Excerpt from

Rouanet H, Bernard J.M, Lecoutre B, Lecoutre M.P, Le Roux B; Foreword by P. Suppes:
"New ways in statistical methodology : from significance tests to Bayesian inference", Bern, Peter Lang.

# Chapter 4

# Introduction to Combinatorial Inference

Henry Rouanet and Marie-Claude Bert

*One could treat of Probability without uttering the word Chance,*
*just as one can treat of Electricity without uttering the word Frog.*
*Paul Valéry*

## Introduction

This chapter is an introduction to Combinatorial Inference, or Set-theoretic Inference (See Section 4.3), an alternative to frequentist inference that our Math & Psy Group has been developing since the early eighties: see Rouanet, Bernard, Lecoutre (1986) and Rouanet, Bernard, Le Roux (1990). Its motivation is to provide researchers with a framework that can be used when the "validity assumptions" of the common procedures are not met. In the making of combinatorial procedures, the dissociation made in earlier chapters between algorithm and statistical framework is put to use. Roughly speaking, the algorithms of combinatorial procedures coincide with those

of conventional tests, while the random framework is discarded. As a result, in data analysis, it will be possible to keep many familiar algorithms, while the conclusions of Combinatorial Inference are stated in terms of new concepts, such as typicality and homogeneity, formalized in a nonprobabilistic way.

We will first present Typicality tests (Section 4.1), and Homogeneity tests (Section 4.2). Then we will outline the making of combinatorial inference and discuss related viewpoints (Section 4.3).

## 4.1 Typicality Tests

In this first section, we present typicality situations (1.1) and characterize the typicality problem (1.2), and, in the elementary context of finite sampling, we present the typicality test for the mean (1.3) and the hypergeometric typicality test for a relative frequency (1.4). We will proceed by making some general comments about combinatorial inference (1.5). Last, the extension to sampling from a distribution will be outlined (1.6).

### 4.1.1 Typicality Situations

Consider the following situations:

*Committee.* Among the members of a club, a committee is appointed. Can the committee be declared to be atypical of the club with respect to the mean age, or the sex ratio, etc.?

*Schoolboys.* At the Louvre station of the Paris metro, a group of 20 schoolboys get off, 7 of whom are red-haired. Assuming that among the French, the percentage of red-haired people is around 10 %, can it be inferred that the group of schoolboys is atypical of French schoolboys?

*Vacation.* A tourist has spent a period of 20 days in August in a resort; 7 of the days were rainy days. An advertisement claims that the percentage of rainy days in August is 10 %. Can the tourist infer that the vacation period was atypical of the advertised climate?

*Gifted children.* In a follow-up study of 5 gifted children, a psychologist has found that for a certain task the mean grade of her group is 30, with an SD of 6, whereas for a reference population of children of the same age the mean is known to be 25. Is she entitled to claim that her group of gifted children is on the average superior to the reference children?

### 4.1.2 The Typicality Problem

The preceding situations exemplify what we call *typicality situations.* In such a situation there is a given group of observations, and there is also a known reference population. Some statistic is considered, such as the mean of a variable of interest. Then the *typicality problem* is raised, intuitively formulated as follows: "Can the group of observations be assimilated to the reference population, or is it atypical of it?"; or more specifically: "How can a typicality level be assessed for the group of observations (with respect to the population, according to some statistic of interest)?"

In typicality situations, it is tempting to do a significance test. Yet the conventional statistical framework is not valid, since no randomness is assumed in the data generating process. In the Committee example, the group under investigation is a subset — but not a *random* subset, as a rule — of the reference population; in the Gifted Children example, the group under investigation is not even a subset of the reference population.

Even for a random sample, the typicality issue may be raised as an issue that is perfectly distinct from randomness. As an example, suppose a trial jury of 9 members that happens to include not a single woman; even if the jury has been lawfully constituted by random sorting, its competence might be questioned on the grounds that it is not typical with respect to the sex-ratio.

In order to offer a solution to the typicality problem, the basic idea will be to compare the group of observations to the samples of the reference population, where samples are simply defined as subsets of the population.

### 4.1.3 Finite Sampling: Typicality Test for the Mean

In the context of *finite sampling*, the term *population* will always refer to a *finite set*. Let $n$ denote the size of the group of observations, and $N$ the size of the reference population. In combinatorial inference, a *sample* of the population will be defined as an $n$-element subset of the reference population, and the set $\mathcal{X}$ of all $\binom{N}{n}$ $n-$element subsets defines the *sample space*. We will now describe the typicality test in the case of a numerical variable, taking the mean as a statistic of interest. The mapping $M$ on $\mathcal{X}$ that, with each sample $x \in \mathcal{X}$, associates its mean $M(x)$, defines the *Mean* as a statistic. Let $m_{obs}$ denote the observed mean of the group of observations, $\mu$ the mean of the reference population, and suppose — to fix our ideas — that $m_{obs} > \mu$. Then we consider the samples whose means are greater than (or equal to) $m_{obs}$, that is, which satisfy the property $(M \geq m_{obs})$. Let

$$\overline{p} = P(M \geq m_{obs}) \text{ (observed upper level)}$$

be the proportion of those samples. This proportion will be taken as defining the *level of typicality* for the mean, with respect to the reference population. The smaller the value of $\overline{p}$, the lower the typicality. For any $\overline{\alpha}$ between 0 and 1/2, if $P(M \geq m_{obs}) \leq \overline{\alpha}$, the group of observations will be declared to be atypical of the reference population, with respect to the mean, upwise at level $\overline{\alpha}$ (one-sided). When $m_{obs} < \mu$, the observed lower level will be similarly considered, and the typicality level defined accordingly. A group of observations will be declared atypical if it is atypical whether upwise or downwise.

**Example.** *Committee with* $N = 9$, $n = 3$ (Rouanet et al, 1990, p. 94). Let a committee of 3 members with mean age $m_{obs} = 69$ to be compared to the club of 9 members with ages (58; 61; 64; 64; 64; 67; 67; 70; 70). There are $\binom{9}{3} = 84$ samples (of size $n = 3$), whose means generate the sampling distribution of the statistic $M$: see Figure 4.1.

