to appear in
The New Handbook of Mathematical Psychology,
to be published by Cambridge University Press.
This version may differ from the published chapter.

Contents

1

Probability, Random Variables, and Selectivity

Ehtibar Dzhafarov[a] and Janne Kujala[b]

1.1 What is it about?

This chapter is about systems with several random outputs whose joint distribution depends on several inputs. More specifically, it is about selectiveness in the dependence of random outputs on inputs. That is, we are concerned with the question of which of the several outputs are influenced by which of the several inputs. A system can be anything: a person, animal, group of people, neural network, technical gadget, two entangled electrons running away from each other. Outputs are responses of the system or outcomes of measurements performed on it. Inputs are entities upon whose values the outputs of the system are conditioned. Even if inputs are random variable in their own right, the outputs are being conditioned upon every specific stimulus. Inputs therefore are always deterministic (not random) entities insofar as their relationship to random outputs is concerned.

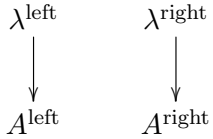
Example 1.1 In a double-detection experiment, the stimulus presented in each trial may consist of two flashes, say, right one and left one, separated by some distance in visual field. Suppose that each flash can have one of two contrast levels, one zero and one (slightly) above zero. These contrasts play the role of two binary inputs, that we can call λ^{left} and λ^{right} (each one with values present/absent). The inputs are used in a completely crossed experimental design: that is, the stimulus in each trial is described by one of four combinations of the two inputs: $\big(\lambda^{\text{left}} = \text{present}, \lambda^{\text{right}} = \text{present}\big)$, $\big(\lambda^{\text{left}} = \text{present}, \lambda^{\text{right}} = \text{absent}\big)$, etc. In response to each such a combination (called a treatment), the participant is asked to say whether the left flash was present (yes/no) and whether the right flash was present (yes/no). These are the two binary outputs, we can denote them A^{left} and A^{right} (each

[a] Purdue University, USA
[b] University of Jyväskylä, Finland

with two possible values, yes/no). The outputs are random variables. Theoretically, they are characterized by joint distributions tied to each of four treatments:

$\left(\lambda^{\text{left}}=i,\lambda^{\text{right}}=j\right)$	$A^{\text{right}}=\text{yes}$	$A^{\text{right}}=\text{no}$
$A^{\text{left}}=\text{yes}$	$p_{\text{yes,yes}}$	$p_{\text{yes,no}}$
$A^{\text{left}}=\text{no}$	$p_{\text{no,yes}}$	$p_{\text{no,no}}$

where i,j stand for "present" or "absent" each. Suppose now that the experimenter hypothesizes that the response to the left stimulus depends only on the contrast of the left stimulus, and the response to the right stimulus depends only on the contrast of the right stimulus,



This hypothesis can be justified, for example, by one's knowledge that the separation between the locations of the flashes is too large to allow for interference, and that subjectively, nothing seems to change in the appearance of the left stimulus as the right one is switched on and off, and vice versa. The meaning of this hypothesis is easy to understand if the two random outputs are known to be stochastically independent, which in this case means that, for every one of the four treatments,

$$p_{\text{yes,yes}}=\Pr\left(A^{\text{left}}=\text{yes},A^{\text{right}}=\text{yes}\right)=\Pr\left(A^{\text{left}}=\text{yes}\right)\Pr\left(A^{\text{right}}=\text{yes}\right).$$

In this case the test of the selectiveness consists in finding out if the distribution of A^{left}, in this case defined by $\Pr\left(A^{\text{left}}=\text{yes}\right)$, remains unchanged as one changes the value of λ^{right} while keeping λ^{left} fixed, and analogously for A^{right}. The experimenter, however, is likely to find out that stochastic independence in such an experiment does not hold: for some, if not all of the four treatments,

$$p_{\text{yes,yes}}\neq\Pr\left(A^{\text{left}}=\text{yes}\right)\Pr\left(A^{\text{right}}=\text{yes}\right).$$

Now the conceptual clarity may be lost. Does the lack of stochastic independence invalidate the hypothesis that the outputs are selectively influenced by the corresponding inputs? Indeed, one might reason that it does, because if A^{left} and A^{right} are not independent, then A^{left} certainly "depends on" A^{right}, whence A^{left} should also depend on anything A^{right} depends on (and this includes λ^{right}). But one might also reason that stochastic relationship

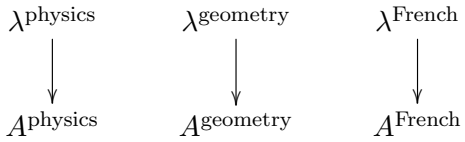
between the two outputs can be ignored altogether. Cannot one declare that the hypothesis in question holds if one establishes that the marginal distributions (i.e., $\Pr\left(A^{\text{left}} = \text{yes}\right)$ and $\Pr\left(A^{\text{right}} = \text{yes}\right)$, taken separately) are invariant with respect to changes in the non-corresponding inputs (here, λ^{right} and λ^{left}, respectively)? We will see in this chapter that stochastic relationship must not be ignored, but that lack of stochastic independence does not by itself rule out selectiveness in the dependence of random outputs on inputs. □

It is easy to generate formally equivalent examples by trivial modifications. For instance, one can replace the two responses of a participant with activity levels of two neurons, determining whether each of them is above or below its background level. The two locations can be replaced with two stimulus features (say, orientation and spatial frequency of a grating pattern) that are hypothesized to selectively trigger the responses from the two neurons.

One can also easily modify any of such examples by increasing the number of inputs and outputs involved, or increasing the number of possible values per input or output. Thus, in the example with double-detection, one can think of several levels of contrast for each of the flashes. Or one can think of responses being multi-level confidence rating instead of the binary yes/no.

Let us consider a few more examples, however, to appreciate the variety in the nature of inputs and outputs falling within the score of our analysis.

Example 1.2 Let a very large group of students have to take three exams, in physics, geometry, and French. Each student prepares for each of the exams, and the preparation times are classified as "short" or "long" by some criteria (which may be different for different exams). The three preparation times serve as the inputs in this example. We denote them by λ^{physics}, $\lambda^{\text{geometry}}$, and λ^{French} (each with possible values short/long). The outputs are scores the students eventually receive: A^{physics}, A^{geometry}, and A^{French} (say, from 0 to 100% each). The hypothesis to be tested is that preparation time for a given subject selectively affects the score in that subject,



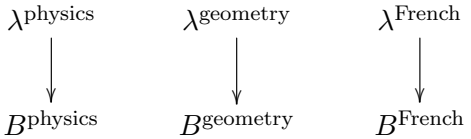
To see if this is the case we subdivide the group of students into eight subgroups, corresponding to the eight combinations of the three preparation

times,

$$\left(\lambda^{\text{physics}} = \text{short/long},\, \lambda^{\text{geometry}} = \text{short/long}\ ,\lambda = \text{short/long}\ \right).$$

Assuming each group is very large, we look at the joint distribution of scores within each of them. The conceptual difficulty here stems from the fact that, for any given treatment, test scores are typically positively correlated rather than stochastically independent. □

Example 1.3 Let us modify the previous example by assigning to each student in each subject a binary grade, "high" or "low," according as the student is, respectively, above or below the median score in this subject received by all student in the same preparation group. Thus, in the preparation group $\left(\lambda^{\text{physics}} = \text{long}\ ,\lambda^{\text{geometry}} = \text{short}\ ,\lambda^{\text{French}} = \text{short}\ \right)$, if the median scores in physics is m, a student gets the grade "high" if her score is above m and "low" if it is not. This defines three outputs that we can call $B^{\text{physics}}, B^{\text{geometry}}, B^{\text{French}}$. The hypothesis represented by the diagram



is more subtle than in the previous example. It says that if one factors out the possible dependence of the median score in physics on all three preparation times (with no selectiveness assumed in this dependence), then whether a student's physics score will or will not fall above the median may only depend on the preparation time for physics, and not on the preparation times for two other subjects. And analogously for geometry and French. Since the grades assigned to students are binary, their theoretical distribution for each of the eight treatments is given by eight joint probabilities

$$\Pr\left(B^{\text{physics}} = \text{high/low},\, B^{\text{geometry}} = \text{high/low},\, B^{\text{French}} = \text{high/low}\right).$$

Again, the conceptual difficulty is in that this probability is not typically equal to 1/8 for all combinations of the high/low values, as it would have to be if the three random variables were independent. Indeed, the marginal (separately taken) probabilities here are, by the definition of median,

$$\Pr\left(B^{\text{physics}} = \text{high}\right) = \Pr\left(B^{\text{geometry}} = \text{high}\right) = \Pr\left(B^{\text{French}} = \text{high}\right) = \frac{1}{2}.$$

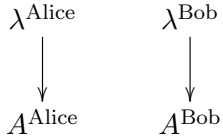
This example also shows why it is not wise to ignore the joint distributions and look at the marginal ones only. If we did this, none of the random outputs $B^{\text{physics}}, B^{\text{geometry}}, B^{\text{French}}$ would be viewed as influenced by any of

the inputs $\lambda^{\text{physics}}, \lambda^{\text{geometry}}, \lambda^{\text{French}}$. But this view would clash with the fact that in different preparation groups the corresponding joint probabilities will typically be different. □

Example 1.4 This example is not from behavioral sciences but from quantum physics. It is not as strange as it may appear to the reader. The fact is, the mathematical formalisms independently developed to study selective influences in psychology turn out to be identical to those developed in quantum physics to study the types of determinism involved in the behavior of so-called entangled particles. Two entangled particles can be thought of as being created as a single particles and then split into two mirror-images running away from each other. Particles possess a property called spin, something that can be measured along differently oriented spatial axes. In the case of so-called spin-1/2 particles, such as electrons, once an axis is chosen the spin can attain one of only two possible values, referred to as "spin-up" and "spin-down." Suppose that two entangled electrons run away from each other towards two observers, Alice and Bob (a traditional way of referring to them in quantum physics), with previously synchronized clocks. At one and the same moment by these clocks Alice and Bob measure spins of their respective electrons along axes they previously chose. The nature of the entanglement is such that if the axes chosen by the two observers are precisely the same, then the spin values recorded will necessarily have opposite values: if Bob records spin-down, Alice will record spin-up. Suppose that Bob always chooses one of two axes, which we will denote $\lambda^{\text{Bob}} = \beta_1$ and $\lambda^{\text{Bob}} = \beta_2$. We view λ^{Bob}, therefore, as one of the two inputs of the system. The other input is the axis chosen by Alice, λ^{Alice}. Let it also have two possible values, $\lambda_1^{\text{Alice}} = \alpha_1$ and $\lambda_2^{\text{Alice}} = \alpha_2$. The outcome of Bob's measurement is the first of two outputs of the system. We denote it by A^{Bob}, with the possible values "spin-up" and "spin-down". The random output A^{Alice}, with the same two values, is defined analogously. The theoretical representation of this situation is given by the joint probabilities

$\left(\lambda^{\text{Alice}} = \alpha_1, \lambda^{\text{Bob}} = \beta_j, \right)$	$A^{\text{Bob}} = \uparrow$	$A^{\text{Bob}} = \downarrow$
$A^{\text{Alice}} = \uparrow$	$p_{\uparrow\uparrow}$	$p_{\uparrow\downarrow}$
$A^{\text{Alice}} = \downarrow$	$p_{\downarrow\uparrow}$	$p_{\downarrow\downarrow}$

where i and j stand for 1 or 2 each. It is reasonable to hypothesize that



In other words, the spin recorded by Alice may depend on which axes she chose, but not on the axis chosen by Bob. And vice versa. But the two outcomes here, for any of the four possible combinations of Alice's and Bob's axes, are not stochastically independent. This makes this situation formally identical to that described in the example with double detection, except that in the entanglement paradigm the invariance of the marginal distributions is guaranteed: $\Pr\left(A^{\text{Bob}} = \uparrow\right)$ is the same no matter what axis was chosen by Alice, and vice versa. In fact, it may very well be the case that these probabilities always remain equal to $1/2$, as in the second example with the three exams. □

Behavioral sciences abound with cases when selective influences are assumed with respect to random variables whose realizations are not directly observable. Rather these random variables are hypothetical entities from which random variables with observable realizations can be derived theoretically. Thus, one may posit the existence of certain unobservable processes selectively influenced by certain experimental manipulations and manifested by their contribution to observable response times. For instance, one may assume the existence of processes called perception and response choice with respective durations A^{percept} and A^{response}, and assume that the observed response time is $A^{\text{percept}} + A^{\text{response}}$. One can further assume that stimulus characteristics selectively influence A^{percept} and instruction versions (such as speed emphasis versus accuracy emphasis) selectively influence A^{response}. The conceptual problem mentioned in the previous examples arises here if the two durations are not assumed to be stochastically independent.

In analyzing "same-different" judgments for pairs of sounds, the observable entities are sounds λ^{first} and λ^{second}, each varying on several levels, and responses "same" or "different" for each pair of these sounds' levels. It is typically postulated, however, that the response is a function (in the mathematical sense of the word) of two unobservable random variables, A^{first} and A^{second}, interpreted as internal representations of the two sounds, their images. For instance, a model may postulate that the response "same" is given if and only if the distance between A^{first} and A^{second} in some metric is less than some epsilon. It is reasonable to hypothesize then that

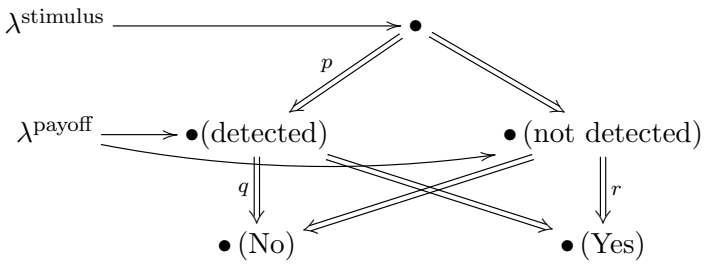
$$\begin{array}{cc} \lambda^{\text{first}} & \lambda^{\text{second}} \\ \downarrow & \downarrow \\ A^{\text{first}} & A^{\text{second}} \end{array}$$

Otherwise, why would one interpret A^{first} and A^{second} as "separate" respec-

tive images of λ^{first} and λ^{second}, rather than speaking of $A = \left(A^{\text{first}}, A^{\text{second}}\right)$ as one image of the compound stimulus $\left(\lambda^{\text{first}}, \lambda^{\text{second}}\right)$?

Stochastic independence of random outputs is, of course, a special case of stochastic relationship. It is clear from our opening examples that this is one case when the issue of defining and testing for selective influences is conceptually transparent. Deterministic outputs are a special case of random outputs, moreover, they can be formally considered stochastically independent. To see that a deterministic output a is influenced by an input λ but not input λ', see if its value changes in response to changes in λ but remains constant if λ' changes with λ fixed. The only reason for mentioning here this obvious consideration is this: there is a wide class of theoretical models which deal with deterministic inputs and and random outputs, but in which selectiveness of influences is formulated as a relationship between deterministic entities, namely, between the inputs and some parameters of the distributions of the random outputs. Parameters of distributions are, by definition, deterministic quantities. Such models require no special theory of selective influences.

Example 1.5 In multinomial processing tree models we see simple examples of random variables related to inputs through parameters describing these variables' distributions. A prototypical example is provided by R. Duncan Luce's (1959) two-state low threshold model of detection,



The processing flow is shown by the double-line arrows: from the root of the tree to the root's children nodes, labeled "detected" and "not detected," and from each of those to their children nodes, labeled "Yes" and "No." The labels p, q, and r are probabilities. The information shown in the processing tree is sufficient for computations, except for one additional constraint: the model stipulates that $qr = 0$ (i.e., when one of the q and r is nonzero the other one must be zero). The inputs $\lambda^{\text{stimulus}}$ and λ^{payoff} are shown on the margins. A single-line arrow pointing at a node of the tree indicates influence on the random variable whose possible values are the children of this node. Stimulus influences the distribution of the (unobservable) binary

random variable called "detection state." It has two values occurring with probabilities p and $1-p$. Payoff is any procedure involving feedback and designed to bias to various degrees the participants towards or against saying "Yes." This input influences the (observable) random variable "response." The point to note here is this: there is no reason to consider the joint distributions of detection state and response for different combinations of stimuli and payoffs; all we need is to declare which of the three parameters of the model, p, q, r depends on which input,

$$p = p\left(\lambda^{\text{stimulus}}\right), q = q\left(\lambda^{\text{payoff}}\right), r = r\left(\lambda^{\text{payoff}}\right).$$

This is simple and clear, even though the outputs "detection state" and "response" are not stochastically independent. □

As it turns out, it is impossible to answer the questions posed in this introductory section without getting "back to basics," to the foundational concepts of probability, random variable, joint distribution, and dependence of joint distributions on deterministic variables. It is simply impossible not to make mistakes and not to get hopelessly confused in dealing with the issues of selective influences if one is only guided by intuitive and informal understanding of these notions. This applies even if the random variables involved are as simple as binary responses. The first part of this chapter (Sections 1.2-1.9) is dedicated to these foundational issues. The reader should be especially attentive when we discuss the fact that not all random variables are jointly distributed, that a set of random variables can always be assigned a joint distribution in the absence of any constraints, but that this may not be possible if the joint distribution should agree with the known distributions of some subsets of this set of random variables. Essentially, the issue of selective influences boils down to establishing whether this is or is not possible in specific cases. We deal with this issue beginning with Section 1.10, as well as the issue of methods by which one can determine whether a particular pattern of selective influences holds. In Section 1.17 we show how the theory of selective influences applies to a classical problem of cognitive psychology, the problem of determining, based on the overall response time, whether certain hypothetical processes involved in the formation of the response are concurrent or serial. The chapter concludes with a brief guide to the relevant literature.

1.2 What is a random variable?

Let us begin with the notion of a *distribution* of a random variable. The formal definition of this notion is as follows: the distribution of a random variable A is a triple

$$\overline{A} = (S, \Sigma, p)\,,$$

where

1. S is some nonempty set, called the *set of possible values* of A;
2. Σ is a *sigma-algebra over* S, which means a collection of subsets of S, each called an *event* or a *measurable set*, such that
   (a) $S \in \Sigma$,
   (b) if $S' \in \Sigma$, then $S - S' \in \Sigma$,
   (c) if $S_1, S_2, \ldots \in \Sigma$ (a finite or countably infinite sequence), then
   $$\bigcup_{i=1,2,\ldots} S_i \in \Sigma;$$
3. p is some function (called *probability measure*) from Σ to $[0,1]$, such that $p(S')$ for $S' \in \Sigma$ is interpreted as the probability with which a value of A falls in (belongs to) event S'; it is assumed that
   (a) $p(S) = 1$,
   (b) (*sigma-additivity*) if $S_1, S_2, \ldots \in \Sigma$ (a finite or countably infinite sequence), and if in this sequence $S_i \cap S_j = \emptyset$ whenever $i \neq j$ (i.e., the subsets in the sequence are pairwise disjoint), then
   $$p\left(\bigcup_{i=1,2,\ldots} S_i\right) = \sum_{i=1,2,\ldots} p(S_i)\,.$$

The following consequences of this definition are easily derived:

1. $\emptyset \in \Sigma$ and $p(\emptyset) = 0$;
2. if $S_1, S_2, \ldots \in \Sigma$, then $\bigcap_{i=1}^{\infty} S_i \in \Sigma$;
3. if $S_1, S_2, \ldots \in \Sigma$ and $S_1 \subset S_2 \subset \ldots$, then
   $$\lim_{i\to\infty} p(S_i) = p\left(\bigcup_{i=1}^{\infty} S_i\right);$$
4. if $S_1, S_2, \ldots \in \Sigma$ and $S_1 \supset S_2 \supset \ldots$, then
   $$\lim_{i\to\infty} p(S_i) = p\left(\bigcap_{i=1}^{\infty} S_i\right);$$

5. if $S_1, S_2 \in \Sigma$ and $S_1 \subset S_2$, then $S_2 - S_1 \in \Sigma$ and

$$p(S_1) + p(S_2 - S_1) = p(S_2);$$

6. if $S_1, S_2 \in \Sigma$, then

$$p(S_1 \cap S_2) \leq \min(p(S_1), p(S_2)) \leq \max(p(S_1), p(S_2)) \leq p(S_1 \cup S_2).$$

Most of these consequences are known as *elementary properties of probability*. It is customary to write $p(S')$ for $S' \in \Sigma$ as $\Pr(A \in S')$, if the distribution of A is known from the context.

We see that in order to know the distribution of a random variable A we have to know its set of possible values S and a set of specially chosen subsets of S, called events. And we should have a procedure "measuring" each event, that is, assigning to it a probability with which a value of A (an element of S) falls within this event (which is also described by saying that the event in question "occurs").

Example 1.6 For a finite S, the sigma-algebra is usually defined as the power set, i.e., the set of all subsets of S. For example, the distribution of the outcome A of a roll of a fair die can be represented by the distribution

$$\overline{A} = (S = \{1,2,3,4,5,6\}, \Sigma = \mathcal{P}(S), p),$$

where $\mathcal{P}(S)$ denotes the power set of S and $p(\{s_1, \ldots, s_k\}) = k/6$ for any set $\{s_1, \ldots, s_k\} \in \Sigma$ of k elements in S. Similarly, the sum of two dice can be represented by the distribution $\overline{A} = (S = \{2,\ldots,12\}, \Sigma = \mathcal{P}(S), p)$, where

$$p(\{s_1, \ldots, s_k\}) = \sum_{i=1}^{k} p(\{s_i\})$$

and $p(\{s\}) = \frac{1}{36}(6 - |7 - s|)$ gives the probability of each singleton (one-element subset) $\{s\}$. □

Example 1.7 Let S be an interval of real numbers, finite or infinite, perhaps the entire set $\mathbb{R}$ of real numbers. For continuous distributions defined on S, at the very least we want to be able to measure the probability of all intervals $(a, b) \subset S$. This requirement implies that our sigma-algebra Σ of events must contain all so-called *Borel subsets* of S. The Borel sets form the smallest sigma-algebra Σ over S that contains all open (or, equivalently, all closed) intervals. One can construct this sigma algebra by the following *recursive procedure*: (1) include in Σ all intervals in S; (2) add to this set of intervals all countable unions of these intervals and of their complements; (3) add to the previously obtained sets all countable unions of these sets of their

complements; (4) and so on. Clearly, these steps are recursive applications of the operations (b) and (c) in the definition of a sigma-algebra. Every Borel set will be obtained at some step of this procedure.

The Borel sigma-algebra is sufficient for most purposes, but often the sigma-algebra is further enlarged by adding to all Borel sets all *null sets*. The latter are sets that can be covered by a countable sequence of intervals with arbitrarily small total length (see Section 1.4). The motivation for this extension is that anything that can be covered by an arbitrarily small length should have its measure equal to zero (and for this it should be measurable). The smallest sigma-algebra containing intervals and null sets is called the *Lebesgue sigma-algebra*.

A continuous distribution on the real line can be defined using a density function $f(a)$. The distribution is given by $\overline{A} = (S, \Sigma, p)$, where Σ is the Lebesgue sigma-algebra, and the probability measure of a set $S_A \in \Sigma$ is given by the integral of the density function f over the set S_A,

$$p(S_A) = \int_{S_A} f(a)\mathrm{d}a.$$

(To be well defined for all Lebesgue-measurable sets S_A, the integral here should be understood in the Lebesgue sense, but we need not go into this.) □

We see that measurability of a subset of S is not a property of the subset itself, but of this subset taken in conjunction with a sigma-algebra Σ. Examples of *non-measurable subsets* of S therefore are easily constructed: choose Σ which is not the entire power set of S, and choose a subset of S which is not in Σ. For instance, if $\Sigma = \{\emptyset, \{1\}, \{2,3\}, \{1,2,3\}\}$ over the set $S = \{1,2,3\}$, then the single-element subset $\{3\}$ is non-measurable. This means that if A is distributed as (S, Σ, p), the probability $p(\{3\})$ with which A falls in $\{3\}$ (or, simply, equals 3) is undefined. This example may seem artificial, as nothing prevents one from complementing Σ with all other subsets of $S = \{1,2,3\}$ (i.e., to assume that p is defined for all of them even if it is only known for some). If S is an interval of reals, however, then there are deeper reasons for not including in Σ all subsets of S.

