

ASSESSING PSYCHOLOGY STUDENTS' DIFFICULTIES WITH CONDITIONAL PROBABILITY AND BAYESIAN REASONING

Carmen Díaz and Inmaculada de la Fuente
University of Granada, Spain
mcdiaz@ugr.es

Conditional probability and Bayesian reasoning are important to psychology students because they are involved in the understanding of classical and Bayesian inference, regression and correlation, linear models, multivariate analysis and other statistical procedures that are often used in psychological research. A study of previous literature showed that there is considerable research on this topic, but no comprehensive questionnaires have been developed to globally assess students' understanding and misconceptions on these topics. At the University of Granada we started building a questionnaire, which takes into account the content of conditional probability taught in the Spanish universities to psychology students, as well as the biases and misconceptions described in the literature. In this work we will describe the process of developing the questionnaire and will report the results from a sample of 206 psychology students.

INTRODUCTION

Conditional probabilities and Bayes theorem are highly relevant in the field of psychology, where these types of reasoning appear in diagnosis, evaluation, decision-making and application of statistical inference in experimental research. However, nowadays there is a tendency to suppress or reduce the teaching of probability in data analysis courses in the training of psychologists. This tendency does not take into account recent research related to conditional probability and Bayesian reasoning that suggests students have frequent difficulties in understanding and applying these concepts. These difficulties include confusing conditional and causal reasoning and the *fallacy of the time axis*, or belief that in $P(A|B)$, event B should always precede in time event A (Falk, 1986, 1989), misconceptions regarding conditional probability (Gras and Totohasina, 1995; Ojeda, 1996), problems in defining the conditioning event (Bar-Hillel and Falk, 1982) and misunderstanding of independence (Sánchez, 1996; Truran and Truran, 1997). Solving conditional probability problems is also affected by the type and format of data even for very elementary problems (Lonjedo and Huerta, 2005).

As regards Bayesian reasoning (see a summary in Koehler, 1996), early research by Tversky and Kahneman (1982) suggested that people do not employ this reasoning intuitively and established the robustness and spread of the *base-rate fallacy* in students and professionals (Bar-Hillel, 1997). More recently, some authors (e.g., Gigerenzer, 1994; Martignon and Wassner, 2002) suggested that Bayesian computations are simpler when information is given in natural frequencies, instead of using probabilities or percentages. Sedlmeier (1999) summarised recent teaching experiments that follow this approach and involve the use of computers.

Although all this previous research pointed to specific difficulties in conditional probability and Bayesian reasoning for different tasks, we did not find a comprehensive instrument that could be used to assess in a reasonable time and in just one application the different biases and misunderstanding of these concepts in the same student. We therefore decided to develop such an instrument to assess understanding of these topics by psychology students at the University of Granada. It is part of a research project that aims to prepare and assess materials to introduce simple ideas of Bayesian reasoning to psychology students. The final goal in building the instrument is using the results of assessing students' difficulties in improving the teaching of conditional probability and, consequently, better preparing our students to start the study of Bayesian inference. Preliminary results were given in Díaz (2004).

METHODS

The building of the questionnaire was based in a rigorous methodological process, including expert judgment, item trialling, validity and reliability assessment and Bayesian estimation of items' psychometric features. All the students in the different samples in this study were in the first year of Psychology and all of them followed an introductory statistics course with

the same programme. The different tests were all given after the students had been taught conditional probability and Bayes theorem. The questionnaires were completed as an activity in the course of data analysis, and the students were asked to study the topic in advance. Below we describe the different steps in the process.

