Acta Psychologica 142 (2013) 370–382

Contents lists available at SciVerse ScienceDirect

Acta Psychologica

journal homepage: www.elsevier.com/locate/actpsy

Sample representativeness affects whether judgments are influenced by base rate or sample size

Natalie A. Obrecht [a,*], Dana L. Chesney [b]

[a] Department of Psychology, William Paterson University, 300 Pompton Road, Wayne, NJ 07470, USA
[b] Department of Psychology, University of Notre Dame, 118 Haggar Hall, Notre Dame, IN 46556, USA

ARTICLE INFO

Article history:
Received 2 August 2012
Received in revised form 14 January 2013
Accepted 18 January 2013
Available online 17 February 2013

PsycINFO classification:
2240 Statistics & Mathematics
2300 Human Experimental Psychology
2340 Cognitive Processes

Keywords:
Base rates
Judgment
Reasoning
Inductive reasoning
Mathematical cognition
Individual differences

ABSTRACT

We investigated how people use base rates and sample size information when combining data to make overall probability judgments. Participants considered two samples from an animal population in order to estimate the probability of that animal being aggressive. Participants' judgments were influenced by subpopulation base rates when they were provided and linked to specific samples. When samples were not identified as coming from different subpopulations, judgments typically reflected sample size information. We conclude that 1) People can use base rates when combining samples to make an inference; 2) People can correctly use sampling information to determine when to use base rates, and 3) People are able to consider base rate and sample size information at the same time. Additionally, we found that individuals' numeracy correlates with the extent to which base rate and sample size information is used.

1. Introduction

1.1. Information integration

When making judgments, people today have access to information from many different sources. If, for example, someone is interested in determining how likely a new restaurant is to have good food, she can visit various websites to see customer reviews. One complication of this strategy is that different sources may provide information about different subgroups within a population. For example, one website might show that only 20% of 520 people like a restaurant, while on another site 80% of 65 people report liking that restaurant. As larger samples typically provide more reliable information (Bernoulli, 1713/2005) it is statistically normative to weight percentage means by their sample sizes when combining data, in this case giving the 20% statistic more weight than the 80%. However, such a practice may not necessarily yield the best estimate of the general population (e.g. the chance that an individual will like a restaurant) when samples might represent different subgroups (see Chesney & Obrecht, 2011, 2012). The ideal of weighting data by sample size assumes that samples have been randomly drawn from the same population. When this assumption holds true, then indeed, larger samples provide more reliable estimates and should be given more weight. If instead samples have been drawn from different subpopulations (e.g. men vs. women), a better estimate of the general population mean will be obtained by weighting sample means in proportion to their subpopulations' prevalence in that general population, i.e., by those subpopulations' base rates.

Here we argue that when samples represent different subpopulations, it is normative to ignore sample size and instead weight data according to their base rates. In the example above, it would be unlikely that random samples from a population would yield diverse means of 20% and 80%, especially given how large each sample is. One might suspect that the two websites cater to different groups of people within the general population. If the base rates of the groups are unknown, it would be reasonable to average the two percentages and estimate a 50% chance that a person would like the restaurant. In contrast, if one knows that the first website with the poorer rating caters to the 5% of Americans who are accustomed to fine dining, while the second website represents the 95% of Americans who are not, one should weight the percentage data accordingly, ignoring sample size. This would result in an estimated 77% chance ([20%∗5%]+[80%∗95%]) that an American would like the restaurant. If the data were instead weighted by sample

∗ Corresponding author at: William Paterson University, Department of Psychology, 300 Pompton Road, Wayne, NJ 07470, USA. Tel.: +1 973 720 2683; fax: +1 973 720 3392.

E-mail addresses: obrechtn@wpunj.edu (N.A. Obrecht), dana.chesney.3@nd.edu (D.L. Chesney).

0001-6918/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.actpsy.2013.01.012

size, the opinions of fine diners would be overweighted relative to their numbers in the general population, yielding a very different estimate of 27% (20% [520/585] + 80% [65/585]). Indeed, if a person is a typical restaurant customer, rather than a fine diner, she could ignore the sample data from the fine diners altogether. However, in the current study we focus on the general case in which a judgment is made about an individual in a population whose subgroup membership is unknown.

1.2. Base rate use

Previous results are mixed regarding whether laypeople integrate base rates into their inferences. Base rate neglect has been demonstrated in the Bayesian reasoning literature in which people answer conditional probability problems (e.g. Bar-Hillel, 1980; Kahneman & Tversky, 1972). For example (Gigerenzer & Hoffrage, 1995):

> "The probability of breast cancer is 1% for women at age forty who participate in routine screening. If a woman has breast cancer, the probability is 80% that she will get a positive mammography. If a woman does not have breast cancer, the probability is 9.6% that she will also get a positive mammography. A woman in this age group had a positive mammography in a routine screening. What is the probability that she actually has breast cancer?___%"

A correct response would integrate the hit rate (80%) with the base rates (1% have cancer; 99% don't have cancer) and false alarm rate (9.6%): (cancer base rate∗hit rate)/(cancer base rate∗hit rate + no-cancer base rate∗false alarm rate) = 7.8%. However, in such problems many participants focus their responses on the hit rate, and estimate, for example, an 80% chance that the woman has cancer (Bar-Hillel, 1980; Eddy, 1982; Gigerenzer & Hoffrage, 1995).