By inspection, it is found that out of the 84 samples, 2 satisfy the property $(M \geq 69)$; hence $\overline{p} = 2/84 = 0.024$. Thus for any $\overline{\alpha} \geq 2/84$, the result is significant. Taking the conventional grid described in Chapter 2, since we have $2/84 < 0.025$ but $2/84 > 0.005$ (significant

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | | 457 | | | | |
| | | | | 456 | | | | |
| | | | | 357 | | | | |
| | | | 345 | 356 | 567 | | | |
| | | | 257 | 347 | 467 | 579 | | |
| | | | 256 | 346 | 459 | 578 | | |
| | | 245 | 247 | 267 | 458 | 569 | | |
| | | 235 | 246 | 259 | 367 | 568 | | |
| | | 234 | 237 | 258 | 359 | 479 | | |
| | | 157 | 236 | 249 | 358 | 478 | | |
| | | 156 | 167 | 248 | 349 | 469 | | |
| | | 147 | 159 | 239 | 348 | 468 | | |
| | 145 | 146 | 158 | 238 | 279 | 379 | 679 | |
| | 135 | 137 | 149 | 179 | 278 | 378 | 678 | |
| 125 | 134 | 136 | 148 | 178 | 269 | 369 | 589 | |
| 124 | 127 | 129 | 139 | 169 | 268 | 368 | 489 | 789 |
| 123 | 126 | 128 | 138 | 168 | 189 | 289 | 389 | 689 |
| *61* | *62* | *63* | *64* | *65* | *66* | *67* | *68* | *69* |

Figure 4.1: Sampling distribution of Mean

result S$^\star$), we conclude that the committee — indeed any group of observations with mean $m_{obs} = 69$ — is atypical of the club with respect to the age, on the upper side, at the level .025 (one-sided). As another example, consider a group of observations with mean $m_{obs} = 68$, together with the same reference population; then we have $\overline{p} = 5/84$. Since $5/84 > 0.025$ (nonsignificant result: NS), we cannot conclude that the group of observations is atypical, at the conventional two-sided .05 level.

**Fundamental typicality property.** The construction of the typicality test extends to any numerical (or simply ordinal) statistic, taking as a typicality index the proportion of samples that are more extreme than (or as extreme as) the data, with respect to this statistic. The typicality test can be applied to every sample of the reference population, viewed as a particular group of observations. Then for any specified $\alpha$, the test will separate out those samples which are atypical at the $\alpha$-level. The fundamental typicality property states that the proportion of these samples is at most $\alpha$ — "at most" rather than "equal to", owing to the discreteness of the sampling distribution.

### 4.1.4 Hypergeometric Typicality Test for a Relative Frequency

With categorized data, combinatorial inference leads to explicit formulas. To illustrate, we will present the typicality test for a relative frequency, in the context of finite sampling.

Suppose that in a group of $n$ observations, $a$ observations possess a character of interest, hence the observed relative frequency of this character $f_{obs} = a/n$. Suppose that in the reference population of size $N$, the corresponding relative frequency is $\varphi_0 = A/N$. The combinatorial test here amounts to *comparing the observed frequency* $f_{obs}$ *to the reference frequency* $\varphi_0$ for a population of size $N$. Let $F$ (statistic) be the mapping on the sample space that, with each sample $x$, associates its relative frequency $F(x)$. The number of samples for which the property $(F = a/n)$ holds is $\binom{A}{a} \times \binom{N-A}{n-a}$; hence the observed upper level $\overline{p}$ is given by

$$\overline{p} = P(F \geq a/n) = \sum_{a'=a}^{n} \binom{A}{a'} \times \binom{N-A}{n-a'} / \binom{N}{n}$$

The algorithm of the typicality test for a relative frequency is the one used in the classical hypergeometric test for frequencies; recasting the latter test in a combinatorial framework amounts to retaining its algorithm while discarding the probabilistic interpretation.

**Example.** *Committee with* $N = 20$, $n = 5$. Among the 20 members of a club, 6 are women; a committee of 5 members is appointed, 4 of whom are women; is the committee atypical of the club with respect to the sex ratio? We have here $n = 5$, $f_{obs} = 4/5 (= .80)$, $\varphi_0 = .30$, $N = 20$. There are $\binom{20}{5} = 15504$ samples, and among them $\binom{6}{4} \times \binom{14}{1} + \binom{6}{5} \times \binom{14}{0} = 216$ for which $F \geq 4/5$. Hence $\overline{p} = 216/15504 = .014$. Taking conventional levels, since $\overline{p}$ lies between .025 and .005, the frequency $f_{obs} = 4/5 = .30$ is *significantly higher* (in a combinatorial sense) than the reference frequency $\varphi_0 = .30$, at the .025 level (one-sided). In typicality terms, it is concluded that the committee — indeed any group of 5 observations with observed frequency $f_{obs} = 4/5$ — is atypical of the club, with respect to the sex ratio, at the one-sided level .025 ($S^\star$). An interpretation of the over-representation of women in the committee is called for.

A more thorough presentation of combinatorial inference for frequencies, including the derivation of *combinatorial confidence limits*, can be found in Rouanet et al. (1986) and Rouanet et al. (1990).

### 4.1.5 Remarks on Combinatorial Inference

**Typicality test and descriptive statistics.** The concept of typicality that we have defined here is in harmony with that of "typical value" of a distribution: mean, median, etc. More importantly, assessing the typicality level of a group of observations appears as the direct extension, for $n \geq 1$, of the natural statistical procedure of assessing the performance of an individual in a given task by means of the proportion of scores exceeding the score of that individual in a reference population. For a group of observations, however, the typicality level depends on the size of the group, therefore, as soon as $n > 1$, it is not a descriptive statistic — in the technical sense that was specified earlier in Chapter 1. Combinatorial inference is thus to be regarded as the first stage of inductive statistics.

