TWENTY FIVE YEARS OF BAYESIAN INFERENCE IN PSYCHOLOGICAL RESEARCH

HENRY ROUANET
Université René Descartes, Paris, France

Paper read at the Sixth World meeting of ISBA (International Society for Bayesian Analysis), 2000, Hersonissos, Crete, Greece.

SUMMARY

This paper is an introduction to the Bayesian work that my colleagues and I have conducted for a quarter of century in connection with psychological research. Using our publications in english–speaking psychological literature as a guideline, I revisit in detail the princeps 1976 Bayesian paper, introducing Bayesian data analysis for asserting largeness/smallness of effects of interest ; then I review the other landmark publications.

KEYWORDS

Largeness/Smallness of effects ; psychological research.

Introduction

For more than a quarter of a century, I have been involved with my colleagues of the *Math & Psy Group* in Paris in Bayesian data analysis, in connection with psychological research. Part of our work has been devoted to Bayesian ANOVA (Dominique Lépine and Bruno Lecoutre), another part to categorized data (Jean–Marc Bernard). Software has been developed that include Bayesian procedures in addition to conventional frequentist procedures.

While most of our work has been published in french language, there have been several publications in english–speaking psychological literature, that I will use as landmarks for the present introduction to our work. I will revisit in some detail the 1976 princeps paper, then I will quickly review the other references.

What led me in the seventies to Bayesian inference was the realization that, for issues of major importance to psychological research, the established statistical methodology was basically inadequate, and the conviction that Bayesian data analysis would definitely overcome the deficiencies of this methodology. The methodology we have arrived at is Bayesian, yet its message does not reduce to “Be Bayesian and go in peace !” ; it preserves what we believe is sound in the conventional methodology.

1. Educational study (Rouanet & al., 1976)

The Data

The data concerned 334 pupils divided into 4 groups (about 90 pupils in each group), defined by the crossing of two factors with 2 levels each, namely *Teaching*

Method (Modern vs Traditional) and *Environnement* (Privileged vs Underprivileged). There were 9 dependent variables, namely verbal IQ, nonverbal IQ, two combinatory tests C1 and C2, two probability tests P1 and P2, two Logic of propositions Tests LP1 and LP2, and a standard Mathematical test.

For each of the nine variables, the four group means were calculated and the following three observed effects of interest were derived : main effect of *Teaching* [difference between marginal means] ; main effect of *Environnement* [difference between marginal means] ; and *Interaction* effect between Teaching and Environnment [difference of differences]. The $9 \times 3 = 27$ observed effects are shown in Table 1. Effects in the Table are standardized, that is, each difference d_{obs} has been divided by the within-group standard deviation s_{obs} ; which renders the effects comparable across variables.

Looking at the *signs* of effects (positive signs have been omitted in the table) : For all variables, Modern teaching is more effective than Traditional (an encouraging finding) ; privileged children are more successful than underprivileged (a not unexpected finding).

Still from a descriptive standpoint, we can look at *sizes of effects.* To fix ideas, let us consider that an effect is *large* whenever $|d_{obs}|/s_{obs}$ is greater than $\ell_{lar} = 1/3$ (lower limit for Largeness) ; and that an effect is *small* whenever $|d_{obs}|/s_{obs}$ is less than $\ell_{sma} = 1/4$ (upper limit for Smallness). Then among the 27 effects, we find 9 large effects and 13 small ones.

TAB. 1: *Standardized observed effects* d_{obs}/s_{obs}

	Teaching	Environment	Interaction
VIQ	0.20 Small	0.59 Large	0.30
NVIQ	0.39 Large	0.51 Large	0.17 Small
C1	0.54 Large	0.36 Large	0.16 Small
C2	0.72 Large	0.40 Large	0.19 Small
P1	0.30	0.28	−0.05 Small
P2	0.20 Small	0.53 Large	−0.16 Small
LP1	0.10 Small	0.23 Small	0.23 Small
LP2	0.25	0.37 Large	0.02 Small
Math	0.08 Small	0.32	0.11 Small

The Largeness/Smallness issue

The Educational Data exemplify the Largeness/Smallness issue, ubiquitous in psychological research ; this issue involves two situations.

