A Brief Introduction to Bayesian Inference

From Tea to Beer

Johnny van Doorn

Invalid Date

Table of contents

Preface

This is a Quarto book.

To learn more about Quarto books visit https://quarto.org/docs/books.

```
1 + 1
```

```
[1] 2
```

1 The Lady Tasting Tea

Over 80 years ago, Sir Ronald Fisher conducted the famous experiment "The Lady Tasting Tea" in order to test whether his colleague, Dr. Muriel Bristol, could taste if the tea infusion or the milk had been added to the cup first (**fisher1937design?**). Dr. Bristol was presented with eight cups of tea and the knowledge that four of these had the milk poured in first. Dr. Bristol was then asked to identify these four cups. Fisher analyzed the results using null hypothesis significance testing:

1. assume the null hypothesis to be true (i.e., Dr. Bristol lacks any ability to discriminate the cups);
2. calculate the probability of encountering results at least as extreme as those observed;
3. if that probability is sufficiently low, consider the null hypothesis discredited.

This probability is now known as the p-value and it features in many statistical analyses across empirical sciences.

1.1 A Bayesian Version

Decades later, Dennis (**Lindley1993?**) used an experimental procedure similar to that of Fisher to highlight some limitations of the p-value paradigm. Specifically, the calculation of the p-value depends on the sampling plan, that is, the *intention* with which the data were collected. Consider the Lindley setup: Dr. Bristol is offered six pairs of cups, where each pair consists of a cup where the tea was poured first, and a cup where the milk was poured first. She is then asked to judge, for each pair, which cup has had the tea added first. A possible outcome is the sequence RRRRRW, indicating that she was right for the first five pairs, and wrong for the last pair. However, as Lindley demonstrated, the original sampling plan is crucial in calculating the p-value because the p-value depends on hypothetical outcomes that are "more extreme."

Was the goal to have the Dr. Bristol taste six pairs of cups –no more, no less– or did she need to continue until she made her first mistake? The observed data could have been the outcome of either sampling plan; yet in the former case, the p-value equals 0.109, whereas in the latter case the p-value equals 0.031. The difference lies in the inclusion of more extreme cases. In the "test six cups" plan, the only more extreme outcome is RRRRRR (i.e., she correctly identified all 6 cups), whereas for the "test until error" plan the more extreme outcomes include sequences

such as RRRRRRW and RRRRRRRW (i.e., 6 and 7 correct responses, before a single incorrect response)[1]. It seems undesirable that the p-value depends on hypothetical outcomes that are in turn determined by the sampling plan. Harold Jeffreys summarized: "What the use of p implies, therefore, is that a hypothesis that may be true may be rejected because it has not predicted observable results that have not occurred. This seems a remarkable procedure." (**Jeffreys1961?**).

This drawback is one of the reasons (for more critique see (**Wasserstein2016?**; **BenjaminEtAl2018?**)) why Bayesian inference has become more popular in the past years as an alternative framework for hypothesis testing and parameter estimation.

1.2 An Alcoholic Version

In this text we revisit Fisher's experimental paradigm to demonstrate several key concepts of Bayesian inference, specifically the prior distribution, the posterior distribution, the Bayes factor, and sequential analysis. Furthermore, we highlight the advantages of Bayesian inference, such as its straightforward interpretation, the ability to monitor the result in real-time, and the irrelevance of the sampling plan. For concreteness, we analyze the outcome of a tasting experiment that featured 57 staff members and students of the Psychology Department at the University of Amsterdam; these participants were asked to distinguish between the alcoholic and non-alcoholic version of the Weihenstephaner Hefeweissbier, a German wheat beer. We analyze and present the results in the open-source statistical software JASP (**JASP2022?**).

[1] A more technical way of describing the difference between the two setups is that data from the "test six cups" follow the binomial distribution, whereas data from the "test until error" follow the negative binomial distribution.