How to Interpret a Bayes Factor

- The Bayes factor BF_{10} is a ratio which quantifies evidence in favour of an effect (represented by "1" in the subscript) vs. no effect (represented by "0" in the subscript).
- $\text{BF}_{10} > 1$ indicates evidence in favour of an effect.
- $0 < \text{BF}_{10} < 1$ indicates evidence in favour of *no* effect.
- $\text{BF}_{10} = 1$ indicates that the data are ambiguous.

by Quentin F. Gronau, Garston Liang, & Alexander Thorpe

1 Bayes factors replace *p*-values

For many years, null-hypothesis significance testing using p-values has been the ubiquitous approach for testing hypotheses such as whether or not a new chronic pain management therapy outperforms an existing therapy. However, this approach has many flaws, which often results in p-values being mis-used and mis-interpreted. Bayesian inference provides an alternative approach which addresses many of these flaws which can in principle be applied in any situation where one would report a p-value. When testing hypotheses using Bayesian inference, p-values are replaced by *Bayes factors*. Bayes factors offer a number of practical benefits compared to p-values, such as enabling the quantification of evidence in favour of the null hypothesis (e.g., two therapy forms being equally effective), and allowing the sequential monitoring of evidence as observations arrive. However, to draw valid conclusions it is crucial that they are interpreted correctly.

2 Bayes factor interpretation

Bayes factors allow researchers and practitioners to assess the evidence that the data provide for an effect being present or absent. Suppose the null hypothesis (H_0) corresponds to a new therapy being equally effective as an existing one. In contrast, the alternative hypothesis (H_1) corresponds to the new therapy outperforming the existing one. The Bayes factor compares how well each hypothesis has predicted the observed data. A Bayes factor of, for instance, $\text{BF}_{10} = 11$, indicates that the observed data are 11 times more likely under H_1, which states that the new therapy is more effective, than under H_0, which states that both therapies are equally effective.

In general, values of BF_{10} larger than 1 indicate evidence in favour of H_1 (i.e., effect present), values between 0 and 1 indicate evidence in favour of H_0 (i.e., no effect). Bayes factors cannot be negative. When interpreting a Bayes factor, it is important to pay attention to the order of the subscripts. Specifically, "1" corresponds to H_1 and "0" corresponds to H_0. Consequently, BF_{10} expresses evidence in favour of H_1 over H_0. However, when BF_{10} is smaller than 1, for instance,

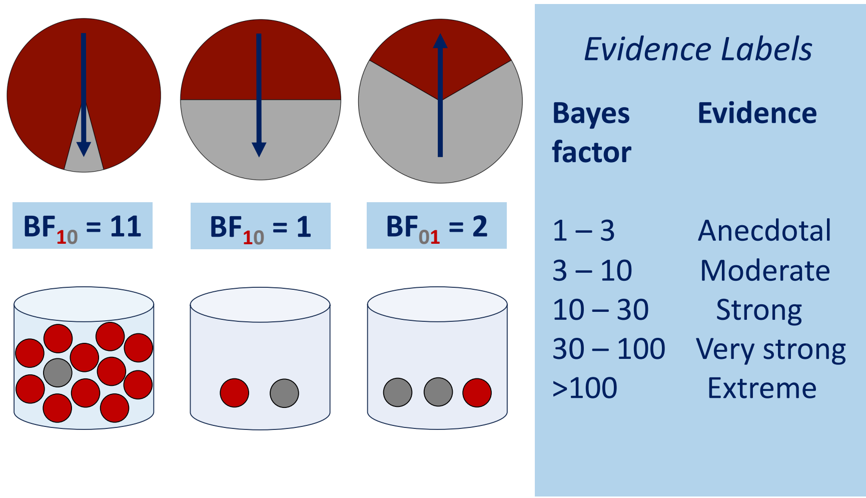

Figure 1: A Bayes factor of $\text{BF}_{10} = 11$ indicates that the data are 11 times more likely under the alternative hypothesis (i.e., effect present, represented by "1" in the subscript) than the null hypothesis (i.e., no effect, represented by "0" in the subscript). This is considered strong evidence for an effect. A Bayes factor of $\text{BF}_{10} = 1$ indicates that the data are equally likely under the alternative hypothesis and the null hypothesis. This is completely ambiguous evidence. A Bayes factor of $\text{BF}_{01} = 2$ indicates that the data are two times more likely under the null hypothesis than the alternative hypothesis. This is considered anecdotal evidence for no effect.

$\text{BF}_{10} = 0.5$, and thus indicates evidence for H_0, it is common practice to re-express the Bayes factor in favour of H_0 as follows: $\text{BF}_{01} = 1/\text{BF}_{10} = 1/0.5 = 2$. This does not change the evidence but is simply a different way of presenting it.

In contrast to p-values, Bayes factors *cannot* be "significant" as they are a continuous measure of evidence. However, to aid interpretation, labels for different degrees of evidence have been proposed (see Figure 1). Furthermore, an intuition for the strength of evidence a Bayes factor provides can be obtained by using the "urn" analogy and the "spinner" analogy (see Figure 1).