Situation 1 (as for Verbal IQ, Environment) : The observed effect is large and *Asserting Largeness* is sought. That is, the researcher wishes to conclude that, allowing for sampling variation, the “true effect” (i.e. population effect) δ is *large* — on the direction of the observed effect, needless to say : Largeness conclusions are naturally oriented. Informally speaking : “There is an effect”.

Situation 2 (as for Math, Interaction) : The observed effect is small and *Asserting Smallness* is sought. This time, the researcher wishes to conclude that the true effect δ is *small* — regardless of direction : Smallness conclusions are naturally nonoriented. Informally speaking : "There is no effect".

Now the common practice in both situations is to proceed to significance testing for the nullity of effects, with the hope of ending up in Situation 1 with a significant effect (hopefully highly significant), and in Situation 2 with a nonsignificant effect (hopefully "largely nonsignificant" so to say). If we do this for the Educational Data, assuming for each effect of interest the usual normal sampling model and testing $\mathcal{H}_0 : \delta = 0$ by means of a t–test (See Appendix), we find 14 significant effects, i.e. for which $p > .05$ (two–sided) ; and 9 "largely nonsignificant" effects, i.e. for which $|t| < 1$; as shown on Table 2.

TAB. 2: *Significance testing of nullity of effects.* $\star$ *:* $p > .05$ *(two–sided). NS :* $|t| < 1$.

	Teaching	Environment	Interaction
VIQ		$\star$	
NVIQ	$\star$	$\star$	NS
C1	$\star$	$\star$	NS
C2	$\star$	$\star$	NS
P1	$\star$	$\star$	NS
P2		$\star$	NS
LP1	NS	$\star$	
LP2	$\star$	$\star$	NS
Math	NS	$\star$	NS

Standard formulations

In research papers, the following formulations are standard :

. For a significant result : "There is evidence of effect".

. For a "largely nonsignificant" result : "There is no evidence of effect".

Though statistically correct, these formulations miss the target researchers have in mind, that is, asserting Largeness or alternatively Smallness of effects. To remind researchers to this sobering fact, all textbooks of Statistics for psychologists spell out the two *ritual Warnings* :

. Warning #1 : *Statistical significance is not psychological significance.*

. Warning #2 : *No evidence of effect is not proof of no effect.*

The Educational Data illustrate Warning #1. There are four significant effects that are not even descriptively large ; which clearly precludes any largeness conclusion for the true effect. Thus there is a conflict between the descriptive conclusions in terms of Largeness vs Smallness and the results of significance tests, as is apparent from the comparison of Tables 1 and 2.

Bayesian Data Analysis

At this point it is obvious for those familiar with Bayesian inference that the Bayesian approach is ideally suited to handle both Situations 1 and 2. In Bayesian inference, a *prior distribution*, expressing uncertainty about parameters independently from the data, is postulated and combined with the sampling distribution and data to yield a *posterior distribution.* The posterior distribution expresses the uncertainty about the parameters, conditionally on data. Therefore, if the bulk of this distribution lies in the region of large effect values, the probability is high that the true effect is large, so largeness of effect can be asserted ; if it lies in the region of small effect values, the probability is high that the true effect is small, so smallness of effect can be asserted.

What prior should we take ? In our practice, we always use noninformative priors, at least as a starting point (see Appendix for the analysis of Educational Data).

Examples in Educational Data

Example 1 : Verbal IQ, Environnemnt. The observed effect 0.59 is large. The major part of the posterior distribution (centered around 0.59) lies in the region of large values. Letting the *credibility level* $\gamma = .90$, we find that $P(\frac{\delta}{s_{obs}} > 0.45) = .90$ hence $P(\frac{\delta}{s_{obs}} > 1/3) > .90$. That is, the true effect δ is greater that $\ell_{lar} = 1/3$ at credibility level $\gamma = .90$. As a conclusion, we can extend the descriptive conclusion at the credibility level .90 ; we assert that the main effect of Environnment on Verbal IQ is large on the side of Data, i.e. of privileged environment.