**From typicality tests to frequentist inference.** On the other hand, the link between combinatorial notions and those of frequentist inference is apparent, when the data set is a sample from the population. The samples from a population in combinatorial inference are just the "unordered samples without replacement" of the frequentist finite sampling theory except that no probabilities are attached to them; for each property of interest, what is assessed instead is simply the proportion of samples for which that property holds. The conventional frequentist framework will be "recovered" if we introduce the additional assumption of *random sampling*, that is — in the finite theory considered here — we now suppose that all samples (subsets) have equal probabilities of being extracted. Then the *conversion property* holds: Under the assumption of random sampling, the *proportion of samples* (subsets) satisfying a certain property becomes the *probability* that a sample will satisfy that property. Thus, for the preceding numerical example, under the random sampling assumption, the *proportion* of samples for which the property

($M \geq 69$) holds — a proportion that we wrote $P(M \geq 69)$ — becomes the *probability* that a randomly extracted sample satisfies this same property ($M \geq 69$) — a probability that we may also write $P(M \geq 69)$, now reading $P$ "probability" instead of "proportion". The conversion property thus transforms combinatorial procedures into frequentist ones. Clearly, it not just "a matter of semantics" to speak of the proportion of samples satisfying a certain property, or of the probability that a sample satisfies this property. The first statement is valid without restriction, whereas the second one demands the additional assumption that all samples are equally probable.

The foregoing discussion applies to the whole of Combinatorial Inference. Whenever a combinatorial procedure technically coincides with the algorithm of a frequentist procedure, the familiar significance formulations can be retained, while qualified as "combinatorial", since no probabilistic interpretation is intended.

*Summarizing:* While conceptually, combinatorial inference is a direct extension of descriptive statistics, technically it involves algorithms of frequentist inference. Combinatorial Inference is thus the first stage of Inductive Data Analysis.

### 4.1.6 Sampling from a Distribution

The typicality test can be extended to sample spaces defined by *sampling from a distribution* (rather than from a population), by generalizing the notion of a sample and making use of mathematical convergence theorems.

**The binomial typicality test.** As a first example, let us take the comparison of an observed relative frequency $f = a/n$ to a reference value $\varphi_0$, when no population size $N$ is specified. We may then consider a sequence of populations, such that the population frequency $\varphi = A/N$ approaches $\varphi_0$ when $N$ tends toward infinity. Then the previously written hypergeometric expression of the upper level converges to the binomial expression

$$\overline{p} = P(F \geq a/n) = \sum_{a'=a}^{n} {\varphi_0}^{a'} \, (1 - \varphi_0)^{n-a'}$$

Even though the number of samples is not finite in the limit, this binomial expression may be taken as defining a *proportion of samples*, when no population size is specified. Intuitively, the procedure may be thought of as an inference for an *arbitrarily large population*. A related approach consists in considering *ordered samples with replacement* in a population. The number of such samples is $N^n$ and the proportion of those that satisfy $(F \geq a/n)$ is $\sum_{a'=a}^{n} (\frac{A}{N})^{a'} (1-\frac{A}{N})^{n-a'}$. When $A/N$ approaches $\varphi_0$, we again obtain the binomial expression.

**Examples.** *Schoolboys, Vacation.* For both examples we have $n = 20$, $f = 7/20$ and $\varphi_0 = .10$. We then find $P(F \geq 7/20) = .0001$, a highly significant result (S$^{\star\star}$), leading to a conclusion of atypicality.

**Sampling from a distribution.** In frequentist inference, samples from *distributions* — e.g. samples form a normal distribution — are considered even though there is no relevant (finite) population, and defined in terms of independent identically distributed (i.i.d.) random variables. In combinatorial inference, a *sample from a distribution* will be defined in measure-theoretic terms, namely as an element of a product-measure space $(\mathcal{U}^n, \Pi^n)$, where $\mathcal{U}$ is a measurable space and $\Pi$ is a positive measure of total mass 1 over $\mathcal{U}$. Even though the *number* of samples may be infinite, the *proportion* of those that satisfy a given property is well-defined by the $\Pi^{(n)}$-measure of the property of the sample space. As a consequence, the intuitive formulations of the finite theory may be carried over (Rouanet et al, 1990, p. 103).

**Student's combinatorial $t$-test.** As an example, let us recast Student's $t$-test — comparing an observed mean to a reference mean value — in combinatorial terms. Taking as a sample space the samples of size $n$ from a normal distribution of mean $\mu$, the classical *Student property*, in combinatorial terms, reads: the proportion of samples for which the ratio $\frac{M-\mu}{S/\sqrt{n}}$ exceeds $t_{\overline{\alpha}}$ is equal to $\overline{\alpha}$; that is, for any $\mu$, we have

$$P(\frac{M-\mu}{S/\sqrt{n}} > t_{\overline{\alpha}}) = \overline{\alpha}$$

Now consider a group of $n$ numerical observations, and a reference mean value $\mu_0$. Let $T = \frac{M-\mu_0}{S/\sqrt{n}}$. Thanks to the Student property, we

may assess the typicality of this group, for the mean, with respect to any normal reference distribution of mean $\mu_0$. Thus for the *Student data* (cf. Chapter 2), we have $n = 10$, $m_{obs} = 1.58$ (observed mean), and $\mu_0 = 0$ (reference mean), and the observed $t$-ratio is $t_{obs} = 4.06$. Hence the one-sided observed level is $\overline{p} = 0.0014$ (S**). In combinatorial inference, the $\overline{p}$-value is interpreted as the typicality level of the group of observations, for the mean, with respect to a normal distribution of mean 0. The conclusion of the combinatorial test is that the group is atypical, for the mean, on the positive side, of a normal distribution of mean 0, at the .005 level (one-sided).