It is obvious that different random variables can have one and the same distribution. For instance, Peter and Paul can flip a fair coin each, and describe the outcomes by one and the same distribution

$$\overline{A} = \left(S = \{0,1\}, \Sigma = \{\emptyset, \{0\}, \{1\}, \{0,1\}\}, p(\Sigma) = \left\{0, \frac{1}{2}, \frac{1}{2}, 1\right\}\right).$$

To distinguish one random variable from another, therefore, it is not sufficient to know its distribution. We should, in addition, have a *label* or *name* for the random variable: for instance, we can identify one random variable as coin_1, distributed as $\overline{A}$, and another as coin_2, also distributed as $\overline{A}$.

Generally speaking, a random variable A can be viewed as a quadruple (ι_A, S, Σ, p), where ι_A is its unique name and $\overline{A} = (S, \Sigma, p)$ is its distribution. We do not need to be that formal, however, as the notation for a random variable, A, also serves as its name. (The reader familiar with the conventional definition of a random variable as a measurable function on a sample space should wait patiently until Sections 1.6 and 1.7. A function may serve as an identifying label too.)

Remark 1.1 Alternatively, one can assume that the name of a random variable is always (implicitly) part of the elements of its domain S. For instance, the domain for one of the two coins mentioned above may be defined as $S^1 = \{(0, \text{coin}_1), (1, \text{coin}_1)\}$ and for another as $S^2 = \{(0, \text{coin}_2), (1, \text{coin}_2)\}$. The sigma-algebras Σ^1 and Σ^2 then have to be (re)defined accordingly. If this approach is followed consistently, every random variable is uniquely determined by its distribution. We do not follow this route in this chapter.

1.3 Jointly distributed random variables

Let A, B, and C be random variables with distributions $\overline{A} = \left(S^1, \Sigma^1, p_1\right)$, $\overline{B} = \left(S^2, \Sigma^2, p_2\right)$, and $\overline{C} = \left(S^3, \Sigma^3, p_3\right)$.

Remark 1.2 We will consistently use numerical superscripts to refer to the domain sets for random variables, to sigma-algebras over these sets, and later to random variables and inputs. Notation S^3, for example, always refers to a domain set of some random variable, *not* to the Cartesian product $S \times S \times S$. This should not cause any difficulties, as we use numerical exponents in this chapter only twice, and both times this is explicitly indicated.

Let $S_A \in \Sigma^1$, $S_B \in \Sigma^2$, and $S_C \in \Sigma^3$ be three events. We know that $p_1(S_A)$ is interpreted as the probability with which a value of A falls in S_A (or, the probability that the event S_A "occurs"); and analogously for $p_2(S_B)$ and $p_3(S_C)$. We also speak of events *occurring jointly*, or *co-occurring*, a concept whose substantive meaning we will discuss in Section 1.7. For now we will take it formally. In order to speak of S_A, S_B, S_C co-occurring and to ask of the probabilities with which they co-occur, we have to introduce a new random variable, denoted D_{ABC}. As any random variable, it is defined

by some unique name (e.g., "D_{ABC}") and a distribution

$$\overline{D_{ABC}} = \left(S^{123}, \Sigma^{123}, p_{123}\right).$$

The set S^{123} of possible values of D_{ABC} is the Cartesian product $S^1 \times S^2 \times S^3$ (the set of all ordered triples with the first components chosen from S^1, the second from S^2, the third from S^3). The sigma-algebra Σ^{123} is denoted $\Sigma^1 \otimes \Sigma^2 \otimes \Sigma^3$ and defined as the *smallest sigma-algebra* containing the Cartesian products $S_A \times S_B \times S_C$ for all $S_A \in \Sigma^1$, $S_B \in \Sigma^2$ and $S_C \in \Sigma^3$. This means that $\Sigma^{123} = \Sigma^1 \otimes \Sigma^2 \otimes \Sigma^3$ is a set of subsets of $S^1 \times S^2 \times S^3$, such that

1. it contains all the Cartesian products $S_A \times S_B \times S_C$ just mentioned;
2. with every subset S' it contains, it also contains the complement $S^{123} - S'$;
3. with every sequence of subsets $S_1, S_2 \ldots$ it contains, it also contains their union, $\bigcup_{i=1,w,\ldots} S_i$;
4. it is included in any other set of subsets of $S^1 \times S^2 \times S^3$ satisfying 1-2-3 above.

The probability measure p_{123} is called a *joint probability measure*. It should satisfy the general requirements of a probability measure, namely:

$$p_{123}\left(S^1 \times S^2 \times S^3\right) = 1,$$

and

$$p_{123}\left(\bigcup_{i=1,2,\ldots} S_i\right) = \sum_{i=1,2,\ldots} p\left(S_i\right)$$

for any sequence of pairwise disjoint elements $S_1, S_2, \ldots$ of Σ^{123}. In addition, p_{123} should satisfy the following *1-marginal probability equations*: for any $S_A \in \Sigma^1$, $S_B \in \Sigma^2$ and $S_C \in \Sigma^3$,

$$p_{123}\left(S_A \times S^2 \times S^3\right) = p_1\left(S_A\right),$$

$$p_{123}\left(S^1 \times S_B \times S^3\right) = p_2\left(S_B\right),$$

$$p_{123}\left(S^1 \times S^2 \times S_C\right) = p_3\left(S_C\right).$$

Example 1.8 Let

$$S = \{0, 1\},\ \Sigma = \{\emptyset, \{0\}, \{1\}, \{0, 1\}\},$$

and let the random variables A, B, and C be distributed as

$$\overline{A} = \left(S, \Sigma, p_1\right),\ \overline{B} = \left(S, \Sigma, p_2\right),\ \overline{C} = \left(S, \Sigma, p_3\right),$$

where

$$p_1(\Sigma) = \{0, 1/2, 1/2, 1\},\ p_2(\Sigma) = \{0, 1/4, 3/4, 1\},\ p_3(\Sigma) = \{0, 1, 0, 1\}.$$

A joint distribution of A, B, C is defined on the product sigma-algebra $\Sigma^{123} = \Sigma \otimes \Sigma \otimes \Sigma$, which is the smallest sigma-algebra containing all Cartesian products $S_A \times S_B \times S_C$ such that $S_A, S_B, S_C \in \Sigma$. As the Cartesian products include those of all singletons (one-element subsets) $\{(a, b, c)\} = \{a\} \times \{b\} \times \{c\}$, and all subsets of $S \times S \times S$ can be formed by finite unions of these, the product sigma algebra $\Sigma \otimes \Sigma \otimes \Sigma$ is the full power set of $S \times S \times S$. One possible joint distribution for A, B, C is given by

$$\overline{D_{ABC}} = \left(S^{123} = S \times S \times S, \Sigma^{123} = \Sigma \otimes \Sigma \otimes \Sigma, p_{123}\right),$$

where

$$p_{123}(S_{ABC}) = \sum_{(a,b,c) \in S_{ABC}} p_{123}(\{(a, b, c)\})$$

and $p_{123}(\{(a, b, c)\})$ is given by the table

a	b	c	$p_{123}(\{(a,b,c)\})$	a	b	c	$p_{123}(\{(a,b,c)\})$
0	0	0	$1/16$	1	0	0	$3/16$
0	0	1	0	1	0	1	0
0	1	0	$7/16$	1	1	0	$5/16$
0	1	1	0	1	1	1	0

Let us verify that this distribution satisfies the 1-marginal probability equations and is thus a proper joint distribution of A, B, C :

$$\begin{aligned}
p_{123}(\{0\} \times S \times S) &= 1/16 + 0 + 7/16 + 0 = 1/2 = p_1(\{0\}),\\
p_{123}(\{1\} \times S \times S) &= 3/16 + 0 + 5/16 + 0 = 1/2 = p_1(\{1\}),\\
p_{123}(S \times \{0\} \times S) &= 1/16 + 0 + 3/16 + 0 = 1/4 = p_2(\{0\}),\\
p_{123}(S \times \{1\} \times S) &= 7/16 + 0 + 5/16 + 0 = 3/4 = p_2(\{1\}),\\
p_{123}(S \times S \times \{0\}) &= 1/16 + 7/16 + 3/16 + 5/16 = 1 = p_3(\{0\}),\\
p_{123}(S \times S \times \{1\}) &= 0 + 0 + 0 + 0 = 0 = p_3(\{1\}).
\end{aligned}$$

For each 1-marginal, it suffices to verify the probabilities of the points 0 and 1 as the probability values for singletons fully determine the discrete distributions. □

The random variable D_{ABC} is commonly called a *vector of the (jointly distributed) random variables* A, B, and C, and it is denoted (A, B, C). We will use this vectorial notation in the sequel. One should keep in mind, however, that any such a vector is a random variable in its own right. Furthermore, one should keep in mind that the distribution $\overline{(A, B, C)}$, called the

joint distribution with respect to the individual random variables A, B, C, is not uniquely determined by these A, B, C. Specifically, although the set $S^{123} = S^1 \times S^2 \times S^3$ and the sigma-algebra $\Sigma^{123} = \Sigma^1 \otimes \Sigma^2 \otimes \Sigma^3$ are uniquely determined by the sets and sigma-algebras in the distributions $\overline{A}$, $\overline{B}$, and $\overline{C}$, there can generally be more than one joint probability measure p_{123}. The individual p_1, p_2, and p_3 only serve as constraints, in the form of the 1-marginal probability equations above.

A, B, and C in (A, B, C) are called *stochastically independent* if, for any $S_A \in \Sigma^1$, $S_B \in \Sigma^2$ and $S_C \in \Sigma^3$,

$$p_{123}\left(S_A \times S_B \times S_C\right) = p_1\left(S_A\right) p_2\left(S_B\right) p\left(S_C\right).$$

This joint probability measure always satisfies the 1-marginal probability equations.

Example 1.9 Let A and B be standard normally distributed random variables. A bivariate normal joint distribution $(A, B)(\rho)$ can be defined with the density function

$$f_{12}(a, b; \rho) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left(-\frac{a^2 + b^2 - 2\rho ab}{2(1-\rho^2)}\right),$$

where $-1 < \rho < 1$ denotes the correlation coefficient. The sigma algebra $\Sigma_{12} = \Sigma_1 \otimes \Sigma_2$ of the joint distribution is the product of two Lebesgue sigma-algebras (called a Lebesgue sigma-algebra itself). The 1-marginal probability equations can be verified by checking that integrating out either a or b yields the standard normal density function with respect to the remaining variable. The probability measure for $C = (A, B)(\rho)$ is obtained as

$$p_{12}(S_C) = \int_{(a,b)\in S_c} f_{12}(a, b; \rho) \mathrm{d}(a, b).$$

Do $C = (A, B)(\rho_1)$ and $D = (A, B)(\rho_2)$ with $\rho_1 \neq \rho_2$ exclude each other? Not in the sense that defining one of them makes the other meaningless. They both can be defined as variables of interest. But C and D cannot be jointly distributed. □

The reverse relationship between joint and marginal distributions is more straightforward: the distribution $\overline{(A, B, C)}$ uniquely determines the distributions and identity of A, B, C, called the *1-marginal random variables* with respect to (A, B, C), as well as the distributions and identity of (A, B), (B, C), and (A, C), called the *2-marginal random variables* with respect to (A, B, C). Thus, in the distribution $\overline{A}$ the set S^1 is the projection Proj_1 of

the set $S^{123} = S^1 \times S^2 \times S^3$, defined by

$$\mathrm{Proj}_1\,(a, b, c) = a.$$

The sigma-algebra Σ^1 consists of the projections Proj_1 of the elements of the sigma-algebra $\Sigma^{123} = \Sigma^1 \otimes \Sigma^2 \otimes \Sigma^3$ having the form $S_A \times S^2 \times S^3$. And the probability measure p_1 is determined by the 1-marginal probability equations. The *2-marginal distributions* $\overline{(A, B)}$, $\overline{(B, C)}$, and $\overline{(A, C)}$ are found analogously. For example, if one defines function Proj_{23} by

$$\mathrm{Proj}_{23}\,(a, b, c) = (b, c)\,,$$

we have

$$\overline{(B, C)} = \left(S^{23}, \Sigma^{23}, p_{23}\right),$$

where

$$S^{23} = \mathrm{Proj}_{23}\left(S^1 \times S^2 \times S^3\right),$$

Σ^{23} consists of the sets of the form

$$\mathrm{Proj}_{23}\left(S^1 \times S_{BC}\right),\ S_{BC} \in \Sigma^2 \otimes \Sigma^3,$$

and

$$p_{23}\,(S_{BC}) = p_{123}\left(S^1 \times S_{BC}\right).$$

The last equality is one of the three *2-marginal probability equations* (the remaining two being for p_{12} and p_{13}).

One can check that

$$S^{23} = S^2 \times S^3,$$

and

$$\Sigma^{23} = \Sigma^2 \otimes \Sigma^3,$$

which is the smallest sigma-algebra containing the Cartesian products $S_B \times S_C$ for all $S_B \in \Sigma^2$ and $S_C \in \Sigma^3$. In other words, the set S^{23} and the sigma-algebra Σ^{23} over it in the 2-marginal distribution are precisely the same as if they were formed for a joint distribution $\overline{(B, C)}$ with respect to the 1-marginal distributions $\overline{B}$ and $\overline{C}$. Moreover, the 2-marginal probability p_{23} is a joint probability satisfying the 1-marginal probability equations

$$p_{23}\left(S_B \times S^3\right) = p_2\,(S_B)\,,$$

$$p_{23}\left(S^2 \times S_C\right) = p_3\,(S_C)\,.$$

Example 1.10 Continuing from Example 1.8, we can derive the following 2-marginals (and 1-marginals shown at the sides of the 2-marginals):

$p_{12}(\{(a,b)\})$	$b=0$	$b=1$	
$a=0$	$1/16$	$7/16$	$1/2$
$a=1$	$3/16$	$5/16$	$1/2$
	$1/4$	$3/4$	

$p_{12}(\{(b,c)\})$	$c=0$	$c=1$	
$b=0$	$1/4$	0	$1/4$
$b=1$	$3/4$	0	$3/4$
	1	0	

$p_{12}(\{(a,c)\})$	$c=0$	$c=1$	
$a=0$	$1/2$	0	$1/2$
$a=1$	$1/2$	0	$1/2$
	1	0	

□

It should be clear now how one should generalize the notion of a joint distribution to an arbitrary number n of random variables, $A^1, \ldots, A^n$, and how to define k-marginal distributions for $k = 1, \ldots, n$ (n-marginal distributions being permutations of the joint one, including itself).

Remark 1.3 For an infinite set of random variables (countable or not) the definition of a joint distribution is less obvious. We will not deal with this notion in this chapter except for mentioning it occasionally, for completeness sake. With little elaboration, let $\left(A^k : k \in K\right)$ be an indexed family of random variables (with an arbitrary indexing set K), each distributed as $\left(S^k, \Sigma^k, p_k\right)$. We say that the random variables in $\left(A^k : k \in K\right)$ are jointly distributed if $A = \left(A^k : k \in K\right)$ is a random variable with the distribution

$$\overline{A} = \left(\prod_{k \in K} S^k, \bigotimes_{k \in K} \Sigma^k, p \right),$$

where

1. $\prod_{k \in K} S^k$ is the Cartesian product of the sets S^k (its elements are functions choosing for each element of K an element of S^k);
2. $\bigotimes_{k \in K} \Sigma^k$ is the smallest sigma-algebra containing sets of the form $S' \times \prod_{k \in K - \{k_0\}} S^k$, for all $k_0 \in K$ and $S' \in \Sigma^{k_0}$;
3. p is a probability measure on $\bigotimes_{k \in K} \Sigma^k$ such that $p\left(S' \times \prod_{k \in K - \{k_0\}} S^k\right) = p_{k_0}(S')$, for all $k_0 \in K$ and $S' \in \Sigma^{k_0}$.

The random variables A^k in $A = \left(A^k : k \in K\right)$ are said to be stochastically independent if any finite subset of them consists of stochastically independent elements.

Remark 1.4 Marginal random variables sometimes have to be defined hierarchically. Consider, for example, $A' = (A, B)$ and $B' = (C, D)$. Then $C' = (A', B')$ has the 1-marginal distributions $\overline{A'} = \overline{(A, B)}$ and $\overline{B'} = \overline{(C, D)}$. And $A' = (A, B)$, in turn, has 1-marginal distributions $\overline{A}$ and $\overline{B}$. It may sometimes be convenient to speak of all of (A, B), (C, D), A, B, C, D as marginal random variables with respect to a random variable $C' = ((A, B), (C, D))$. Note that $((A, B), (C, D))$, $((A, B, C), D)$, $(A, (B, (C, D)))$, etc. are all distributed as (A, B, C, D), because the Cartesian product $S^1 \times S^2 \times S^3 \times S^4$ and the product sigma algebra $\Sigma^1 \otimes \Sigma^2 \otimes \Sigma^3 \otimes \Sigma^3$ are associative. The random variables $((A, B), (C, D))$, $((A, B, C), D)$, $(A, (B, (C, D)))$, etc. differ in their labeling only. (In the infinite case (Remark 1.3) the formal definition is rather straightforward, but it involves potentially more than a finite number of hierarchical steps. We will assume that the notion is clear and a formal definition may be skipped.)

1.4 Random variables in the narrow sense

The concept of a random variable used in this chapter is very general, with no restrictions imposed on the sets and sigma-algebras in their distributions. Sometimes such random variables are referred to as *random entit*ies, *random element*s, or *random variables in the broad sense*, to distinguish them from *random variables in the narrow sense*. The latter are most important in applications. In particular, all our example involve random variables in the narrow sense. They can be defined as follows. Let A be distributed as $\overline{A} = (S, \Sigma, p)$.

(i) If S is countable, Σ is the power set of S (the set of all its subsets), then A is a random variable in the narrow sense;

(ii) if S is an interval of real numbers, Σ is the *Lebesgue sigma-algebra* over S (as defined in Example 1.7), then A is a random variable in the narrow sense;

(iii) if $A_1, \ldots, A_n$ are random variables in the narrow sense, then any jointly distributed vector $(A_1, \ldots, A_n)$ is a random variable (also referred to as a *random vector*) in the narrow sense.

Random variables satisfying (i) are called *discrete*. The distribution of such a random variable is uniquely determined by the probabilities assigned to its singleton (one-element) subsets. These probabilities can also be viewed as assigned to the elements themselves, in which case they form a *probability mass function*. An example of a discrete random variable is given in Example 1.6. But S may also be countably infinite.

Example 1.11 Let S be the set of positive integers $\{1, 2, \ldots, n, \ldots\}$, and let $p(\{n\}) = \alpha^{n-1}(1-\alpha)$, where α is a constant in $[0, 1]$. This defines a discrete random variables interpreted as the number of independent trials n with binary outcomes (success/failure) until the first failure. It is customary to replace (or even confuse) $p(\{n\})$ with the probability mass function function $p^*(n) = p(\{n\})$. □

Random variables satisfying (ii) are called *continuous* (see Example 1.7). Any such a variable can be viewed as having S extended to the entire set of reals, and its distribution is uniquely determined by the *distribution function*

$$F(x) = p((-\infty, x]),$$

for every real x. The function $F(x)$ has the following properties:

1. it is nondecreasing;
2. as $x \to -\infty$, $F(x) \to 0$;
3. as $x \to \infty$, $F(x) \to 1$;
4. for any real x_0, as $x \to x_0+$, $F(x) \to F(x_0)$ (right-continuity);
5. for any real x_0, as $x \to x_0+$, $F(x)$ tends to a limit.

$F(x)$ generally is not left-continuous: as $x \to x_0-$, the limit of $F(x)$ need not coincide with $F(x_0)$, the function may instead "jump" from the value of $\lim_{x \to x_0-} F(x)$ to $F(x_0)$. The difference $F(x_0) - \lim_{x \to x_0-} F(x)$ equals $p(\{x_0\})$, so the jumps occur if and only if $p(\{x_0\}) > 0$. A distribution function cannot have more than a countable set of jump points. For any two reals $x_1 \le x_2$,

$$F(x_2) - F(x_1) = p((x_1, x_2]).$$

Example 1.12 A discrete random variable can always be redefined as a continuous one. Thus, the variable in the previous example can be redefined into a random variable X whose distribution is given by

$$F(x) = \begin{cases} 0 & \text{for} \quad x < 1 \\ \alpha^{n-1}(1-\alpha) & \text{for} \quad \lfloor x \rfloor = n \ge 1, \end{cases}$$

where $\lfloor x \rfloor$ is the floor function (the largest integer not exceeding x). □

The Lebesgue sigma-algebra over the reals, as defined in Example 1.7, is the smallest sigma-algebra including all intervals and all *null sets*. A subset S' of reals is a null set if, for any $\varepsilon > 0$, however small, S' is contained within a union of open intervals $S_1, S_2, \ldots$ whose overall length is less than ε. An empty set is, obviously a null set, and so is a single point, and a countable set of points.

Remark 1.5 Let us prove that a countable set of points is a null set, to better understand the definition. Enumerate this set as $x_1, x_2, \ldots$, choose an $\varepsilon > 0$, and enclose each x_i into interval $\left]x - \frac{\varepsilon}{2^{i+1}}, x + \frac{\varepsilon}{2^{i+1}}\right[$. The length of this interval is $\frac{\varepsilon}{2^i}$, whence the overall length of the system of such intervals cannot exceed

$$\sum_{i=1,2,\ldots} \frac{\varepsilon}{2^i} \leq \varepsilon.$$

We conclude that a countable subset of S is a null set. There are uncountable null sets.

As should be clear from our discussion of jumps and Example 1.12, a null set may have a nonzero probability. If this does not happen, i.e., if $F(x)$ has no jumps, the distribution of the random variable is called *absolutely continuous*.

Finally, the combination rule (iii) allows one to form vectors of discrete, continuous, and mixed jointly distributed random variables using the construction discussed in Section 1.3.

1.5 Functions of random variables

Let A be a random variable with distribution $\overline{A} = \left(S^1, \Sigma^1, p_1\right)$, let S^2 be some set, and let $f : S^1 \to S^2$ be some function. Consider some sigma algebra Σ^2 of events over S^2. For every $S_B \in \Sigma^2$ one can determine the subset of all elements of S^1 that are mapped by f into S_B,

$$f^{-1}(S_B) = \left\{a \in S^1 : f(a) \in S_B\right\}.$$

This subset, $f^{-1}(S_B)$, does not have to be an event in Σ^1. If it is, for every $S_B \in \Sigma^2$, then f is said to be a *measurable function* (or $\Sigma^1 \to \Sigma^2$-*measurable function*, to be specific). Measurability of a function therefore is not a property of the function itself, but of the function taken in conjunction with two sigma-algebras. In particular, given S^1 and Σ^1, any *onto* function $f : S^1 \to S^2$ (one with $f\left(S^1\right) = S^2$) will be measurable if we agree to define $\Sigma^2 = f\left(\Sigma^1\right)$, the set of all f-images of the elements of Σ^1; it is easy to prove that $f\left(\Sigma^1\right)$ is a sigma-algebra over $f\left(S^1\right)$, for any f.

Example 1.13 Let $S^1 = S^2 = \{1, 2, 3\}$,

$$\Sigma^1 = \{\emptyset, \{1\}, \{2, 3\}, \{1, 2, 3\}\},$$

and

$$\Sigma^2 = \{\emptyset, \{1\}, \{2\}, \{3\}, \{1, 2\}, \{2, 3\}, \{1, 3\}, \{1, 2, 3\}\}.$$

Then the function $f : \Sigma^1 \to \Sigma^2$ defined by $f(a) = a$ is not measurable, because $\{2\} \in \Sigma^2$ but $f^{-1}(\{2\}) = \{2\} \notin \Sigma^1$. However, one can easily verify that $f(a) = \min(a, 2)$ is a $\Sigma^1 \to \Sigma^2$-measurable function. □

Of course, with finite S^1, S^2, one can always define the sigma-algebras as full power sets and then all functions between these sets will be measurable.

Why is the notion of a measurable function important? Because measurable functions can be used to obtain new random variables from existing ones. Given a random variable A and a $\Sigma^1 \to \Sigma^2$-measurable function $f : S^1 \to S^2$, one can define a random variable $B = f(A)$ distributed as $\overline{B} = \left(S^2, \Sigma^2, p_2\right)$ by putting, for any $S' \in \Sigma^2$,

$$p_2\left(S'\right) = p_1\left(f^{-1}\left(S'\right)\right).$$

In other words, the probability with which the new variable B falls in an event belonging to Σ^2 is defined as the probability with which A falls in the f-preimage of this event in Σ^1 (which probability is well defined because f is measurable). Of course, the notation $B = f(A)$ serves as a unique identification of B once we agree that A is uniquely identified.