Example 2 : Nonverbal IQ, Teaching. Here again the observed effect is large, but this time it is not the case that the major part of the posterior distribution covers large values. We find that $P(\frac{\delta}{s_{obs}} > 0.26) = .90$; hence $P(\frac{\delta}{s_{obs}} > 1/3) < .90$. At the credibility level .90, we cannot extend the descriptive conclusion of a large effect of Teaching on Nonverbal IQ.

Example 3 : Math test, Teaching. The observed effect is small, and we find : $P(\frac{|\delta|}{s_{obs}} < 0.22) = .90$, hence $P(\frac{|\delta|}{s_{obs}} < 1/4) > .90$. We assert that the main effect of Teaching on Math test is small.

Example 4 : Probability P2, Teaching. The observed effect is small, but we find : $P(\frac{|\delta|}{s_{obs}} < 1/4) < .90$. The descriptive conclusion cannot be extended.

The overall set of Bayesian results is shown on Table 3. “Large” is underlined (Large) when Largeness is asserted at the credibility level .90, i.e. when $P(\frac{\delta}{s_{obs}} > 1/3) > .90$. “Small” is underlined (Small) when Smallness is asserted : $P(\frac{|\delta|}{s_{obs}} < 1/4) > .90$. Signs of effects appear only for Largeness conclusions. Comparing Table 3 with Table 1 of observed effects, it is apparent that not only there is no conflict with descriptive conclusions, but that the Bayesian data analysis is a natural extension of the descriptive analysis.

The conventional methodology revisited

The comparaison of the Bayesian results with the Significance Tests ones is striking. Among the 14 significant results (S*), only 9 largeness conclusions are reached ; among the 9 largely nonsignificant results (NS), only 2 smallness

TAB. 3: *Synopsis of Bayesian results (Credibility level .90)*

	Teaching	Environment	Interaction
VIQ		+Large	
NVIQ		+Large	
C1	+Large		
C2	+Large		
P1			
P2		+Large	
LP1	Small		
LP2			
Math	Small		

conclusions are reached. Of course, those numbers depend on the conventions adopted for the limits of largeness and smallness, as well as for the levels of credibility (and significance !). Yet the two patterns of results cannot be "reconciled" by juggling with those limits. The insurmountable character of this difficulty is shown by the Bayesian reinterpretation of significance level in the elementary case with noninformative priors. If p denotes the two–sided observed level (p value), then $1 - p/2$ is simply the probability that the effect δ has the same sign as the observed effect d_{obs}, and $1 - p$ is the probability that δ lies between 0 and $2d_{obs}$ (see e.g. Rouanet, 1996).

Consequently, if an observed effect is large— a proviso not to be forgotten — finding a significant result can indeed be regarded as providing some support to the descriptive conclusion, in so far as it means that, at least, the *existence* of effect — and for a 1 d.f. effect its *direction* — is established, and this can be viewed as a necessary condition for asserting largeness. On the other hand, if an observed effect is small, finding a NS result is not sufficient to assert smallness (as exemplified in the Educational Data) ; it is not necessary either, since for large samples, trivially small effects can be significant — that is, ascertained to be nonzero. As a conclusion, in order to assert smallness, significance testing is irrelevant, and would be better avoided even as a first step.

Summarizing : When the issue at stake is Largeness/Smallness of effects, the significance testing of effects is basically unsound. When correctly interpreted, it answers the wrong question ; when incorrectly interpreted, it is misguided. Here is what we wrote in the 1976 paper :

> There are several ways of misusing significance tests. One is to superstitiously stick to the sacred levels (.05 or .01, etc.) Another one is to ignore validity assumptions. The present line of criticism is more basic in character ; a significance test may be perfectly *valid* in a given situation, yet at the same time absolutely *irrelevant* for the objectives of the study ; important as they may be, considerations of validity should be subordinated to those of relevance. In other words, before thinking of performing a significance test, the first question to ask ought *not* to be : "*May I* do this test ?", but "*Should I* do this test ?"

A strategy for asserting Largeness /Smallness

For dealing with the Largeness/Smallness issue, the statistical strategy we have arrived at can be sketched as follows (notice the asymmetrical way significance testing is incorporated). The order of successive phases is mandatory. At each phase, if the answer to the question is "No", the process comes to an end.