The semantic difference with the frequentist $t$-test is apparent. In the frequentist test, normality is an *assumption*, whereas in the combinatorial one, it is a *reference*. In the combinatorial test, there is no *validity* issue, even though there is a *relevance* isssue, since the choice of a particular distribution as a reference may be more or less appropriate. In this connection, the combinatorial $t$-test will often be better justified than the frequentist one, because the normal distribution is a privileged reference in many situations. Let us take for instance the *Gifted Children* example of Section 1.1, with $n = 9$, $m_{obs} = 30$, $s = 6$, $\mu_0 = 25$ (mean score of reference children), and hence $t_{obs} = 2.5$. The result is significant at the 0.025 level (one-sided) (S$^\star$). In so far as the distribution for reference children is based on normalized scores — a widespread psychometric technique — the psychologist is entitled to claim that her group of gifted children is on the average superior to the reference children. This example illustrates how the combinatorial framework may offer real, interesting, and plausible settings — to paraphrase Freedman et al. (1991, p. A20) — for common procedures such as the $t$-test.

## 4.2 Homogeneity Tests

In this section, we will outline homogeneity tests along an approach similar to the one used for typicality. We first present homogeneity situations and state the homogeneity problem (2.1). Then we will

present the homogeneity tests for two basic structures (2.2), and related combinatorial tests (2.3). Then we will expose the passage from Combinatorial inference to frequentist inference (2.4) and to the Bayesian framework (2.5).

### 4.2.1 Homogeneity Situations

Consider the following situations.

*Summer school.* Participants in a summer school are allocated to several teaching groups. At the end of the course, an exam is given to the participants, revealing substantial differences among the mean scores of the groups. Can it be said that the groups are heterogeneous with respect to their mean scores?

*Wage modification* (adapted from Faverge, 1956, p. 88). A modification in the wage system is introduced in a workshop of a factory. For the 12 workers in the workshop ("subjects" $s1$ through $s12$), the outputs (number of items per hour) are the following ($a$ after modification, $b$ before):

$s1$: $a$ 220, $b$ 203 $s2$: $a$ 226, $b$ 222 $s3$: $a$ 254, $b$ 246 $s4$: $a$ 246, $b$ 221
$s5$: $a$ 296, $b$ 287 $s6$: $a$ 222, $b$ 224 $s7$: $a$ 293, $b$ 275 $s8$: $a$ 247, $b$ 246
$s9$: $a$ 240, $b$ 246 $s10$: $a$ 269, $b$ 258 $s11$: $a$ 236, $b$ 216 $s12$: $a$ 199, $b$ 197

The mean of the individual output differences ("after" − "before") is 8.92, and the (corrected) S.D. 9.59. Thus descriptively, there is a substantial mean increase (0.67 times the S.D.). Are the two groups of scores ("before" and "after") heterogeneous?

**The homogeneity problem.** The preceding situations exemplify what we call *homogeneity situations.* There are several groups of observations, and some statistic of interest is considered. The *homogeneity problem* is raised, intuitively formulated as "Can the groups be merged, or are they heterogeneous?"; "Can a level of homogeneity be assessed?" As in the typicality problem, it is tempting to do some conventional significance test. Yet again, no randomness is assumed in the data generating process. To get combinatorial homogeneity tests, one may, like for typicality, take significance tests that were originally devised within a frequentist framework and just

retain their algorithms. For our purpose here, we will take the classical *permutation tests*, or *Fisher-Pitman tests*, initiated by Fisher and by Pitman (1937); for a brief historical account, see Edgington (1987, p. 17-21).

**Permutation tests.** The familiar nonparametric tests, such as the sign test, rank tests, Fisher's exact test for a $2 \times 2$ table, etc. are variants of permutation tests, for which explicit formulas can be derived and tables can be constructed. This has rendered those tests applicable before the computer era. By contrast, for the basic Fisher-Pitman tests, a considerable amount of computation is required, even for modest sample sizes. This formidable computation obstacle, which has long hindered the full use of permutation tests, is being overcome nowadays. For small data sets, exact combinatorial computations can be carried out. For intermediate sizes, Monte Carlo procedures, that is, computer sampling from permutation distributions, can be used. For large data sets, approximate methods involving classical distributions are often available.

Leaving aside the computational obstacle, the justification of permutation tests in the frequentist framework is intricate and often elusively treated in textbooks. Readers who are not too clear about permutation tests will find it advantageous to get acquainted with them through the combinatorial framework, whose logic is straightforward and will lighten the slippery paths leading to frequentist interpretations.

### 4.2.2 Homogeneity Tests for Two Basic Structures

To say that several groups are homogeneous amounts to saying that the subdivision into groups may be ignored, that is, any observation belonging to a group *might have belonged as well* to any one of the groups. This *exchangeability principle* leads us to consider the *baseline data set* obtained by disregarding the subdivision into groups, and then to construct all possible data sets obtained by reallocating the observations of this baseline data set to the groups in all possible ways. Technically, this amounts to applying a *permutation group* to

the observed data set, thus generating a set of possible data sets — all of the same structure as the observed one — or *protocol space*, against which the observed data set (observed protocol) will be situated. The remainder of the test procedure will be the same as for typicality tests, replacing "sample space" by "protocol space" and typicality by homogeneity. For each protocol in the protocol space, the statistic of interest is calculated, and the proportion of protocols for which this statistic is more extreme than (or as extreme as) the observed value defines the *level of homogeneity* of the groups.

The permutation group used to generate the protocol space depends on the *design structure*. Hereafter we describe homogeneity tests, first for the structure of two independent groups (Summer School), then for that of two matched groups (Wage modification).

**Independent group design** *(Nesting structure)* (Rouanet et al., 1990, p. 116). Consider several independent groups of observations, that is the design where Subjects are nested within a Group factor. Disregarding this Group factor, the derived baseline data set is the pool of the groups. Hereafter we describe in detail the case of *two groups*, $g1$ and $g2$, of sizes $n_1$ and $n_2$; the derived baseline data set is the pool of the $n_1 + n_2$ observations. The protocol space is generated by reallocating $n_1$ of the pool of $n_1 + n_2$ observations to $g1$ and the other $n_2$ to $g2$; it thus comprises $\binom{n_1+n_2}{n_1}$ protocols.

