THE BAYESIAN APPROACH TO THE REJECTION OF OUTLIERS

BRUNO DE FINETTI
UNIVERSITY OF ROME

1. Summary

The aim of this paper is twofold: first, to contribute to the subject in question; second, in so doing to illustrate some general features and techniques of the Bayesian (or neo-Bayesian) subjectivistic approach to problems of statistics.

What may be said about the "true value" X of a quantity after observations have yielded measurements $x_1, \cdots, x_n$? The variable X or "unknown parameter" X, as the usual terminology calls it, is in fact a *random number* in the subjectivistic probability theory. For in this theory, probability is nothing else than the expression of beliefs about unknown facts; for example, this may be the probability that X is greater than some given value x. We need then merely to show how our final probability distribution for X, after the observations, comes from the initial one and from the conditional distributions of the measurements $x_1, \cdots, x_n$ given $X = x$, according to Bayes' theorem. No estimation problem per se is acknowledged to exist from such a viewpoint. X has a final distribution, and if such an expression as "estimated value" is used at all, it should be conceived of as a measure of location of the final distribution, suitably chosen for some practical purpose.

Everything is embodied in this over-all formulation. A rational answer to the question of how and why to attach less or no "confidence" to some "outlying" observations can arise from nothing else.

To fill the gap between such an abstract (or, as some might perhaps say, "philosophical") formulation, and a realistic detailed analysis of practical situations, we need only consider some set of more or less "reasonable" and interesting hypotheses about our opinions concerning the process of taking observations.

Three cases, all concerning an "error distribution," will be studied: (a) the errors are independent; (b) the errors are exchangeable; (c) the errors are partially exchangeable.

(a) *Independence* means "independence with known error distribution"; if the distribution is not normal, the combination of the observations is no longer so simple, and particular problems arise for "outlying" observations. This case has been considered by Poincaré [5] and others. More recently, in a paper of mine [1] at the 1957 Meeting of the Societá Italiana di Statistica, the case of mixtures of normal centered distributions has been particularly stressed.

(b) *Exchangeability* translates "independence with unknown error distribu-

tion" into language consistent with the subjectivistic theory, and means "mixtures of independence schemes (in the sense of the first case)." The problems become somewhat more complex. Now, the scatter of the observed measurements reacts on the initial opinion about the possible error distribution itself, and the notion of "outlying" depends on that opinion. The interest in applying to this problem the notion of exchangeability (or "equivalence" or "symmetry": de Finetti, Khinchin, Fréchet, Hewitt and Savage) was pointed out by W. H. Kruskal in the discussion of the lecture in which I reported the results of my above-mentioned paper at the University of Chicago (1957).

(c) *Partial exchangeability* translates, for the same reason and in a similar fashion, "independence with an unknown conditional error related to visible features of the individual observations" (for example, the observer, the instrument, the temperature). The information given by the observations may then suggest a dependence of the error distribution, and hence of outlying measurements, upon the said features. The notion of partial exchangeability (or partial equivalence) presented in a communication of mine at the "Colloque de Genève" in 1937, has been restated and developed in [2].

In each of these cases, as in all other problems of inference, it is not assumed, by the subjectivistic theory, that exactly this or that initial distribution ought to be selected as corresponding to the real opinion of a given person. It is maintained, nevertheless, that an initial distribution and Bayes' theorem give a good portrayal of inference, even though, in practical cases, as in all practical applications of mathematics, we must rely on initial, conditional and final distributions that are only *vaguely specified.* Again, any analysis of different attitudes toward a situation like the rejection of outliers consists only in recognizing corresponding differences in the initial opinions or in hypotheses and features concerning them. This is illustrated by cases (a), (b), and (c) or by even more complex possibilities.

2. Introduction

First, let us write down the formula expressing, for our general problem, the Bayesian approach, in order to fix the notation and to explain how the formula will be applied to the three particular cases to be considered (independence, exchangeability, partial exchangeability). To avoid complications irrelevant to the present discussion, we shall confine attention to the case where all distributions admit of a density.