For each question of interest :

. Derive a specific relevant data set.
. Devise an index of importance of effect.

- *Situation 1, Largeness*
  Descriptive Phase. Is Observed Effect Large ?
  Significance Test. Can Existence and /or (1 d.f.) Direction of Effect be ascertained ?
  Bayesian Phase. Can Largeness of Effect (on the side of data) be asserted ?
- *Situation 2, Smallness*
  Descriptive Phase. Is Observed Effect Small ?
  Bayesian Phase. Can Smallness be asserted ?
  In Situation 2, testing null effect is irrelevant.

For the Educational Data, this strategy generates the results of Table 3.

At the credibility level $\gamma = .90$, the following conclusions can be asserted for main effects. Taking 1/3 as a criterion for largeness of standardized effects, Modern Teaching can be asserted to be *largely more effective* than Traditional for variables C1 and C2 ; and privileged Environment largely more effective than underprivileged for variables VIQ, NVIQ and P2. Taking 1/4 as a criterion for smallness, the main effects of Teaching can be asserted to be *small* for variables LP1 and Math.

2. Other Landmarks

Validating Models (Rouanet & al., 1978)

In the context of model validation, the blind practice of goodness-of-fit statistics, with no concern for descriptive statistics of model appraisal, is especially questionable. "Experimental evidence is consistent with the model" : This statement, typically found in research papers when the goodness–of–fit is non-significant, albeit in itself statistically correct, is highly misleading, as soon as in the ensuing line of argumentation the model's validity is taken for granted.

In the Validating Models paper, this issue was discussed in connection with a model classical in experimental psychology, namely the model of additive stages in reaction times, and a Bayesian solution was fully developed.

Specific Inference (Rouanet & Lecoutre, 1983)

Psychological experiments often involve complex designs ; an important part of our work has been to develop Bayesian ANOVA procedures applicable to complex designs. One crucial step in this direction has been to devise the methodological approach that we call *specific inference.* In this approach — as opposed to the conventional General Linear Model — a relevant data set is derived for each question of interest, a *specific model* is put on this data set, and the inference is made through this specific model. In the Bayesian framework, specific inference can readily be formalized in terms of of partial sufficiency.

In complex designs, the distributional assumptions involved in a specific inference are simply those of the corresponding specific model, regardless of the complexity of the design. For any source of variation for which there is a valid F–test, there is a corresponding valid specific Bayesian procedure based on the same two sums of squares. In practice, this means that if you know how to build the ANOVA table and F ratios, Bayesian ANOVA extensions are readily available.

Psychological Bulletin (Rouanet, 1996)

In spite of the developments of Bayesian techniques, the established "statistics for psychologists" has remained virtually unchanged for decades. To challenge such an astonishing state of facts, I wrote a paper for the *Psychological Bulletin* — the first Bayesian paper published in this journal since 1969 ! — emphasizing the advances brought by Bayesian data analyis.

Peter Lang book (Rouanet et al., 1998)

The Peter Lang book, also directed to researchers and statisticians in behavioral and social sciences, is a comprehensive presentation of the work done in our *Math & Psy* research group on statistical inference. Following a Foreword by Patrick Suppes (Stanford University), the chapters are : *Statistics for researchers*, *Statistical practice revisited* (Henry Rouanet) ; *What about the researcher's point of view ?* (Marie-Paule Lecoutre) ; *Introduction to combinatorial inference* (Henry Rouanet & Marie-Claude Bert) ; *From significance tests to Fiducial Bayesian inference* (Bruno Lecoutre) ; *Bayesian inference for categorized data* (Jean-Marc Bernard) ; *Geometric Data : from Euclidean clouds to Bayesian* MANOVA (Henry Rouanet, Brigitte Le Roux, Jean-Marc Bernard & Bruno Lecoutre).

Additional Comment

(In connection with the discussion following the talk) :

(i) What we object to current statistical practice is that with the moderate sample sizes commonly used, too many significant effects are unduly interpreted as large effects, *and also* too many nonsignificant effects are unduly interpreted as small effects. Thus our criticism is two–fold and certainly does not reduce to the claim that current practice produces too many significant results.

