1 Selected concepts from probability

Hans Colonius

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.

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 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

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

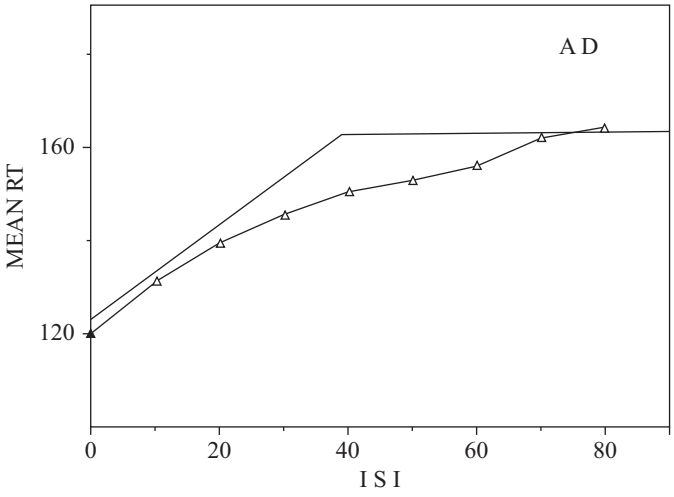

Figure 1.1 *Bimodal (mean) reaction time to light and sound with interstimulus interval (ISI) and sound following light, $RT_V = 160$ ms, $RT_A = 120$ ms. Upper graph: prediction in absence of interaction, lower graph: observed mean RTs; data from Diederich and Colonius (1987).*

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 amount of facilitation, and the discussion of how the effect should be modeled is ongoing, 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. The 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

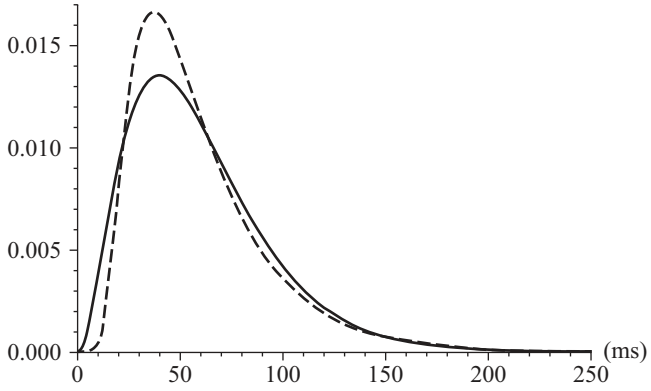

Figure 1.2 *Inverse gaussian (dashed line) and gamma densities with identical mean (60 ms) and standard deviation (35 ms).*

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 notoriously 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.

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.

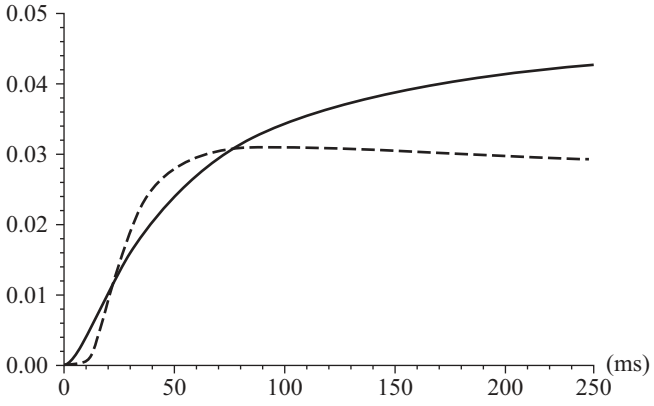

Figure 1.3 *Hazard functions of the inverse gaussian (dashed line) and gamma distributions corresponding to the densities of Figure 1.2.*

We hope to convey in this chapter that even seemingly simple situations, 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 has not yet found the role that it arguably deserves.

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

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.

probability space for previously unrelated random variables (or, more general random entities). Although it is 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 Kujala (see also Chapter 2 by Dzhafarov and Kujala in this volume).

While coupling strives to construct a joint probability space, the 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]

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 2 for a more general approach).

Let Ω be an arbitrary set, often referred to as the *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. Because 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 1.1 (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 $\{A_n, n \geq 1\}$,

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

Then P is called the *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. $A_1 \subset A_2 \rightarrow P(A_1) \leq P(A_2)$.

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

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

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.

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 1.3 (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 1.1 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 1.2 (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

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

Definition 1.4 (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

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

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, $P = Q$.

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, $\Re \cup \{+\infty\} \cup \{-\infty\}$.

Definition 1.5 (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:

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

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

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

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 1.3 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).

However, sometimes it is important to realize that two random variables 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, for example, it is meaningless to ask whether two events are independent (see Section 1.3.5).

"The equality" of random variables can be interpreted in different ways.

Remark 1.4 Random variables X and Y are *equal in distribution* iff they are governed by the same probability measure:

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

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

Example 1.4 (Gut, 2013, p. 27) Toss a fair coin once and set

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

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