The Bayesian principle may be written

Final distribution $\propto$ *initial distribution* $\times$ *likelihood,* and it will here be used in the form

Final density $\propto$ *initial density* $\times$ *likelihood,* where the initial and final distributions are those of X (the "true value") as evaluated by someone in his state of information *before* and *after* knowledge of the result of the set of observations considered. That is,

$$f_n(x|x_1, x_2, \cdots, x_n) = Kf_0(x)f(x_1, x_2, \cdots, x_n|x), \tag{1}$$

where $f_n(x|\cdots)$ is the density of the probability distribution of X after the n observations have given the values $x_1, x_2, \cdots, x_n$; K is the normalization constant; $f_0(x)$ is the density of the initial probability distribution of X, that is, before the said observations; $f(\cdots|x)$ is the density of the (n-dimensional) probability distribution of the set of observed values $X_1, X_2, \cdots, X_n$ (or of the "errors" $Y_1 = X_1 - X, \cdots, Y_n = X_n - X$) under the condition $X = x$; that is the likelihood.

Although the extension to the case where X (and hence the X_h and the errors Y_h) is a vector would require but obvious formal changes, it seems preferable here to consider only the one-dimensional case (X is a scalar), because it is sufficient and more suitable for our explanatory purposes.

When speaking of "errors" it is generally understood that their distribution is independent of the "true value" (each $Y_h = X_h - X$ is independent of X). More precisely, the validity of the following translation property will be postulated,

$$\begin{aligned} f(x_1, x_2, \cdots, x_n|x) &= f(x_1 - x, x_2 - x, \cdots, x_n - x|0) \\ &= f(y_1, y_2, \cdots, y_n|0). \end{aligned} \tag{2}$$

This condition is satisfied by each of the three cases which we shall consider. All that is required is the specification of the appropriate form of the likelihood function in (1), namely,

(a) in the case of *independence*,

$$f(x_1, x_2, \cdots, x_n|x) = f(x_1 - x)f(x_2 - x) \cdots f(x_n - x), \tag{3}$$

where $f(y)$ is the density of the probability distribution of each of the mutually independent errors Y_h.

(b) In the case of *exchangeability*,

$$f(x_1, x_2, \cdots, x_n) = \int f_\theta(x_1 - x)f_\theta(x_2 - x) \cdots f_\theta(x_n - x)\, d\Phi(\theta). \tag{4}$$

We have here a mixture of distributions of the kind (3) of case (a). Its meaning is immediate if there really exists a random number T that is an objectively definable but at present unknown quantity, such that conditionally on each possible value θ of T the errors become independent with distribution density $f_\theta(y)$. Then $\Phi(\theta)$ is the distribution function of T. The same holds if T is a random element in any abstract space Θ. Here this extension is essential, because only if Θ is such that all distributions $f(y)$ belong to the set $\{f_\theta(y): \theta \in \Theta\}$ will the mixtures give all, not merely some, of the possible cases of exchangeability. This problem is studied under general conditions in Hewitt and Savage [3].

However, it is by no means necessary that such an interpretation by a random element T hold in any given problem. It is only essential to know that exchangeability implies the mixture formula (4), and it may be useful to remember that all further developments hold *as if* the interpretation based on "the unknown element T'' applies (no matter whether that is true or not). This remark, often referred to later on, is more thoroughly discussed elsewhere, as in [2], sections 5(3) and 7.

Before considering the last case, (c) *partial exchangeability,* some preliminary statement is necessary and some exemplification may be useful.

The meaning of (1) must be interpreted, in this case, in a wider sense; it must be understood that the likelihood f may depend also on other circumstances concerning the observations, or, to make it explicit, it is better to write

$$f(x_1, x_2, \cdots, x_n|x; \lambda_1, \lambda_2, \cdots, \lambda_n), \tag{5}$$

where λ_h summarizes the whole of the circumstances, relating to the hth observation, which are supposedly able to influence it. A particular case (conditional independence given the λ) is that where

$$f(\cdots|x; \cdots) = f(x_1|x; \lambda_1)f(x_2|x; \lambda_2) \cdots f(x_n|x; \lambda_n). \tag{6}$$

In order to make clear by examples what kind of circumstances are meant to be considered and represented by λ, the λ may give information about

(1) the person having taken the observation (set $\lambda = 1, 2, \cdots, r$ according as the person was $I_1, I_2, \cdots, I_r$);

(2) the instrument used for the observation (again $\lambda = 1, 2, \cdots, s$);

(3) the pair person-instrument [then $\lambda = (h, k)$, the pairs with $h = 1, 2, \cdots, r$ and $k = 1, 2, \cdots, s$];