(ii) The far–ranging character of the methodology we are proposing will be appreciated when it is realized that the situations that are (by far) the most prevalent in actual research are those where the hypothesis of no effect is merely a "dividing hypothesis" (in Cox' phrase) and where enlarged hypotheses (largeness/smallness of effects) are actually at stake; as opposed to the situations — often dealt with in frequentist but also Bayesian statistical work — involving sharp null hypotheses (zero vs nonzero effect).

(iii) I thank Jean–Marc Bernard for our stimulating Cretan discussions.

REFERENCES

Rouanet, H. (1996). Bayesian methods for assessing importance of effects. *Psychological Bulletin*, 119, 1, 149-158.

Rouanet, H. & Lecoutre, B. (1983). Specific inference in ANOVA. *British Journal of Mathematical and Statistical Psychology*, 36, 252-268.

Rouanet, H., Lépine, D. & Holender, D. (1978). Model acceptability and the use of Bayes–fiducial methods for validating models. In J. Requin (Ed.), *Attention and performance VII*, 687-701. Hillsdale, NJ : Erlbaum.

Rouanet, H., Lépine, D. & Pelnard–Considère J. (1976). Bayes–fiducial procedures as practical substitutes for misplaced significance testing : An application to educational data. In D.N.M. de Gruijter & al. (Eds.), *Advances in psychological and educational measurement*, 33-48. New York : Wiley.

Rouanet, H., Bernard, J.M., Bert, M.C., Lecoutre, B., Lecoutre, M.P., Le Roux, B. (1998). *New Ways in Statistical Methodology : From Significance Tests to Bayesian Inference.* Bern, Switzerland : Peter Lang.

Appendix : Technical Details

We assume for each effect of interest the usual normal sampling model :

$$d|\delta, \sigma \sim \mathcal{N}(\delta, \tfrac{\sigma^2}{n}) \text{ hence under } \mathcal{H}_0 : \tfrac{d}{s} \sim t(0, \tfrac{1}{n})$$

where $\widetilde{n}$ is a number homogeneous to a sample size. We have $\widetilde{n} = 90.73$ for main effects, and $90.73/4 = 22.67$ for interactions. Owing to the large number of d.f. for s, the t–distribution amounts to the normal one. Effects are significant at $p < 0.05$ (two–sided) (written S($\star$)) if $d_{obs}/s_{obs} > 1.96/\sqrt{\widetilde{n}}$; that is, for mean effects if $d_{obs}/s_{obs} > 0.21$, and for interaction effects if $d_{obs}/s_{obs} > 0.41$. By a "largely nonsignificant" effect we mean $\sqrt{\widetilde{n}}\, d/s < 1$, hence $d/s < 0.10$ for main effects and $d/s < 0.21$ for interaction.

Assuming for each effect the standard noninformative priors of classical elementary normal theory, the posterior distribution of "true effect" δ is a scaled t–distribution :

$$\tfrac{\delta}{s_{obs}} \sim t_q(d_{obs}, \tfrac{1}{n}) \text{ (practically here a normal distribution)}$$

Then taking a *credibility level* γ, we can assert Largeness (for $d > 0$) if $P(\frac{\delta}{s_{obs}} > 1/3) > \gamma$; and we can assert Smallness if $P(\frac{|\delta|}{s_{obs}} < 1/4) > \gamma$. For the educational data we took $\gamma = .90$ —- a more reasonable choice, we think, than simply taking the complementary value of the familiar .05 significance level, because asserting largeness or Smallness is a more demanding task than asserting significance.

With those conventions, for the Educational Data, mean positive effects can be asserted to be large at the credibility $\gamma = .90$ if $d_{obs}/s_{obs} > 0.47$, and positive interaction effects if $d_{obs}/s_{obs} > 0.39$. Mean effects can be asserted to be small if $|d_{obs}|/s_{obs} < 0.15$; no interaction effect can be asserted to be small, because $\widetilde{n} = 22.67$ is too small.