Example 1.14 Let S^1 and S^2 be two intervals of reals, and let Σ^1 and Σ^2 be the Borel sigma-algebras over them (see Example 1.7). A function $f : S^1 \to S^2$ which is $\Sigma^1 \to \Sigma^2$-measurable is called a *Borel-measurable function*. If in this definition Σ^1 is the Lebesgue sigma algebra over S^1 while Σ^2 continues to be the Borel sigma-algebra over S^2 (note the asymmetry), then f is a *Lebesgue-measurable function*. It is sufficient to require in these two definitions that for any interval $(a, b) \subset S^2$, its preimage $f^{-1}((a, b))$ be a Borel-measurable (respectively, Lebesgue-measurable) subset of S^1. It is easy to prove that if f is monotone or continuous, then it is Borel-measurable (hence also Lebesgue-measurable).

Let now A be a random variable with distribution $\overline{A} = \left(\mathbb{R}, \Sigma^1, p\right)$, where Σ^1 is the Lebesgue sigma-algebra over $\mathbb{R}$. The function $F(x) = p((-\infty, x])$ is called the *distribution function* for A. It is monotonically non-decreasing and maps into $S^2 = [0, 1]$. If we define Σ^2 to be the Borel sigma-algebra over $[0, 1]$, then F (being nondecreasing) is Lebesgue-measurable. If we apply F to A, the resulting random variable $B = F(A)$ is distributed on $[0, 1]$. If furthermore F is a continuous function, then the distribution of $B = F(A)$ on $[0, 1]$ is uniform. That is, its distribution is $\overline{B} = \left([0, 1], \Sigma^2, q\right)$, where $q((a, b)) = b - a$ for any $(a, b) \subset [0, 1]$. □

Let A be distributed as $\overline{A} = \left(S^1, \Sigma^1, p_1\right)$, and let $B = f(A)$ and $C = g(A)$ be two random variables with distributions $\overline{B} = \left(S^2, \Sigma^2, p_2\right)$ and $\overline{C} =$

$\left(S^3, \Sigma^3, p_3\right)$. This implies that both f and g are measurable functions in the sense of, respectively, $\Sigma^1 \to \Sigma^2$ and $\Sigma^1 \to \Sigma^3$. For every $S_B \in \Sigma^2$ and every $S_C \in \Sigma^3$ we have

$$p_2\left(S_B\right) = p_1\left(f^{-1}\left(S_B\right)\right), \text{ and } p_3\left(S_C\right) = p_1\left(g^{-1}\left(S_C\right)\right).$$

A value b of B falls in S_B if and only if $b = f(a)$ for some $a \in f^{-1}\left(S_B\right)$. A value c of C falls in S_C if and only if $c = g(a)$ for some $a \in g^{-1}\left(S_C\right)$. This suggests a way of defining the notion of a *joint occurrence* of these events, S_B and S_C: they occur jointly if and only if a in the previous two sentences is one and the same. In other words, a value b of B falls in S_B and, jointly, a value c of C falls in S_C if and only if, for some $a \in f^{-1}\left(S_B\right) \cap g^{-1}\left(S_C\right)$, $b = f(a)$ and $c = g(a)$. Since $f^{-1}\left(S_B\right) \cap g^{-1}\left(S_C\right)$ is Σ^1-measurable in (belongs to Σ^1), the probability

$$p_{23}\left(S_B \times S_C\right) = p_1\left(f^{-1}\left(S_B\right) \cap g^{-1}\left(S_C\right)\right)$$

is well defined, and we can take it as the joint probability of S_B and S_C.

We now can construct the joint distribution of (B, C),

$$\overline{(B, C)} = \left(S^2 \times S^3, \Sigma^2 \otimes \Sigma^3, p_{23}\right),$$

where the set and the sigma-algebra are defined as required by the general notion of a joint distribution (Section 1.3). The joint probability measure p_{23} defined above for $S_B \times S_C$-type sets is extended to all other members of $\Sigma^2 \otimes \Sigma^3$ by using the basic properties of a probability measure (Section 1.2). Equivalently, the joint probability measure p_{23} can be defined by

$$p_{23}\left(S'\right) = p\left(\left(f, g\right)^{-1}\left(S'\right)\right),$$

for any $S' \in \Sigma^2 \otimes \Sigma^3$. The notation $\left(f, g\right)^{-1}\left(S'\right)$ designates the set S_A of all $a \in S$, such that $\left(f(a), g(a)\right) \in S'$. It can be shown that $S_A \in \Sigma^1$, that is, (f, g) is a measurable function.

It can easily be checked that p_{23} satisfies the 1-marginal probability equations,

$$p_{23}\left(S_B \times S^3\right) = p_1\left(f^{-1}\left(S_B\right) \cap g^{-1}\left(S^3\right)\right) = p_1\left(f^{-1}\left(S_B\right)\right) = p_2\left(S_B\right),$$

$$p_{23}\left(S^2 \times S_C\right) = p_1\left(f^{-1}\left(S^2\right) \cap g^{-1}\left(S_C\right)\right) = p_1\left(g^{-1}\left(S_C\right)\right) = p_3\left(S_C\right),$$

where we used the fact that

$$g^{-1}\left(S^3\right) = f^{-1}\left(S^2\right) = A.$$

We see that if two random variables are formed as functions of another random variable, their joint distribution is uniquely determined.

Example 1.15 A simple but instructive example is the joint distribution of a random variable A and itself. Let A be distributed as (S, Σ, p). (A, A) is a random variable both components of which are functions of one and the same random variable, $A = \mathrm{id}\,(A)$, where id is the identity function defined by $\mathrm{id}(a) = a$. Let the distribution of (A, A) be $(S \times S, \Sigma \otimes \Sigma, p_2)$. By the general theory, for any $S' \in \Sigma$ we have $S' \times S' \in \Sigma \otimes \Sigma$ and

$$p_2\left(S' \times S'\right) = p\left(\mathrm{id}^{-1}\left(S'\right) \cap \mathrm{id}^{-1}\left(S'\right)\right) = p\left(S'\right),$$

as it should be. It is not always true, however, that the probability measure p_2 of the set of pairs

$$\mathrm{diag}_{S\times S} = \{(a, a) : a \in S\}$$

equals 1, because this set is not necessarily an event in $\Sigma \otimes \Sigma$. As an example, $\{(1,1), (2,2), (3,3), (4,4)\}$ is not such an event if $\Sigma = \{\emptyset, \{1,2\}, \{3,4\}, \{1,2,3,4\}\}$. If, however, $\mathrm{diag}_{S\times S} \in \Sigma \otimes \Sigma$, then

$$p_2\left(\mathrm{diag}_{S\times S}\right) = p\left((\mathrm{id}, \mathrm{id})^{-1}\left(\mathrm{diag}_{S\times S}\right)\right) = p\left(S\right) = 1.$$

□

The generalization to several functions of a random variable A is trivial. Thus, we can form a joint distribution not just of B, C but of A, B, C (for symmetry, we can consider A the identity function of A). In particular, the joint probability of $S_B \in \Sigma^2$, $S_C \in \Sigma^3$, and $S_A \in \Sigma^1$ is defined here as

$$p_{23}\left(S_A \times S_B \times S_C\right) = p_1\left(S_A \cap f^{-1}\left(S_B\right) \cap g^{-1}\left(S_C\right)\right).$$

One of the important classes of measurable functions of random variables are projections. We have already dealt with them in Section 1.3, when discussing marginal distributions. More generally, a vector of jointly distributed random variables $A^1, A^2, \ldots, A^n$ is a random variable with a distribution

$$\left(S^1 \times \ldots \times S^n, \Sigma^1 \otimes \ldots \otimes \Sigma^n, p_{1\ldots n}\right),$$

where the notation should be clear from the foregoing. A projection function $\mathrm{Proj}_{i_1\ldots i_k}$, where $k \leq n$ and $i_1, \ldots, i_k$ is a set of k distinct numbers chosen from $(1, \ldots, n)$, is defined by

$$\mathrm{Proj}_{i_1\ldots i_k}\left(a_1, \ldots, a_n\right) = \left(a_{i_1}, \ldots, a_{i_k}\right).$$

Without loss of generality, let $(i_1, \ldots, i_k) = (1, \ldots, k)$; if this is not the case, one can always make it so by renumbering the original set of n random variables. The function $\mathrm{Proj}_{1\ldots k}$ creates a *k-marginal random variable*

$$\mathrm{Proj}_{1\ldots k}\left(A^1, \ldots, A^n\right) = \left(A^1, \ldots, A^k\right),$$

with the k-marginal distributions

$$\left(S^1 \times \ldots \times S^k, \Sigma^1 \otimes \ldots \otimes \Sigma^k, p_{1\ldots k}\right).$$

where, for any measurable even $S^{'}$ in $\Sigma^1 \otimes \ldots \otimes \Sigma^k$,

$$p_{1\ldots k}\left(S'\right) = p_{1\ldots n}\left(S' \times S^{k+1} \times \ldots \times S^n\right).$$

1.6 Random variables as measurable functions

We have seen that if $A^1, \ldots, A^n$ are all functions of one and the same random variable R, then they posses a joint distribution. To recapitulate, if

$$A^1 = f_1(R), \ldots, A^n = f_n(R),$$

$$\overline{R} = (S^*, \Sigma^*, p_*),$$

and

$$\overline{A^i} = \left(S^i, \Sigma^i, p_i\right),\ i = 1, \ldots, n,$$

then

$$\overline{(A^1, \ldots, A^n)} = \left(S^1 \times \ldots \times S^n, \Sigma^1 \otimes \ldots \otimes \Sigma^n, p_{1\ldots n}\right),$$

where

$$p_{1\ldots n}\left(S'\right) = p_*\left((f_1, \ldots, f_n)^{-1}\left(S'\right)\right),$$

for any $S' \in \Sigma^1 \otimes \ldots \otimes \Sigma^n$. In particular,

$$p_{1\ldots n}(S_1 \times \ldots \times S_n) = p_*\left(\bigcap f_i^{-1}(S_i)\right),$$

for all

$$S_1 \in \Sigma^1, \ldots, S_n \in \Sigma^n.$$

It is easy to see that the reverse of this statement is also true: if $A^1, \ldots, A^n$ have a joint distribution, they can be presented as functions of one and the same random variable. Indeed, denoting the random variable $\left(A^1, \ldots, A^n\right)$ by R, we have

$$A^1 = f_1(R), \ldots, A^n = f_n(R),$$

where

$$f_i \equiv \mathrm{Proj}_i.$$

These two simple observations constitute a proof of an important theorem.

Theorem 1.1 *A vector* $\left(A^1, \ldots, A^n\right)$ *of random variables possesses a joint distribution if and only if there is a random variable* R *and a vector of functions* $\{f_1, \ldots, f_n\}$*, such that* $A^1 = f_1(R), \ldots, A^n = f_n(R)$*.*

Note that we need not specify here that the functions are measurable, because both A^i and R in $A^i = f_i(R)$ are random variables (implying that f_i is measurable).

Although we do not deal in this chapter with infinite sets of jointly distributed random variables, it must be mentioned that Theorem 1.1 has the following *generalized formulation* (see Remark 1.3).

Theorem 1.2 *A family* $\left(A^k : k \in K\right)$ *of random variables possesses a joint distribution if and only if there is a random variable* R *and a family of functions* $(f_k : k \in K)$ *such that* $A^k = f_k(R)$ *for all* $k \in K$*.*

In probability textbooks, consideration is almost always confined to random variables that are jointly distributed. This enables what we may call the *traditional conceptualization of random variables*. It consists in choosing some distribution

$$\overline{R} = \left(S^*, \Sigma^*, p_*\right),$$

calling it a *sample (probability) space*, and identifying any random variable A as a $(\Sigma^* \to \Sigma^1)$-measurable function $f : S^* \to S^1$. The set and sigma-algebra pair $\left(S^1, \Sigma^1\right)$ being chosen, the probability measure p_1 satisfying, for every $S' \in \Sigma^1$,

$$p_1\left(S'\right) = p_*\left(f^{-1}\left(S'\right)\right),$$

is referred to as an i*nduced probability measure*, and the distribution $\overline{A} = \left(S^1, \Sigma^1, p_1\right)$ as an *induced (probability) space*.

The sample space $\overline{R}$ is the distribution of some random variable R; in the language just presented R should be defined as the identity function $\mathrm{id} : S^* \to S^*$ (one that maps each element into itself) on the sample space $\overline{R}$; its induced probability space is, obviously, also $\overline{R}$. In our conceptual framework we simply define R by its distribution $\overline{R}$ and some unique identifying label (such as "R"). Note that the traditional language, too, requires an identifying label and a distribution (using our terminology) in order to define the sample space itself.

Remark 1.6 The traditional language does not constitute a different approach. It is a terminological variant of the conceptual set-up adopted in this chapter and applied to a special object of study: a class $\mathcal{A}$ of random

variables that can be defined as functions of some "primary" random variable R. In accordance with Theorem 1.2, $\mathcal{A}$ can also be described without mentioning R, as a class of random variables such that, for any indexed family of random variables $\left(A^k : k \in K\right)$ with $A^k \in \mathcal{A}(R)$ for all $k \in K$, there is a random variable $A = \left(A^k : k \in K\right)$ that also belongs to $\mathcal{A}$.

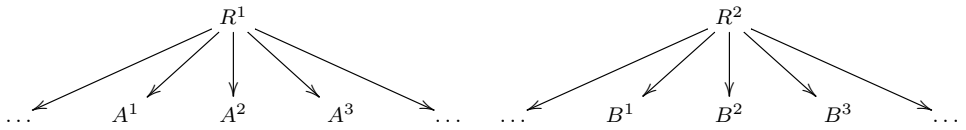
1.7 Unrelated random variables and coupling schemes

There are two considerations to keep in mind when using the traditional language of random variables as measurable functions on sample spaces.

One of them is that sample spaces $\overline{R}$ (or "primary" random variables R) are more often than not nebulous: they need not be and usually are not explicitly introduced when dealing with collections of jointly distributed random variables, and they often have no substantive interpretation if introduced. Consider an experiment in which a participant is shown one of two stimuli, randomly chosen, and is asked to identify them by pressing one of two keys as soon as possible. In each trial we record two random variables: stimulus presented and response time observed, RT. The joint distribution of stimuli and response times is well defined by virtue of pairing them trial-wise. But what would the "primary" random variable R be of which stimulus and RT would be functions? No one would normally attempt determining one, and it is difficult if one tries, except for the trivial choice $R =$(stimulus, RT) or some one-to-one function thereof. The stimulus and RT then would be projections (i.e., functions) of R, but this hardly adds insights to our understanding of the situation. Moreover, as soon as one introduces a new random variable in the experimental design, say, "response key," indicating which of the two keys was pressed, the "primary" random variable R has to be redefined. It may now be the jointly distributed triple $R =$(stimulus, response key, RT).

The second consideration is that there can be no such thing as a single "primary" random variable R allowing one to define all conceivable random variables as its functions. This is obvious from the cardinality considerations alone: the set S^* in $\overline{R}$ would have to be "larger" than the set of possible values for any conceivable random variable (which can, of course, be chosen arbitrarily large). It is a mathematical impossibility. The universe of all conceivable random variables should necessarily include random variables that are not functions of a common "primary" one. In view of Theorem 1.2, this means that there must be random variables that do not possess a joint distribution. The situation should look like in the diagram below, with

$A^1, A^2, \ldots$ being functions of some R^1, $B^1, B^2, \ldots$ being functions of some R^2, but R^1 and R^2 being *stochastically unrelated*, with no joint distribution.



It is true that, as explained below, once R^1 and R^2 are introduced (by their distributions and identifying labels), there is always a way to introduce a new random variable $\left(H^1, H^2\right)$ (whose components are functions of some random variables) such that H^1 has the same distribution as R^1 and H^2 has the same distribution as R^2. But there is no way of conceiving all random variables in the form of functions of a single "primary" one.

Examples of random variables that normally are not introduced as jointly distributed are easy to find. If RTs in an experiment with two stimuli (say, "green" and "red") are considered separately for stimulus "green" and stimulus "red", we have two random variables: RT^{green} and RT^{red}. What "natural" stochastic relationship they might have? The answer is, none: the two random variables occur in mutually exclusive conditions, so there is no privileged way of coupling realizations of RT^{green} and RT^{red} and declaring them co-occurring. Once these random variables are introduced, one can impose a joint distribution on them. For example, one may consider them stochastically independent, essentially forcing on them the coupling scheme in which each realization of RT^{green} considered as if it co-occurred with every realization RT^{red}. But it is also possible to couple them differently, for instance, by the common quantile ranks, so that the qth quantile of RT^{red} is paired with and only with the qth quantile of RT^{green}. The two random variables then are functions of the quantile rank, which is a random variable uniformly distributed between 0 and 1. The point is, neither of these nor any of the infinity of other coupling schemes for the realizations of RT^{green} and RT^{red} is privileged, and none is necessary: one need not impose any joint distribution on RT^{green} and RT^{red}.

It can be shown that stochastic independence can be imposed on any set of pairwise *stochastically unrelated* random variables.

Theorem 1.3 *For any vector* $\left(R^1, \ldots, R^n\right)$ *(more generally, any family* $\left(R^k : k \in K\right)$*) of random variables that are pairwise stochastically unrelated there is a random variable* $H = \left(H^1, \ldots, H^n\right)$ *(generally,* $H =$

$\big(H^k : k \in K\big)$*) with stochastically independent* H^k*, such that* $\overline{H^k} = \overline{R^k}$ *for all* $k \in K$*.*

H is called the *independent coupling* for $\big(R^k : k \in K\big)$. In general, a *coupling* for a family of random variables $\big(R^k : k \in K\big)$, is any random variable $H = \big(H^k : k \in K\big)$ whose every 1-marginal random variable H^k is distributed as R^k.

Theorem 1.3 must not be interpreted to mean that one can take all pairwise stochastically unrelated random variables and consider them stochastically independent. The reason for this is that this class is not a well defined set, and cannot be therefore indexed by any set. Indeed, if it were possible to present it as $\big(R^k : k \in K\big)$, then one could form a new random variable $R = \big(R^k : k \in K\big)$ whose distribution is the same as $\overline{(H^k : k \in K)}$ in Theorem 1.3, and it would follow that the set contains itself as an element (which is impossible for a set).

Summarizing, in practice random variables are often well defined without their joint distribution being well defined. There is nothing wrong in dealing with stochastically unrelated random variables without trying to embed them in jointly distributed system. When such an embedding is desirable, the joint distribution is "in the eyes of the beholder," in the sense of depending on how one wishes to couple the realizations of the variables being interrelated.

1.8 On sameness, equality, and equal distributions

We have to distinguish two different meanings in which one can understand the equality of random variables, $A = B$.

One meaning is that A and B are different notations for one and the same variable, that is, that A and B have the same identifying label and the same distribution. This meaning of equality is implicit when we say "let D be (A, B, C), jointly distributed" or "there is a random variable $A = \big(A^k : k \in K\big)$."

The other meaning of $A = B$ is that

1. these random variables have (or may have) different identifying labels (i.e., they are not or may not be the same);
2. they are identically distributed, $\overline{A} = \overline{B} = (S, \Sigma, p_1)$;
3. they are jointly distributed, and their joint distribution has the form $(S \times S, \Sigma \otimes \Sigma, p_2)$;

4. for any $S' \in \Sigma$,

$$p_2\left(S' \times S'\right) = p_1\left(S'\right).$$

In some cases, if $\mathrm{diag}_S = \{(a, a) : a \in S\}$ is a measurable set (i.e., it belongs to $\Sigma \otimes \Sigma$), one can replace the last property with

$$p_2\left(\mathrm{diag}_S\right) = 1,$$

which can also be presented as

$$\Pr\left(A = B\right) = 1.$$

If A and B about which we know that $A = B$ are represented as functions of some random variable R, then it is usually assumed that $\mathrm{diag}_S \in \Sigma \otimes \Sigma$, and the two functions representing A and B are called *equal with probability 1* (or *almost surely*). Of course, if A and B are merely different notations for one and the same random variable, they are always jointly distributed and equal in the second sense of the term (see Example 1.15).

The equality of random variables, in either sense, should not be confused with the equality of distributions, $\overline{A} = \overline{B}$. The random variables A and B here may but do not have to be jointly distributed. They may very well be stochastically unrelated. We will use the symbol $\sim$ in the meaning of "has the distribution" or "has the same distribution as." Thus, $A \sim \overline{A}$ always, $A \sim B$ if and only if $\overline{A} = \overline{B}$, and $A = B$ always implies $A \sim B$.

An important notational consideration applies to random variables with imposed on them or redefined joint distributions. One may write (A, B) either as indicating a pair of stochastically unrelated random variables, or some random variable $C = (A, B)$. The two meanings are distinguished by context. Nothing prevents one, in principle, from considering the same A and B as components of two differently distributed pairs, $C = (A, B)$ and $C' = (A, B)$, or as components of a $C = (A, B)$ possessing a joint distribution and a pair (A, B) of stochastically unrelated random variables. Doing this within the same context, however, will create conceptual difficulties. For one thing, we would lose the ability of presenting A and B as functions of some R (based on their joint distribution in C).

There is a simple and principled way of avoiding this inconvenience: use different symbols for random variables comprising different pairs (more generally, vectors or indexed families), considering them across the pairs (vectors, families) as equally distributed stochastically unrelated random variables. In our example, we can write $C = (A, B)$ and $C' = (A', B')$, where $A \sim A'$ and $B \sim B'$, with C and C' being stochastically unrelated. The same principle was applied in the formulation of Theorem 1.3 and more

generally, in the definition of a coupling: rather than saying that given a family of stochastically unrelated $\left(R^k : k \in K\right)$, its coupling is any random variable $H = \left(R^k : k \in K\right)$ whose components are jointly distributed (e.g., independent), the definition says that a coupling is a random variable $H = \left(H^k : k \in K\right)$ such that $H^k \sim R^k$ for all $k \in K$. This means, in particular, that every vector of random variables is stochastically unrelated to any of its couplings.

1.9 Random outputs depending on inputs

Let a random variable be distributed as (S, Σ, p_ϕ), where ϕ stands for some deterministic variable taking values in a set Φ. This means that the probability measure on Σ (the entire function) is generally different for different values of Φ. One could also write $p\,(\phi)$ instead of p_ϕ, but one should keep in mind that this is not a function from Φ to a set of values of p (real numbers between 0 and 1) but rather a function from Φ to the set of all possible probability measures on Σ. The dependence of p_ϕ on ϕ means that the distribution (S, Σ, p_ϕ) of the random variable in question depends on ϕ. We can present it as $\overline{A_\phi}$, and the random variable itself as A_ϕ. One can say that the random variable A *depends on* ϕ, which is equivalent to saying that there is an indexed family of random variables $(A_\phi : \phi \in \Phi)$.

Let ϕ_1 and ϕ_2 be two different elements of Φ. We will assume throughout the rest of the chapter that the corresponding random variables A_{ϕ_1} and A_{ϕ_2} always have different identifying labels (such as "A at $\phi = \phi_1$" and "A at $\phi = \phi_2$"), that is, they are never one and the same variable. But they may have one and the same distribution function, if $p_{\phi_1} \equiv p_{\phi_2}$. If A is a vector of jointly distributed random variables $\left(A^1, \ldots, A^n\right)$, then its dependence on ϕ can be shown as $A_\phi = \left(A^1, \ldots, A^n\right)_\phi$ or $A_\phi = \left(A^1_\phi, \ldots, A^n_\phi\right)$.

In the following, ϕ always represents mutually exclusive conditions under which A is observed, and the indexed family $(A_\phi : \phi \in \Phi)$ abbreviated by A consists of pairwise stochastically unrelated random variables. The elements of Φ are referred to as *treatments*, the term being used in the same way as in the analysis of variance: a combination of values of different *factors*, or *inputs*. We will use the latter term. An input is simply a variable λ with a set of possible values Λ. If the number of inputs considered is m, a treatment is a vector

$$\phi = \left(\lambda^1, \ldots, \lambda^m\right),$$

with $\lambda^1 \in \Lambda^1$, ..., $\lambda^m \in \Lambda^m$. The set of treatments is therefore

$$\Phi \subset \Lambda^1 \times \ldots \times \Lambda^m.$$

Remark 1.7 As it is commonly done in mathematics, we will use the same symbol to denote a variable and its specific values. For example, in $\lambda^1 \in \Lambda^1$ the symbol λ^1 refers to a value of λ^1, whereas in the sentence "A^1 depends on λ^1" the same symbol refers to the variable as a whole. This ambiguity is possible to avoid by using Λ^1 in place of λ^1 when referring to the entire variable, and using a pair $\left(\lambda^1, \Lambda^1\right)$ when referring to an input value as that of a given input. We do not use this rigorous notation here, assuming context will be sufficient for disambiguation.