(4) the temperature t at which the observation has been taken (set $\lambda = t$, $t_1 \leqq t \leqq t_2$);

(5) a complex of m values (integers or real numbers or both) such as person-instrument-temperature [$\lambda = (h, k, t)$, with $h = 1, 2, \cdots, r$; $k = 1, 2, \cdots, s$; $t_1 \leqq t \leqq t_2$] or with other or more data such as pressure, dampness, or illumination);

(6) the shape of the temperature variation during the experiment (in cases when the observation requires a rather long time). In this case λ is a point in a function space, in fact $\lambda \equiv \lambda(\cdot)$, if $t = \lambda(\tau)$, with $0 \leqq \tau \leqq l$, is the function giving the temperature t at the time τ from the beginning ($\tau = 0$) to the end ($\tau = l$) of the observation; and so on.

Here too it would be inappropriate to confine attention to the most elementary and easy cases, such as examples (1) to (4), although they are obviously the most practical for a first explanation.

The change in the expression of the likelihood f from the preceding case of exchangeability in formula (4) to the present one, lies in the fact that the value of the parameter θ (which is unique in the former case) may now vary from observation to observation, according to the circumstances summarized by λ. The distribution thus no longer concerns a single element θ of Θ but, in the most general case, a complex of n elements $\theta_1, \theta_2, \cdots, \theta_n$ in the product space Θ^n.

To avoid too abstract a generality, let us specify separately the expression of f in the two simplest cases: that of λ admitting a finite number of values, as in examples (1), (2), (3), and that of λ being a continuous variable as in example 4; similarly for two or more continuous variables.

In the finite case, let us suppose for clarity that the possible values are only $\lambda = 1, 2, 3$. That is, that the set of n observations x_h may be partitioned into three groups of $n_1 + n_2 + n_3 = n$ (say) observations with the conditions $\lambda = 1$, $\lambda = 2$, and $\lambda = 3$ respectively. With obvious conventions we shall have

$$(7) \qquad f(x_1^{(1)}, \cdots, x_{n_1}^{(1)}; x_1^{(2)}, \cdots, x_{n_2}^{(2)}; x_1^{(3)}, \cdots, x_{n_3}^{(3)} | x) \\ = \int f_{\theta_1}(x_1^{(1)} - x) \cdots f_{\theta_1}(x_{n_1}^{(1)} - x) f_{\theta_2}(x_1^{(2)} - x) \cdots f_{\theta_2}(x_{n_2}^{(2)} - x) \\ f_{\theta_3}(x_1^{(3)} - x) \cdots f_{\theta_3}(x_{n_3}^{(3)} - x) \, d\Phi(\theta_1, \theta_2, \theta_3).$$

Here $\Phi(\theta_1, \theta_2, \theta_3)$ is a distribution in Θ^3; it is $\theta = \theta_1$ for all observations taken in the condition $\lambda = 1$, and similarly for $\theta = \theta_2$ and $\theta = \theta_3$.

Exchangeability arises as the particular case in which the distribution Φ is concentrated on the diagonal $\theta_1 = \theta_2 = \theta_3$; if it is concentrated in a point, we have conditional independence or independence (the latter if the point belongs to the diagonal). For the continuous case we may write

$$(8) \quad f(x_1, \cdots, x_n; \lambda_1, \cdots, \lambda_n) \\ = \int f_{\theta_1}(x_1 - x) f_{\theta_2}(x_2 - x) \cdots f_{\theta_n}(x_n - x) \, d\Phi(\theta_1, \theta_2, \cdots, \theta_n | \lambda_1, \lambda_2, \cdots, \lambda_n),$$

which may be interpreted as a limiting case of the former [see (7)] when the space Λ is subdivided into an indefinitely increasing number of parts.

3. The problem of the outliers from the Bayesian point of view

According to the Bayesian point of view, *there exist no observations to be rejected.* In fact, the Bayesian solution shows in all cases the final distribution determined *on the ground of all the observations taken.* In such cases and under those hypotheses where usually the procedure of "rejecting some observations" is adopted and justified, the Bayesian method leads automatically to the similar (but exact) result in which the influence of those observations on the final distribution is weak or practically negligible.

Nevertheless, it may be worth while to investigate which observations should have little weight, and why, both for theoretical and practical reasons. Theoretically, in order to have a better insight about the effective meaning of the question and about the role of different hypotheses on the results; practically, in order to find easier procedures, if there are any, for the rejection of some observations. Of course, that could be justified only as an approximation to the exact Bayesian rule (under well-specified hypotheses), but never by empirical ad hoc reasoning.

