Stopping Rules and Their Irrelevance for Bayesian Inference: Online Appendix to "A Practical Solution to the Pervasive Problems of p–Values", to appear in *Psychonomic Bulletin & Review*

Eric-Jan Wagenmakers
University of Amsterdam

Correspondence concerning this online appendix should be addressed to:
Eric–Jan Wagenmakers
University of Amsterdam, Department of Psychology
Roetersstraat 15
1018 WB Amsterdam, The Netherlands
Ph: (+31) 20–525–6876
Fax: (+31) 20-639-0279
E-mail may be sent to EJ.Wagenmakers@gmail.com.

This appendix provides a definition of a stopping rule, and then shows why a "noninformative" stopping rule is irrelevant for Bayesian statistical inference. Finally, it is shown why a Bayesian cannot be mislead by optional stopping. The following discussion relies extensively on Berger and Wolpert (1988), Bernardo and Smith (1994) and Raiffa and Schlaifer (1961), and the reader is referred to these monographs for further details.

Consider an experiment in which data are sampled sequentially, and data $x^n = (x_1, x_2, ..., x_n)$ have been observed. With probability $\tau_n(x^n)$, sampling stops. With the complementary probability $1 - \tau_n(x^n)$, sampling continues and one collects the additional observation x_{n+1}. A stopping rule $\boldsymbol{\tau}$ is *proper* if the experiment is guaranteed to stop at some finite n. When $\tau_n(x^n) \in \{0, 1\}$, so that for any given data x^n there is no uncertainty whether to stop or to continue, the stopping rule is *deterministic*. Otherwise, the stopping rule is said to be *randomized*. When the stopping rule does not depend on the parameters of the model, and when the stopping rule is furthermore *a priori* independent of the parameters of the model, the stopping rule is said to be *noninformative* (for examples see Raiffa & Schlaifer, 1961).

Every experiment can be characterized by the generated data x^n and the stopping process $\boldsymbol{\tau}$ that resulted in a sample size of n observations. Hence, the likelihood function is generally given by $Pr(n, x^n|\boldsymbol{\tau}, \boldsymbol{\theta})$, where $\boldsymbol{\theta}$ denotes the parameters of a given model. Almost always, the above likelihood is simplified to $Pr(x^n|\boldsymbol{\theta})$, de facto assuming that the number of observations was fixed in advance. This simplification is also appropriate for experiments that use all kinds of different stopping rules, as long as they are noninformative.

Example 1 Irrelevance of Biased Stopping for Bayesian Inference (cf. Bernardo & Smith, 1994, pp. 251–255). Consider a sequence of independent questions. Each observation x_i equals 1 (i.e., a correct answer) with probability θ, and 0 (i.e., an incorrect answer) otherwise. Define a deterministic stopping rule as follows: when the first question is answered correctly, the experiment stops (i.e., $\tau_1(1) = 1$). When the first question is answered incorrectly, a second question is asked (i.e., $\tau_1(0) = 0$), after which the experiment stops (i.e., $\tau_2(x_1, x_2) = 1$).

One may intuit that this sampling scheme causes the estimate of θ to be higher than it would have been in case the number of observations was fixed in advance. A more detailed analysis shows that this intuition is false. First, suppose the first question is answered correctly (i.e., $x_1 = 1$). The probability of this happening is

$$\begin{aligned} Pr(n = 1, x_1 = 1|\boldsymbol{\tau}, \theta) &= Pr(x_1 = 1|n = 1, \boldsymbol{\tau}, \theta)Pr(n = 1|\boldsymbol{\tau}, \theta) \\ &= 1 \times Pr(n = 1|\boldsymbol{\tau}, \theta) = Pr(x_1 = 1|\theta). \end{aligned} \quad (1)$$