Table 1: Primary and secondary content assessed by each item

	Content units	Items including this content	
		Specific to assess this content	Content included implicitly
Conceptual understanding	1. Defining conditional probability	I1	(*)
	2. Recognising that conditional probability involves a restriction in the sample space	I2	
	3. Distinguishing a conditional probability and its inverse	I10	I3
	4. Distinguishing conditional, simple and joint probabilities	I3	(*)
	5. Conjunction fallacy	I9	
	6. Distinguishing independent and mutually exclusive events	I4	I6, I8, I14, I15
Procedural understanding	7. Computing conditional probabilities in a simple experiment	I6	I13
	8. Solving conditional probability problems in a sampling with replacement setting	I8	I2
	9. Solving conditional probability problems in a sampling without replacement setting	I5	I14, I15
	10. Computing conditional probabilities from joint and compound probabilities	I7	I3, I12, I13, I17, I18
	11. Solving conditional probability problems when the time axis is reversed	I18	I3, I12, I16, I17
	12. Distinguishing conditional, causal and diagnosis situations	I11	I14, I12
	13. Solving conditional probability problems in a diachronic setting (sequential experiments)	I12	I5, I8, I17
	14. Solving conditional probability problems in a synchronic setting (simultaneous experiments)	I14	I6, I7, I11, I16
	15. Solving compound probability problems by applying the product rule to independent events	I13	I4
	16. Solving compound probability problems by applying the product rule to dependent events	I15	I3, I11, I16
	17. Solving total probability problems	I11	I3, I16
	18. Solving compound Bayes problems	I3, I16	

(*) Whenever conditional probability is involved

1. *Semantic definition of the variable:* In order to achieve objectivity, the conceptual and procedural content units included in Table 1 were defined after content analysis of 19 text books used in the teaching of data analysis to psychologists. Content analysis is based on the idea that units in a text can be classified in a reduced number of categories that serve to carry out inferences about the text-specific characteristics (Ghiglione and Matalón, 1989). To select the books, the list of references recommended in these courses was requested from the 31 Faculties of Psychology in Spain. All the textbooks recommended in data analysis courses by at least 4 different Faculties were analysed, after discarding those in which conditional probability was not included. The analysis also took into account our review of research on conditional reasoning in psychology and mathematical education, which identified a bank of $n=49$ different items used in this research.
2. *Final selection of items (Study 1):* This took into account the items' psychometric characteristics (item difficulty and discrimination), which were estimated from the answers by different samples of psychology students (between 49 and 117 students depending on the item), as well as expert judgment (Osterlind, 1989) by 10 researchers in statistics education from 5 different countries (Brazil, Colombia, Mexico, Spain, and Venezuela) who have themselves carried out research related to conditional probability or independence. These researchers were asked to rate (on a 5 point scale) the adequacy of the content units for

- understanding conditional probability, as well as the suitability of each item to assess understanding of a specific content unit.
3. *Formatting and revising items.* Two different formats of item were used: a) Multiple choice items with 4 possible responses were used to allow quick evaluation in the sample of some of the most pervasive biases described in the previous literature; b) Open-ended items were also used to assess strategies in problem solving and gain a deeper picture of students' conditional reasoning. Successive revisions of the items were made to improve its understanding.
 4. *Pilot trial of the instrument* took place in the academic year 2003-2004 with a small sample of 57 Psychology students (Study 2). The aim was to make an initial estimate of questionnaire reliability and validity, and to compare the estimation of difficulty indexes by classic and Bayesian inferential procedures. A second sample of 37 students majoring in Mathematics served to compare performances in the two groups and identify items with and without discriminative properties.
 5. *Revising the Pilot questionnaire.* After discarding those items with bad psychometric features, further expert judgement served to improve the wording of the items. Thirteen expert methodology instructors were given 3 alternative wordings for each item (all of them with identical content and context) and asked to select the best version, considering methodology standards, as well as give the reasons for their choice. For each item the version preferred by the majority was selected and additional suggestions by the methodology instructors were used to improve readability further.
 6. *Trial of the revised instrument (Study 3).* Finally the questionnaire was given to 206 students in the same group and academic year as those who participated in Studies 1 and 2.

There was a total of 18 items in the final questionnaire, some examples of which are shown in Figure 1. Some items were multiple-choice (e.g., items 3, 9, 10 and 17), and some had two parts (e.g., item 17), which are scored independently. Other items were open ended (e.g., items 1, 2, 8, 15 and 16) and asked for a definition, examples or the solution to a problem. In these cases scoring is a measure of the completeness of the response. In Table 1 we present the content of the items. The complete questionnaire can be requested from the author.

SOME RESULTS

Below we report results from Study 3. For each item the answers were analysed. In case of open-ended items the scoring took into account the completeness of response (For example, in item 15 students were given a point if they identified correctly the problem data, 2 points if they correctly built a tree diagram and identified conditional probability in the response, 3 points if they correctly computed total probability in the denominator of Bayes' formula, and 4 points for the complete solution.) The maximum possible scoring in the questionnaire was 34 points. The empirical distribution of scores ranged between 3 and 30 with an average of 19.12 points and standard deviation of 5.91 points.

