Online Appendix for "Why Psychologists Must Change the Way They Analyze Their Data: The Case of Psi": A Robustness Analysis

Eric–Jan Wagenmakers, Ruud Wetzels, Denny Borsboom, & Han van der Maas
University of Amsterdam

Abstract

In this online appendix we study the robustness of the Bayesian t-test, that is, we examine the extent to which the default settings yield potentially misleading results. The results show that any other setting would not have changed the qualitative conclusions that were drawn based on the default settings. Hence, our earlier conclusions (based on the default prior) are robust against alternative prior specifications.

In our manuscript "Why psychologists must change the way they analyze their data: The case of psi" we presented a Bayesian re-analysis of the data from Bem (in press). In particular, we analyzed each of Bem's experiments using the default Bayesian t-test (Rouder, Speckman, Sun, Morey, & Iverson, 2009). The results showed that there was no evidence for precognition to speak of. Table 1 shows the results.

As explained in our main manuscript, the Bayes factor BF_{01} quantifies the evidence for H_0 (i.e., no precognition) versus H_1 (i.e., precognition). In order to calculate this Bayes factor, we need to specify a probability distribution for effect size, given H_1. That is, what effect sizes do we expect, should precognition really exist?

In our main manuscript, we used the default option that reflects a lack of knowledge about precognition—a Cauchy distribution on effect size that is centered around zero with scale parameter or probable error $r = 1$, that is, $\delta \sim \text{Cauchy}(0, 1)$. This distribution is shown as the red line in Figure 1.[1]

However, one might argue that this default distribution is not appropriate, or, at least, that is sensible to examine other prior distributions on effect size as well. This

[1] See also `http://en.wikipedia.org/wiki/Cauchy_distribution`.

This research was supported by Vidi grants from the Dutch Organization for Scientific Research (NWO). Correspondence concerning this article may be addressed to Eric–Jan Wagenmakers, University of Amsterdam, Department of Psychology, Roetersstraat 15, 1018 WB Amsterdam, the Netherlands. Email address: ej.wagenmakers@gmail.com. We thank Rogier Kievit and Jan de Ruiter for constructive discussions.

Table 1: The results of 10 crucial tests for the experiments reported in Bem (in press), reanalyzed using the default Bayesian t-test.

Exp	df	$\lvert t \rvert$	p	BF_{01}	Evidence category
					(in favor of $H_.$)
1	99	2.51	0.01	0.61	Anecdotal (H_1)
2	149	2.39	0.009	0.95	Anecdotal (H_1)
3	96	2.55	0.006	0.55	Anecdotal (H_1)
4	98	2.03	0.023	1.71	Anecdotal (H_0)
5	99	2.23	0.014	1.14	Anecdotal (H_0)
6	149	1.80	0.037	3.14	Substantial (H_0)
6	149	1.74	0.041	3.49	Substantial (H_0)
7	199	1.31	0.096	7.61	Substantial (H_0)
8	99	1.92	0.029	2.11	Anecdotal (H_0)
9	49	2.96	0.002	0.17	Substantial (H_1)

was suggested independently by Patrizio Tressoldi (by Email) and Eric Kvaalen (on `www.newscientist.com`). In particular, one might argue that previous work has shown effect sizes in precognition and psi to be relatively small (e.g., Storm, Tressoldi, & Di Risio, 2010). Therefore, one could argue that instead of assuming $\delta \sim \text{Cauchy}(0, 1)$, we might want to assume a Cauchy distribution that is more narrowly peaked, for instance $\delta \sim \text{Cauchy}(0, 0.5)$, a distribution shown as the dotted line in Figure 1. Naturally, one might then wonder whether and to what extent a change in the scale parameter of the Cauchy distribution fundamentally alters our conclusions.

In order to examine this possibility we conducted a robustness analysis in which we systematically varied the scale parameter r from 0 to 3 to quantify the effect that this has on the Bayes factor BF_{01}. The results are shown in Figure 2.

Note that Figure 2 plots the Bayes factor such that the scale of evidence in favor of H_0 is visually equivalent to the scale of evidence in favor of H_1. Also note that when $r = 0$, $H_0 = H_1$, and the Bayes factor indicates that the evidence is perfectly ambiguous (i.e., $BF_{01} = 1$).

The different panels in Figure 2 indicate that our choice for the default prior does not affect our conclusions. In fact, the red dot—the result of our default test—seems to provide a relatively accurate summary of the evidence. Yes, it is true that for very small values of r the evidence is occasionally in favor of H_1, but—and this is the crucial point—only for the bottom right panel is the evidence clearly in favor of H_1. That is, in the bottom right panel the maximum Bayes factor is almost 1/10, meaning that the observed data are about 10 times more likely under H_1 than they are under H_0, given of course that the prior scale parameter r is chosen a posteriori, something that greatly biases the Bayes factor in favor of H_1.