When the first question is answered incorrectly (i.e., $x_1 = 0$), another question is asked, the outcome of which is denoted by x. The probability of observing $x_1 = 0$ followed by $x_2 = x$ is given by

$$\begin{aligned} & Pr(n = 2, x_1 = 0, x_2 = x|\boldsymbol{\tau}, \theta) \\ &\quad = Pr(x_1 = 0, x_2 = x|n = 2, \boldsymbol{\tau}, \theta)Pr(n = 2|\boldsymbol{\tau}, \theta) \\ &\quad = Pr(x_1 = 0|n = 2, \boldsymbol{\tau}, \theta)Pr(x_2 = x|x_1 = 0, n = 2, \boldsymbol{\tau}, \theta)Pr(n = 2|\boldsymbol{\tau}, \theta) \\ &\quad = 1 \times Pr(x_2 = x|x_1 = 0, \theta)Pr(x_1 = 0|\theta) \\ &\quad = Pr(x_2 = x, x_1 = 0|\theta). \end{aligned} \quad (2)$$

Hence, for all data that can be observed in this experiment, $Pr(n, x^n|\boldsymbol{\tau}, \theta) = Pr(x^n|\theta)$, which means that the stopping rule does not affect our inference about θ.

More generally, for noninformative stopping rules the probability of observing x^n and terminating the sampling process is given by

$$\begin{aligned} Pr(x^n, n|\boldsymbol{\tau}, \boldsymbol{\theta}) &= Pr(x^n|\boldsymbol{\tau}, \boldsymbol{\theta})Pr(n|x^n, \boldsymbol{\tau}, \boldsymbol{\theta}) \\ &= Pr(x^n|\boldsymbol{\theta})Pr(n|x^n, \boldsymbol{\tau}). \end{aligned} \quad (3)$$

In this derivation, $Pr(x^n|\boldsymbol{\tau}, \boldsymbol{\theta}) = Pr(x^n|\boldsymbol{\theta})$ because the specific values for the observed data x^n do not depend on the stopping rule $\boldsymbol{\tau}$. Also, from the definition for noninformative stopping rules, it follows that $Pr(n|x^n, \boldsymbol{\tau}, \boldsymbol{\theta}) = Pr(n|x^n, \boldsymbol{\tau})$. Equation 3 shows that a noninformative stopping rule does not affect the *kernel* of the likelihood function (i.e., the part that contains the parameters $\boldsymbol{\theta}$) but only affects the *residue* of the likelihood function. A comparison of the likelihood functions for the binomial and negative binomial sampling schemes confirms this fact. Because frequentist inference for $\boldsymbol{\theta}$ partly depends on data that could have been observed but were not, it partly relies on the residue of the likelihood function. In contrast, Bayesian inference for $\boldsymbol{\theta}$ only relies on the kernel of the likelihood function, as it only considers the data that were actually observed.

For noninformative stopping rules, $\boldsymbol{\theta}$ and $\boldsymbol{\tau}$ are independent *a priori*, that is, $Pr(\boldsymbol{\tau}, \boldsymbol{\theta}) = Pr(\boldsymbol{\tau})Pr(\boldsymbol{\theta})$. It then follows that

$$\begin{aligned} Pr(\boldsymbol{\tau}, \boldsymbol{\theta}|x^n, n) &= \frac{Pr(x^n, n|\boldsymbol{\tau}, \boldsymbol{\theta})Pr(\boldsymbol{\tau}, \boldsymbol{\theta})}{Pr(x^n, n)} \\ &\propto Pr(x^n, n|\boldsymbol{\tau}, \boldsymbol{\theta})Pr(\boldsymbol{\tau}, \boldsymbol{\theta}) \\ &\propto Pr(x^n|\boldsymbol{\theta})Pr(\boldsymbol{\theta})Pr(n|x^n, \boldsymbol{\tau})Pr(\boldsymbol{\tau}). \end{aligned} \tag{4}$$

After integrating out $\boldsymbol{\tau}$ one obtains $Pr(\boldsymbol{\theta}|x^n) \propto Pr(x^n|\boldsymbol{\theta})Pr(\boldsymbol{\theta})$. Thus, the posterior distribution for $\boldsymbol{\theta}$ is not affected by uninformative stopping rules $\boldsymbol{\tau}$ (Raiffa & Schlaifer, 1961).