The urn analogy works as follows. Suppose the Bayes factor is $\text{BF}_{10} = 11$. This is analogous to putting eleven red balls corresponding to H_1 and one grey ball corresponding to H_0 into an urn. Suppose you draw a ball at random and end up drawing the single grey ball. Your level of surprise provides an intuition for the strength of evidence. In this case, you might be quite surprised, indicating that a Bayes factor of 11 provides strong evidence. In contrast, suppose $\text{BF}_{10} = 0.5$. Since this Bayes factor is smaller than 1, it is easier to express it in favour of the null hypothesis: $\text{BF}_{01} = 1/0.5 = 2$. This is analogous to placing two grey balls corresponding to H_0 and one red ball corresponding to H_1 into an urn. If you end up drawing the single red ball you may not be very surprised, indicating that this Bayes factor does not provide much evidence.

The "spinner" analogy works in a similar manner and is relevant because it is used in the output of statistical software such as JASP (`jasp-stats.org`). Instead of placing balls into an urn, one creates a pie chart with the ratio of red and grey area in line with the observed Bayes factor. For instance, for $\text{BF}_{10} = 11$, one creates a pie chart where the ratio of the area coloured in red vs. grey is 11 to 1. Next, imagine attaching a spinner to this pie chart and spinning it. Suppose the spinner ends up in the smaller, in this case grey, area. Your level of surprise provides you with an intuition for the strength of evidence. Similarly, for $\text{BF}_{01} = 2$ one constructs a pie chart where the ratio of the area coloured in grey vs. red is 2 to 1 and imagines the spinner ends up again in the smaller, in this case red, area.

In sum, Bayes factors are easily interpretable and, in contrast to p-values, enable one to assess the evidence in favor of an effect being present or absent. To highlight this, suppose a t-test comparing two therapies yields $p = .03$. This p-value is significant ($p < .05$) however, only a Bayes factor can tell how much evidence there really is. Suppose the number of participants being exposed to each therapy was 20. In this case $\mathrm{BF}_{10} = 2.18$ indicating only anecdotal evidence in favour of an effect. Alternatively, suppose the number of participants being exposed to each therapy was 5,000. In this case $\mathrm{BF}_{01} = 1/\mathrm{BF}_{10} = 4.22$. Therefore, although the p-value is significant, the Bayes factor indicates that there is in fact moderate evidence in favour of *no* effect.

3 Common misconceptions

When interpreting Bayes factors there are a number of common misconceptions that one should be aware of:

Misconception 1: A Bayes factor of about 1 indicates evidence for the null hypothesis (i.e., no effect).

A Bayes factor of about 1 indicates that the data are about equally likely under H_1 and H_0: the evidence is ambiguous. In contrast, if BF_{10} is smaller than 1 or, equivalently, BF_{01} is larger than 1, this indicates evidence for the null hypothesis. One strength of Bayes factors is that this way, they can disentangle evidence of absence (i.e., evidence in favour of H_0) from absence of evidence (i.e., ambiguous evidence where the Bayes factor is about 1). This is not possible with p-values.

Misconception 2: Treating Bayes factors larger vs. smaller than 10 as significant vs. non-significant.

While it is true that Bayes factors larger than ten are considered "strong" evidence, a Bayes factor of $\mathrm{BF}_{10} = 9.9$ does not provide fundamentally different evidence from $\mathrm{BF}_{10} = 10.1$. The Bayes factor is a continuous measure of evidence, in contrast to dichotomous "significant" vs. "non-significant" p-values. The evidence labels merely serve as interpretation aid.

Misconception 3: Interpreting Bayes factors close to 1 as strong evidence.

While it might sound at first glance impressive that the data are, for instance, two times more likely under H_1 than under H_0, considering analogies such as the "urn" or "spinner" and the evidence labels highlight that such Bayes factors do not provide much evidence.

4 Further reading

Below are a few resources that provide a great introduction to Bayesian inference in general and Bayesian hypothesis testing using Bayes factors in particular:

- Wagenmakers, E.-J., Marsman, M., Jamil, T., Ly, A., Verhagen, A. J., Love, J., ... & Morey, R. D. (2018). Bayesian inference for psychology. Part I: Theoretical advantages and practical ramifications. *Psychonomic Bulletin & Review*, *25*, 35–57.
- Etz, A., Gronau, Q. F., Dablander, F., Edelsbrunner, P. A., & Baribault, B. (2018). How to become a Bayesian in eight easy steps: An annotated reading list. *Psychonomic Bulletin & Review*, *25*, 219–234.
- Wagenmakers, E.-J., Love, J., Marsman, M., Jamil, T., Ly, A., ... & Morey, R. D. (2018). Bayesian inference for psychology. Part II: Example applications with JASP. *Psychonomic Bulletin & Review*, *25*, 58–76.