An example of a procedure, leading to an outliers rejection rule approaching an exact one, is the following. Consider an "estimated value" $\hat{x}$, which, on the basis of the exact (Bayesian) method, is expressed as a weighted average of the observed values,

$$\hat{x} = \rho_1 x_1 + \rho_2 x_2 + \cdots + \rho_n x_n. \tag{9}$$

If the weights ρ_h depend in too complicated a way on the observations $x_1 \cdots x_n$, but there exists a simpler rule giving $\hat{x}$ approximately on the basis of equal weights for all x_h except for some small weights which are set equal to 0, we could say that the corresponding observations are outliers in regard to this method. It is clear however that such a notion is vague and rather arbitrary. In order to avoid ambiguity and to deal with a handy expression, we shall in the sequel confine ourselves to the estimate given by the expectation of the final distribution,

$$\hat{x} = E\{X|x_1, x_2, \cdots, x_n\} = \int x f_n(x|x_1, x_2, \cdots, x_n)\, dx. \tag{10}$$

A further simplification occurs on considering the particular case of a "constant" initial density $f_0(x)$. That, of course, is no reason for giving credit to any a priori argument for a constant initial density, but we are often justified in taking it constant in so far as we suppose the distribution to be *diffuse* in the interval under consideration and the conclusions to be sufficiently well determined by such a hypothesis. A way to give this notion of diffuseness a precise and suitable meaning is indicated by Savage in [6], section 5.

Even dealing only with the example of the $\hat{x}$ = expectation and with $f_0(x)$ constant, there is no unambiguous principle for giving a meaning to the concept which we tried to suggest by calling it the "weight" of each observation x_h in determining $\hat{x}$. To seek for a direct method, several ideas, like the following two, could give rise to some significant investigation: (i) to compare the $\hat{x}$ resulting from all observations with that resulting from all but x_h, (ii) to explore how $\hat{x}$ varies with $\hat{x}_h$. In the latter case the partial derivatives

$$\rho_h = \frac{\partial \hat{x}}{\partial x_h} \tag{11}$$

are, in a sense, meaningful. Moreover, their sum is one [provided the translative property (2) holds, and $f_0(x)$ = constant]. Nevertheless, they are not a system of "weights" in the sense of (9).

To arrive at a system of weights in this sense, indirect methods seem more successful, although their validity is limited to special classes of error distributions $f(y)$ and their significance is tied to a particular interpretation of such $f(y)$ as mixtures of distributions of some simple kind.

As a first example, let us consider the mixtures of rectangular centered distributions, say those with constant density $1/2\gamma$ in $(-\gamma, +\gamma)$. The mixtures are obviously all the distributions with decreasing symmetrical density from 0 to $\pm\infty$,

$$f(y) = \int_{|y|}^{+\infty} \alpha(\gamma)\frac{1}{2\gamma}\, d\gamma, \qquad \alpha(\gamma) = -2\gamma f'(\gamma), \qquad \int_0^{\infty} \alpha(\gamma)\, d\gamma = 1. \tag{12}$$

The case we will work out is that of the mixtures of normal centered distributions,

$$(13) \qquad f(y) = \int_0^\infty \alpha(\gamma)\left(\frac{\gamma}{2\pi}\right)^{1/2} e^{-1/2\gamma y^2}\,d\gamma, \qquad \int_0^\infty \alpha(\gamma)\,d\gamma = 1,$$

including the limiting case of a sum or series,

$$(14) \qquad f(y) = \sum_i \alpha_i \left(\frac{\gamma_i}{2\pi}\right)^{1/2} e^{-1/2\gamma_i y^2}, \qquad \sum_i \alpha_i = 1.$$

As an example of a distribution of the form (13), let us remark that the Cauchy distribution appears on putting $\alpha(\gamma) = a(2\pi\gamma)^{-1/2}\exp(-a^2\gamma/2)$; as an examample of (14), the simplest case (with two values for γ, $\gamma = 1$ and $\gamma = 1/k^2$) is $(2\pi)^{-1/2}[(1-\alpha)\exp(-y^2/2) + (\alpha/k)\exp(-y^2/2k^2)]$. Something more about these cases is in [1]; the latter has been considered by Poincaré [5].