For instance, suppose there are two groups of sizes 4 and 5, with the following numerical data set: $g1 : 3, 8, 10, 10$; $g2 : 1, 1, 2, 5, 5$. Taking the difference of means $D$ as the statistic of interest, the observed value of this statistic is $d_{obs} = 7.75 - 2.8 = 4.95$. The pool of $g1$ and $g2$ is the group of 9 observations (written in increasing values): $g1\_g2$: $1, 1, 2, 3, 5, 5, 8, 10, 10$. Applying the permutation group, $9!/5!\,4! = 126$ protocols are constructed. Thus starting with $g1$: $3, 8, 10, 10$; $g2$: $1, 1, 2, 5, 5$ (observed data set), and permuting the first observations of $g1$ and $g2$, we get the protocol: $g1$: $1, 8, 10, 10$; $g2$: $3, 1, 2, 5, 6$, etc. For each protocol the value of $D$ is calculated: thus 4.95 (for the data set), then 4.05, etc. Then, by inspection, it is found that out of the 126 protocols, there are 3 for which the dif-

ference of means is greater than or equal to the observed difference; hence: $P(D \geq d_{obs}) = 3/126 = .024$. Since 3/126 lies between .025 and .005, it is concluded that the two groups are heterogeneous — $g1$ being higher than $g2$ — at level .025 (one-sided) (S$^\star$).

The homogeneity test for two independent groups is seen to be equivalent to a typicality test, taking the pool of the two groups (baseline data set) as a "reference population" and one of the two groups as a "sample."

Of special interest are *extremal data sets*, that is, data sets that are more extreme than all other protocols. For an extremal data set, the two groups are "separated," in the sense that all observations of one group exceed all observations of the other. Then the homogeneity level is simply $1/\binom{n_1+n_2}{n_1}$. Taking for simplicity two groups of equal sizes, it is readily seen that for $n_1 = n_2 \leq 3$, two separated groups cannot be said to be heterogeneous (at the two-sided level .05); that for $n_1 = n_2 = 4$, they are heterogeneous at the one-sided level .025 (S$^\star$); and that for $n_1 = n_2 \geq 5$, they are heterogeneous at the one-sided level .005 (S$^{\star\star}$).

**Matched-group design (*crossing structure*)** (Rouanet et al., 1990, p. 121). Now consider the $S*T$ design, where $n$ subjects are crossed with a factor $T$ ("treatments", or "trials", etc.). Each experimental unit is nested in — indeed is confounded with — the crossing of factors $S$ and $T$. Therefore, disregarding factor $T$, the derived baseline data set is characterized by the sole structure of the nesting of units within factor $S$ (restricted exchangeability). Hereafter we deal with the case of a two-level factor $T$, i.e. the *matched-pair design*; then the group of permutations is defined by permuting the observations within each pair in all possible ways. The protocol space thus comprises $2^n$ protocols.

For instance, for the Wage modification data, there are $2^{12} = 4096$ protocols. Let $D$ denote the mean of the individual output differences "after − before" (statistic of interest). For the observed data set we have $d_{obs} = 8.92$. The baseline data set is the set of 12 unordered pairs (written in increasing value order):

$s1$: 203, 220 $s2$: 222, 226 $s3$: 246, 254 $s4$: 221, 246
$s5$: 287, 296 $s6$: 222, 224 $s7$: 275, 293 $s8$: 246, 247
$s9$: 240, 246 $s10$: 258, 269 $s11$: 216, 236 $s12$: 197, 199

Applying the permutation group, starting with the observed data set, we get, by permuting the two observations of subject $s1$ (hereafter written in boldface characters):

$s1$: $a$ **220**, $b$ **203** $s2$: $a$ 226, $b$ 222 $s3$: $a$ 254, $b$ 246 etc.

Then, among the 4096 protocols, the number for which $D$ is greater than or equal to 8.92 is easily found — using a computer progam such as the INFER program described in Rouanet et al. (1990) — to be 20, hence the proportion $P(D \geq d_{obs}) = 0.0049$ (one-sided). At the .005 level (one-sided), it is concluded that the matched pairs are heterogeneous (S$^{\star\star}$), "after" being higher than "before".

Here again, of special interest are extremal data sets, here, those for which all individual differences have the same sign; then the homogeneity level is simply $1/2^n$. It is readily seen that for $n \leq 5$, the matched pairs of an extremal data set cannot be said to be heterogeneous (at the two-sided level .05); that for $n = 6$ and $n = 7$, they are heterogeneous at the one-sided level .025 (S$^{\star}$); and that for $n \geq 8$, they are heterogeneous at the one-sided level .005 (S$^{\star\star}$).

An example of an extremal data set is provided by the classical *Student data*, for which 9 differences are strictly positive, and one is null, hence $P(D \geq d_{obs}) = 1/2^9 = .0020$. The conclusion of heterogeneity is attained at the one-sided level .005 (S$^{\star\star}$). It may be noticed that the homogeneity level .0020 differs from the value .0014 found for the typicality level with respect to a normal distribution, obtained by Student's $t$-test. Such a discrepancy is not surprising, since the two tests answer different questions.

### 4.2.3 Related Combinatorial Tests

**Structured data.** The approach of homogeneity tests extends to various sorts of *structured data* commonly encountered in planned experimentation or observation. In order to investigate a factor of interest, the general principle remains the same: construct the baseline data set by removing this factor from the structure, then gener-

ate the space of all protocols sharing the original structure, by means of a permutation group associated with that structure.

**Combinatorial independence test.** The combinatorial approach also applies to a problem akin to homogeneity, namely the *independence problem* (Rouanet et al., 1990, p. 125-130), in connection with the bivariate structure in observations.

As an example, let us consider the following *Sex bias* situation (from Freedman & Lane, 1983). In the 1973-74 academic year, at one of the largest departments of the U.C. at Berkeley, there were 191 men and 393 women who applied for admission to graduate school; 54 men and 94 women were admitted, hence an appreciable difference in percentages (28% for men *vs* 24% for women). Can it be suspected that there was a sex bias in the University's admission policy? In terms of independence, the problem reads: "Can the two attributes Sex and Admission be said to be independent or associated?"