In computing difficulty indexes 'complete or almost complete' responses were scored 'correct' and all other responses were scored 'error' and we carried out both Frequentist and Bayesian estimates (see Table 2). In the Frequentist estimate only the final results in the questionnaire ($n=206$) are taken into account, because in a Frequentist perspective you do not take into account previous information about the parameters. We computed the 95% confidence intervals with the usual normal approximation to the binomial. The difficulty indexes (see Table 2) ranged between 0.15 (I14, distinguishing independent and mutually exclusive events) and 0.88 (I6A, computing single probabilities from a 2×2 contingency table). Some difficult items with less than 30% percent of correct responses were I4 (distinguishing mutually exclusive and independent events), I7 (computing conditional probability from single and compound probabilities), I9 (conjunction fallacy) and I10 (distinguishing a conditional probability and its inverse). Easier items with more than 60 % correct responses were I5 and I18 (computing compound probabilities in a sampling without replacement setting), I11 (computing total probability), I12 (conditional probability, diachronic setting) and I6A (computing simple probability from a 2×2 table); I2 (sample space) and I15 (product rule).

- Item 1.* Define simple and conditional probability and provide an example of each one.
- Item 2.* Complete the sample space in the following random experiments: a) Observing gender (male/female) of children in a three children family (e.g., MFM,...); b) Observing gender (male/female) of children in a three children family when two or more children are males.
- Item 3.* A witness sees a crime involving a taxi in Carborough. The witness says that the taxi is blue. It is known from previous research that witnesses are correct 80% of the time when making such statements. The police also know that 15% of the taxis in Carborough are blue, the other 85% being green. What is the probability that a blue taxi was involved in the crime?
 a) $80/100$; b) $15/100$; c) $(15/100) \times (80/100)$; d) $\frac{15 \times 80}{85 \times 20 + 15 \times 80}$
- Item 8.* In throwing two dice the product of the two numbers was 12. Which is the probability that none of the two numbers is a six (we differentiate the order of numbers in the two dice).
- Item 9.* Suppose a tennis star goes to the Roland Garros final in 2005. He has to win 3 out of 5 sets to win. Which of the following events is more likely? a) The player wins the first set; b) He wins the first set but loses the match; c) Both events a) and b) are equally likely
- Item 10.* A cancer test was given to all the residents in a large city. A positive result was indicative of cancer and a negative result of no cancer. Which of the following results is more likely? a) That a person had cancer if they got a positive result; b) Having a positive test if the person had cancer; c) The two events are equally likely
- Item 15.* According a recent survey, 91% of the population in a city usually lie and 36% of them usually lie about important matters. If we pick a person at random from this city, which is the probability that the person usually lies about important matters?
- Item 16.* Two machines M1 and M2 produce balls. Machine M1 produces 40 % and M2 60% of balls. 5% out of the balls produced by M1, and 1% out of those produced by M2 are defective. We take a ball at random and it is defective. Which is the probability that that ball was produced by machine M1?
- Item 17.* Two black marbles and two white marbles are put in an urn. We pick a white marble from the urn. Then, without replacing the white marble in the urn, we pick a second marble at random from the urn.
- Which is the probability that the second marble is white if the first marble was white? $P(W_2/W_1)$
 a) $1/2$; b) $1/6$; c) $1/3$; d) $1/4$
 - Which is the probability that the first marble is white if the second marble was white? $P(W_1/W_2)$
 a) $1/3$; b) Cannot be computed; c) $1/6$; d) $1/2$