It should be acknowledged that some Bayesian statisticians recommend that prior distributions depend on the sampling scheme (e.g., Box & Tiao, 1973, p. 46; Bernardo & Smith, 1994, p. 253). Obviously, this is under the assumption that the stopping rule is uninformative with respect to the likelihood, but not with respect to the prior, so that $Pr(\boldsymbol{\tau}, \boldsymbol{\theta}) \neq Pr(\boldsymbol{\tau})Pr(\boldsymbol{\theta})$. However, this implies that one's knowledge, or ignorance, of a quantity depends on the experiment being used to determine it (paraphrased from Lindley, 1972, p. 71). A discussion of this issue would take us too far afield.

Example 2 A Discussion on Optional Stopping (cf. Berger & Berry, 1988; Berger & Wolpert, 1988; Bernardo & Smith, 1994; Edwards, Lindman, & Savage, 1963; Jennison & Turnbull, 1990; Kadane, Schervish, & Seidenfeld, 1996a; Royall, 1997). The above derivation shows that rational, coherent inference does not depend on noninformative stopping rules. Frequentist procedures do depend on noninformative stopping rules, and quite critically so. In particular, when the stopping rule is ignored, a researcher that continues to collect data until the frequentist p–value reaches some desired level of significance is guaranteed to achieve her goal eventually.

An important issue, first raised by Armitage (1961) in the context of monitoring the outcome of clinical trials, is whether or not the Bayesian analysis is similarly affected by the mechanism of optional stopping.[1] Suppose that a researcher obtains data through optional stopping. Will the Bayesian statistician be fooled into always reporting evidence in favor of the alternative hypothesis? The short answer is a resounding "no" (cf. Kadane et al., 1996a; Kadane, Schervish, & Seidenfeld, 1996b; Kerridge, 1963; Royall, 2000).

A Bayesian analysis is misleading when it assigns a relatively low posterior probability to a true hypothesis. It can be shown that there is a limit as to how often one may encounter such misleading evidence. For instance, assume that H_0 and H_1 are equally plausible a priori. Then, at the termination of sampling, the frequency with which the posterior probability of the true hypothesis is less than or equal to x cannot exceed $x/(1-x)$ (Kerridge, 1963).

To illustrate, suppose H_0 is true. The frequency with which H_0 will produce a posterior probability as least as low as 1/100 is no more than 1/99. Note that this bound on the frequency of reporting misleading evidence is quite independent of the stopping rule that is used. This is perhaps contrary to intuition, which may falsely suggest that if one monitors the posterior probability of H_0, and stops the experiment whenever $Pr(H_0|D) < .01$, one

[1] For an application of Bayesian inference in clinical trials see Berry (1989), Carlin, Kadane, and Gelfand (1998), and Kadane and Vlachos (2002).

will eventually reach this goal. Instead, as pointed out by Edwards et al. (1963, p. 239): "(...) if you set out to collect data until your posterior probability for a hypothesis which unknown to you is true has been reduced to .01, then 99 times out of 100 you will never make it, no matter how many data you, or your children after you, may collect (...)".

In contrast to Bayesian hypothesis testing, Bayesian parameter estimation appears to be more vulnerable to the effects of optional stopping. For instance, assume that data that are normally distributed with known standard deviation, e.g., $x^n \sim N(\mu, 1)$. One might continue to collect data until the sample mean is k standard deviations from zero, say $k = 3$. In this case, the 95% Bayesian confidence interval does not contain zero. As the Bayesian parameter estimation procedure ignores the stopping rule, this result is guaranteed to hold for any experiment. Has the experimenter succeeded in misleading the Bayesian statistician into believing that μ is not zero? Not really. If the value $\mu = 0$ is special, one needs to assign some probability mass to that value. This is exactly what happens in Bayesian hypothesis testing. Cornfield (1966, p. 22) explains