The baseline data set, obtained by removing the bivariate structure, consists here of the two derived sets of 584 observations pertaining to each one of the separate attributes Sex and Admission. In the combinatorial independence test, the protocol space will consist of all possible matchings between those two sets. For two dichotomous attributes, the algorithm of the independence test amounts to Fisher's classical exact test, and in turn, when the number of observations is large, to the familiar $\chi^2$-test. In the present example, the value of the $\chi^2$ statistic, for the corresponding $2 \times 2$ table, is found to be $\chi^2_{obs} = 1.29$, hence $P(\chi^2 > \chi^2_{obs}) = 0.26$. The observed level is not low enough to be declared significant (at conventional levels). In combinatorial terms, the conclusion is that it cannot be inferred that there is an association between Sex and Admission. The U.C. at Berkeley cannot be charged with Sex bias.

### 4.2.4 From Combinatorial to Frequentist Inference

The preceding discussion reinforces the view of combinatorial inference as the first stage of inductive data analysis. In some situations, the conclusions reached through combinatorial inference may be felt

to be sufficient. Or alternatively, it may be wished to prolong them by probabilistic conclusions. Taking homogeneity situations once again, we are going to discuss how, starting with a combinatorial conclusion, *frequentist tests* can be constructed.

With the notion of homogeneity we may associate a *null hypothesis* expressing, in intuitive terms, that the factor of interest "has no effect." Then, in order to make a statement about this hypothesis, we will try to build — as we did for typicality tests — a frequentist framework entailing a *conversion property*, that is, for that matter, transforming proportions of protocols into (frequentist) probabilities. For homogeneity tests, things are not as straightforward as they are for typicality tests. To begin with, more than one single framework may be devised. Below we sketch *two frameworks* — both classical — for comparing two independent groups, that share the same algorithm but rest on different assumptions, and lead to different interpretations of the notion of "no effect".

**Random sampling and conditional test.** In this framework, a *random sampling* model of the conventional frequentist kind is assumed for each group, that is, each group is assumed to be a random sample from some unknown continuous parent distribution — the continuity assumption being made to dispose of the problem of ties. The null hypothesis states that the two parent distributions are identical. Under the null hypothesis, the pool of the two samples — our baseline data set defined in Section 2.2 — can be regarded as a single sample (of size $n_1 + n_2$) from the common parent distribution. Therefore, *conditionally to the baseline data set,* all $\binom{n_1+n_2}{n_1}$ protocols generated by permutation are equally probable.

Thus, for the Summer School data, we have a combinatorial conclusion of heterogeneity. Under the random sampling model (conditional test), the null hypothesis tested is that the two groups are samples from two identical parent distributions, and the combinatorial conclusion becomes the frequentist conclusion that this null hypothesis is not compatible with the data (at level .025, one-sided, that is: $S^\star$).

Formally, a conditional test can be devised for any homogeneity situation — as well as for independence situations: Fisher's exact test for a $2 \times 2$ table is classically justified as a conditional test. Random sampling, however, may not be a realistic assumption. Furthermore, in many homogeneity situations, the question of interest does not really pertain to some conjectural parent populations, but rather to the experimental units at hand. Thus, for the Summer School, the real question is to investigate whether or not the Group Factor — specifically, the sources of variation linked with the division into groups: different teachers, etc. — has had an effect on the performance of the participants. Similarly, in the Sex bias example (Section 2.3), the independence question is raised about the 584 students under consideration, rather than to some conjectural population from which these students would be supposed to be extracted. There is a broad range of situations where the random sampling assumption either is unrealistic or induces the wrong question.

**Randomization tests.** The concern just mentioned is undoubtly taken up in the *randomization model*, in which no underlying parent distribution is assumed, and the inference sought only pertains to the units that appear in the experiment. The null hypothesis now states that for each unit, the two observations that can be made do not depend on which condition is applied to that unit. Under this null hypothesis, all $\binom{n_1+n_2}{n_1}$ protocols are again equiprobable. For instance, in the Summer School example, the parameters are now the unknown scores that the participants (units) would have obtained if they had been assigned to the other group rather than to the group to which they were actually assigned. The primary frequentist justification of the test, now, is the *physical act of randomization*, by which conditions have been allocated to experimental units. Thus in the Summer School example, suppose the participants have been assigned to groups by means of a random device. Then the null hypothesis considered will be that all 9 participants would have obtained the same scores in the group to which they were not assigned. Then, from the heterogeneity conclusion of the combinatorial test, it

may be inferred that this null hypothesis is not compatible with the data. Conditional and randomization tests are further discussed by Cox and Hinkley (1974, p. 179-204).

**Status of randomization.** The methodological status of randomization as an experimental procedure has been matter of debate. In sensitive domains like medical research, randomization raises immense ethical problems that are beyond the scope of this book. Confining ourselves to statistical issues, it is a fact that experimental randomization generates a consensus about the probability of observables under privilegied null hypotheses, and this is often a definite advantage, in research areas where knowledge is limited or controversial. This statistical advantage is sometimes erected as a *principle*, along which — when random sampling is lacking — randomization is a *must* for statistical inference that might be drawn from data. We do *not* adhere to this principle, if only because there are too many situations that are not amenable to randomization and for which statistical inference still appears desirable. One such situation is the nesting structure when the groups do not pertain to "conditions" — to which units may be allocated or not — but are *natural groups*, such as boys and girls in a classroom, etc. Other situations are the crossing structure such as the before and after design (see next subsection), the bivariate structure (leading to the combinatorial independence test), etc. In such situations, should one renounce statistical inference just because randomization is out of the question? We think not. We firstly propose combinatorial inference, as a nonprobabilistic statistical inference that is applicable in any case. We then suggest that probabilistic inference might be rethought along the line we sketch below.

### 4.2.5 Toward the Bayesian Framework

In this subsection, we submit reflections and tentative suggestions aiming at overcoming the limitations of frequentist inference, when randomness assumptions (random sampling or randomization) are not met.