Figure 1: Sample items

Table 2: Classical and Bayesian Estimates of Difficulty Indexes

Item	Difficulty Index	95% Confidence interval		Bayesian estimate of difficulty index		
	(n=206)	Inf.	Sup	Mean value	Credibility interval (95%)	
I1	0.37	0.30	0.43	0.52	0.46	0.57
I2	0.85	0.80	0.89	0.85	0.80	0.89
I3	0.51	0.44	0.57	0.42	0.37	0.47
I4	0.19	0.14	0.25	0.27	0.23	0.32
I5	0.85	0.80	0.90	0.79	0.75	0.83
I6A	0.88	0.83	0.92	0.81	0.76	0.85
I6B	0.58	0.51	0.65	0.60	0.54	0.65
I6C	0.52	0.46	0.59	0.54	0.49	0.60
I6D	0.47	0.40	0.54	0.54	0.49	0.60
I7	0.29	0.22	0.35	0.39	0.34	0.44
I8	0.35	0.28	0.41	0.43	0.38	0.48
I9	0.26	0.20	0.32	0.38	0.33	0.44
I10	0.30	0.23	0.36	0.38	0.34	0.43
I11	0.69	0.63	0.76	0.48	0.43	0.52
I12	0.66	0.60	0.72	0.71	0.66	0.76
I13	0.49	0.42	0.56	0.51	0.46	0.57
I14	0.15	0.10	0.20	0.29	0.24	0.34
I15	0.67	0.61	0.73	0.62	0.57	0.68
I16	0.57	0.51	0.64	0.60	0.54	0.65
I17A	0.60	0.53	0.66	0.63	0.58	0.68
I17B	0.25	0.19	0.31	0.35	0.30	0.40
I18	0.77	0.71	0.82	0.75	0.71	0.79

To compute Bayesian estimates and credibility intervals in Study 3 we used the information from Studies 1 and 2 in three steps since Bayesian analysis allows use of information from previous studies. Even when the items in these previous studies had been slightly reworded, the variations were minimal (identical content, context and distracters; similar students), it was unreasonable to assume complete ignorance about items' difficulties. Firstly, we used a non-informative prior distribution for proportion of correct answers in each item and deduced the posterior distribution of difficulty indexes after the item trials (Study 1). Then, for each item the posterior distribution is Beta ($s+0.5, f+0.5$), where s and f are the successes and failures in Study 1. These posterior distributions were used as prior distributions of proportions for each item in Study 2 ($n=57$), getting a new posterior distribution for each index difficulty Beta ($s+s'+0.5, f+f'+0.5$), where s' and f' are the successes and failures in the item in Study 2.

In the last step (Study 3), we used this posterior distribution as a prior distribution of index difficulty to compute the final distribution and credibility intervals we are presenting in Table 2. We can see in this table that Bayesian estimation often is different from the Frequentist estimate, and this is true in particular when the sample in the earlier trials of this item was relatively large. Bayesian estimates also provide more precise intervals, since they combine results from the current sample with the previous information available. Finally, these intervals give a range of values for the population difficulty index (Serrano Angulo, 2003).

Our first approach to the reliability of the instrument was carried out by computing the coefficient Alpha, that gave a moderate value (Alpha=0.721), which is reasonable, given that the questionnaire is multidimensional, in that it assesses a wide range of knowledge (see Table 1), so a particular student can understand some of these concepts and not understand others. A second estimation of reliability used test-retest in a sample of 106 students taking part in Study 3. These students completed the questionnaire twice (with a month between testing times) and provided a reliability coefficient of 0.859, which is reasonably high. A theoretical analysis of the questionnaire content as well as the results from experts' judgement served to justify content validity comparing the content evaluated by each item to the semantic units included in the semantic definition.

A more complete analysis of the incorrect responses to the items is included in Díaz (2004), where we describe examples of students' responses and found the following percentages of students with problems in understanding conditional probability: not correctly restricting the sample space in conditional probability (26.3%), base-rate fallacy (16.3%), incorrect identification of Bayes' formula (29.1%), confusion of independence with mutual exclusiveness (30.1 % in item 7, 57% in item 10), only considering independence in diachronic experiments (18.4%), difficulties in reading conditional or joint probabilities from a 2x2 table (between 30 and 40% in the different parts of Item 6), confusing a conditional probability with its inverse (33%, conjunction fallacy (7%), and difficulties in computing probabilities when the time axis is inverted (80% in items 14 and 17b).

DISCUSSION

Given the relevance of understanding and correctly applying conditional probability to progress in the study of inference, regression, correlation and multivariate analysis, all these results give good reason for us to continue this research. At the same time, it shows the utility of the CPR (conditional probabilistic reasoning) questionnaire that can be used by lecturers and researchers to assess with just an instrument understanding of conceptual and procedural knowledge as regards conditional probability, as well as a variety of related misconceptions. We are now analysing the questionnaire responses to open-ended items in this sample, as well as analysing new results from giving the questionnaire to other samples of psychology students in the Universities of Huelva, Jaén and Murcia (250 new students). We are also starting to prepare didactic units that allow psychology students to overcome the described problems, and that serve to introduce them to some basic ideas of Bayesian inference.