Example 1.16 Let ϕ describe a stimulus presented to a participant. Let it attain eight possible values formed by combinations of three binary attributes, such as

$$\lambda^1 \in \Lambda^1 = \{\text{large}, \text{small}\}, \lambda^2 \in \Lambda^2 = \{\text{bright}, \text{dim}\}, \lambda^3 \in \Lambda^3 = \{\text{round}, \text{square}\}.$$

Let the participant respond by identifying (correctly or incorrectly) these attributes, by saying A^1 = "large" or "small", A^2 = "bright" or "dim", and A^3 = "round" or "square". The response therefore is a vector of three binary random variables $\left(A^1, A^2, A^3\right)_\phi$ that depends on stimuli $\phi = \left(\lambda^1, \lambda^2, \lambda^3\right)$. Equivalently, we can say that there are eight triples of random variables, one for each treatment, $\left(A^1, A^2, A^3\right)_{\phi_1}, \ldots, \left(A^1, A^2, A^3\right)_{\phi_8}$. □

The set of all treatments Φ may be equal to $\Lambda^1 \times \ldots \times \Lambda^m$, but it need not be. Some of the logically possible combinations of input values may not be physically realizable or simply may not be of interest. The elements of Φ therefore are referred to as *allowable treatments*. We will see later that this notion is important in pairing inputs with random outputs.

Example 1.17 Suppose Λ^1 and Λ^2 denote the sets of possible lengths of two line segments presented side by side in the visual field of an observer. Let A^1 and A^2 denote the observer's numerical estimates of the two lengths. If the goal of the experiment is to study perceptual discrimination, it may be reasonable (and time-saving) to exclude the pairs with large values of $\left|\lambda^1 - \lambda^2\right|$. For example, if $\Lambda^1 = \Lambda^2 = \{5, 6, 7, 8, 9\}$, the set of allowable treatments may be defined as

$$\Phi = \{(\lambda^1, \lambda^2) \in \Lambda^1 \times \Lambda^2 : |\lambda^1 - \lambda^2| \leq 2\}.$$

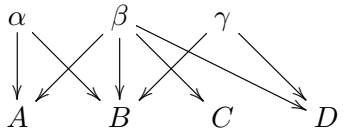
This set contains only 19 treatments of the 25 logically possible combinations. □

As explained in the introductory section, inputs may very well be random variables themselves, but only their possible values rather than their distributions are relevant in our analysis: the distributions of random outputs are always *conditioned* upon particular treatments. All inputs therefore are always treated as deterministic quantities.

1.10 Selectiveness in the dependence of outputs on inputs

We are interested in the relationship between (deterministic) inputs and random outputs. Specifically, we are interested in the selectiveness in this relationship: which input may and which may not influence a given output. Such selectiveness can be presented in the form of a *diagram of influences*, where an arrow from an input λ to a random output A means that λ *influences* A (note that the meaning of "influence" has not been as yet defined). The absence of an arrow from an input λ to a random output A excludes λ from the set of inputs that influence A.

Consider, for example the following arrow diagram

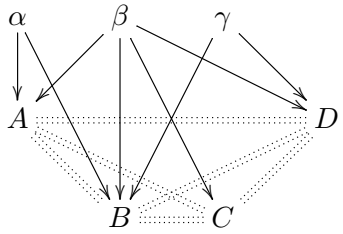


This diagram can be interpreted by saying that:

1. the random outputs (A, B, C, D) are jointly distributed, and their joint distribution (specifically, joint probability measure) depends on the inputs (α, β, γ); in other words, (A, B, C, D) is in fact $(A, B, C, D)_{\alpha\beta\gamma}$, or $(A_{\alpha\beta\gamma}, B_{\alpha\beta\gamma}, C_{\alpha\beta\gamma}, D_{\alpha\beta\gamma})$.
2. output A is influenced by inputs α, β but not by input γ;
3. output B is influenced by all inputs, α, β, γ;
4. output C is influenced by input β but not by inputs α, γ;
5. output D is influenced by inputs β and γ, but not by α.

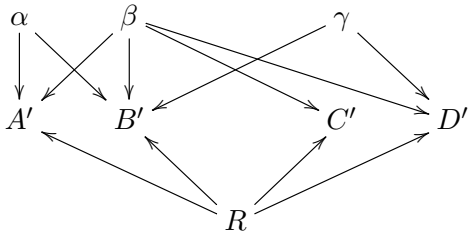
The first thing to do here is to ask the question we asked in the introductory section: does this even make sense? It certainly does if $(A, B, C, D)_{\alpha\beta\gamma}$, for every treatment (α, β, γ), is a vector of *independent* random variables. Then the points 2,3, and 4, above simply translate into the statements: the marginal distribution of A depends on α, β but not on γ; the marginal distribution of B depends on α, β, γ; etc. But does the selectiveness make sense if the random outputs are not stochastically independent? Look at the

diagram below, the same as above, but with added point lines indicating stochastic interdependences.



We see, for instance, that output A is influenced by α, and output C is stochastically dependent on A. In what sense then one can say that α does not influence C? The output B is influenced by all inputs, and every other output is stochastically dependent on B. Does not this mean that every output is influenced by every input?

This seemingly compelling line of reasoning is a conceptual confusion. It confuses two types of relations, both of which can be described using the word "dependence." Stochastic dependence and dependence of outputs on inputs are different in nature. This is easy to understand if we consider the following diagram:



In this diagram, every random variable is a function of all the arguments from which the arrows leading to this random variable initiate:

$$\begin{aligned} A'_{\alpha\beta\gamma} &= f_1(\alpha, \beta, R), \\ B'_{\alpha\beta\gamma} &= f_2(\alpha, \beta, \gamma, R), \\ C'_{\alpha\beta\gamma} &= f_3(\beta, R), \\ D'_{\alpha\beta\gamma} &= f_4(\alpha, \beta, R). \end{aligned}$$

For every value of R and for every treatment (α, β, γ), the values of $(A', B', C', D')_{\alpha\beta\gamma}$ are determined uniquely. Suppose now that we have, for every treatment,

$$\left(A', B', C', D'\right)_{\alpha\beta\gamma} \sim (A, B, C, D)_{\alpha\beta\gamma}.$$

This assumption explains the coexistence of the stochastic relationship between the random outputs and the selectiveness in their dependence on the inputs. *For any given treatment*, the components of $(A, B, C, D)_{\alpha\beta\gamma}$ are generally stochastically interdependent because they are distributed as functions of one and the same random variable R (of course, as a special case, they may also be stochastically independent). At the same time, *for any fixed value r of R*, the value $a = f_1(\alpha, \beta, r)$ of the output $A'_{\alpha\beta\gamma}$ cannot depend on γ, the value $c = f_3(\beta, r)$ of the output $C'_{\alpha\beta\gamma}$ cannot depend on anything but β, etc. And since the distributions of $(A', B', C', D')_{\alpha\beta\gamma}$ and $(A, B, C, D)_{\alpha\beta\gamma}$ are the same, they share the same selectiveness pattern.

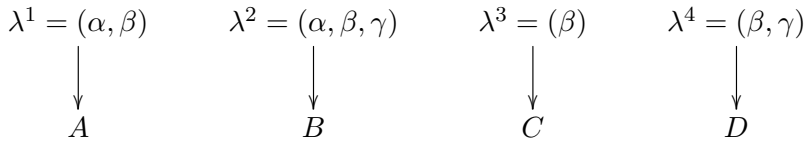
This consideration leads us to a rigorous definition of what it means for a vector of random outputs $(A, B, C, D)_{\alpha\beta\gamma}$ to satisfy the pattern of selective influences represented in the opening diagram of this section: this pattern is satisfied if and only if the equations above are satisfied for some choice of a random variable R and function f_1, f_2, f_3, f_4. This definition can be generalized to an arbitrary family of random outputs and an arbitrary family of inputs. However, we will confine our attention to the case when these families are finite vectors. And we will use a special (re-)arrangement of the inputs to make the definition especially simple.

Remark 1.8 It should be kept in mind that the meaning of "λ influences A" includes, as a special case the possibility of λ not influencing A. There is an asymmetry in saying that, in the example used in this section, C depend on β, and saying that C does not depend on α. The latter is a definitive statement: α is not within the list of arguments in the function $c = f_3(\beta, r)$. The dependence on β means that β is within this list. But a constant function is a special case of a function. So $c = f_3(\beta, r)$ may, as a special case, be constant at all values of R, or at all values of R except on a subset of measure zero. For instance, if R is uniformly distributed between 0 and 1 (we will see below that this choice is possible in a wide class of cases) and $c = f_3(\beta, r)$ is a non-constant function of β only at rational r, then C does not depend on β with probability 1 (because the set of all rational points is countable, hence its Lebesgue measure is zero). This shows that the terms "depends on" and "influences" should generally be understood as "may depend on" and "may influence."

1.11 Selective Influences in a canonical form

Continuing with the same example, let us consider the random outputs one by one, and for each of them group together all inputs that influence it. We

get

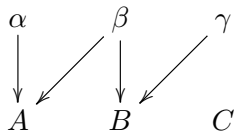


Let us assume that each of the inputs α, β, γ has three possible values, crossed in all possible ways to form 27 treatments. Each of the newly formed groups of inputs can be viewed as a new input in its own right. Thus, λ^1 and λ^4 are inputs whose sets of possible values Λ^1 and Λ^4 have nine possible values each, λ^2 is an input with 27 possible values in Λ^2, and λ^3 is an input with three values in Λ^3.

Such a rearrangement is always possible, whatever the original pattern of influences, and it achieves a one-to-one correspondence between random outputs and inputs. We call a diagram with such one-to-one correspondence a *canonical diagram of influences*. (The term "canonical" is used in mathematics to refer to a standard representation into which a variety of other representations can be transformed.) The problem of selectiveness with a canonical diagram acquires a simple form: is every random output selectively influenced by its corresponding input?

When dealing with canonical diagrams it is especially important to keep in mind that *allowable treatments* are generally just a subset of the Cartesian product of the sets of input values. In our example, this Cartesian product is $\Lambda^1 \times \Lambda^2 \times \Lambda^3 \times \Lambda^4$ and it consists of $9 \times 27 \times 3 \times 9$ elements. But, obviously, only 27 combinations of new inputs' values are allowable, corresponding to the 27 treatments formed by the completely crossed original inputs. Thus, if $\lambda^2 = (\alpha, \beta, \gamma)$, then the only allowable treatment containing this value of λ^2 also contains $\lambda^1 = (\alpha, \beta)$, $\lambda^3 = (\beta)$, and $\lambda^4 = (\beta, \gamma)$.

Another consideration related to the canonical diagrams of influences is that in order to ensure one-to-one correspondence between inputs and random outputs, we may need to allow for "dummy" inputs, with a single possible value. Consider the following example:



Not being influenced by any inputs (as it is the case with the output C) is a special case of selectiveness, so this situation falls within the scope of our

analysis. Presented in the canonical form, this diagram becomes

$$\begin{array}{ccc} \lambda^1 = (\alpha, \beta) & \lambda^2 = (\beta, \gamma) & \lambda^3 = () \\ \downarrow & \downarrow & \downarrow \\ A & B & C \end{array}$$

The new input λ^3 represents an empty subset of original inputs. Therefore λ^3 does not change, and should formally viewed as an input whose set of possible values Λ^3 contains a single element, that we may denote arbitrarily.

We are ready now to give a formal definition of selective influences. Let $\left(\lambda^1, \ldots, \lambda^n\right)$ be a vector of inputs, with values belonging to nonempty sets $\left(\Lambda^1, \ldots, \Lambda^n\right)$, respectively. Let $\Phi \subset \Lambda^1 \times \ldots \times \Lambda^n$ be a nonempty set of allowable treatments. Let $\left(A^1_\phi, \ldots, A^n_\phi\right)$ be a vector of random variables jointly distributed for every $\phi \in \Phi$. (Recall that for $\phi \neq \phi'$, the random variables $\left(A^1_\phi, \ldots, A^n_\phi\right)$ and $\left(A^1_{\phi'}, \ldots, A^n_{\phi'}\right)$ are stochastically unrelated.) We say that the dependence of $\left(A^1_\phi, \ldots, A^n_\phi\right)$ on ϕ satisfies the (canonical) diagram of influences

$$\begin{array}{ccc} \lambda^1 & \ldots & \lambda^n \\ \downarrow & & \downarrow \\ A^1 & \ldots & A^n \end{array}$$

if and only if one can find a random variable R and functions $f_1, \ldots, f_n$ such that

$$\left(A^1_\phi, \ldots, A^n_\phi\right) \sim \left(f_1\left(\lambda^1, R\right), \ldots, f_n\left(\lambda^n, R\right)\right)$$

for every $\left(\lambda^1, \ldots, \lambda^n\right) = \phi \in \Phi$.

Remark 1.9 There is no implication of uniqueness in this definition: below, in the discussion of the linear feasibility test, we will reconstruct R explicitly, and we will see that it can, as a rule, be chosen in infinitely many ways. Theorem 1.6 below shows the non-uniqueness of R by another argument.

Instead of drawing diagrams, in the sequel we will present the same pattern of selective influences as

$$\left(A^1, \ldots, A^n\right) \leftarrow\!\!\!\!\circ \left(\lambda^1, \ldots, \lambda^n\right),$$

and say that $A^1, \ldots, A^n$ are *selectively influenced* by $\lambda^1, \ldots, \lambda^n$ (respectively). If it is known that for a given vector of input-output pairs the definition above is not satisfied whatever R and $f_1, \ldots, f_n$ one chooses, then we

write

$$\left(A^1,\ldots,A^n\right) \not\leftrightarrow \left(\lambda^1,\ldots,\lambda^n\right).$$

Note that for this schematic notation to make sense, context in which it is used should specify the sets of input values, the distributions of $\left(A^1_\phi,\ldots,A^n_\phi\right)$, and the set of allowable treatments.

Example 1.18 Let $R=(R_1,R_2,R_3)$ denote a vector of independent standard normal random variables, and suppose the input factors Λ^1 and Λ^2 are some subsets of $\mathbb{R}$. Then, the binary random variables

$$A^1_{(\lambda^1,\lambda^2)} = \begin{cases} 1 & \text{if } R_1 < \lambda^1 + R_3, \\ 0 & \text{otherwise,} \end{cases}$$

$$A^2_{(\lambda^1,\lambda^2)} = \begin{cases} 1 & \text{if } R_2 < \lambda^2 + R_3, \\ 0 & \text{otherwise,} \end{cases}$$

are selectively influenced by respectively $\lambda^1 \in \Lambda^1$ and $\lambda^2 \in \Lambda^2$, because A^1 depends only on (λ^1, R) and A^2 depends only on (λ^2, R). For any given $\left(\lambda^1,\lambda^2\right)$, the random variables $A^1_{(\lambda^1,\lambda^2)}$ and $A^2_{(\lambda^1,\lambda^2)}$ are not stochastically independent because $R_1 - R_3$ and $R_2 - R_3$ have a nonzero correlation. □

1.12 Joint Distribution Criterion

Let us begin by making sure that the simplest special case, when $\left(A^1_\phi,\ldots,A^n_\phi\right)$ are mutually independent random variables at every allowable treatment ϕ, falls within the scope of the general definition. We expect, if our general definition is well constructed, that in this case selectiveness of influences, $\left(A^1,\ldots,A^n\right) \leftrightarrow \left(\lambda^1,\ldots,\lambda^n\right)$, follows from the fact that the distribution of A^k_ϕ (for $k=1,\ldots,n$) depends only on λ^k. In order not to deal with infinite indexed families, let us assume that λ^k has a finite number of values, enumerated as $1,\ldots,m_k$. Consider the random variable

$$H = \left(H^1_1,\ldots,H^1_{m_1},\ldots,H^k_1,\ldots,H^k_{m_k},\ldots,H^n_1,\ldots,H^n_{m_n}\right)$$

with stochastically independent components, such that, for all $i=1,\ldots,m_k$ and $k=1,\ldots,n$,

$$H^k_i \sim A^k_\phi$$

whenever $\lambda^k = i$ is in ϕ. In other words, if the treatment ϕ contains the ith value of the input λ^k, then we pick A^k_ϕ, and change its identifying label with its distribution intact to create H^k_i. Clearly, the H^k_i will be the same

(provided we always use the same label) irrespective of which ϕ contains $\lambda^k = i$. The variable H above always exists by Theorem 1.3. Let us define function f_k for $k = 1, \ldots, n$ by

$$f_k\left(i, h_1^1, \ldots, h_{m_k}^1, \ldots, h_1^k, \ldots, h_{m_k}^k, \ldots, h_1^n, \ldots, h_{m_k}^n\right) = h_i^k.$$

This can be understood as the "first-level" kth projection that selects from the range of the arguments the subrange $h_1^k, \ldots, h_{m_k}^k$, followed by the "second-level" ith projection that selects from this subrange the argument h_i^k. It is obvious then that, for every $\phi \in \Phi$,

$$A_\phi^k \sim f_k\left(i, H\right)$$

whenever ϕ contains $\lambda^k = i$. But then

$$\left(A_\phi^1, \ldots, A_\phi^n\right) \sim \left(f_1\left(\lambda^1, H\right), \ldots, f_n\left(\lambda^n, H\right)\right)$$

whenever $\left(\lambda^1, \ldots, \lambda^n\right) = \phi \in \Phi$, as it is required by the general definition.

The vector H constructed in this analysis is a special case of the *reduced coupling vector* introduced next. As it turns out, the existence of such a vector, with one random variable per each value of each input is the general criterion for selective influences. A criterion for a statement is another statement which is equivalent to it. Put differently, a criterion is a condition which is both necessary and sufficient for a given statement.

Consider the statement that $A^1, \ldots, A^n$ are selectively influenced by $\lambda^1, \ldots, \lambda^n$, respectively. By definition, for this to be true, there should exist functions $f_1, \ldots, f_n$ and a random variable R such that

$$\left(A_\phi^1, \ldots, A_\phi^n\right) \sim \left(f_1\left(\lambda^1, R\right), \ldots, f_n\left(\lambda^n, R\right)\right)$$

for every $\left(\lambda^1, \ldots, \lambda^n\right) = \phi \in \Phi$. We continue to assume that every input λ^k has a finite number of values, enumerated $1, \ldots, m_k$. (Recall, from the discussion of dummy inputs, that $m_k = 1$ is allowed.)

For each k and every value of λ^k, denote

$$H_{\lambda^k}^k = f_k\left(\lambda^k, R\right).$$

As λ^k runs from 1 to m_k and k runs from 1 to n, this creates $m_1 + \ldots + m_n$ random variables, one random variable per each value of each input, jointly distributed due to being functions of one and the same R. We have therefore a random variable

$$H = \left(H_1^1, \ldots, H_{m_1}^1, \ldots, H_1^k, \ldots, H_{m_k}^k, \ldots, H_1^n, \ldots, H_{m_n}^n\right).$$

If follows from the definition of selective influences that if $\left(A^1,\ldots,A^n\right) \leftharpoonup \left(\lambda^1,\ldots,\lambda^n\right)$, then, for every allowable treatment $\phi = \left(\lambda^1,\ldots,\lambda^n\right)$,

$$\left(A^1_\phi,\ldots,A^n_\phi\right) \sim \left(H^1_{\lambda^1},\ldots,H^n_{\lambda^n}\right).$$

In other words, the existence of a jointly distributed vector of random variables H with this property is a necessary condition for $\left(A^1,\ldots,A^n\right) \leftharpoonup \left(\lambda^1,\ldots,\lambda^n\right)$.

Let us now assume that a vector H with the above property exists. Let us define functions as we did it in the case with stochastic independence,

$$f_k\left(i,h^1_1,\ldots,h^1_{m_1},\ldots,h^k_1,\ldots,h^k_{m_k},\ldots,h^n_1,\ldots,h^n_{m_n}\right) = h^k_i.$$

Then

$$\left(A^1_\phi,\ldots,A^n_\phi\right) \sim \left(H^1_{\lambda^1},\ldots,H^n_{\lambda^n}\right) = \left(f_1\left(\lambda^1,H\right),\ldots,f_n\left(\lambda^n,H\right)\right)$$

for every $\left(\lambda^1,\ldots,\lambda^n\right) = \phi \in \Phi$. This means that the existence of H is a sufficient condition for $\left(A^1,\ldots,A^n\right) \leftharpoonup \left(\lambda^1,\ldots,\lambda^n\right)$.

Summarizing, we have proved the following theorem.

Theorem 1.4 (Joint Distribution Criterion) *Let $\left(\lambda^1,\ldots,\lambda^n\right)$ be a vector of inputs, with $\lambda^k \in \Lambda^k = \{1,\ldots,m_k\}$ ($m_k \geq 1$, $k = 1,\ldots,n$). Let $\Phi \subset \Lambda^1 \times \ldots \times \Lambda^n$ be a nonempty set of allowable treatments. Let $\left(A^1_\phi,\ldots,A^n_\phi\right)$ be a set of random variables jointly distributed for every $\phi \in \Phi$. Then*

$$\left(A^1,\ldots,A^n\right) \leftharpoonup \left(\lambda^1,\ldots,\lambda^n\right)$$

if and only if there exists a vector of jointly distributed random variables

$$H = \left(\overbrace{H^1_1,\ldots,H^1_{m_1}},\ldots,\overbrace{H^k_1,\ldots,H^k_{m_k}},\ldots,\overbrace{H^n_1,\ldots,H^n_{m_n}}\right),$$

(one variable per each value of each input) such that

$$\left(A^1_\phi,\ldots,A^n_\phi\right) \sim \left(H^1_{\lambda^1},\ldots,H^n_{\lambda^n}\right)$$

for every $\left(\lambda^1,\ldots,\lambda^n\right) = \phi \in \Phi$.

The vector H in this theorem is called a *reduced coupling vector* for the family $\left(\left(A^1_\phi,\ldots,A^n_\phi\right) : \phi \in \Phi\right)$ (or for a given pattern of selective influences).

Remark 1.10 According to the general definition of a coupling (Section

1.7), a coupling for the family $\left(\left(A^1_\phi,\ldots,A^n_\phi\right):\phi\in\Phi\right)$ is any random variable

$$H^*=\left(\left(H^1_\phi,\ldots,H^n_\phi\right):\phi\in\Phi\right)$$

such that, for all $\phi\in\Phi$,

$$\left(A^1_\phi,\ldots,A^n_\phi\right)\sim\left(H^1_\phi,\ldots,H^n_\phi\right).$$

The vector H of Theorem 1.4 is obtained from such a coupling by imposing on it additional constraints: for any $k=1,\ldots,n$ and any $\phi,\phi'\in\Phi$ sharing the same value of input λ^k,

$$H^k_\phi=H^k_{\phi'}.$$

These constraints allow one to reduce all different occurrences of H^k in H to one occurrence per each value of factor λ^k. Hence the adjective "reduced" in the name for this special coupling. (In the literature on selective influences the reduced coupling was also called a *joint distribution vector*, and a *Joint Distribution Criterion vector*. We will not use these terms here.)

Theorem 1.4 is much more important than it may be suggested by its simple proof (essentially, by means of renaming functions of a random variable into random variables and vice versa). The reasons for its importance are two:

1. it is often easier to determine whether a coupling vector exists than whether one can find certain functions of a single random variable (unless the latter is taken to be the reduced coupling vector and the functions to be its projections);
2. even when a reduced coupling vector is not explicitly constructed, its existence provides insights into the nature of the random variable R in the definition of selective influences.