**The randomization paradox.** In the Wage Modification example, the question of interest is to assess the effectiveness of wage modification for the group of the 12 workers in the workshop. Now, if we take the randomization principle seriously, the lack of randomization in the before and after design precludes interpreting heterogeneity in terms of some "no effect" hypothesis pertaining to the group of 12 workers. Now instead of the before and after design, we might have randomly divided the 12 workers into two groups $g1$ and $g2$ of 6 workers each, and proceeded to make "before" observations only on the 6 workers belonging to $g1$ and "after" observations only on the 6 workers belonging to $g2$. We would then have two independent groups of 6 observations each, to which an unobjectionable randomization test might be applied and allow one, in the case of a significant result, to assess the effectiveness of the wage modification. Equivalently, starting with the full matched-pair data set at hand, we may randomly sample 6 "before" observations and 6 "after" ones and confine our statistical analysis to these 12 observations. Now the before and after design, where subjects are their own controls, is surely better than the preceding "randomization design." It thus seems paradoxical that using all available information should preclude a sort of conclusion that would be authorized using only partial information.

**The "no effect" hypothesis.** Leaving aside randomness assumptions of frequentist models — random sampling and randomization alike — let us take a new look at the "no effect" hypothesis in homogeneity situations, starting with the remark that whatever formal meaning is given to this hypothesis, the baseline data set does not contain information about this hypothesis. Now suppose an individual is shown the set of protocols generated from the baseline data set, and asked to guess which one of the protocols is the observed data set. If this individual believes that there is no effect, then all protocols will be equiprobable for that individual — and under the belief that there is an effect they will presumably not be equiprobable — hence a conversion property from proportions to probabilities, valid regardless of any randomness assumption. We submit this

conversion property to be taken in all situations as the operational probabilistic characterization of the null hypothesis of "no effect". We hope that readers will feel with us that this characterization of the null hypothesis is natural. The reason for which it is not classical is that the probabilities involved may not interpretable as long-run frequencies; they basically express degrees of belief in particular situations. In Bayesian terms, those probabilities are *predictive* and *conditional* upon the null hypothesis. The Bayesian framework is often presented — as in the late chapters of this book — as an enlargement of the frequentist one, that is, as a superstructure that is added to a frequentist model. The foregoing discussion suggests a *direct way* from combinatorial inference to the Bayesian framework, bypassing the intricacies of the frequentist framework(s).

**Chance formulations.** When the "no effect" hypothesis is compatible with the data (nonsignificant result), it is commonly said that the result "might have occurred by chance" — i.e. as a matter of coincidence, or luck, fortuitousness, fluke, etc., suggesting, by implication, that attempting to interpret the effect any further would be fruitless. When on the contrary the "no effect" hypothesis is not compatible with data (significant result), it is commonly said that the result "is not due to chance," which means that attempting interpretation is in order. Such formulations have a long-standing history that goes back to Laplace, at least. As a Laplace-inspired example, suppose that a child using a typewriter for the first time composes the following 12-character sequence: KINDERGARTEN. The reason that leads us to think that this arrangement is not due to chance, Laplace would explain, cannot be the fact that, physically speaking, it is less probable than the others, because, if the word KINDERGARTEN were not in use in any language, this arrangement would be neither less nor more probable, and we would then not suspect any particular cause in connection with it. But as the word is in use among us, it is incomparably more probable that the arrangement of characters is intentional rather than due to chance (Laplace, 1825/1986, p. 229). Such Laplacian comments again point to the Bayesian framework.

## 4.3 The Making of Combinatorial Inference

### 4.3.1 Frequencies and Probabilities

Both probabilities and relative frequencies are isomorphic, that is, they obey the same formal rules of a more general *calculus of proportions*. Yet the semantics of probabilities refers to uncertainty, and that of frequencies, to observed statistical data. To confuse two isomorphic entities is to commit a *structural fallacy*[1].

In Appendix 2 of Chapter 1, we discussed the fallacious assimilation of probabilities to frequencies. In Rouanet (1982), we discussed the fallacious converse assimilation, which is conveyed when the probabilistic language is used to introduce theoretical distributions, such as the normal distribution. The first step toward Combinatorial Inference thus consists in characterizing such distributions as "stylized" frequency distributions, instead of "probability" distributions. Along this line, the notation $P(Z > 1.96) = 0.025$ is interpreted as "the proportion of standard scores greater than 1.96 is 2.5%." There are indeed some statistical textbooks that adopt such a nonprobabilistic presentation, above all, those written in the psychometric tradition, such as Faverge (1956). In our statistical teaching, we have constantly adhered to this tradition, as reflected in Lecoutre and Lecoutre (1979), and then in Rouanet, Bernard, Le Roux (1990, chapters 2 and 3).

### 4.3.2 The Crucial Step

Admittedly, nonprobabilistic formulations of statistical inference are occasionally found in textbooks. For instance, the sentence "95 percent of calculated confidence intervals will cover the parameter's value" is commonly found. Nonetheless, such sentences appear in isolation, and the basic combinatorial structures of statistical inference are masked by the probabilistic phraseology. Virtually all sta-

1. As an example of structural fallacy discussed by Jeffreys (1961): Heat and vapor obey the same differential equations, but it does not follow from this that heat is a vapor.

tistical textbooks stress the probabilistic framework and randomness assumptions. The randomness habit is so rooted that "sample" is often used as a synonym of "random sample"! Incidentally, such an insistence on randomness is further evidence that the change from probabilities to proportions is not just a "matter of semantics".

To arrive at combinatorial inference, the crucial step is the second one, which consists in stripping the concept of a sample of its "randomness" character, and replace the probabilistic formulations by the formulations in terms of "proportions of samples." We took this step in the early eighties, when we started teaching introductory statistical inference.