For 7 out of the remaining 9 other panels, even the *maximum* Bayes factor indicates only “anecdotal” evidence (i.e., evidence worth “no more than a bare mention”, that is, the data are less than 3 times more likely under H_1 than under H_0). This leaves the top-left two panels, for which the maximum Bayes factor does reach the criterion for “substantial”

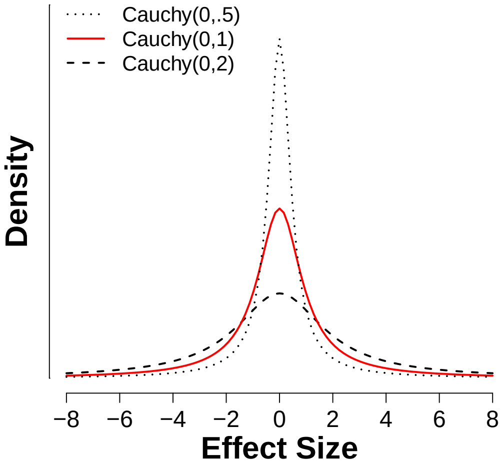

Figure 1. Three examples of a Cauchy distribution. The red line indicates the prior that underlies the default Bayesian t-test.

evidence; however, it does so only just, and only for very specific values of the scale parameter. Again, the default test (indicated by the red dot) seems to provide a reasonable indication of the evidence.

In sum, we conclude that our results are robust to different specifications of the scale parameter for the effect size prior under H_1. This reinforces our general argument that p-values may strongly overstate the evidence against H_1.

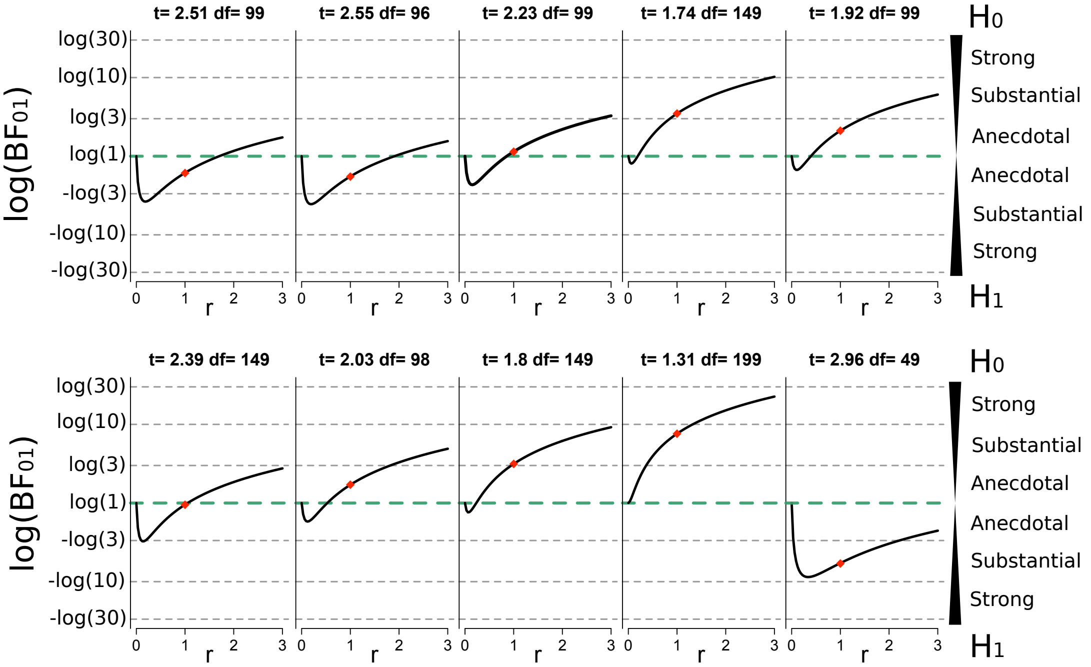

Figure 2. A robustness analysis for the data from Bem (in press). The Bayes factor BF_{01} is plotted as a function of the scale parameter r of the Cauchy prior for effect size under H_1. The red dot indicates the result from the default prior, the horizontal green line indicates complete ambiguous evidence, and the horizontal grey lines demarcate the different qualitative categories of evidence (see our main manuscript). Importantly, the results in favor of H_1 are never compelling, except perhaps for the bottom right panel.

References

Bem, D. J. (in press). Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of Personality and Social Psychology.*

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t–tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, *16*, 225–237.

Storm, L., Tressoldi, P. E., & Di Risio, L. (2010). Meta–analysis of free–response studies, 1992–2008: Assessing the noise reduction model in parapsychology. *Psychological Bulletin*, *136*, 471–485.