The first of these reasons is yet another illustration of the fact that jointly distributed random variables are not, as a rule, introduced as functions of a single random variable (see Section 1.7). Take a simple example, when there are two binary inputs λ^1,λ^2 (with values 1,2 each) paired with two binary outputs (with values 1,2 each). Let the set of allowable treatments consist of all four combinations,

$$\left(\lambda^1=1,\lambda^2=1\right),\left(\lambda^1=1,\lambda^2=2\right),\left(\lambda^1=2,\lambda^2=1\right),\left(\lambda^1=2,\lambda^2=2\right).$$

Note that 1 and 2 as values for the inputs are chosen merely for convenience. We could replace them by any numbers or distinct symbols (say, $\boxtimes,\boxplus$ for λ^1,

and $\nearrow, \swarrow$ for λ^2). The existence of the jointly distributed vectors $\left(A^1_\phi, A^2_\phi\right)$ means that for each of the four treatments ϕ we are given four probabilities of the form

$$\begin{aligned} &\Pr(A^1_\phi = 1, A^2_\phi = 1), \quad \Pr(A^1_\phi = 1, A^2_\phi = 2), \\ &\Pr(A^1_\phi = 2, A^2_\phi = 1), \quad \Pr(A^1_\phi = 2, A^2_\phi = 2). \end{aligned}$$

Of course, the four probabilities sum to 1. Again, the use of 1 and 2 for values here is arbitrary, other symbols, generally different for A^1_ϕ and A^2_ϕ, would do as well. According to the Joint Distribution Criterion, $\left(A^1, A^2\right) \leftrightharpoons \left(\lambda^1, \lambda^2\right)$ means the existence of four jointly distributed random variables

$$H = \left(H^1_1, H^1_2, H^2_1, H^2_2\right),$$

with H^1_1 corresponding to the first value of input λ^1, H^1_2 to the second value of input λ^1, etc., such that

$$\begin{aligned} &\left(A^1, A^2\right)_{\lambda^1=1,\lambda^2=1} \sim \left(H^1_1, H^2_1\right), \quad \left(A^1, A^2\right)_{\lambda^1=1,\lambda^2=2} \sim \left(H^1_1, H^2_2\right), \\ &\left(A^1, A^2\right)_{\lambda^1=2,\lambda^2=1} \sim \left(H^1_2, H^2_1\right), \quad \left(A^1, A^2\right)_{\lambda^1=2,\lambda^2=2} \sim \left(H^1_2, H^2_2\right). \end{aligned}$$

This implies, of course, that $H^1_1, H^1_2, H^2_1, H^2_2$ are all binary random variables, with values 1 and 2 each.

What is the meaning of saying that they are jointly distributed? The meaning is that for any of the $2 \times 2 \times 2 \times 2$ possible combinations of values for $H^1_1, H^1_2, H^2_1, H^2_2$ we can find a probability,

$$\Pr\left(H^1_1 = i, H^1_2 = i', H^2_1 = j, H^2_2 = j'\right) = p_{ii'jj'},$$

where $i, j, i', j' \in \{1, 2\}$. It does not matter what these probabilities $p_{ii'jj'}$ are, insofar as they

(i) are legitimate probabilities, that is, they are nonnegative and sum to 1 across the 16 values of H;

(ii) satisfy the 2-marginal constraints

$$\left(A^1, A^2\right)_{\lambda^1=i,\lambda^2=j} \sim \left(H^1_i, H^2_j\right),$$

for all $i, j \in \{1, 2\}$.

The latter translates into

$$\begin{aligned} p_{i1j1} + p_{i1j2} + p_{i2j1} + p_{i2j2} &= \Pr\left(H^1_1 = i, H^2_1 = j\right) = \Pr\left(A^1 = i, A^2 = j\right)_{\lambda^1=1,\lambda^2=1}, \\ p_{i11j} + p_{i12j} + p_{i21j} + p_{i22j} &= \Pr\left(H^1_1 = i, H^2_2 = j\right) = \Pr\left(A^1 = i, A^2 = j\right)_{\lambda^1=1,\lambda^2=2}, \\ p_{1ij1} + p_{1ij2} + p_{2ij1} + p_{2ij2} &= \Pr\left(H^1_2 = i, H^2_1 = j\right) = \Pr\left(A^1 = i, A^2 = j\right)_{\lambda^1=2,\lambda^2=1}, \\ p_{1i1j} + p_{1i2j} + p_{2i1j} + p_{2i2j} &= \Pr\left(H^1_1 = i, H^2_1 = j\right) = \Pr\left(A^1 = i, A^2 = j\right)_{\lambda^1=2,\lambda^2=2}. \end{aligned}$$

This is a simple system of four linear equations with 16 unknowns, subject to being legitimate probabilities (i.e., being non-negative and summing to 1). We will discuss this algebraic structure in the next section, but it should be clear that this is a much more transparent task than the one of finding a random variable R and some functions, or proving that they cannot be found.

Example 1.19 Let A^1, A^2 have values in {1,2} and depend on the factors $\lambda^1 \in \Lambda^1 = \{1, 2\}$ and $\lambda^2 \in \Lambda^2 = \{1, 2\}$. Let all four possible treatments be allowable. Suppose we observe the following joint distributions of A^1, A^2 for these treatments:

λ^1	λ^2	A^1	A^2	Pr
1	1	1	1	.140
		1	2	.360
		2	1	.360
		2	2	.140

λ^1	λ^2	A^1	A^2	Pr
1	2	1	1	.198
		1	2	.302
		2	1	.302
		2	2	.198

λ^1	λ^2	A^1	A^2	Pr
2	1	1	1	.189
		1	2	.311
		2	1	.311
		2	2	.189

λ^1	λ^2	A^1	A^2	Pr
2	2	1	1	.460
		1	2	.040
		2	1	.040
		2	2	.460

The question of whether $\left(A^1, A^2\right) \nleftarrow\!\rho \left(\lambda^1, \lambda^2\right)$ now reduces to finding a solution for the system of linear equations mentioned above. Let us substitute the above observed probabilities into the system:

$$
\begin{array}{ll}
p_{1111} + p_{1112} + p_{1211} + p_{1212} = 0.140, & p_{1111} + p_{1121} + p_{1211} + p_{1221} = 0.198, \\
p_{1121} + p_{1122} + p_{1221} + p_{1222} = 0.360, & p_{1112} + p_{1122} + p_{1212} + p_{1222} = 0.302, \\
p_{2111} + p_{2112} + p_{2211} + p_{2212} = 0.360, & p_{2111} + p_{2121} + p_{2211} + p_{2221} = 0.302, \\
p_{2121} + p_{2122} + p_{2221} + p_{2222} = 0.140, & p_{2112} + p_{2122} + p_{2212} + p_{2222} = 0.198,
\end{array}
$$

$$
\begin{array}{ll}
p_{1111} + p_{1112} + p_{2111} + p_{2112} = 0.189, & p_{1111} + p_{1121} + p_{2111} + p_{2121} = 0.460, \\
p_{1121} + p_{1122} + p_{2121} + p_{2122} = 0.311, & p_{1112} + p_{1122} + p_{2112} + p_{2122} = 0.040, \\
p_{1211} + p_{1212} + p_{2211} + p_{2212} = 0.311, & p_{1211} + p_{1221} + p_{2211} + p_{2221} = 0.040, \\
p_{1221} + p_{1222} + p_{2221} + p_{2222} = 0.189, & p_{1212} + p_{1222} + p_{2212} + p_{2222} = 0.460.
\end{array}
$$

The values (found using the simplex linear programming algorithm)

$$
\begin{array}{llll}
p_{1111} = 0.067, & p_{1211} = 0, & p_{2111} = 0.122, & p_{2211} = 0.04, \\
p_{1112} = 0, & p_{1212} = 0.073, & p_{2112} = 0, & p_{2212} = 0.198, \\
p_{1121} = 0.131, & p_{1221} = 0, & p_{2121} = 0.14, & p_{2221} = 0, \\
p_{1122} = 0.04, & p_{1222} = 0.189, & p_{2122} = 0, & p_{2222} = 0
\end{array}
$$

satisfy these equations, and as they are nonnegative and sum to one, they represent a probability distribution. Thus, according to the Joint Distribution Criterion, the observed joint distributions satisfy selective influences. □

To illustrate the second reason for the importance of Theorem 1.4, we consider the following question. By the definition of selective influences, the proposition $\left(A^1,\ldots,A^n\right) \leftrightharpoons \left(\lambda^1,\ldots,\lambda^n\right)$ means the existence of a random variable R and functions $f_1,\ldots,f_n$ such that

$$\left(A^1_\phi,\ldots,A^n_\phi\right) \sim \left(f_1\left(\lambda^1,R\right),\ldots,f_n\left(\lambda^n,R\right)\right)$$

for every $\left(\lambda^1,\ldots,\lambda^n\right)=\phi\in\Phi$. This definition says nothing about the nature and complexity of R and the functions involved, even for the simplest observable random variables $\left(A^1,\ldots,A^n\right)_\phi$. In most applications $\left(A^1,\ldots,A^n\right)_\phi$ are random variables in the narrow sense (Section 1.4). It seems intuitive to expect that in such cases R, if it exists, is also a random variable in the narrow sense. But this does not follow from the definition of selective influences. Even if one manages to prove that for a given family of random variables $\left(A^1,\ldots,A^n\right)_\phi$ in the narrow sense this definition is satisfied by no random variable R in the narrow sense, we still do not know whether this means that the selectiveness $\left(A^1,\ldots,A^n\right) \leftrightharpoons \left(\lambda^1,\ldots,\lambda^n\right)$ is ruled out. What if there is a random variable R of a much greater complexity (say, a random function or a random set) for which one can find functions $f_1,\ldots,f_n$ as required by the definition?

The Joint Distribution Criterion, however, allows one to rule out such a possibility. Since the reduced coupling vector

$$H=\left(H^1_1,\ldots,H^1_{m_1},\ldots,H^n_1,\ldots,H^n_{m_n}\right),$$

if it exists, should satisfy

$$\left(A^1_\phi,\ldots,A^n_\phi\right) \sim \left(H^1_{\lambda^1},\ldots,H^n_{\lambda^n}\right),$$

it follows that, for any k and λ^k,

$$H^k_{\lambda^k} \sim A^k_\phi,$$

whenever the treatment ϕ contains λ^k. But this means that each $H^k_{\lambda^k}$ is a random variable in a narrow sense, and from Section 1.4 we know then that H is a random variable in the narrow sense. This constitutes a proof of the following theorem, a simple corollary to the Joint Distribution Criterion.

Theorem 1.5 *Let $\left(\lambda^1,\ldots,\lambda^n\right)$, Φ, and $\left(A^1_\phi,\ldots,A^n_\phi\right)$ be the same as in Theorem 1.4 Let, in addition, $\left(A^1_\phi,\ldots,A^n_\phi\right)$ be random variables in the narrow sense. Then*

$$\left(A^1,\ldots,A^n\right) \leftrightharpoons \left(\lambda^1,\ldots,\lambda^n\right)$$

if and only if there is a random variable R in the narrow sense and functions $f_1, \ldots, f_n$ such that

$$\left(A_\phi^1, \ldots, A_\phi^n\right) \sim \left(f_1\left(\lambda^1, R\right), \ldots, f_n\left(\lambda^n, R\right)\right)$$

for every $\left(\lambda^1, \ldots, \lambda^n\right) = \phi \in \Phi$.

If one feels dissatisfied with considering vectors of random variables on a par with "single" random variables, this dissatisfaction is not well-grounded. The fact is, the dimensionality of vectors of random variables in the narrow sense is not essential. Consider, for example, the reduced coupling vector

$$H = \left(H_1^1, H_2^1, H_1^2, H_2^2\right),$$

constructed earlier for two binary random variables selectively influenced by two binary inputs. Clearly, in all considerations this four-component vector of binary random variables can be replaced with a single 16-valued random variable, H'. Let these 16 values be $0, \ldots, 15$. The two variables are equivalent if one puts

$$\begin{aligned}&\Pr\left(H_1^1 = i, H_2^1 = i', H_1^2 = j, H_2^2 = j'\right)\\ &= \Pr\left(H' = (i-1)\,2^3 + \left(i'-1\right)2^2 + (j-1)\,2 + \left(j'-1\right)\right).\end{aligned}$$

In particular, any functions of H can be presented as functions of H'.

In the case of continuous random variables the situation is, in a sense, even simpler, although we will have to omit the underlying justification. It follows from the theory of Borel-equivalent spaces (which is part of descriptive set theory), that any vector of continuous random variables

$$R = \left(R^1, \ldots, R^k\right),$$

can be presented as a function of any continuous variable R' with an *atomless* distribution on an interval of real numbers. The "atomlessness" means that the sigma-algebra of R' contains no null-set whose probability measure is not zero. Simple examples are uniformly and normally distributed random variables. If the vector is discrete, the previous statement applies with no modifications (although we know that in this case one can also choose a discrete R'). It follows that the statement also applies to mixed vectors, containing both discrete and continuous random variables (or vectors thereof, or vectors of vectors thereof, etc.)

We can complement, therefore, Theorem 1.5 with the following statement.

Theorem 1.6 *Under the conditions of Theorem 1.5, the random variable*

R can always be chosen to be any continuous random variable with an atomless distribution on an interval of real numbers. If all the random variables $A^1_\phi, \ldots, A^n_\phi$ are discrete (in particular, have finite numbers of values), then R can be chosen to be discrete (respectively, have finite number of values).

We have quite a bit more specificity now than based on the initial definition of selective influences. And it is achieved due to the Joint Distribution Criterion almost "automatically."

Theorem 1.4 is not restricted to finite-valued inputs. Nor is it restricted to a finite number of inputs, or to outputs of a specific kind. It is completely general. For the reader's convenience, we formulate here the general version of this theorem, avoiding all elaborations.

Theorem 1.7 (Joint Distribution Criterion (general version)) *Let $\left(\lambda^k : k \in K\right)$ be an indexed family of inputs, with $\lambda^k \in \Lambda^k \neq \emptyset$, for all $k \in K$. Let $\Phi \subset \prod_{k\in K} \Lambda^k$ be a nonempty set of allowable treatments. Let $\left(A^k_\phi : k \in K\right)$ be a family of random variables jointly distributed for every $\phi \in \Phi$. Then*

$$\left(A^k : k \in K\right) \hookleftarrow \left(\lambda^k : k \in K\right)$$

if and only if there exists an indexed family of jointly distributed random variables

$$H = \left(H^k_{\lambda^k} : \lambda^k \in \Lambda^k, k \in K\right),$$

(one variable per each value of each input) such that

$$\left(A^k_\phi : k \in K\right) \sim \left(H^k_{\lambda^k} : k \in K\right)$$

for every $\left(\lambda^k : k \in K\right) = \phi \in \Phi$.

1.13 Properties of selective influences and tests

Certain properties of selective influences (in the canonical form) are immediately obvious.

The first one is *nestedness with respect to input values*: if random outputs $A^1, \ldots, A^n$ are selectively influenced by inputs $\lambda^1, \ldots, \lambda^n$, with sets of possible values $\Lambda^1, \ldots, \Lambda^n$, then the same random outputs are selectively influenced by inputs $\lambda'^1, \ldots, \lambda'^n$ whose sets of possible values are $\Lambda'^1 \subset \Lambda^1, \ldots, \Lambda'^n \subset \Lambda^n$. Every variable is essentially the set of its possible values. Inputs are no exception. In fact, in a more rigorous development λ would be reserved for input values, whereas input themselves, considered as variables, would be identified by Λ (see Remark 1.7). When a set of an

input's values changes, the input is being replaced by a new one. The nestedness property in question tells us that if the change consists in removing some of the possible values of some of the inputs, the selectiveness pattern established for the original inputs cannot be violated. This does not, of course, work in the other direction: if we augment $\Lambda^1, \ldots, \Lambda^n$ by adding to them new elements, then the initial pattern of selectiveness may very well disappear.

The second property is *nestedness with respect to inputs and outputs* (in a canonical diagram they are in a one-to-one correspondence): if a vector of random outputs is selectively influenced by a vector of inputs, then any subvector of the random outputs is selectively influenced by the corresponding subvector of the inputs. In symbols, if

$$\left(A^1, \ldots, A^n\right) \leftarrow\!\!\!\!\rho\, \left(\lambda^1, \ldots, \lambda^n\right)$$

and $i_1, \ldots, i_k \in \{1, \ldots, n\}$, then

$$\left(A^{i_1}, \ldots, A^{i_k}\right) \leftarrow\!\!\!\!\rho\, \left(\lambda^{i_1}, \ldots, \lambda^{i_k}\right).$$

Note that the set of allowable treatments has to be redefined whether we eliminate certain input-output pairs or certain input values. In the latter case, the new set of allowable treatments is the largest $\Phi' \subset \Lambda'^1 \times \ldots \times \Lambda'^n$, such that $\Phi' \subset \Phi$. In the case we drop input-output pairs, the new set of allowable treatments is the largest $\Phi'' \subset \Lambda^{i_1} \times \ldots \times \Lambda^{i_k}$, such that every $\phi'' \in \Phi''$ is a part of some $\phi \in \Phi$.

Both these nestedness properties follow from the fact that any subset of random variables that are components of a reduced coupling vector

$$H = \left(H_1^1, \ldots, H_{m_1}^1, \ldots, H_1^n, \ldots, H_{m_n}^n\right),$$

are also jointly distributed. When we eliminate an ith value of input k, we drop from this vector H_i^k. When we eliminate an input k, we drop the subvector $H_1^k, \ldots, H_{m_k}^k$. In both cases the resulting H' is easily checked to be a reduced coupling vector for the redefined sets of treatments and outputs.

By similar arguments one can establish that a pattern of selective influences is well-behaved in response to all possible groupings of the inputs, with or without a corresponding grouping of outputs: thus, if

$$\left(A^1, \ldots, A^k, \ldots, A^l, \ldots, A^n\right) \leftarrow\!\!\!\!\rho\, \left(\lambda^1, \ldots, \lambda^k, \ldots, \lambda^l, \ldots, \lambda^n\right),$$

then

$$\left(A^1, \ldots, A^k, \ldots, A^l, \ldots, A^n\right) \leftarrow\!\!\!\!\rho\, \left(\lambda^1, \ldots, \left(\lambda^k, \lambda^l\right), \ldots, \left(\lambda^k, \lambda^l\right), \ldots, \lambda^n\right)$$

and

$$\left(A^1,\ldots,\left(A^k,A^l\right),\ldots,\left(A^k,A^l\right),\ldots,A^n\right)$$
$$\hookleftarrow\left(\lambda^1,\ldots,\left(\lambda^k,\lambda^l\right),\ldots,\left(\lambda^k,\lambda^l\right),\ldots,\lambda^n\right).$$

We omit the details related to redefinitions of allowable treatments.

A simple consequence of the nestedness with respect to input-output pairs turns out to be of a great importance for determining if a selectiveness pattern is present. This consequence is called *complete marginal selectivity*: if $\left(A^1,\ldots,A^n\right)\hookleftarrow\left(\lambda^1,\ldots,\lambda^n\right)$ and $i_1,\ldots,i_k\in\{1,\ldots,n\}$, then the distribution of $\left(A^{i_1}_\phi,\ldots,A^{i_k}_\phi\right)$ depends only on $\left(\lambda^{i_1},\ldots,\lambda^{i_k}\right)$. In other words, if ϕ and ϕ' include the same subset $\left(\lambda^{i_1},\ldots,\lambda^{i_k}\right)$,

$$\left(A^{i_1}_\phi,\ldots,A^{i_k}_\phi\right)\sim\left(A^{i_1}_{\phi'},\ldots,A^{i_k}_{\phi'}\right).$$

In particular (*simple marginal selectivity*),

$$A^i_\phi\sim A^i_{\phi'}$$

for any ϕ and ϕ' that share a value of λ^i $(i=1,\ldots,n)$. The importance of marginal selectivity is that it is easy to check, ruling out selective influences whenever it is found violated.

Example 1.20 Let A^1,A^2 have values in $\{1,2\}$ and depend on the external factors $\lambda^1\in\Lambda^1=\{1,2\}$ and $\lambda^2\in\Lambda^2=\{1,2\}$. Let the joint distribution of A^1,A^2 for each treatment (all four being allowable) be as follows:

$\lambda^1=1,\ \lambda^2=1$	$A^2=1$	$A^2=2$	
$A^1=1$	.2	.2	.4
$A^1=2$	.3	.3	.6
	.5	.5	

$\lambda^1=1,\ \lambda^2=2$	$A^2=1$	$A^2=2$	
$A^1=1$	.3	.1	.4
$A^1=2$	.2	.4	.6
	.5	.5	

$\lambda^1=2,\ \lambda^2=1$	$A^2=1$	$A^2=2$	
$A^1=1$	.4	.3	.7
$A^1=2$	.1	.2	.3
	.5	.5	

$\lambda^1=2,\ \lambda^2=2$	$A^2=1$	$A^2=2$	
$A^1=1$	.3	.4	.7
$A^1=2$	.1	.2	.3
	.4	.6	

Marginal selectivity here is violated because the marginal distribution of A^2 changes when $\lambda^2=2$ and λ^1 changes from 1 to 2. □

Marginal selectivity is strictly weaker than selective influences. The latter do imply marginal selectivity, but marginal selectivity can very well hold in the absence of selective influences.

Example 1.21 Consider the following joint distributions:

$\lambda^1=1,\ \lambda^2=1$	$A^2=1$	$A^2=2$	
$A^1=1$	.5	0	.5
$A^1=2$	0	.5	.5
	.5	.5	

$\lambda^1=1,\ \lambda^2=2$	$A^2=1$	$A^2=2$	
$A^1=1$	.5	0	.5
$A^1=2$	0	.5	.5
	.5	.5	

$\lambda^1=2,\ \lambda^2=1$	$A^2=1$	$A^2=2$	
$A^1=1$	.5	0	.5
$A^1=2$	0	.5	.5
	.5	.5	

$\lambda^1=2,\ \lambda^2=2$	$A^2=1$	$A^2=2$	
$A^1=1$	0	.5	.5
$A^1=2$	.5	0	.5
	.5	.5	

Marginal selectivity is trivially satisfied as all marginals are uniform. However, $\left(A^1,A^2\right) \not\leftrightarrows \left(\lambda^1,\lambda^2\right)$ in this case. The joint distribution criterion would require the existence of a jointly distributed vector H whose components satisfy $\left(A^1_{ij},A^2_{ij}\right) \sim \left(H^1_i,H^2_j\right)$ for $i,j\in\{1,2\}$. But combining this with the above joint distributions, we obtain

$$H^1_1=H^2_1, \qquad H^1_1=H^2_2, \qquad H^1_2=H^2_1, \qquad H^1_2=3-H^2_2,$$

which yields the contradiction

$$3-H^2_2=H^2_2.$$

□

Another property of selective influences is that if $\left(A^1,\ldots,A^n\right) \leftrightarrows \left(\lambda^1,\ldots,\lambda^n\right)$, and if, for all $\phi=\left(\lambda^1,\ldots,\lambda^n\right)\in\Phi$,

$$B^1_\phi = g_1\left(\lambda^1,A^1_\phi\right),\ldots,B^n_\phi = g_n\left(\lambda^n,A^n_\phi\right),$$

then $\left(B^1,\ldots,B^n\right) \leftrightarrows \left(\lambda^1,\ldots,\lambda^n\right)$. The functions $g_1,\ldots,g_n$ are referred to as *input-value-specific transformations of random outputs*. The property in question therefore is the invariance of selective influences, if established, with respect to such transformations.

Let us make sure that this property is true. According to the general definition, we have a random variable R and functions $f_1,\ldots,f_n$ such that

$$\left(A^1_\phi,\ldots,A^n_\phi\right) \sim \left(f_1\left(\lambda^1,R\right),\ldots,f_n\left(\lambda^n,R\right)\right),$$

for every $\phi=\left(\lambda^1,\ldots,\lambda^n\right)\in\Phi$. But then

$$\left(B^1_\phi,\ldots,B^n_\phi\right) \sim \left(g_1\left(\lambda^1,f_1\left(\lambda^1,R\right)\right),\ldots,g_n\left(\lambda^n,f_n\left(\lambda^n,R\right)\right)\right),$$

and every $g_k\left(\lambda^k,f_k\left(\lambda^k,R\right)\right)$ is some function $f^*_k\left(\lambda^k,R\right)$. The vectors $\left(B^1_\phi,\ldots,B^n_\phi\right)$ therefore satisfy the definition too.

