and the law of practice. Pages 1–55 of: Anderson, J. R. (ed), *Cognitive skills and their acquisition*. Erlbaum.

O'Hagen, A:, and Forster, J. 2004. *Bayesian inference*. 2nd edn. Kendall's Advanced Theory of Statistics Vol. 2B. London, UK: Arnold.

Palmer, E.M., Horowitz, T.S., Torralba, A., and Wolfe, J.M. 2011. What are the shapes of response time distributions in visual search? *Journal of Experimental Psychology: Human Perception and Performance*, **37**(1), 58–71.

Peterson, A.V. 1976. Bounds for a joint distribution function with fixed subdistribution functions: application to competing risks. *Proceedings of the National Academy of Sciences USA*, **73**, 11–13.

Pfeifer, D., and Nešlehová, J. 2004. Modeling and generating dependent risk processes for IRM and DFA. *ASTIN Bulletin*, **34**(2), 333–360.

Pollard, D. 2002. *A user's guide to measure theoretic probability*. New York, NY: Cambridge University Press.

Raab, D.H. 1962. Statistical facilitation of simple reaction time. *Transactions of the New York Academy of Sciences*, **24**, 574–590.

Ross, S.M. 1983. *Stochastic processes.* New York, NY: John Wiley & Sons.

Ross, S.M. 1996. *Stochastic processes.* Second edn. New York, NY: John Wiley & Sons.

Rüschendorf, L. 2013. *Mathematical risk analysis.* New York, NY: Springer-Verlag.

Schmitz, V. 2004. Revealing the dependence structure between $X_{(1)}$ and $X_{(n)}$. *Journal of Statistical Planning and Inference*, **123**, 41–47.

Schwarz, W. 2001. The ex-Wald distribution as a descriptive model of response times. *Behavior Research Methods, Instruments, & Computers*, **33**(4), 457–469.

Schwarz, W. 2002. On the convolution of inverse Gaussian and exponential random variables. *Communications in Statistics, Theory and Methods*, **31**(12), 2113–2121.

Schwarz, W. 2003. Stochastic cascade processes as a model of multi-stage concurrent information processing. *Acta Psychologica*, **113**, 231–261.

Schwarz, W., and Miller, J. 2012. Response time models of delta plots with negative-going slopes. *Psychomic Bulletin & Review*, **19**, 555–574.

Schweickert, R., Fisher, D.L., and Sung, K. 2012. *Discovering cognitive architecture by selectively influencing mental processes.* Advanced Series on Mathematical Psychology, vol. 4. Singapore: World Scientific.

Schweizer, B., and Sklar, A. 1974. Operations of distribution functions not derivable from operations on random variables. *Studia Math.*, **52**, 43–52.

Severini, T.A. 2005. *Elements of distribution theory.* New York, NY: Cambridge University Press.

Shaked, M., and Shantikumar, J.G. 2007. *Stochastic orders.* New York, NY: Springer Science+Business Media.

Shea, G.A. 1983. Hoeffding's lemma. Pages 648–649 of: Kotz, S., Johnson, N.L., and Read, C.B. (eds), *Encyclopedia of Statistical Science*, vol. 3. New York, NY: Wiley.

Smith, P.L., and Van Zandt, T. 2000. Time-dependent Poisson counter models of response latency in simple judgment. *British Journal of Mathematical and Statistical Psychology*, **53**, 293–315.

Speckman, P.L., Rouder, J.N., Morey, R.D., and Pratte, M.S. 2008. Delta plots and coherent distribution ordering. *The American Statistician*, **62**(3), 262–266.

Strassen, V. 1965. The exisistence of probability measures with given marginals. *Annals of Mathematical Statistics*, **36**, 423–439.

Szekli, R. 1995. *Stochastic ordering and dependence in applied probability.* New York, NY: Springer Verlag.

Thorisson, H. 2000. *Coupling, stationarity, and regeneration.* New York, NY: Springer Verlag.

Townsend, J.T. 1976. Serial and within-stage independent parallel model equivalence on the minimum completion time. *Journal of Mathematical Psychology*, **14**, 219–238.

Townsend, J.T. 1990. Truth and consequences of ordinal differences in statistical distributions: toward a theory of hierarchical inference. *Journal of Mathematical Psychology*, **108**(551–567).

Townsend, J.T., and Ashby, F.G. 1983. *The stochastic modeling of elementary psychological processes.* Cambridge University Press.

Townsend, J.T., and Colonius, H. 2005. Variability of the max and min statistic: a theory of the quantile spread as a function of sample size. *Psychometrika*, **70**(4), 759–772.

Townsend, J.T., and Eidels, A. 2011. Workload capacity spaces: A unified methodology for response time measures of efficiency as workload is varied. *Psychonomic Bulletin and Review*, **18**, 659–681.

Townsend, J.T., and Nozawa, G. 1995. Spatio-temporal properties of elementary perception: an investigation of parallel, serial, and coactive theories. *Journal of Mathematical Psychology*, **39**(4), 321–359.

Townsend, J.T., and Wenger, M.J. 2004. A theory of interactive parallel processing: new capacity measures and predictions for a response time inequality series. *Psychological Review*, **111**(4), 1003–1035.

Tsiatis, A.A. 1975. A nonidentifiability aspect of the problem of competing risks. *Proceedings of the National Academy of Sciences USA*, **72**, 20–22.

Ulrich, R., and Miller, J. 1994. Effects of truncation on reaction time analysis. *Journal of Experimental Psychology: General*, **123**(1), 34–80.

Van Zandt, T. 2000. How to fit a response time distribution. *Psychonomic Bulletin and Review*, **7**, 424–465.

Van Zandt, T. 2002. *Stevens' Handbook of Experimental Psychology*. Vol. 4. New York, NY: Wiley. Chap. Analysis of response time distributions, pages 461–516.

von Mises, R. 1936. La distribution de la plus grande de n valeurs. *Rev. Math. Union Interbalcanique*, **1**, 141–160. Reproduced in Selected Papers of Richard von Mises, *Am. Math. Soc. II*, (1964), 271–294.

Index