The transformation from α to f is nothing but a modified Laplace transformation. It is thus easy to write the inverse transformation from f to α, and to formulate the condition for $f(y)$ to be such a mixture (the derivatives $g^{(h)}(t)$ of $f(\sqrt{t}) = g(t)$, say, must be $\geqq 0$ or $\leqq 0$ for $0 \leqq t \leqq \infty$, according as h is even or odd).

The interpretation of $f(y)$ as a mixture may have real significance or not, and, consequently, the conclusion based on it may be itself significant or simply formally valid. Formula (13) has real significance if, for instance, we know that each observation is taken using an instrument with normal error, but each time chosen at random from a collection of instruments of different precisions, the distribution of the precisions being that indicated by $\alpha(\gamma)$. The same hypothesis, but taking the error distribution to be rectangular (as is the case with some rounding-off error), leads to an effective interpretation of formula (12).

4. The case of independence

The case of independence is the starting point for the other cases too. If the $f(y)$ assumes the form (13), the likelihood (3) takes the form

$$\begin{aligned}
(15) \qquad & f(x_1, x_2, \cdots, x_n | x) \\
&= K \int_0^\infty \int_0^\infty \cdots \int_0^\infty \alpha(\gamma_1)\alpha(\gamma_2)\cdots\alpha(\gamma_n)(\gamma_1\gamma_2\cdots\gamma_n)^{1/2} \\
&\qquad \exp\left[-\frac{1}{2}\sum_h \gamma_h(x_h - x)^2\right] d\gamma_1\,d\gamma_2\cdots d\gamma_n \\
&= K \int_0^\infty \int_0^\infty \cdots \int_0^\infty \alpha(\gamma_1)\alpha(\gamma_2)\cdots\alpha(\gamma_n)\left(\frac{\gamma_1\gamma_2\cdots\gamma_n}{\gamma}\right)^{1/2} \\
&\qquad \exp\left\{-\frac{1}{2}\left[\sum_h \gamma_h x_h^2 - \frac{1}{\gamma}\left(\sum_h \gamma_h x_h\right)^2\right]\right\} \\
&\qquad \gamma^{1/2}\exp\left[-\frac{1}{2}\gamma\left(x - \frac{1}{\gamma}\sum_h \gamma_h x_h\right)^2\right] d\gamma_1\,d\gamma_2\cdots d\gamma_n, \\
&\qquad\qquad \gamma = \gamma_1 + \gamma_2 + \cdots + \gamma_n.
\end{aligned}$$

If we suppose that $f_0(y)$ is approximately constant, the final distribution (1) $f_n(X|\cdots)$ itself is identical with (15). In order to compute the expectation (10) we need only to multiply (15) by x and integrate over $-\infty \leqq x \leqq +\infty$. The only factor in (15) depending on x is the last one,

$$\gamma^{1/2} \exp\left\{-\frac{\gamma}{2}\left[x - \frac{1}{\gamma}\sum_h \gamma_h x_h\right]^2\right\}, \tag{16}$$

and its integral is simply the weighted average of the x_h with weights γ_h,

$$\frac{1}{\gamma}\sum_h \gamma_h x_h. \tag{17}$$

The constant $(2\pi)^{1/2}$ is neglected because it is absorbed by normalization. The integral of (15) multiplied by $x\,dx$, that is, the expectation $\hat{x}$, is then $\hat{x} = \sum_h \rho_h x_h$, with

$$\rho_h = K \iint \cdots \int \gamma_h T(\gamma_1, \gamma_2, \cdots, \gamma_n)\, d\gamma_1\, d\gamma_2 \cdots d\gamma_n, \tag{18}$$

where $T(\cdots)$ indicates all factors remaining in the last form of (15).