Bayesian reasoning is improved, however, when data are presented as unstandardized natural frequencies that express probability in terms of subsets within a greater super set (Brase, 2008; Gigerenzer & Hoffrage, 1995; Obrecht, Anderson, Schulkin, & Chapman, 2012). For example, analogous to the problem above:

> "10 out of every 1,000 women at age forty who participate in routine screening have breast cancer. 8 of every 10 women with breast cancer will get a positive mammography. 95 out of every 990 women without breast cancer will also get a positive mammography. Here is a new representative sample of women at age forty who got a positive mammography in routine. How many of these women do you expect to actually have breast cancer?" (Gigerenzer & Hoffrage, 1995).

Such natural frequency formats simplify the conditional probability computation and make set-subset relationships clear which results in greater use of base rates (Evans, Handley, Perham, Over, & Thompson, 2000; Fiedler, Brinkmann, Betsch, & Wild, 2000; Girotto & Gonzalez, 2001; Macchi, 2000; Neace, Michaud, Bolling, Deer, & Zecevic, 2008; Yamagishi, 2003).

In another variation of base rate studies, participants are given base rate information about a group made up of two kinds of individuals (e.g. there are 70 engineers and 30 lawyers) and also a personality description of an individual that they are told was randomly drawn from the group. As an example, a personality description might state "Tom W. is of high intelligence, is quite self-confident, and tends to be argumentative…". The participants' task is to judge the profession of the individual (e.g. engineer or lawyer). The personality description is designed to be stereotypical of one of the subgroups. Here Tom's description may sound typical of a lawyer, but base rate data suggests that lawyer are less common in the group (out of 100 men, 70 are engineers and 30 are lawyers). The classic outcome is that personality descriptions trump base rates; most people say that the individual belongs to the group that matches the personality, even when that group has the lower base rate (Tversky & Kahneman, 1974).

Again, although this study casts doubt, other findings are more optimistic about humans' abilities to use base rates when making judgments. The order in which base rates are presented, relative to individual personality descriptions, like Tom's above, has been shown to influence judgments (Krosnick, Li, & Lehman, 1990). When base rates are presented after, rather than before, individual descriptions, judgments more closely reflect the base rate information (Krosnick et al., 1990; Obrecht, Chapman, & Gelman, 2009). Also, base rate use increases when sampling procedures are shown to be random (Gigerenzer, Hell, & Blank, 1988), and under a variety of other conditions (e.g. Bar-Hillel & Fischhoff, 1981; Schwarz, Strack, Hilton, & Naderer, 1991). Further, relating causal mechanisms to base rates in Bayesian problems increases base rate use (Tversky & Kahneman, 1980; also see Krynski & Tenenbaum, 2007).

A recent study by Pennycook and Thompson (2012) further tempers the conclusions which can be drawn from Tversky and Kahneman's (1974) seminal base rate neglect finding that stereotypical personality descriptions trump base rates when judging group membership. Pennycook and Thompson found that when participants were given just personality descriptions, without any base rate information, and were asked to judge the chances that a person belonged to a group (e.g. lawyer), responses were quite variable, typically falling between 50 and 100%. However, when base rates were also given, and were congruent with personality data (i.e. when personality descriptions favored the large base rate group) participants gave highly consistent probability estimates regarding the chances of a person belonging to the larger base rate group (i.e. most responses were between 90 and 100%). This decrease in the variability of responses indicates that when data are consistent with one another, people can integrate both base rates and personality descriptions together to make judgments. Moreover, as these patterns were seen even when participants were under time pressure, this suggests that "reasoning with base rates is…relatively effortless" (Pennycook & Thompson, 2012). However, the response pattern seen when personality and base rate information were consistent sharply contrasts to the bimodal responses that were seen when base rates and personality descriptions conflicted; here some responses appeared to reflect the base rate data, and others the personality information (Pennycook & Thompson, 2012). Overall, the literature on both the Bayesian and personality problems demonstrates that people use base rates under some conditions, such as when problems are simplified (e.g. Gigerenzer & Hoffrage, 1995), although this factor is sometimes underweighted relative to normative standards (see Kahneman & Tversky, 1996).

1.3. Sample size

Another factor that is of interest to our current study is sample size. As with base rates, the literature is somewhat mixed in regard to people's ability to use sample size when making judgments. When people use a sample to make an inference to a population, they are generally more confident in their judgments when the provided sample is comprised of a larger, rather than smaller, number of items (e.g. Irwin, Smith, & Mayfield, 1956; Jacobs & Narloch, 2001; Nisbett, Krantz, Jepson, & Kunda, 1983). For example, when asked to make an inference about the average value of a deck of cards, people are more confident in their judgment after viewing a sample of 20, rather than 10, cards (Irwin et al., 1956). This appreciation for larger samples is also seen in tasks where people compare data from two populations in order to judge which has the higher mean value (Fouriezos, Rubenfeld, & Capstick, 2008; Masnick & Morris, 2008; Obrecht, Chapman, & Gelman, 2007; Obrecht, Chapman, & Suárez, 2010). In contrast, some studies have shown that the weight given to sample size is overly small or inconsistent (Bar-Hillel, 1979; Obrecht et al., 2007), and other research has shown sample size to be almost entirely neglected or