> "If one is seriously concerned about the probability that a stopping rule will certainly result in the rejection of a true hypothesis, it must be because some possibility of the truth of the hypothesis is being entertained. In that case it is appropriate to assign a non–zero prior probability to the hypothesis. If this is done, differing from the hypothesized value by k standard errors will not result in the same posterior probability for the hypothesis for all values of n. In fact for fixed k the posterior probability of the hypothesis monotonically approaches unity as n increases, no matter how small the prior probability assigned, so long as it is non–zero, and how large the k, so long as it is finite. Differing by k standard errors does not therefore necessarily provide any evidence against the hypothesis and disregarding the stopping rule does not lead to an absurd conclusion. The Bayesian viewpoint thus indicates that the hypothesis is certain to be erroneously rejected – not because the stopping rule was disregarded – but because the hypothesis was assigned zero prior probability and that such assignment is inconsistent with concern over the possibility that the hypothesis will certainly be rejected when true."

For an extended discussion, I refer the interested reader to Basu (1975), Berger and Berry (1988), and Berger and Wolpert (1988, pp. 80–83).

References

Armitage, P. (1961). Comment on "consistency in statistical inference and decision" by Cedric A. B. Smith. *Journal of the Royal Statistical Society B*, *23*, 30–31.

Basu, D. (1975). Statistical information and likelihood. *Sankhya A*, *37*, 1–71.

Berger, J. O., & Berry, D. A. (1988). The relevance of stopping rules in statistical inference. In S. S. Gupta & J. O. Berger (Eds.), *Statistical decision theory and related topics: Vol. 1* (pp. 29–72). New York: Springer Verlag.

Berger, J. O., & Wolpert, R. L. (1988). *The likelihood principle (2nd ed.).* Hayward (CA): Institute of Mathematical Statistics.

Bernardo, J. M., & Smith, A. F. M. (1994). *Bayesian theory.* New York: Wiley.

Berry, D. A. (1989). Monitoring accumulating data in a clinical trial. *Biometrics*, *45*, 1197–1211.

Box, G. E. P., & Tiao, G. C. (1973). *Bayesian inference in statistical analysis.* Reading: Addison–Wesley.

Carlin, B. P., Kadane, J. B., & Gelfand, A. E. (1998). Approaches for optimal sequential decision analysis in clinical trials. *Biometrics*, *54*, 964–975.

Cornfield, J. (1966). Sequential trials, sequential analysis, and the likelihood principle. *The American Statistician*, *20*, 18–23.

Edwards, W., Lindman, H., & Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psychological Review*, *70*, 193–242.

Jennison, C., & Turnbull, B. W. (1990). Statistical approaches to interim monitoring of medical trials: A review and commentary. *Statistical Science*, *5*, 299–317.

Kadane, J. B., Schervish, M. J., & Seidenfeld, T. (1996a). Reasoning to a foregone conclusion. *Journal of the American Statistical Association*, *91*, 1228–1235.

Kadane, J. B., Schervish, M. J., & Seidenfeld, T. (1996b). When several Bayesians agree that there will be no reasoning to a foregone conclusion. *Philosophy of Science*, *63*, S281–S289.

Kadane, J. B., & Vlachos, P. K. (2002). Hybrid methods for calculating optimal few–stage sequential strategies: Data monitoring for a clinical trial. *Statistics and Computing*, *12*, 147-152.

Kerridge, D. (1963). Bounds for the frequency of misleading Bayes inferences. *The Annals of Mathematical Statistics*, *34*, 1109–1110.

Lindley, D. V. (1972). *Bayesian statistics, a review.* Philadelphia (PA): SIAM.

Raiffa, H., & Schlaifer, R. (1961). *Applied statistical decision theory.* Cambridge (MA): The MIT Press.

Royall, R. (2000). On the probability of observing misleading statistical evidence (with discussion). *Journal of the American Statistical Association*, *95*, 760-780.

Royall, R. M. (1997). *Statistical evidence: A likelihood paradigm.* London: Chapman & Hall.