$T(\cdots)$ is a function of the γ as explicitly indicated and of the observations x_h, which are in a given problem constant parameters. It is interesting to remark that there is a dependence on the x_h only through the function

$$z = \gamma \sum_h \gamma_h x_h^2 - (\sum_h \gamma_h x_h)^2, \tag{19}$$

which is precisely, for our purposes, a quadratic form in the γ_h with coefficients which are quadratic functions of the x_h. In a better form it is

$$z = (\sum_i \gamma_i)(\sum_h \gamma_h x_h^2) - (\sum_i \gamma_i x_i)(\sum_h \gamma_h x_h) = \sum_{ih} \gamma_i \gamma_h (x_h^2 - x_i x_h), \tag{20}$$

or, interchanging i and h and taking the half sum,

$$z = \frac{1}{2}\sum_{ih} \gamma_i \gamma_h (x_i - x_h)^2 = \sum_{i<h} \gamma_i \gamma_h (x_i - x_h)^2. \tag{21}$$

The dependence of T on the x_h consists in having as a factor

$$e^{-z/2\gamma}, \tag{22}$$

that is, the penultimate factor in the last form of (15), where, however, the x_h are present in the last factor also.

Our formula (18) is too complex to give directly an insight into interesting qualitative conclusions, although the form (21) for the z of factor (19) seems suitable for computations. I had hoped to present here the results of some investigations and computations on a number of examples, possibly giving some rough notion about the qualitative behavior of our problem according to' different circumstances. Unfortunately, there was a delay in installing the computer at the Institute. A few tests seem to confirm that the program is promising but it would be reckless and premature to anticipate any probable feature.

Let us mention only the very obvious and usual case of a single observation at great distance from all others. All the really interesting problems are thus left out, in order to illustrate on this very simple example how the notions considered do work in the Bayesian approach, at least inasmuch as we take seriously the interpretation of formula (13) or (14). If each observation is taken with an instrument with normal error but unknown precision, the weight ρ_h of the hth observation is proportional to a suitable mean value of the unknown weight λ_h which would belong to it if the precision of the instrument chosen and used on this occasion were known. This mean value is based on the information after the observations ("a posteriori" in the old terminology). That is, the probability distribution of the precision of the instrument used in taking a given observation is evaluated on the basis of the whole set of observations. It is clear that a value far from the others is likely to be produced by an instrument of low precision and will therefore be given a low weight. If one seeks a rough rule, such an observation may be neglected and called an *outlier*. In order to specify the conditions and the sense in which a practical rule of this kind is justified, it is necessary to work out, by the above method or one similar to it, some conclusions about the ρ_h.

5. The case of exchangeability

To begin with, let us illustrate this case with reference to a very particular example, namely, the case where θ is just a scaling parameter. That is, we consider only one error distribution with density $f(y)$, supposing however that the scale is "unknown," so that we may have any distribution $f_\theta(y) = 1/\theta f(y/\theta)$. The parameter θ may be any number $(0 < \theta < +\infty)$ with probability density $\phi(\theta)$. The real meaning of all that, from a consistently subjective point of view (for which the framework of unknown parameters is meaningless unless it comes out of a realistic description), is again that the error distribution is a mixture of the f_θ, but, unlike the mixtures over γ, the parameter is the same for all the observations, that is, the likelihood is no longer a product of mixtures but a mixture of products.

Precisely, the likelihood (4) is given by

$$(23) \qquad f(x|\cdots) = K\int_0^\infty f\left(\frac{x_1 - x}{\theta}\right)f\left(\frac{x_2 - x}{\theta}\right)\cdots f\left(\frac{x_n - x}{\theta}\right)\theta^{-n}\phi(\theta)\,d\theta,$$

and, if we suppose $f_0(x) =$ constant, the same holds for the final distribution.

If $f(y)$ is the normal distribution, this is nothing but the Bayesian approach to the classical problem of a normal distribution with unknown mean and variance. For such an approach, see [6], section 7(3). In the normal case, as is well known, no phenomenon of the kind of the outlier appears. It appears however fŏr all other distributions f, so that we need only, as an example and in order to make use of our result so far, consider again for $f(y)$ a mixture of normal distributions of the form (13).

In this case, formula (23) may be developed in the same way as (13). Everything follows the same lines, with only one more factor in the exponents, namely $1/\theta^2$, and one more integration in $d\theta$ (the factor θ^n is split for convenience into $\theta^{n-1}\theta$),

$$(24)\qquad K \int_0^\infty \frac{\phi(0)\,d\theta}{\theta^{n-1}} \int_0^\infty \cdots \int_0^\infty \alpha(\gamma_1) \cdots \alpha(\gamma_n) \left(\frac{\gamma \cdots \gamma_n}{\gamma}\right)^{1/2} \exp\left\{-\frac{1}{2\theta^2}\left[\sum_h \gamma_h x_h^2 - \frac{1}{\gamma}(\sum_h \gamma_h x_h)^2\right]\right\} \frac{\gamma^{1/2}}{\theta} \exp\left[\frac{\gamma}{2\theta^2}\left(x - \frac{1}{\gamma}\sum_h \gamma_h x_h\right)^2\right] d\gamma_1 \cdots d\gamma_n.$$