### 4.3.3 Teaching Motivations

The difficulties of teaching statistical inference are well-known, and indeed, the teaching motivations have been strong in our making of combinatorial inference. In the early eighties, the idea emerged in the reflections of our colleagues and ourselves that the algorithms of the elementary inference procedures could be taught immediately following descriptive statistics, dropping the traditional "probability prerequisites". Such a strategy, we felt, would allow students to concentrate first on computational aspects, without being prematurely concerned with the conceptual difficulties of probabilistic interpretations. We started teaching proportion formulations, and devising interpretations in terms of typicality and homogeneity. The phrase *Set-theoretic Inference* was coined to refer to the new approach, and a first presentation of it was made at the International Conference on Teaching Statistics held in Sheffield (England), with the provocative title "Teaching statistical inference without probability prerequisites" (Rouanet et al, 1992). A more detailed paper followed: Rouanet et al. (1986), and then the reference book Rouanet, Bernard, Le Roux (1990), with its companion teaching software INFER. At the University René Descartes, Combinatorial Inference has been taught continuously since 1982, both to psychology students with no previous knowlege of either probability or statistical infer-

ence, and to mathematical students as a complement of the standard statistical curriculum[2].

### 4.3.4 Combinatorial Data Analysis

A growing trend in statistics in the last few years has been Combinatorial Data Analysis, which emphasizes algorithms instead of probabilistic models. This trend has been especially active in the area of Classification and is seen to naturally include all those techniques such as half-split, jacknife, bootstrap, etc. where the probabilistic phraseology is often misleading. It soon became clear that "Set-theoretic inference" was part of this trend — a point well taken by Arabie et al. (1996, p. 5) and others. The Rouanet et al. book (1990) thus appears to be the first Introduction to statistical inference written along the line of Combinatorial Data Analysis. In order to emphasize this connection, Ove Frank suggested we called the approach "Combinatorial Inference", and we have now definitely adopted this welcome suggestion.

### 4.3.5 Toward Recognition of Combinatorial Inference

A variety of reasons concur that should facilitate the acceptance — paving the way to recognition — of Combinatorial Inference by the community of researchers. Firstly, statistical procedures are often used in situations where the frequentist "validity assumptions" are not met. By providing assumption-free interpretations, combinatorial inference makes sense of common practice. Here is a revealing comment made by a psychologist: "But this is just what I have always done!"[3] Secondly, the terms of typicality and homogeneity are so natural that they are spontaneously adopted. Thirdly, following

2. Similarly, at the University René Descartes, an introduction to Bayesian inference has been taught since 1993, as an extension of classical significance testing, in the line of the last chapter of Rouanet, Bernard, Le Roux (1990) and of the subsequent chapters of the present book.
3. This comment curiously echoes the one (by a statistician) reported by Freedman and Lane (1982): "This is just what I have always thought!"

Rosch's work, cognitive psychologists have been deeply interested in typicality; from this viewpoint, statistical typicality appears as the application of the general notion to *collective objects*. Fourthly, there is the current prestige of *exact tests* in statistics. The magic of "exactness" must be qualified, of course. Student's $t$-test was (and still is) an exact test too! Rather than striking up the "exactness cant", we prefer to stress that combinatorial inference does not require unverifiable assumptions.

### 4.3.6 Related Viewpoints

Once Combinatorial Inference had taken shape, we started inquiring about related approaches. Then, leaving aside the abundant *technical* developments about permutation tests, Monte-Carlo procedures, etc., we discovered that there have been really few publications developing conceptual viewpoints akin to combinatorial inference. In what follows we sketch three such significant contributions[4].

**Maurice Allais and nonprobabilistic models.** Allais' name is familiar to econometric statisticians for his famous 1954 paradox in Decision Theory — also, perhaps, for the Nobel prize he got in 1989. Now in the early eighties, Maurice Allais vigorously denounced the confusion of frequencies with probabilities in the current interpretation of econometric models. Here is what we read in Allais (1983): "The so-called mathematical theories of probability could all be presented without ever using the words chance, probable, random, or any similar term... All the fundamental theorems of the so-called Probability theory, the Bernoulli law of large numbers[5], or the central limit theorem of convergence to the normal law, the law of iter-

4. Other references are also worth mentioning, such as Matheron (1989), a reference that did not escape Shafer's (1994) attention.

5. In Rouanet et al. (1990), we state the Bernoulli law of large numbers in terms of the *limit proportion of central paths*. We are not aware of any single reference — other than Allais — that would suggest (even remotely) that such combinatorial formulations of standard probability theorems are not only possible but highly meaningful.

ated logarithm, the arcsine law, etc. are *only asymptotic properties of frequency distributions based on calculations of combinatorial techniques."* (Author's italics). To enhance his claim, Allais exhibits a quasi-periodic model — hence fully deterministic — whose predictions could typically (and fallaciously) be interpreted in terms of a stochastic model. In spite of the author's notoriety, Allais' message went virtually unnoticed — and unchallenged.

**Edgington and nonrandom samples.** Starting from the fact that nonrandom samples are widely used in experimentation, and building on the distinction between random sampling and randomization — cf. Section 2.4 — Eugene Edgington, in numerous publications such as Edgington (1987, 1995), has cogently and valiantly defended the position that frequentist inference may be performed in nonrandom samples, whenever randomization is available. We are basically in full sympathy with a position which stresses a statistical framework — i.e. randomization — which is badly neglected. As a counterpart, the physical act of randomization seems to be for Edgington a necessary requirement (along the "randomization principle"), and this departs from our viewpoint, as we discussed in Section 2.5.

**David Freedman and nonstochastic settings.** In Freedman and Lane (1982, 1983), the authors consider the following problem. "Data are obtained in a nonstochastic [i.e. nonrandom] setting, and for some attribute of this data, the question is raised: can this attribute be dismissed as an artifact, or does it require a more substantial explanation?" The solution suggested by the authors, and illustrated through examples — such as the Sex Bias example (Section 2.3) — comes very close in spirit to Combinatorial Inference. One may regret that those thought-provoking papers have not been followed by systematic developments, and that the introductory statistical book written by David Freedman and his colleagues (Freedman et al, 1991) — in all respects a most commendable book — is confined to the frequentist viewpoint.