ACKNOWLEDGEMENTS

This work was supported by the project BSO2002-0334-7 and grant AP2003-5130, MEC, Madrid, Spain.

REFERENCES

- Bar-Hillel, M. (1987). The base rate fallacy controversy. In R. W. Scholz (Ed.), *Decision Making Under Uncertainty*, (pp. 39 – 61) Amsterdam: North Holland.
- Bar-Hillel, M., and Falk, R. (1982). Some teasers concerning conditional probabilities. *Cognition*, 11, 109-22.
- Díaz, C. (2004). *Elaboración de un instrumento de evaluación del razonamiento condicional. Un estudio preliminar*. (Building an instrument to assess conditional reasoning). Master's Thesis. University of Granada, Spain.
- Gras, R. and Totohasina, A. (1995). Chronologie et causalité, conceptions sources d'obstacles épistémologiques à la notion de probabilité. *Recherches en Didactique des Mathématiques*. 15(1), 49 – 95.
- Falk, R. (1986). Conditional probabilities: insights and difficulties. In R. Davidson and J. Swift (Eds.), *Proceedings of the Second International Conference on Teaching Statistics*, Victoria, BC: University of Victoria, (pp. 292–297). Voorburg, The Netherlands: International Statistical Institute.
- Falk, R. (1989). Inference under uncertainty via conditional probabilities. In R. Morris (Ed.), *Studies in Mathematics Education: Vol. 7: The Teaching of Statistics*, (pp. 175-184). Paris: UNESCO,
- Ghiglione, R. and Matalón, B. (1991). *Les enquêtes sociologiques. Théories et pratique*. (Theory and practice of sociological surveys). Paris: Armand Colin.
- Gigerenzer, G. (1994). Why the distinction between single-event probabilities and frequencies is important for psychology (and vice-versa). In G. Wright and P. Ayton (Eds.). *Subjective Probability*, (pp. 129 – 161). Chichester: Wiley.
- Koehler, J. J. (1996). The base rate fallacy reconsidered: Descriptive, normative, and methodological challenges. *Behavior and Brain Sciences*, 19, 1-54.
- Lonjedo, M. A. and Huerta, P. (2005). The nature of the quantities in a conditional probability problem. Its influence in the problem resolution. Paper presented at the CERME IV Conference, <http://cerme4.crm.es/Papers%20definitius/5/wg5litofpapers.htm>.
- Martignon, L. and Wassner, C. (2002). Teaching decision making and statistical thinking with natural frequencies. In B. Phillips (Ed.), *Proceedings of the Sixth International Conference on Teaching of Statistics*, Cape Town. Voorburg, The Netherlands: International Statistical Institute.
- Ojeda, A. M. (1996). Contextos, representaciones y la idea de probabilidad condicional (Contexts, representations and conditional probability). In F. Hitt (Ed.), *Investigaciones en Matemáticas Educativas*, (pp. 291-310). México: Grupo Editorial Iberoamericano.
- Osterlind, S. J. (1989). *Constructing Test Items*. Boston: Kluwer.
- Sánchez, E. (1996). Dificultades en la comprensión del concepto de eventos independientes (Difficulties in understanding independent events). In F. Hitt (Ed.), *Investigaciones en Matemática Educativa*, (pp. 389–404). México: Grupo Editorial Iberoamericano.
- Sedlmeier, P. (1999). *Improving Statistical Reasoning. Theoretical Models and Practical Implications*. Mahwah, NJ: Erlbaum.
- Serrano Angulo, J. (2003). *Iniciación a la estadística Bayesiana* (Introduction to Bayesian statistics). Madrid: La Muralla.
- Truran, J.M. and Truran, K.M. (1997). Statistical independence: One concept or two? In B. Phillips (Ed.), *Papers from Statistical Education Presented at Topic Group 9, ICME 8*, Seville, (pp. 87–100). Hawthorn, Victoria: Swinburne University of Technology.
- Tversky, A. and Kahneman, D. (1982). Evidential impact of base rates. In D. Kahneman, P. Slovic and A. Tversky (Eds.), *Judgment Under Uncertainty: Heuristics and Biases*, (pp. 153-160). New York: Cambridge University Press.