As a special case, the transformation may not depend on input values,

$$B_\phi^1 = g_1\left(A_\phi^1\right), \ldots, B_\phi^n = g_n\left(A_\phi^n\right).$$

This would include all possible *renamings* and *groupings* of the values of the random outputs: a pattern of selective influences is preserved under all such transformations. For instance, one can rename values $1, 2$ of a binary output into $\sqcup, \sqcap$, or one can group values $1, 2, 3, 4$ into "cruder" values, by means of a transformation like

$$1 \mapsto \sqcup, 2 \mapsto \sqcup, 3 \mapsto \sqcap, 4 \mapsto \sqcap.$$

The meaning of the input-value-specificity is this. We choose a $k \in \{1, \ldots, n\}$ and assume, for simplicity, that λ^k has discrete values, $1, 2, \ldots$. Let A_ϕ^k be transformed into random variables $B_{1,\phi}^k$, $B_{2,\phi}^k$, etc., all sharing the same set of possible values and the same sigma-algebra. We know that one can replace A^k in

$$\left(A^1, \ldots A^k, \ldots, A^n\right) \leftarrow\!\!\!\!\!\rho\ \left(\lambda^1, \ldots, \lambda^k, \ldots, \lambda^n\right)$$

with any of these new random variables,

$$\left(A^1, \ldots B_1^k, \ldots, A^n\right) \leftarrow\!\!\!\!\!\rho\ \left(\lambda^1, \ldots, \lambda^k, \ldots, \lambda^n\right),$$
$$\left(A^1, \ldots B_2^k, \ldots, A^n\right) \leftarrow\!\!\!\!\!\rho\ \left(\lambda^1, \ldots, \lambda^k, \ldots, \lambda^n\right),$$
$$\text{etc.}$$

The input-value-specificity is involved if one forms a random variable

$$B_\phi^k = \begin{cases} B_{1,\phi}^k & \text{if} \quad \lambda^k = 1 \\ B_{2,\phi}^k & \text{if} \quad \lambda^k = 2 \\ & \text{etc.} \end{cases}$$

The invariance property says that this random variable, too, can replace A^k in a pattern of selective influences,

$$\left(A^1, \ldots B^k, \ldots, A^n\right) \leftarrow\!\!\!\!\!\rho\ \left(\lambda^1, \ldots, \lambda^k, \ldots, \lambda^n\right).$$

Note that the property in question works in one direction only: if $\left(A^1, \ldots, A^n\right) \leftarrow\!\!\!\!\!\rho\ \left(\lambda^1, \ldots, \lambda^n\right)$ then $\left(B^1, \ldots, B^n\right) \leftarrow\!\!\!\!\!\rho\ \left(\lambda^1, \ldots, \lambda^n\right)$. It is perfectly possible (if we use grouping of values) that $\left(A^1, \ldots, A^n\right) \not\leftarrow\!\!\!\!\!\rho\ \left(\lambda^1, \ldots, \lambda^n\right)$ but following an input-value-specific transformation, $\left(B^1, \ldots, B^n\right) \leftarrow\!\!\!\!\!\rho\ \left(\lambda^1, \ldots, \lambda^n\right)$. However, if the transformation $B_\phi^1 = g_1(\lambda^1, A_\phi^1), \ldots, B_\phi^n = g_n(\lambda^n, A_\phi^n)$, is reversible, that is, there exist another transformation $A_\phi^1 = h_1(\lambda^1, B_\phi^1), \ldots, A_\phi^n = h_n(\lambda^n, B_\phi^n)$

back to the original variables, then $\left(A^1, \ldots, A^n\right) \leftarrow\!\!\!\rho \left(\lambda^1, \ldots, \lambda^n\right)$ if and only if $\left(B^1, \ldots, B^n\right) \leftarrow\!\!\!\rho \left(\lambda^1, \ldots, \lambda^n\right)$.

Example 1.22 Consider the random variables A^1, A^2 with values in $\{1, 2\}$, depending on the input factors $\lambda^1 \in \{1, 2\}$, $\lambda^2 \in \{1, 2\}$, and having the following joint distributions at the four possible treatments:

$\lambda^1 = 1, \lambda^2 = 1$	$A^2 = 1$	$A^2 = 2$	
$A^1 = 1$	0.3	0.4	0.7
$A^1 = 2$	0.1	0.2	0.3
	0.4	0.6	

$\lambda^1 = 1, \lambda^2 = 2$	$A^2 = 1$	$A^2 = 2$	
$A^1 = 1$	0.35	0.35	0.7
$A^1 = 2$	0.15	0.15	0.3
	0.5	0.5	

$\lambda^1 = 2, \lambda^2 = 1$	$A^2 = 1$	$A^2 = 2$	
$A^1 = 1$	0.32	0.48	0.8
$A^1 = 2$	0.08	0.12	0.2
	0.4	0.6	

$\lambda^1 = 2, \lambda^2 = 2$	$A^2 = 1$	$A^2 = 2$	
$A^1 = 1$	0.45	0.35	0.8
$A^1 = 2$	0.05	0.15	0.2
	0.5	0.5	

We will see in the next section that $\left(A^1, A^2\right) \leftarrow\!\!\!\rho \left(\lambda^1, \lambda^2\right)$ is satisfied in this case. Let us define the input value specific transformations $B^1 = g_1(\lambda^1, A^1)$ and $B^2 = g_2(\lambda^2, A^2)$, where

$$g_1(1, \{1, 2\}) = \{+1, -1\}, \quad g_2(1, \{1, 2\}) = \{7, 3\},$$
$$g_1(2, \{1, 2\}) = \{-1, +1\}, \quad g_2(2, \{1, 2\}) = \{3, 7\}.$$

As we see, $A^1 = 1$ is mapped into $B^1 = +1$ or $B^1 = -1$ according as λ^1 is 1 or 2, $A^2 = 1$ is mapped into $B^2 = 7$ or $B^2 = 3$ according as λ^2 is 1 or 2, etc. We obtain the following joint distributions

$\lambda^1 = 1, \lambda^2 = 1$	$B^2 = 7$	$B^2 = 3$	
$B^1 = +1$	0.3	0.4	0.7
$B^1 = -1$	0.1	0.2	0.3
	0.4	0.6	

$\lambda^1 = 1, \lambda^2 = 2$	$B^2 = 7$	$B^2 = 3$	
$B^1 = +1$	0.35	0.35	0.7
$B^1 = -1$	0.15	0.15	0.3
	0.5	0.5	

$\lambda^1 = 2, \lambda^2 = 1$	$B^2 = 7$	$B^2 = 3$	
$B^1 = +1$	0.08	0.12	0.2
$B^1 = -1$	0.32	0.48	0.8
	0.4	0.6	

$\lambda^1 = 2, \lambda^2 = 2$	$B^2 = 7$	$B^2 = 3$	
$B^1 = +1$	0.15	0.05	0.2
$B^1 = -1$	0.35	0.45	0.8
	0.5	0.5	

We know that the transformed variables satisfy $\left(B^1, B^2\right) \leftarrow\!\!\!\rho \left(\lambda^1, \lambda^2\right)$ because $\left(A^1, A^2\right) \leftarrow\!\!\!\rho \left(\lambda^1, \lambda^2\right)$. □

In the subsequent sections we will consider several *tests of selective influences*. Such a test is always a statement whose truth value (whether it is true or false) determines whether a given pattern of selective influences holds or does not hold. The truth value of the test statement must be determinable from the distributions of $\left(A^1_\phi, \ldots, A^n_\phi\right)$ for all allowable ϕ. If its truth implies $\left(A^1, \ldots, A^n\right) \leftarrow\!\!\!\rho \left(\lambda^1, \ldots, \lambda^n\right)$, then the test provides a sufficient condition for

selective influences; if its falsity implies $\left(A^1,\ldots,A^n\right) \not\leftrightarrows \left(\lambda^1,\ldots,\lambda^n\right)$, then the test provides a necessary condition for selective influences. If the test provides both necessary and sufficient condition, it is a criterion.

The distribution of $\left(A^1_\phi,\ldots,A^n_\phi\right)$, if the random variables are known from their observed realizations, cannot be known precisely, because probabilities are never observable. All our tests require that the distributions of $\left(A^1_\phi,\ldots,A^n_\phi\right)$, or at least some parameters thereof, be known precisely. Therefore they can only be applied to empirical observations if the latter are replaced by theoretical distributions. This can be done based on statistical considerations, outside the scope of the tests themselves. In particular, if all sample sizes are sufficiently large, theoretical distributions can be assumed to be so close to the empirical ones that their difference cannot affect the outcome of a test.

As follows from the discussion above, the most basic and obvious test of selective influences is the (complete) *marginal selectivity test*. This is a necessary condition for selective influences: if, at least for one pair of distinct treatments ϕ and ϕ' that include one and the same subvector $\left(\lambda^{i_1},\ldots,\lambda^{i_k}\right)$, the distributions of the k-marginal random variables $\left(A^{i_1}_\phi,\ldots,A^{i_k}_\phi\right)$ and $\left(A^{i_1}_{\phi'},\ldots,A^{i_k}_{\phi'}\right)$ are not the same, then $\left(A^1,\ldots,A^n\right) \not\leftrightarrows \left(\lambda^1,\ldots,\lambda^n\right)$.

1.14 Linear Feasibility Test

In this section we will discuss a test which is both a necessary and sufficient condition for the selective influences in the case when the number of input-output pairs, the set of values of each input, and the set of possible values of each random output are all finite. Let us enumerate, for $k = 1,\ldots,n$, the values of each input λ^k as $1,\ldots,m_k$, and the values of each random output A^k as $1,\ldots,v_k$. In Section 1.12 we discussed the case $n = 2$, $m_1 = m_2 = 2$, and $v_1 = v_2 = 2$. We determined there that the question of whether $\left(A^1,A^2\right) \leftrightarrows \left(\lambda^1,\lambda^2\right)$ translates into a question of whether certain linear equations have a solution subject to certain constraints. We will see that this is the case generally.

The observable distributions of $\left(A^1_\phi,\ldots,A^n_\phi\right)$ are represented by the probabilities of the events that can be described as

$$\left(\overbrace{A^1=a_1,\ldots,A^k=a_k,\ldots,A^n=a_n}; \overbrace{\lambda^1=l_1,\ldots,\lambda^k=l_k,\ldots,\lambda^n=l_n}\right),$$

where $a_k \in \{1,\ldots,v_k\}$ (output values) and $l_k \in \{1,\ldots,m_k\}$ (input values).

Let us form a matrix M whose rows are enumerated (labeled) by all such vectors. We only consider the vectors with allowable treatments,

$$\phi = \left(\lambda^1 = l_1, \ldots, \lambda^k = l_k, \ldots, \lambda^n = l_n\right) \in \Phi.$$

If the number of the allowable treatments is t (between 1 and $m_1 \times \ldots \times m_n$), then the number of the rows in M is $t \times v_1 \times \ldots \times v_n$.

The columns of the matrix M are enumerated (labeled) by the vectors of the form

$$\left(\overbrace{H_1^1 = h_1^1, \ldots H_{m_1}^1 = h_{m_1}^1}, \ldots, \overbrace{H_1^n = h_1^n, \ldots, H_{m_n}^n = h_{m_n}^n}\right),$$

where $h_i^k \in \{1, \ldots, v_k\}$. Such vectors represent events whose probabilities define the distribution of a reduced coupling vector H (if one exists). The number of such events, hence the number of the columns in M is $(v_1)^{m_1} \times \ldots \times (v_n)^{m_n}$ (where the superscripts represent conventional exponents).

We also form a column vector P whose elements are labeled in the same way and in the same order as the rows of the matrix M, and a column vector Q whose elements are labeled in the same way and in the same order as the columns of the matrix M.

Let us now fill in the entries of the vectors P, Q, and the matrix M. The matrix M is Boolean: it is filled with 1's and 0's. Consider a cell (I, J) belonging to the column labeled

$$J = \left(\overbrace{H_1^1 = h_1^1, \ldots H_{m_1}^1 = h_{m_1}^1}, \ldots, \overbrace{H_1^n = h_1^n, \ldots, H_{m_n}^n = h_{m_n}^n}\right)$$

and to the row labeled

$$I = \left(\overbrace{A^1 = a_1, \ldots, A^k = a_k, \ldots, A^n = a_n}; \overbrace{\lambda^1 = l_1, \ldots, \lambda^k = l_k, \ldots, \lambda^n = l_n}\right).$$

In the vector-label J pick the entries

$$H_{l_1}^1 = h_{l_1}^1, \ldots, H_{l_k}^k = h_{l_k}^k, \ldots, H_{l_n}^n = h_{l_n}^n$$

corresponding to the values of $\left(\lambda^1, \ldots, \lambda^n\right)$ indicated in the vector-label I. If

$$\left(h_{l_1}^1, \ldots, h_{l_k}^k, \ldots, h_{l_n}^n\right) = (a_1, \ldots, a_k, \ldots, a_n)$$

then the cell (I, J) should be filled with 1; otherwise its value is 0.

The vector P is filled with the probabilities

$$\Pr\left(A^1 = a_1, \ldots, A^n = a_n\right)_{\phi=(\lambda^1=l_1,\ldots,\lambda^n=l_n)}.$$

For any allowable ϕ, the probabilities across all possible combinations of $(a_1, \ldots, a_n)$ sum to 1. These probabilities are assumed to be known.

The vector Q is filled with the probabilities

$$\Pr\left(H_1^1 = h_1^1, \ldots, H_{m_1}^1 = h_{m_1}^1, \ldots, H_1^n = h_1^n, \ldots, H_{m_n}^n = h_{m_n}^n\right),$$

which sum to 1 across all possible values of $\left(h_1^1, \ldots, h_{m_1}^1, \ldots, h_1^n, \ldots, h_{m_n}^n\right)$. These probabilities are not known, they have to be found or determined not to exist.

Example 1.23 Let us now apply these general definitions to the simplest nontrivial case $n = 2$, $m_1 = m_2 = 2$, $v_1 = v_2 = 2$ considered in Section 1.12. The matrix M filled with binary values is (replacing 0 with "·" for better legibility)

		H_1^1	1	1	1	1	1	1	1	2	2	2	2	2	2	2	2
		H_2^1	1	1	1	1	2	2	2	2	1	1	1	1	2	2	2	2
		H_1^2	1	1	2	2	1	1	2	2	1	1	2	2	1	1	2	2
		H_2^2	1	2	1	2	1	2	1	2	1	2	1	2	1	2	1	2
$\lambda^1 = 1, \lambda^2 = 1$	$A^1 = 1, A^2 = 1$		1	1	·	·	1	1	·	·	·	·	·	·	·	·	·	·
	$A^1 = 1, A^2 = 2$		·	·	1	1	·	·	1	1	·	·	·	·	·	·	·	·
	$A^1 = 2, A^2 = 1$		·	·	·	·	·	·	·	·	1	1	·	·	1	1	·	·
	$A^1 = 2, A^2 = 2$		·	·	·	·	·	·	·	·	·	·	1	1	·	·	1	1
$\lambda^1 = 1, \lambda^2 = 2$	$A^1 = 1, A^2 = 1$		1	·	1	·	1	·	1	·	·	·	·	·	·	·	·	·
	$A^1 = 1, A^2 = 2$		·	1	·	1	·	1	·	1	·	·	·	·	·	·	·	·
	$A^1 = 2, A^2 = 1$		·	·	·	·	·	·	·	·	1	·	1	·	1	·	1	·
	$A^1 = 2, A^2 = 2$		·	·	·	·	·	·	·	·	·	1	·	1	·	1	·	1
$\lambda^1 = 2, \lambda^2 = 1$	$A^1 = 1, A^2 = 1$		1	1	·	·	·	·	·	·	1	1	·	·	·	·	·	·
	$A^1 = 1, A^2 = 2$		·	·	1	1	·	·	·	·	·	·	1	1	·	·	·	·
	$A^1 = 2, A^2 = 1$		·	·	·	·	1	1	·	·	·	·	·	·	1	1	·	·
	$A^1 = 2, A^2 = 2$		·	·	·	·	·	·	1	1	·	·	·	·	·	·	1	1
$\lambda^1 = 2, \lambda^2 = 2$	$A^1 = 1, A^2 = 1$		1	·	1	·	·	·	·	·	1	·	1	·	·	·	·	·
	$A^1 = 1, A^2 = 2$		·	1	·	1	·	·	·	·	·	1	·	1	·	·	·	·
	$A^1 = 2, A^2 = 1$		·	·	·	·	1	·	1	·	·	·	·	·	1	·	1	·
	$A^1 = 2, A^2 = 2$		·	·	·	·	·	1	·	1	·	·	·	·	·	1	·	1

The vector P consists of the observed probabilities corresponding to the row labels of the matrix, and the vector Q consists of the joint probabilities of the coupling vector $H = \left(H_1^1, H_2^1, H_1^2, H_2^2\right)$ as indicated in the column labels of the matrix. Using the observed probabilities of Example 1.22 we obtain

$$P = [.3, .4, .1, .2, .35, .35, .15, .15, .08, .12, .32, .48, .15, .05, .35, .45]^T.$$

□

Theorem 1.8 *If the sets of values for all n inputs and all n random*

outputs are finite, then, using the notation of this section,

$$\left(A^1, \ldots, A^n\right) \nleftrightarrow \left(\lambda^1, \ldots, \lambda^n\right)$$

holds if and only if the system of linear equations

$$MQ = P$$

has a solution $Q \geq 0$ *(the inequality meaning that the elements of* Q *are non-negative).*

Without the non-negativity constraint, the system $MQ = P$ always has solutions, because the number of the unknowns (elements of Q) equals or exceeds the rank of the matrix M, which can be shown to never exceed

$$\left(m_1 \left(v_1 - 1\right) + 1\right) \times \ldots \times \left(m_n \left(v_n - 1\right) + 1\right).$$

Moreover, the structure of the matrix M is such that that any solution for Q should automatically have its elements summing to 1. The latter therefore is not a constraint. However, it is not guaranteed that $Q \geq 0$: it is possible that all solutions for Q have some of the elements negative, in which case our test establishes that $\left(A^1, \ldots, A^n\right) \not\nleftrightarrow \left(\lambda^1, \ldots, \lambda^n\right)$.

Let us introduce a function

$$\mathrm{Sol}\left(M, P\right)$$

that attains two values: "True," if $MQ = P$ has a non-negative solution, and "False," if such a solution does not exist. Note that M is an argument that is determined uniquely by the format of the problem: the number of input-output pairs and number of possible values for inputs and outputs. The task of computing $\mathrm{Sol}\left(M, P\right)$ is a standard *feasibility problem* of the area of linear algebra called linear programming. Due to this term, the test in question is called the *linear feasibility test*,

$$\left(A^1, \ldots, A^n\right) \nleftrightarrow \left(\lambda^1, \ldots, \lambda^n\right) \text{ if and only if } \mathrm{Sol}\left(M, P\right) = \text{True}.$$

It is known from linear programming that $\mathrm{Sol}\left(M, P\right)$ can always be computed.

Example 1.24 Let us apply the linear feasibility test to the matrix M and vector P of Example 1.23. Using the simplex linear programming algorithm, we obtain the solution

$$Q = [.03, 0, 0, 0, 0, .27, .32, .08, 0, .05, .12, 0, 0, .05, .03, .05]^T \geq 0$$

satisfying $MQ = P$. This means that $\mathrm{Sol}(M, P) =$ "True", hence $\left(A^1, A^2\right) \nleftrightarrow \left(\lambda^1, \lambda^2\right)$.

The (complete) marginal selectivity test mentioned in the previous section is part of the linear feasibility test. If the former is violated, so will also the latter. It follows from the structure of the matrix M, as explained in the following example.

Example 1.25 Consider the matrix of Example 1.23. If $\left(A^1, A^2\right) \nleftrightarrow \left(\lambda^1, \lambda^2\right)$ is satisfied for a given vector P of observed probabilities, then we know that there exists a vector $Q \geq 0$ such that $MQ = P$. The marginal probabilities of A^1 and A^2 within each treatment are obtained by summing certain elements of P. However, as $MQ = P$, we can obtain these marginal probabilities also by summing certain rows of M and then multiplying these summed rows by Q. Thus, if we sum the rows of M corresponding to the same value of A^1 within each treatment, we obtain

$\lambda^1 = 1, \lambda^2 = 1$	$A^1 = 1$	1	1	1	1	1	1	1	1	·	·	·	·	·	·	·	·
	$A^1 = 2$	·	·	·	·	·	·	·	·	1	1	1	1	1	1	1	1
$\lambda^1 = 1, \lambda^2 = 2$	$A^1 = 1$	1	1	1	1	1	1	1	1	·	·	·	·	·	·	·	·
	$A^1 = 2$	·	·	·	·	·	·	·	·	1	1	1	1	1	1	1	1
$\lambda^1 = 2, \lambda^2 = 1$	$A^1 = 1$	1	1	1	1	·	·	·	·	1	1	1	1	·	·	·	·
	$A^1 = 2$	·	·	·	·	1	1	1	1	·	·	·	·	1	1	1	1
$\lambda^1 = 2, \lambda^2 = 2$	$A^1 = 1$	1	1	1	1	·	·	·	·	1	1	1	1	·	·	·	·
	$A^1 = 2$	·	·	·	·	1	1	1	1	·	·	·	·	1	1	1	1

As the rows corresponding to the marginal probabilities of A^1 are identical between the treatments with $\lambda^1 = 1$ and between the treatments with $\lambda^1 = 2$, we see that the marginal distribution of A^1 does not depend on λ_2. If we then sum the rows of M corresponding to the same value of A^2 within each treatment, we obtain

$\lambda^1 = 1, \lambda^2 = 1$	$A^2 = 1$	1	1	·	·	1	1	·	·	1	1	·	·	1	1	·	·
	$A^2 = 2$	·	·	1	1	·	·	1	1	·	·	1	1	·	·	1	1
$\lambda^1 = 1, \lambda^2 = 2$	$A^2 = 1$	1	·	1	·	1	·	1	·	1	·	1	·	1	·	1	·
	$A^2 = 2$	·	1	·	1	·	1	·	1	·	1	·	1	·	1	·	1
$\lambda^1 = 2, \lambda^2 = 1$	$A^2 = 1$	1	1	·	·	1	1	·	·	1	1	·	·	1	1	·	·
	$A^2 = 2$	·	·	1	1	·	·	1	1	·	·	1	1	·	·	1	1
$\lambda^1 = 2, \lambda^2 = 2$	$A^2 = 1$	1	·	1	·	1	·	1	·	1	·	1	·	1	·	1	·
	$A^2 = 2$	·	1	·	1	·	1	·	1	·	1	·	1	·	1	·	1

and we can see that the marginal distribution of A^2 does not depend on λ^1. Thus, linear feasibility test includes the test for marginal selectivity, so if the latter is violated, the former fails. □

One may feel that $\mathrm{Sol}\,(M, P)$ is not a "true" function, as it requires a computer algorithm to be computed, and it is not presented in an analytic form. Such a misgiving is not well-founded. An analytic (or closed-form) solution is merely one that can be presented in terms of familiar functions

and operations. For example, if a solution of a problem involves the standard normal integral

$$N(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{t} \exp\left(-z^2/2\right) \mathrm{d}z,$$

the solution may or may not be called analytic depending on how familiar and easily computable this function is. In the past, $N(t)$ could be viewed as "less analytic" than $\exp(x)$, and in Napier's time $\exp(x)$ would be viewed as "less analytic" than x^2. Familiarity is not a mathematical category, and the existence of a rigorous definition of a function combined with an algorithm allowing one to compute it to a desired level of precision is all one needs to use it in a solution to a problem. The computational complexity, of course, may be a concern. In our case, however, it is known that as the size of the matrix M increases, the computational time required to compute $\mathrm{Sol}(M, P)$ increases only as a polynomial function of this size (rather than exponentially or even faster). This makes the linear feasibility test practical.

It still may be of interest to see whether the linear feasibility test could be formulated in terms of a system of equalities and inequalities involving the entries of the vector P alone. This can always be achieved, with every linear feasibility problem. These equalities and inequalities, in fact, can be generated by a computer algorithm (called a *facet enumeration algorithm*).

Example 1.26 Geometrically, the linear feasibility test checks if P is within the *convex polytope* determined by points MQ such that $Q \geq 0$, $\sum Q = 1$. The columns of M correspond to the vertices of this polytope. A facet enumeration algorithm transforms this *vertex representation* of the polytope to the so-called *half-plane representation*, that is, to a representation of the form

$$M_1 P \geq Q_1,\ M_2 P = Q_2,$$

where M_1, M_2 are matrices and Q_1, Q_2 are vectors. For our 16×16 example matrix, this yields

$$M_1 = \begin{bmatrix}
1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & -1 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & -1 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & -1 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & -1 & 0 & 0 & 0 \\
1 & 1 & 1 & 0 & -1 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\
1 & 1 & 0 & 1 & -1 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 1 & 0 & 0 & 0 \\
1 & 0 & 1 & 1 & 0 & 0 & -1 & 0 & -1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\
0 & 1 & 1 & 1 & 0 & 0 & -1 & 0 & 0 & -1 & 0 & 0 & 1 & 0 & 0 & 0
\end{bmatrix} \qquad Q_1 = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

and

$$M_2 = \begin{bmatrix} 1 & 1 & 0 & 0 & -1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & -1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & -1 & -1 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & -1 & 0 & -1 & 0 & -1 & -1 & 0 & 0 & 1 & 0 & 0 & -1 \end{bmatrix} \qquad Q_2 = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 0 \end{bmatrix}.$$

The equations $M_2P = Q_2$ of this representation always correspond to the marginal selectivity constraints. Thus, a vector P of observed probabilities satisfying marginal selectivity satisfies selective influences if and only if $M_1P \geq Q_1$. Assuming marginal selectivity, $M_1P \geq Q_1$ can in this case also be simplified into the four double-inequalities

$$0 \leq p_{i\cdot} + p_{\cdot j} + p_{i'j'} - p_{ij'} - p_{i'j'} - p_{i'j} \leq 1, \qquad i \neq i',\, j \neq j',$$

where we denote

$$\begin{aligned} p_{i\cdot} &= \Pr(A^1 = 1)_{\phi=(\lambda^1=i,\lambda^2=\cdot)}, \\ p_{\cdot j} &= \Pr(A^2 = 1)_{\phi=(\lambda^1=\cdot,\lambda^2=j)}, \\ p_{ij} &= \Pr(A^1 = 1, A^2 = 1)_{\phi=(\lambda^1=i,\lambda^2=j)} \end{aligned}$$

(the definition of $p_{i\cdot}$ and $p_{\cdot j}$ presupposes marginal selectivity). These are known as the *Bell/CHSH/Fine inequalities* in quantum mechanics. □

In the same way, the representation as inequalities can be obtained for any linear feasibility test matrix M. It should be noted, however, that the number of the inequalities increases explosively as the size of the matrix M increases. Thus, for three pairs of completely crossed binary inputs and three binary random outputs, the number of independent equalities representing marginal selectivity is 42, and the number of inequalities is 53792. From a practical point of view, therefore, computing $\mathrm{Sol}\,(M, P)$ directly is a better approach in all but the simplest cases.