With the same integration used to obtain (18) from (15), we arrive at a similar expression for the expectation $\hat{x}$ (of the final distribution, supposing the initial one to have constant density). The last factor of (24) gives as before $(1/\gamma) \sum \gamma_h x_h$, so that $\hat{x}$ too is still the weighted mean (9) of the x_h with suitable weights ρ_h. These weights are again averages of the γ_h as in (18) with only one more integration in $d\theta$,

$$(25)\qquad \rho_h = K \int_0^\infty \frac{\phi(\theta)\,d\theta}{\theta^{n-1}} \int_0^\infty \cdots \int_0^\infty \alpha(\gamma_1) \cdots \alpha(\gamma_n) \left(\frac{\gamma_1 \cdots \gamma_n}{\gamma}\right)^{1/2} \gamma_h \exp\left(\frac{-z}{2\gamma\theta^2}\right) d\gamma_1 \cdots d\gamma_n$$

$$= K \int_0^\infty \cdots \int_0^\infty \alpha(\gamma_1) \cdots \alpha(\gamma_n) \left(\frac{\gamma_1 \cdots \gamma_n}{\gamma}\right)^{1/2} \gamma_h\, d\gamma_1 \cdots d\gamma_n \int_0^\infty \frac{\phi(\theta)}{\theta^{n-1}} \exp\left(\frac{-z}{2\gamma\theta^2}\right) d\theta$$

$$= K \int_0^\infty \cdots \int_0^\infty \gamma_h \Psi_n\left[\frac{1}{\gamma} z(\gamma_1 \cdots \gamma_n)\right] \alpha(\gamma_1) \cdots \alpha(\gamma_n) \left(\frac{\gamma_1 \cdots \gamma_n}{\gamma}\right)^{1/2} d\gamma_1 \cdots d\gamma_n,$$

where z is the quadratic form of (20) or (21), and the Ψ_n are the transforms of ϕ defined by

$$(26)\qquad \Psi_n(u) = \int_0^\infty \theta^{1-n} e^{-u/2\theta^2} \phi(\theta)\,d\theta.$$

The result is not much more complex than in the case of independence. It shows how even on the present hypotheses the dependence on the observations occurs through the same form z, (20) or (21). The meaning too, qualitatively, is still clear, and essentially the same as the case of independence with only one change, but an important one. To judge whether an observed value is *far* from the other values and should be considered an outlier, the distance is measured on a fixed scale in the case of independence, because the precision is that of the $f(y)$ assumed. In the case of exchangeability, on the other hand, we use a relative scale. The scale is relative to the distances between the observed values, in so far as the range of the distribution $\phi(\theta)$ allows. With a narrow range there is a very little alteration from a hypothesis of independence; the relativity becomes almost absolute for distributions approaching the degenerate cases of Jeffreys [4].

Of course in such conclusions the features essentially concerning the case of exchangeability are intermixed with those concerning only the particular example of a scaling parameter. In order to see what happens in a far more general case, a few remarks are sufficient, and the case we shall consider is indeed the most general one to which the notions used here are applicable, that is, the case where all functions f_0 are mixtures of normal distributions, of the form (13) or (14). Assuming they are of the form (13), we need only consider the functions $\alpha(\gamma, \theta)$ corresponding to f_θ and write down formulas (15) and (18) simply replacing the $\alpha(\gamma_1) \cdots \alpha(\gamma_n)$ with $\alpha(\gamma_1, \theta) \cdots \alpha(\gamma_n, \theta)$, multiplying by $\phi(\theta)\, d\theta$ and integrating over Θ (a one- or many-dimensional field, or any abstract space). Note that even now the dependence through the z only still holds. The particular case of (18) is the new one with $\alpha(\gamma, \theta) = \alpha_\theta(\gamma) = \theta^2\alpha(\gamma\theta^2)$.

6. The case of partial exchangeability

Let us give a very simple illustration, namely the case of formula (7) with the further simplifying assumptions that θ is simply a scaling parameter as in (23), and $f(y)$ is a mixture of normal distributions of the form (13). The situation is thus as follows, if we agree to explain it in terms of unknown parameters despite the unfitness of this interpretation from the subjective point of view. We distinguish three types of observations, such as those taken by three different persons, supposing that there might be some difference between the types. Precisely, we suppose that a parameter θ may have different values, θ_1, θ_2, and θ_3, for the three persons; the joint (initial) probability distribution has a density $\phi(\theta_1, \theta_2, \theta_3)$; if the values θ_i were known, the error distribution would be $f_i(y) = (1/\theta_i)f(y/\theta_i)$, where $f(y)$ is a given function for the observations of type i, and all errors of any kind would be independent.

The expression corresponding to (23) for this case, under the same assumption that $f_0(x)$ is nearly constant, becomes

$$(27)\quad f(x|\cdots)$$
$$= K \int_0^\infty \int_0^\infty \int_0^\infty f\left(\frac{x_1^{(1)} - x}{\theta_1}\right) \cdots f\left(\frac{x_{n_1}^{(1)} - x}{\theta_1}\right) f\left(\frac{x_1^{(2)} - x}{\theta_2}\right) \cdots$$
$$f\left(\frac{x_{n_2}^{(2)} - x}{\theta_2}\right) f\left(\frac{x_1^{(3)} - x}{\theta_3}\right) \cdots f\left(\frac{x_{n_3}^{(3)} - x}{\theta_3}\right) \phi(\theta_1, \theta_2, \theta_3)\theta_1^{n_1}\theta_2^{n_2}\theta_3^{n_3}\, d\theta_1\, d\theta_2\, d\theta_3.$$

Let us note two trivial cases. If the θ_i are independent,

$$\phi(\theta_1, \theta_2, \theta_3) = \phi_1(\theta_1)\phi_2(\theta_2)\phi_3(\theta_3),$$

the multiple integral is simply the product of the simple integrals, and the influence of the three groups of observations are independent. If the whole distribution is concentrated on the diagonal $\theta_1 = \theta_2 = \theta_3$, we are back to the case of exchangeability. What is worthy of remark is that the really interesting cases of partial exchangeability are chiefly those close to such a degenerate case.

In fact the practical situation for which partial exchangeability offers a satisfactory model is chiefly that arising when a little improvement of the model of exchangeability seems desirable in order to take into account some weak doubts about the possible differentiation between observations taken in different circumstances. The peculiar facts concern just the case of a weak doubt, because it is interesting to investigate the correct form of reasoning which must guide us in the evolution of our judgment about significance of differences in the frequency when there is a large number of possible classifications.

7. Conclusions

This paper provides no conclusions in the form of formulas for direct and general application. Rather, it purports to show that no conclusions of this nature are possible. The whole problem, in fact, depends not only on the different types of hypotheses but, within each hypothesis, on the particular form of the distribution of errors. The author is confident that further research along the lines of this paper, while throwing more light on the differences arising from the several possible cases, would not find some compromise solution acceptable in all circumstances.

It may happen, however, that a few particular instances provide a practically satisfactory model for a wider class of cases of interest; if such simplified or standardized forms of conclusions exist, it would certainly be worthwhile to find them and to investigate their properties thoroughly. An analogy may be the fact that we often consider it reasonable to employ a normal distribution (or a rectangular, or an exponential distribution) even though its similarity to the actual distribution is only qualitative or is accepted for the sake of simplicity.

REFERENCES

[1] B. de Finetti, "Sulla combinazione di osservazioni," *Atti XVII Riunione Soc. Ital. Statist.* (1957), Rome, 1958.

[2] ———, "La probabilità e la statistica nei rapporti con l'induzione, secondo i diversi punti di vista," *Induzione e Statistica* (C.I.M.E., Varenna), Rome, Cremonese, 1959.

[3] E. Hewitt and L. J. Savage, "Symmetric measures on Cartesian products," *Trans. Amer. Math. Soc.*, Vol. 80 (1955), pp. 470–501.

[4] H. Jeffreys, *Theory of Probability*, Oxford, Clarendon Press, 1948 (2nd ed.).

[5] H. Poincaré, *Calcul des Probabilités*, Paris, Gauthier-Villars, 1912 (2nd ed.).

[6] L. J. Savage, "La probabilità soggettiva nei problemi pratici della statistica," *Induzione e Statistica* (C.I.M.E., Varenna), Rome, Cremonese, 1959.