1.15 Distance Tests

Let us establish some general terminology. A *pseudo-quasi-metric* (or *p.q.-metric*, for short) on a nonempty set X is defined as a function $d : X \times X \to \mathbb{R}^+$ (set of non-negative real numbers), such that, for any $x, y, z \in X$,

(1) (zero property) $d\,(x, x) = 0$,

(2) (triangle inequality) $d(x,y) + d(y,z) \geq d(x,z)$.

A p.q.-metric that satisfies, in addition,

(3) (symmetry) $d(x,y) = d(y,x)$,

is called a *pseudo-metric.* A p.q.-metric that satisfies

(4) (positivity) if $x \neq y$, then $d(x,y) > 0$,

is called a *quasi-metric.* Finally, a p.q.-metric that satisfies both (3) and (4) is called a *metric.* The terminology is not well-established and varies from one area or application to another.

Remark 1.11 To refer to the value $d(x,y)$ of a metric, pseudo-metrics, quasi-metrics, or a p.q.-metric at a specific pair of points (x,y), one usually uses the generic term "distance," adding the corresponding prefixes (pseudo, quasi, or p.q.) only if it is required for disambiguation. Thus, the value of a p.q.-metric for a specific pair (x,y) can be called the distance from x to y, or the p.q.-distance from x to y. (For pseudo-metrics, "from x to y" can be replaced with "between x and y.") The term "distance" can also be used (with or without the prefixes) to refer to the functions themselves. Therefore "p.q.-metric tests" below can also be referred to as "distance tests" or "p.q.-distance tests."

The nature of the set X in the definition is entirely arbitrary. We are interested in a set of jointly distributed random variables, that is, those representable as functions of one and the same random variable. A p.q.-metric on such a set is a function d mapping pairs of random variables into non-negative real numbers, such that $d(R,R) = 0$ and $d\left(R^1,R^2\right) + d\left(R^2,R^3\right) \geq \mathrm{d}\left(R^1,R^3\right)$, for any random variables R^1,R^2,R^3 in the set. We assume that $d\left(R^1,R^2\right)$ is entirely determined by the joint distribution of $\left(R^1,R^2\right)$. In other words, it does not depend on the identifying label of the pair (or on how R^1 and R^2 are presented as functions of a common random variable).

An immediate consequence (and generalization) of the triangle inequality is the following *chain inequality*: if $R^1,\ldots,R^l$ are elements of X ($l \geq 3$), not necessarily distinct, then

$$d\left(R^1,R^l\right) \leq \sum_{i=2}^{l} d\left(R^{i-1},R^i\right).$$

This inequality, as it turns out, can be utilized to construct tests of selective influences.

Suppose that the random outputs $\left(A_\phi^1, \ldots, A_\phi^n\right)$ across all $\phi \in \Phi$ belong to a certain type, or class of random variables (e.g., those in the narrow sense, or with a finite number of values, etc.). We continue to consider, for simplicity, inputs with finite number of values each. We know that $\left(A^1, \ldots, A^n\right) \leftharpoonup \left(\lambda^1, \ldots, \lambda^n\right)$ if and only if there exists a reduced coupling vector H. Assuming that it does exist, its elements are of the same type, or class, as $\left(A_\phi^1, \ldots, A_\phi^n\right)$, and any $l \geq 3$ of these elements,

$$H_{j_1}^{k_1}, H_{j_2}^{k_2}, \ldots, H_{j_l}^{k_l}$$

can be used to form a chain inequality,

$$d\left(H_{j_1}^{k_1}, H_{j_l}^{k_l}\right) \leq \sum_{i=2}^{l} d\left(H_{j_{i-1}}^{k_{i-1}}, H_{j_i}^{k_i}\right).$$

Let us choose these elements of H so that $\lambda^{k_1} = j_1$ and $\lambda^{k_l} = j_l$ belong to some allowable treatment ϕ_{1k}, and each pair $\lambda^{k_{i-1}} = j_{i-1}, \lambda^{k_i} = j_i$ belongs to some allowable treatment $\phi_{i-1,i}$ $(i = 2, \ldots, l)$. The allowable treatments $\phi_{1k}, \phi_{12}, \ldots, \phi_{l-1,l}$ need not be pairwise distinct. Such a sequence of input values,

$$\lambda^{k_1} = j_1, \lambda^{k_2} = j_2, \ldots, \lambda^{k_l} = j_l$$

is called *treatment-realizable*. This choice ensures that

$$\left(H_{j_1}^{k_1}, H_{j_l}^{k_l}\right) \sim \left(A_{\phi_{1k}}^{k_1}, A_{\phi_{1l}}^{k_l}\right)$$

and

$$\left(H_{j_{i-1}}^{k_{i-1}}, H_{j_i}^{k_i}\right) \sim \left(A_{\phi_{i-1,i}}^{k_{i-1}}, A_{\phi_{i-1,i}}^{k_i}\right), \text{ for } i = 2, \ldots, l.$$

But then

$$d\left(H_{j_1}^{k_1}, H_{j_l}^{k_l}\right) = d\left(A_{\phi_{1k}}^{k_1}, A_{\phi_{1l}}^{k_l}\right)$$

and

$$d\left(H_{j_{i-1}}^{k_{i-1}}, H_{j_i}^{k_i}\right) = d\left(A_{\phi_{i-1,i}}^{k_{i-1}}, A_{\phi_{i-1,i}}^{k_i}\right), \text{ for } i = 2, \ldots, l,$$

whence the chain inequality can be rewritten using only observable pairwise distributions,

$$d\left(A_{\phi_{1k}}^{k_1}, A_{\phi_{1l}}^{k_l}\right) \leq \sum_{i=2}^{l} d\left(A_{\phi_{i-1,i}}^{k_{i-1}}, A_{\phi_{i-1,i}}^{k_i}\right).$$

This inequality is a necessary condition for the existence of H. If it is found violated for at least one treatment-realizable sequence of input values, then

the existence of H is ruled out, and one should conclude that $\left(A^1, \ldots, A^n\right) \not\wp \left(\lambda^1, \ldots, \lambda^n\right)$.

There are numerous ways of constructing p.q.-metrics for jointly distributed random variables. We will confine our consideration to only two examples.

If all random outputs have one and the same set of possible values S, then one way of creating a p.q.-metric on a set X of such random variables is to use any p.q.-metric D on S and put, for any random variables $Q, R \in X$,

$$d(Q, R) = \mathrm{E}\left[D(Q, R)\right].$$

The right-hand expression is the expected value of the random variable $D(Q, R)$. The underlying assumption is, of course, that this random variable is well-defined (that is, D is a measurable function from $S \times S$ to non-negative real numbers), and that its expectation is finite. It can easily be proved then that d is a p.q.-metric on X.

As a simple example, consider the p.q.-metric

$$D(x, y) = \begin{cases} |x - y|^p & \text{if } x < y \\ 0 & \text{otherwise} \end{cases}$$

on the set of real numbers, with $0 \leq p \leq 1$ (a power exponent). It is a p.q.-metric because $D(x, x) = 0$, and

$$D(x, y) + D(y, z) \geq D(x, z),$$

as one can prove by considering various arrangements of numbers x, y, z. Using D one can construct a p.q.-metric for any set X of random variables whose (common) set of possible values is a subset of reals. Let this set be a subset of integers. Then the p.q.-metric on X derived from D is

$$d_p(Q, R) = \sum_{q<r} |q - r|^p \, p(q, r),$$

where

$$p(q, r) = \Pr(Q = q, R = r).$$

Example 1.27 Let the outputs A^1, A^2 have the following distributions for

treatments in $\Lambda^1 \times \Lambda^2 = \{1,2\} \times \{1,2\}$:

$\lambda^1 = 1$, $\lambda^2 = 1$

	$A^2=0$	$A^2=1$	$A^2=2$
$A^1=0$	.24	.07	0
$A^1=2$	.07	.24	.07
$A^1=4$	0	.07	.24

$\lambda^1 = 1$, $\lambda^2 = 2$

	$A^2=0$	$A^2=1$	$A^2=2$
$A^1=0$	.24	.07	0
$A^1=2$	.07	.24	.07
$A^1=4$	0	.07	.24

$\lambda^1 = 2$, $\lambda^2 = 1$

	$A^2=0$	$A^2=1$	$A^2=2$
$A^1=0$	.24	.07	0
$A^1=2$	.07	.24	.07
$A^1=4$	0	.07	.24

$\lambda^1 = 2$, $\lambda^2 = 2$

	$A^2=0$	$A^2=1$	$A^2=2$
$A^1=0$	0	.07	.24
$A^1=2$	.07	.24	.07
$A^1=4$	.24	.07	0

Let us put $p = 1$ and compute the values of the d_1-p.q.-metric. For any λ^1, λ^2 here,

$$d_1(A^1, A^2) = \sum_{a_1<a_2} |a_1 - a_2|^1 \, p\,(a_1, a_2) = |1-0|^1 \, p\,(0,1) + |2-0|^1 \, p\,(0,2)$$

and

$$\begin{aligned} d_1(A^2, A^1) &= \sum_{a_2<a_1} |a_1 - a_2|^1 \, p\,(a_1, a_2) \\ &= |2-0|^1 \, p\,(2,0) + |4-0|^1 \, p\,(4,0) + \ldots + |4-2|^1 \, p\,(4,2)\,. \end{aligned}$$

The calculations yield the following distances:

	$\lambda^1=1,\ \lambda^2=1$	$\lambda^1=1,\ \lambda^2=2$	$\lambda^1=2,\ \lambda^2=1$	$\lambda^1=2,\ \lambda^2=2$
$d_{p=1}(A^1, A^2)$	.07	.07	.07	.55
$d_{p=1}(A^2, A^1)$	1.07	1.07	1.07	1.55

Using this table, all possible distance test inequalities are of the form $a \leq b + c + d$, where a, b, and d belong to one row and c to another, provided all four values are in distinct columns. It is easy to see that all the inequalities are passed. □

P.q.-metrics can be introduced directly in probabilistic terms rather derived from "deterministic" metrics on sets of possible values. Consider, as an example, the following construction. Let $\left(S^1, \Sigma^1\right), \ldots, \left(S^m, \Sigma^m\right)$ be the sets of possible values with sigma-algebras for random variables $R^1, \ldots, R^m$, respectively, and let us partition each S^k into $l_k > 1$ measurable subsets $S^{1k}, \ldots, S^{l_k k} \in \Sigma^k$. It follows that the joint probabilities of any pair $S^{ik}, S^{i'k'}$,

$$\Pr\left(R^k \in S^{ik}, R^{k'} \in S^{i'k'}\right),$$

are well defined. It can easily be proved that the function

$$d_{class}\left(R^k, R^{k'}\right) = \sum_{i<i'} \Pr\left(R^k \in S^{ik}, R^{k'} \in S^{i'k'}\right)$$

is a p.q.-metric. It is called a *classification p.q.-metric*, and it can be applied to all types of random variables without restrictions.

Example 1.28 Consider the case with two real-valued random variables R^1, R^2 and define the partition of $S^1 = \mathbb{R}$ and $S^2 = \mathbb{R}$ as, respectively,

$$S^{11} = (-\infty, x),\ S^{21} = [x, \infty)$$

and

$$S^{12} = (-\infty, y),\ S^{22} = [y, \infty).$$

Then, the classification distance is simply

$$d_{class}(R^1, R^2) = \Pr\left(R^1 \in S^{11}, R^2 \in S^{22}\right) = \Pr\left(R^1 < x, R^2 \geq y\right).$$

Different choices of x, y give us different classification distances. □

Remark 1.12 A classification p.q.-metric can also be viewed as a limit case of the metric d_p introduced above, provided we first map by a measurable function f_k each S^k into a set $\{1, \ldots, l_k\}$, and then define all the transformed random variables $f_k\left(R^k\right)$ as distributed on $\{1, \ldots, l\}$, with $l = \max(l_1, \ldots, l_m)$. The latter is always possible by assigning to the "redundant" integers probability zero. Following this transformation and equalization of domains, d_{class} is obtained as $d_{p=0}$. Another way of introducing the classification metric is as a special case of an *order-distance*. Without elaborating, the latter involves a relation of strict order $\prec$ between values of one random variable and values of another. The order-distance is defined as

$$d_{ord}(Q, R) = \Pr(Q \prec R).$$

Recall that a sequence $\lambda^{k_1} = j_1, \lambda^{k_2} = j_2, \ldots, \lambda^{k_l} = j_l$ of input values is treatment-realizable if $\left\{\lambda^{k_1} = j_1, \lambda^{k_k} = j_k\right\}$ and $\left\{\lambda^{k_{i-1}} = j_{i-1}, \lambda^{k_i} = j_i\right\}$ for $i = 2, \ldots, l$ belong to allowable treatments. If the elements of all these pairs are distinct, and if these pairs are the only subsequences of more than one element that have the property of being a subset of an allowable treatment, then the sequence is called *irreducible*. It turns out that one only has to check the chain inequalities for irreducible sequences: these inequalities are satisfied for all treatment-realizable sequences if and only if they are satisfied for all irreducible ones.

The set of irreducible sequences may be significantly smaller than the set of all treatment-realizable sequences. Thus, it can be shown that if the set Φ consists of all possible combinations of input values, then the only irreducible sequences are quadruples of the form

$$\lambda^k = j_1, \lambda^{k'} = j_2, \lambda^k = j_3, \lambda^{k'} = j_4,$$

with $k \neq k'$, $j_1 \neq j_3$ and $j_2 \neq j_4$. The only inequalities to check then are of the form,

$$d\left(A^k_{\phi_{14}}, A^{k'}_{\phi_{14}}\right) \leq d\left(A^k_{\phi_{12}}, A^{k'}_{\phi_{12}}\right) + d\left(A^{k'}_{\phi_{23}}, A^k_{\phi_{23}}\right) + d\left(A^k_{\phi_{34}}, A^{k'}_{\phi_{34}}\right),$$

where $\phi_{14}, \phi_{12}, \phi_{23}, \phi_{34}$ are any allowable treatments that contain, respectively,

$$\left\{\lambda^k = j_1, \lambda^{k'} = j_4\right\},\ \left\{\lambda^k = j_1, \lambda^{k'} = j_2\right\},$$
$$\left\{\lambda^{k'} = j_2, \lambda^k = j_3\right\},\ \left\{\lambda^k = j_3, \lambda^{k'} = j_4\right\}.$$

1.16 (Non)Invariance of tests with respect to transformations

In this section we introduce another class of tests of selective influences, called *cosphericity tests*. Prior to introducing them, however, we should discuss an important issue.

We know from Section 1.13 that if $\left(A^1, \ldots, A^n\right) \leftarrow\!\!\!\!\!\rho\ \left(\lambda^1, \ldots, \lambda^n\right)$, then $\left(B^1, \ldots, B^n\right) \leftarrow\!\!\!\!\!\rho\ \left(\lambda^1, \ldots, \lambda^n\right)$, where the B's are input-value-specific transformations of the A's, that is,

$$B^1_\phi = g_1\left(\lambda^1, A^1_\phi\right), \ldots, B^n_\phi = g_n\left(\lambda^n, A^n_\phi\right),$$

for all $\phi = \left(\lambda^1, \ldots, \lambda^n\right) \in \Phi$. It follows that if a test provides a necessary condition for selective influences, then its failure for any of the input-value-specific transformations of $\left(A^1, \ldots, A^n\right)_\phi$ establishes $\left(A^1, \ldots, A^n\right) \not\leftarrow\!\!\!\!\!\rho\ \left(\lambda^1, \ldots, \lambda^n\right)$. If the outcome of a test is not invariant with respect to some of such transformations, this consideration automatically expands this test into a multitude of tests, one for each of these transformations. This may enormously increase the ability of a test to detect violations of selective influences. This might sound paradoxical, or at least unexpected, but this is generally true for any test that provides a necessary but not sufficient condition for a tested proposition: the lack of invariance in the test's outcome with respect to transformations that preserve the tested proposition is an advantage rather than a drawback.

Remark 1.13 If a test provides a sufficient condition for $\left(A^1,\ldots,A^n\right) \hookleftarrow\!\!\!\!\!\rho\ \left(\lambda^1,\ldots,\lambda^n\right)$, and it is not invariant with respect to input-value-specific transformations, then one should apply it to a variety of $\left(B^1,\ldots,B^n\right)_\phi$ from which $\left(A^1,\ldots,A^n\right)_\phi$ can be obtained by such a transformation. At the time this is written (end of 2012), we do not have nontrivial tests that provide sufficient but not necessary conditions. If a test is a criterion when applied to input-output pairs of a particular type, then its (non)invariance with respect to transformations is immaterial for establishing or rejecting selective influences for original random variables (although transformed ones may be of interest for their own sake).

Of the two distance tests considered in the previous section, d_p-test is not invariant (for any fixed p) with respect to numerical transformations of the random outputs.

Example 1.29 Continuing Example 1.27, let us transform the outputs A^1, A^2 as $B^1 = g_1(A^1)$, $B^2 = g_2(A^2)$, where g_1 is given by $0 \mapsto 2$, $2 \mapsto 1$, $4 \mapsto 1$ and g_2 is given by $0 \mapsto 2$, $1 \mapsto 1$, $2 \mapsto 1$. We get the joint distributions

$\lambda^1=1,\lambda^2=1$	$B^2=1$	$B^2=2$
$B^1=1$	.62	.07
$B^1=2$	.07	.24

$\lambda^1=1,\lambda^2=2$	$B^2=1$	$B^2=2$
$B^1=1$	.62	.07
$B^1=2$	.07	.24

$\lambda^1=2,\lambda^2=1$	$B^2=1$	$B^2=2$
$B^1=1$	.62	.07
$B^1=2$	.07	.24

$\lambda^1=2,\lambda^2=2$	$B^2=1$	$B^2=2$
$B^1=1$	.38	.31
$B^1=2$	.31	0

and the corresponding $d_{p=1}$ distances are

	$\lambda^1=1,\ \lambda^2=1$	$\lambda^1=1,\ \lambda^2=2$	$\lambda^1=2,\ \lambda^2=1$	$\lambda^1=2,\ \lambda^2=2$
$d_{p=1}(B^1,B^2)$	.07	.07	.07	.31
$d_{p=1}(B^2,B^1)$	.07	.07	.07	.31

Now the distance test inequality $.31 \leq .07 + .07 + .07 = .21$ fails implying $\left(B^1,B^2\right) \not\hookleftarrow\!\!\!\!\!\rho\ \left(\lambda^1,\lambda^2\right)$ which in turn implies $\left(A^1,A^2\right) \not\hookleftarrow\!\!\!\!\!\rho\ \left(\lambda^1,\lambda^2\right)$. Thus, the d_p-test is not invariant with respect to transformations of the variables. □

The second distance test considered in the previous section, d_{class}-test, is invariant (for any given partition scheme) with respect to any transformations of the possible values of random outputs. The obvious proviso for this statement is that a transformed value is always classified into a partition with the same number as the original value. If this proviso is violated, it would amount to changing the partition scheme for the original outputs. The power of the d_{class}-test to detect violations of selective influences does not come from different transformations. Rather it comes from complete

flexibility in the partitioning scheme. Another way of looking at this test (see Remark 1.12) is that a transformation of the random outputs (different mappings into natural numbers) is built into the identity of the test. If the transformation changes, we apply a different test.

Example 1.30 Consider the system (A^1, A^2) of Example 1.27. Let us partition S^1 into $S^{11} = \{0\}$, $S^{21} = \{2, 4\}$, and S^2 into $S^{12} = \{0, 1\}$, $S^{22} = \{2\}$. We obtain the following joint probabilities for the partition memberships $A^k \in S^{ik}$:

$\lambda^1 = 1, \lambda^2 = 1$	$A^2 \in S^{21}$	$A^2 \in S^{22}$
$A^1 \in S^{11}$	.31	0
$A^1 \in S^{12}$	.38	.31

$\lambda^1 = 1, \lambda^2 = 2$	$A^2 \in S^{21}$	$A^2 \in S^{22}$
$A^1 \in S^{11}$	.31	0
$A^1 \in S^{12}$	.38	.31

$\lambda^1 = 2, \lambda^2 = 1$	$A^2 \in S^{21}$	$A^2 \in S^{22}$
$A^1 \in S^{11}$	.31	0
$A^1 \in S^{12}$	.38	.31

$\lambda^1 = 2, \lambda^2 = 2$	$A^2 \in S^{21}$	$A^2 \in S^{22}$
$A^1 \in S^{11}$	.07	.24
$A^1 \in S^{12}$	.62	.07

This yields the classification distances

	$\lambda^1 = 1, \lambda^2 = 1$	$\lambda^1 = 1, \lambda^2 = 2$	$\lambda^1 = 2, \lambda^2 = 1$	$\lambda^1 = 2, \lambda^2 = 2$
$d_{class}(A^1, A^2)$	0	0	0	.24
$d_{class}(A^2, A^1)$	.38	.38	.38	.62

which can be seen to satisfy all distance test inequalities, as in Example 1.27.

Consider now the partitioning of S^1 into $S^{11} = \{0, 2\}$, $S^{21} = \{4\}$, and of S^2 into $S^{12} = \{0, 1\}$, $S^{22} = \{2\}$. The partition membership indicator B^k (given by $B^k = i$ when $A^k \in S^{ik}$) corresponds to the transformed variables B^k of Example 1.29. As a result, we get the same joint distribution tables as there. We know that d_{class} corresponds to $d_{p=0}$ (see Remark 1.12), and it is easy to see that $d_{p=0}$ is identical to $d_{p=1}$ when the sets are partitioned into only two classes each. Therefore d_{class} distance table we obtain is identical to the $d_{p=1}$ table shown in Example 1.29, and we conclude that the d_{class}-distance test fails, implying $\left(A^1, A^2\right) \not\leftrightarrow \left(\lambda^1, \lambda^2\right)$. □

We conclude this section by presenting a test based on pairwise correlation between random outputs. It is called the *cosphericity test*, and confined to random variables for which conventional correlations can be computed. These are all variables that are defined (or can be redefined) on the set of real numbers with the Lebesgue sigma-algebra. Discrete random variables can always be redefined to fall within this category.

The primary application of the cosphericity test is to two input-output pairs, with two values per input, and all four treatments allowable. That is, we test the assumption $\left(A^1, A^2\right) \leftrightarrow \left(\lambda^1, \lambda^2\right)$, with $\Lambda^1 = \{1, 2\}$, $\Lambda^2 = \{1, 2\}$,

and allowable treatments $\phi_{11} = \left(\lambda_1^1, \lambda_1^2\right)$, $\phi_{12} = \left(\lambda_1^1, \lambda_2^2\right)$, etc. The use of the test for larger designs will be discussed later.

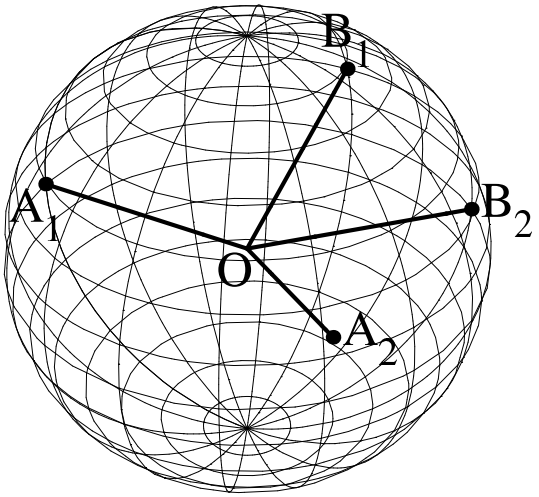
Denote the correlation between $A^1_{\phi_{ij}}$ and $A^2_{\phi_{ij}}$ (as the two are jointly distributed) by ρ_{ij}, $i, j \in \{1, 2\}$. The cosphericity test is the proposition: if $\left(A^1, A^2\right) \leftrightharpoons \left(\lambda^1, \lambda^2\right)$, then

$$|\rho_{11}\rho_{12} - \rho_{21}\rho_{22}| \leq \sqrt{1-(\rho_{11})^2}\sqrt{1-(\rho_{12})^2} + \sqrt{1-(\rho_{21})^2}\sqrt{1-(\rho_{22})^2}.$$

Superscript 2 here indicates squaring. If this inequality is violated, then the initial assumption $\left(A^1, A^2\right) \leftrightharpoons \left(\lambda^1, \lambda^2\right)$ should be rejected.

The explanation for the name "cosphericity" is this: the inequality above holds if and only if one can place four points, $\mathbf{A}_1, \mathbf{A}_2, \mathbf{B}_1, \mathbf{B}_2$, on the surface of a unit sphere (in the Euclidean three-dimensional space) centered at point $\mathbf{O}$, so that

$$\begin{aligned} &\cos\angle\mathbf{A}_1\mathbf{O}\mathbf{B}_1 = \rho_{11}, \quad \cos\angle\mathbf{A}_1\mathbf{O}\mathbf{B}_2 = \rho_{12}, \\ &\cos\angle\mathbf{A}_2\mathbf{O}\mathbf{B}_1 = \rho_{21}, \quad \cos\angle\mathbf{A}_2\mathbf{O}\mathbf{B}_2 = \rho_{22}. \end{aligned}$$



Example 1.31 Consider the following output distributions of A^1, A^2 for

the treatments in $\Lambda^1 \times \Lambda^2 = \{1,2\} \times \{1,2\}$:

$\lambda^1 = 1,\ \lambda^2 = 1$

	$A^2=0$	$A^2=1$	$A^2=5$
$A^1=0$	.24	.07	0
$A^1=1$	.07	.24	.07
$A^1=5$	0	.07	.24

$\lambda^1 = 1,\ \lambda^2 = 2$

	$A^2=0$	$A^2=1$	$A^2=5$
$A^1=0$	.24	.07	0
$A^1=1$	.07	.24	.07
$A^1=5$	0	.07	.24

$\lambda^1 = 2,\ \lambda^2 = 1$

	$A^2=0$	$A^2=1$	$A^2=5$
$A^1=0$	.24	.07	0
$A^1=1$	.07	.24	.07
$A^1=5$	0	.07	.24

$\lambda^1 = 2,\ \lambda^2 = 2$

	$A^2=0$	$A^2=1$	$A^2=5$
$A^1=0$	0	.07	.24
$A^1=1$	.07	.24	.07
$A^1=5$	.24	.07	0

The correlations coefficients of the four distributions are $\rho_{11} = \rho_{12} = \rho_{21} \approx .7299$ and $\rho_{22} \approx -.6322$. Substituting these in the cosphericity test, we obtain

$$\begin{aligned} .9942 &\approx |.7299 \cdot .7299 - .7299(-.6322)| \\ &\leq \sqrt{1-.7299^2}\sqrt{1-.7299^2} + \sqrt{1-.7299^2}\sqrt{1-.6322^2} \approx .9969, \end{aligned}$$

so the test is passed. □

Correlation between two random variables is not invariant with respect to any but affine transformations of the random variables. This allows one to expand the single cosphericity test into a potential infinity of tests, corresponding to different nonlinear input-value-specific transformations $g_1\left(\lambda^1, A^1_\phi\right)$ and $g_2\left(\lambda^2, A^2_\phi\right)$. An interesting fact is that if, by means of some *reversible* transformations g_1, g_2 the random variables $\left(A^1, A^2\right)_\phi$ can be made bivariate-normally distributed at all four treatments, then the cosphericity test performed on thus transformed random outputs provides both a necessary and sufficient condition for $\left(A^1, A^2\right) \leftrightharpoons \left(\lambda^1, \lambda^2\right)$.

Example 1.32 The system of Example 1.31 passed the coshpericity test. However, if we apply the nonlinear transformation $B^1 = g(A^1)$, $B^2 = g(A^2)$,

where g is given by $0 \mapsto 0$, $1 \mapsto 1$, $5 \mapsto 2$, we get

$\lambda^1 = 1, \lambda^2 = 1$

	$B^2=0$	$B^2=1$	$B^2=2$
$B^1=0$	.24	.07	0
$B^1=1$	.07	.24	.07
$B^2=2$	0	.07	.24

$\lambda^1 = 1, \lambda^2 = 2$

	$B^2=0$	$B^2=1$	$B^2=2$
$B^1=0$	.24	.07	0
$B^1=1$	.07	.24	.07
$B^2=2$	0	.07	.24

$\lambda^1 = 2, \lambda^2 = 1$

	$B^2=0$	$B^2=1$	$B^2=2$
$B^1=0$	.24	.07	0
$B^1=1$	.07	.24	.07
$B^2=2$	0	.07	.24

$\lambda^1 = 2, \lambda^2 = 2$

	$B^2=0$	$B^2=1$	$B^2=2$
$B^1=0$	0	.07	.24
$B^1=1$	.07	.24	.07
$B^2=2$	.24	.07	0

and the correlations for these joint distributions are $\rho_{11} = \rho_{12} = \rho_{21} \approx .7742$ and $\rho_{22} \approx -.7742$. Substituting these in the cosphericity test, we obtain

$$\begin{aligned} 1.1988 &\approx |.7742 \cdot .7742 - .7742(-.7742)| \\ &\leq \sqrt{1-.7742^2}\sqrt{1-.7742^2} + \sqrt{1-.7742^2}\sqrt{1-.7742^2} \approx .8012. \end{aligned}$$

We see that the cosphericity test is not passed for the transformed variables. As a result selective influences are ruled out for the original variables as well. □

The cosphericity test can also be applied to more than two input-output pairs. If we assume that $\left(A^1, \ldots, A^n\right) \leftharpoonup \left(\lambda^1, \ldots, \lambda^n\right)$, then, by the nestedness property for input-output pairs, for any two of them, $\left(A^k, \lambda^k\right)$ and $\left(A^{k'}, \lambda^{k'}\right)$, we should have $\left(A^k, A^{k'}\right) \leftharpoonup \left(\lambda^k, \lambda^{k'}\right)$. The test only applies if there are two values i and i' of λ^k and two values j and j' of $\lambda^{k'}$ such that, for some allowable treatments $\phi_{ij}, \phi_{ij'}, \phi_{i'j}, \phi_{i'j'}$,

$$\lambda^k = i, \lambda^{k'} = j \in \phi_{ij}, \lambda^k = i, \lambda^{k'} = j' \in \phi_{ij'}, \text{ etc.}$$

In other words, the inputs and their values should be chosen so that $\left\{\lambda^k = i, \lambda^k = i'\right\}$ and $\left\{\lambda^{k'} = j, \lambda^{k'} = j'\right\}$ form a completely crossed subdesign within the set of allowable treatments. By the nestedness property for input values, we have $\left(A^k, A^{k'}\right) \leftharpoonup \left(\lambda^k, \lambda^{k'}\right)$ with the input values restricted to $\{i, i'\}$ and $\{j, j'\}$ and the new set of allowable treatments consisting of all four possible combinations. If this cosphericity inequality is violated for all least one combination of k, k', i, i', j, j', then the initial assumption $\left(A^1, \ldots, A^n\right) \leftharpoonup \left(\lambda^1, \ldots, \lambda^n\right)$ should be rejected.

1.17 Conditional determinism and conditional independence of outcomes

The definition of selective influences (in the canonical form) requires the existence of a random variable R and functions $f_1, \ldots, f_n$ such that, for all allowable treatments ϕ,

$$\left(A_\phi^1, \ldots, A_\phi^n\right) \sim \left(f_1\left(\lambda^1, R\right), \ldots, f_n\left(\lambda^n, R\right)\right).$$

One obvious consequence of this definition is that, conditioned on any value r of R, the outputs become (equal to) deterministic functions of the corresponding factors,

$$f_1\left(\lambda^1, r\right), \ldots, f_n\left(\lambda^n, r\right).$$

It is sometimes easy to deal with these deterministic quantities, derive certain inequalities that hold for every value of r, and then show that they are preserved as R randomly varies. It is an especially useful approach if the distributions of $\left(A_\phi^1, \ldots, A_\phi^n\right)$ at allowable treatments ϕ are not known, and instead we know distributions of certain functions of these random variables, such as their sums or maxima.

Let us discuss this on an example from studies of mental architectures. This is a traditional area of psychology dealing with decomposing performance of a task into a network of subprocesses when we only know the distributions of the overall performance time (referred to as response time) at different treatments. Let us assume that we observe response times T in an experiment with two factors, λ^1, λ^2, manipulated at two levels each, denoted in both cases by 1 and 2. All four treatments are allowable. Let us postulate that there are two processes involved, with their durations A^1 and A^2 being random variables, and that $\left(A^1, A^2\right) \leftrightarrows \left(\lambda^1, \lambda^2\right)$. We want to determine which of the three "architectures," or composition schemes, is being employed:

1. serial, $T_\phi = A_\phi^1 + A_\phi^2$
2. parallel-OR, $T_\phi = \min\left(A_\phi^1, A_\phi^2\right)$, or
3. parallel-AND, $T_\phi = \max\left(A_\phi^1, A_\phi^2\right)$.

One tool traditionally used for this purpose is the *interaction contrast*,

$$c(t) = \Pr\left(T_{11} \leq t\right) + \Pr\left(T_{22} \leq t\right) - \Pr\left(T_{12} \leq t\right) - \Pr\left(T_{21} \leq t\right),$$

where t is any non-negative number, and T_{ij} abbreviates $T_{\phi=(i,j)}$.

We do not know the joint distribution of A^1_ϕ, A^2_ϕ at any of the four treatments, but we can write

$$\left(A^1_{ij}, A^2_{ij}\right) \sim \left(f_1\left(\lambda^1 = i, R\right), f_2\left(\lambda^2 = j, R\right)\right) = \left(g^1_i\left(R\right), g^2_j\left(R\right)\right),\ i,j \in \{1,2\}.$$

We need one additional assumption: that R can be chosen in such a way that, for any of its possible values r,

$$g^1_1\left(r\right) \leq g^1_2\left(r\right), \quad g^2_1\left(r\right) \leq g^2_2\left(r\right).$$

In other words, switching either factor from level 1 to level 2 prolongs the corresponding processing time. We call this assumption *prolongation constraints*. Various analogues of this assumption are common in studies of mental architectures.

Deterministic real-valued quantities can be viewed as random variables with Heaviside distribution functions:

$$\Pr\left(g^k_l\left(r\right) \leq t\right) = \begin{cases} 0 & \text{if} \quad t < g^k_l\left(r\right), \\ 1 & \text{if} \quad t \geq g^k_l\left(r\right). \end{cases}$$

Analogously,

$$\Pr\left(\mathrm{comp}\left(g^1_i\left(r\right), g^2_j\left(r\right)\right) \leq t\right) = \begin{cases} 0 & \text{if} \quad t < \mathrm{comp}\left(g^1_i\left(r\right), g^2_j\left(r\right)\right), \\ 1 & \text{if} \quad t \geq \mathrm{comp}\left(g^1_i\left(r\right), g^2_j\left(r\right)\right), \end{cases}$$

where comp stands for one of the three composition rules of interest, plus, maximum, or minimum. This allows us to form the *conditional interaction contrast*,

$$c^*\left(t,r\right) = \Pr\left(t_{11}\left(r\right) \leq t\right) + \Pr\left(t_{22}\left(r\right) \leq t\right) - \Pr\left(t_{12}\left(r\right) \leq t\right) - \Pr\left(t_{21}\left(r\right) \leq t\right),$$

where

$$t_{ij} = \mathrm{comp}\left(g^1_i\left(r\right), g^2_j\left(r\right)\right).$$

It is easy to see that

$$\Pr\left(T_{ij} \leq t\right) = \int_{S_R} \Pr\left(t_{ij}\left(r\right) \leq t\right) \mathrm{d}p_R\left(r\right)$$

and

$$c\left(t\right) = \int_{S_R} c^*\left(t,r\right) \mathrm{d}p_R\left(r\right),$$

where the Lebesgue integral is taken over the entire domain S_R of R, and p_R is the probability measure in the distribution of R. (The reader not familiar with Lebesgue integrals can think of $\mathrm{d}p_R\left(r\right)$ above as a generalized version of $f_R\left(r\right)\mathrm{d}r$, where f_R is the density function of R over the set of real numbers.)

Using this observation we can easily establish that if the composition rule is min (parallel-OR architecture), then $c(t) \leq 0$, for all t, because $c^*(t,r) \leq 0$ at any t and any fixed r. Indeed, consider all possible arrangements of $g_1^1(r), g_2^1(r), g_1^2(r), g_2^2(r)$ keeping in mind the prolongation constraints and assuming, with no loss of generality, that $g_1^1(r) \leq g_1^2(r)$. These possible arrangements are

$$\begin{array}{ll} \text{(i)} & g_1^1(r) \leq g_2^1(r) \leq g_1^2(r) \leq g_2^2(r), \\ \text{(ii)} & g_1^1(r) \leq g_1^2(r) \leq g_2^1(r) \leq g_2^2(r), \\ \text{(iii)} & g_1^1(r) \leq g_1^2(r) \leq g_2^2(r) \leq g_2^1(r). \end{array}$$

Thus, for (ii), we have

$$\begin{array}{l} t_{11}(r) = \min\left(g_1^1(r), g_1^2(r)\right) = g_1^1(r), \\ t_{12}(r) = \min\left(g_1^1(r), g_2^2(r)\right) = g_1^1(r), \\ t_{21}(r) = \min\left(g_2^1(r), g_1^2(r)\right) = g_1^2(r), \\ t_{22}(r) = \min\left(g_2^1(r), g_2^2(r)\right) = g_2^1(r). \end{array}$$

Then, substituting for the numerical values

$$\begin{aligned} c^*(t,r) &= \Pr(t_{11}(r) \leq t) + \Pr(t_{22}(r) \leq t) - \Pr(t_{12}(r) \leq t) - \Pr(t_{21}(r) \leq t) \\ &= \begin{cases} 0+0-0-0=0 & \text{if } t < g_1^1(r), \\ 1+0-1-0=0 & \text{if } g_1^1(r) \leq t < g_1^2(r), \\ 1+0-1-1<0 & \text{if } g_1^2(r) \leq t < g_2^1(r), \\ 1+1-1-1=0 & \text{if } g_2^1(r) \leq t < g_2^2(r), \\ 1+1-1-1=0 & \text{if } t \geq g_2^2(r). \end{cases} \end{aligned}$$

In the same way one proves that $c^*(t,r)$ is never positive in cases (i) and (iii).

By analogous reasoning we can show that if the composition rule is max (parallel-AND architecture), then $c(t) \geq 0$, for all t, because $c^*(t,r) \geq 0$ at any t and any fixed r.

For the serial architecture (the composition rule +) $c^*(t,r)$ does not preserve its sign, but the analysis of the arrangements shows that, for any t and r,

$$\int_0^t c^*(t,r)\,\mathrm{d}t \geq 0$$

and

$$\int_0^\infty c^*(t,r)\,\mathrm{d}t = 0.$$

Then the same properties should hold for $c(t)$, because

$$\int_0^t c(t)\,\mathrm{d}t = \int_0^t \left(\int_{S_R} c^*(t,r)\,\mathrm{d}p_R(r)\right)\mathrm{d}t = \int_{S_R}\left(\int_0^t c^*(t,r)\,\mathrm{d}t\right)\mathrm{d}p_R(r).$$

However, dealing with deterministic quantities is not always convenient. If a deterministic quantity changes as a function of r, the probability with which it falls within a given measurable subset may jump from 0 to 1 or vice versa. In some cases it may be desirable to deal with "well-behaved" distributions only, with associated probabilities that change continuously or even sufficiently smoothly. (The term "smooth" refers to the highest order of continuous derivative a function possesses.) To make this desideratum achievable in the context of selective influences, we begin by stating the following equivalence.

Theorem 1.9 $\left(A^1,\ldots,A^n\right) \leftrightharpoons \left(\lambda^1,\ldots,\lambda^n\right)$ *if and only if one can find stochastically independent random variables* $R, R^1,\ldots,R^n$ *and functions* $w_1,\ldots,w_n$, *such that*

$$\left(A^1_\phi,\ldots,A^n_\phi\right) \sim \left(w_1\left(\lambda^1,R,R^1\right),\ldots,w_n\left(\lambda^n,R,R^n\right)\right)$$

for all allowable treatments $\phi = \left(\lambda^1,\ldots,\lambda^n\right)$.

By analogy with factor analysis, we can call $R^1,\ldots,R^n$ *specific sources of variability*, and call R a *common source of variability*. The proof of the theorem is very simple. If a representation

$$\left(A^1_\phi,\ldots,A^n_\phi\right) \sim \left(f_1\left(\lambda^1,R\right),\ldots,f_n\left(\lambda^n,R\right)\right)$$

exists, one can choose arbitrary $R^1,\ldots,R^n$ (combined together and with R by an independent coupling) and put $w_k(\lambda,r,r') = f_k(\lambda,r)$, $k=1,\ldots,n$. If a representation stated in the theorem exists, then define $R^* = \left(R,R^1,\ldots,R^n\right)$ and put $f_k\left(\lambda,\left(r,r^1,\ldots,r^n\right)\right) = w_k\left(\lambda,r,\mathrm{Proj}_k\left(r^1,\ldots,r^n\right)\right)$.

The consequences of this simple theorem are significant. Once the possibility of splitting a single source of randomness into a common and specific components has been established, it becomes possible that in certain situations this split can be more than a formal redefinition of a single source. It follows from the theorem that conditioned upon any value r of R, the random variables $w_1\left(\lambda^1,r,R^1\right),\ldots,w_n\left(\lambda^n,r,R^n\right)$ are stochastically independent. One can hypothesize now, that these independent random variables have distributions with desired properties. For example, if all random variables $\left(A^1_\phi,\ldots,A^n_\phi\right)$ are real-valued and continuous, $w_1\left(\lambda^1,r,R^1\right),\ldots,w_n\left(\lambda^n,r,R^n\right)$

may be assumed to possess densities, or have the property that the probability with which $w_k\left(\lambda^k, r, R^k\right)$ falls within any interval of reals is a continuously differentiable function of r. Such assumptions may be important in studying mental architectures or random variables underlying comparisons of stimuli.

1.18 Related literature

There are many textbooks treating measure theory and probability (e.g., Chung, 1974). However, the reader should be aware that (a) older textbooks usually deal with random variables in the narrow sense only; (2) in most textbooks random variables are defined as measurable functions on a sample space, restricting thereby the consideration to jointly distributed random variables. For random variables that need not be jointly distributed and the associated theory of coupling them into jointly distributed entities, see Thorisson (2000). The earliest explicit discussions of selective influences in psychology can be found in Sternberg (1969) and Townsend (1984). Marginal selectivity for two random variables was first mentioned in Townsend and Schweickert (1989). Other historical details and relations can be found in Dzhafarov (2003a), where the theory of selective influences presented in this chapter was first proposed. In this earlier work (and its elaboration in Dzhafarov and Gluhovsky (2006) the "is distributed as" relation in the defining representation for selective influences,

$$\left(A_\phi^1, \ldots, A_\phi^n\right) \sim \left(f_1\left(\lambda^1, R\right), \ldots, f_n\left(\lambda^n, R\right)\right)$$

was somewhat carelessly replaced with equality. For a mathematically rigorous and maximally general version of the definition and Joint Distribution Criterion, see Dzhafarov and Kujala (2010). The tests of selective influences were first introduced in Kujala and Dzhafarov (2008). They included the cosphericity tests and a special form of distance tests. A general version of distance tests (p.q.-metric tests) was introduced in Dzhafarov and Kujala (2013). The linear feasibility test is described in Dzhafarov and Kujala (2012b). For applications of the theory of selective influences to discrimination judgments and to mental processing architectures, see Dzhafarov (2003b,c) and Dzhafarov et al. (2004). The parallels between the theory of selective influences and the analysis of determinism in the so-called Bohmian version of the Einsten-Podolsky-Rosen entanglement paradigm of quantum physics are described in Dzhafarov and Kujala (2012b,a). The history there dates back to Bell's (1964) epoch-making inequalities, and then to their elborations in Clauser et al. (1969) and Fine (1982). Mathematically, this line

of work is subsumed by the linear feasibility test, whose most general version in quantum physics is described in Basoalto and Percival (2003).

1.19 Acknowledgments

This work was supported by NSF grant SES-1155956. We are grateful to Jing Chen, Shree Frazier, Nicole Murchison, Alison Schroeder, and Ru Zhang for pointing out numerous typos and imprecisions in the original draft of the chapter.

References

Basoalto, R. M. and Percival, I. C. (2003). BellTest and CHSH experiments with more than two settings. *Journal of Physics A: Mathematical & General*, 36(26):7411–7423.

Bell, J. S. (1964). On the Einstein Podolsky Rosen paradox. *Physics*, 1(3):195–200.

Chung, K. L. (1974). *A Course in Probability Theory*. Academic Press, New York.

Clauser, J. F., Horne, M. A., Shimony, A., and Holt, R. A. (1969). Proposed experiment to test local hidden-variable theories. *Physical Review Letters*, 23:880–884.

Dzhafarov, E. N. (2003a). Selective influence through conditional independence. *Psychometrika*, 68(1):7–25.

Dzhafarov, E. N. (2003b). Thurstonian-type representations for "same-different" discriminations: Deterministic decisions and independent images. *Journal of Mathematical Psychology*, 47:184–204.

Dzhafarov, E. N. (2003c). Thurstonian-type representations for "same-different" discriminations: Probabilistic decisions and interdependent images. *Journal of Mathematical Psychology*, 47:205–219.

Dzhafarov, E. N. and Gluhovsky, I. (2006). Notes on selective influence, probabilistic causality, and probabilistic dimensionality. *Journal of Mathematical Psychology*, 50:390–401.

Dzhafarov, E. N. and Kujala, J. V. (2010). The joint distribution criterion and the distance tests for selective probabilistic causality. *Frontiers in Psychology*, 1:151.

Dzhafarov, E. N. and Kujala, J. V. (2012a). Quantum entanglement and the issue of selective influences in psychology: An overview. In Busemeyer, J. R., Dubois, F., Lambert-Mobiliansky, A., and Melucci, M., editors, *Quantum Interaction*, volume 7620 of *Lecture Notes in Computer Science*, pages 184–195. Springer.

Dzhafarov, E. N. and Kujala, J. V. (2012b). Selectivity in probabilistic causality: Where psychology runs into quantum physics. *Journal of Mathematical Psychology*, 56:54–63.

Dzhafarov, E. N. and Kujala, J. V. (2013). Order-distance and other metric-like functions on jointly distributed random variables. *Proceedings of the American Mathematical Society*, 141:3291–3301.

Dzhafarov, E. N., Schweickert, R., and Sung, K. (2004). Mental architectures with selectively influenced but stochastically interdependent components. *Journal of Mathematical Psychology*, 48:51–64.

Fine, A. (1982). Joint distributions, quantum correlations, and commuting observables. *Journal of Mathematical Physics*, 23:1306–1310.

Kujala, J. V. and Dzhafarov, E. N. (2008). Testing for selectivity in the dependence of random variables on external factors. *Journal of Mathematical Psychology*, 52:128–144.

Sternberg, S. (1969). The discovery of processing stages: Extensions of donders' method. *Acta Psychologica*, 30:276–315.

Thorisson, H. (2000). *Coupling, Stationarity, and Regeneration*. Springer, New York.

Townsend, J. T. (1984). Uncovering mental processes with factorial experiments. *Journal of Mathematical Psychology*, 28(4):363–400.

Townsend, J. T. and Schweickert, R. (1989). Toward the trichotomy method of reaction times: Laying the foundation of stochastic mental networks. *Journal of Mathematical Psychology*, 33(3):309 – 327.