Applying Bayesian Inference to Sequential Analysis

Ben Christensen

May 9, 2021

1 Abstract

Bayesian Inference has the potential to improve models of prediction by incorporating human expertise into formal statistical models. Sequential Analysis provides an opportunity to incorporate prior belief with a statistical model. In this article I show that using Bayesian Inference in Sequential Analysis improves accuracy and shortens the time required to come to a decision. Although Bayesian Inference can decrease accuracy relative to traditional Sequential Analysis, the possibility of eliciting expert judgment that fairs better than random chance in constructing prior distributions leads to potential lasting benefits in applying Bayesian Inference.

2 Introduction to Bayesian Inference

The most common approach to statistical analyses is the frequentist method. In this camp resides the hypothesis testing method with confidence level α often set at 0.05. It makes statements about the expected frequency of a random process over an increasing number of trials. But another common meaning of probability is used in Bayesian Statistics, it is "a degree of belief" as Etz and Vanderckhove (2017) write, "it is a number between zero and one that quantifies how strongly we should think something to be true based on the relevant information we have. In other words, probability is a mathematical language for expressing our uncertainty. This kind of probability is inherently subjective—because it depends on the information that you have available—and reasonable people may reasonably differ in the probabilities that they assign to events (or propositions)" (p. 6).

Wagenmakers et. al (2018) have encouraged the use of Bayesian hypothesis testing because of the benefit of monitoring evidence as data come in. Kersten et. al (2004) show the use Bayesian Inference has in modeling human image recognition. In this paper, I show the benefit of Bayesian Inference in updating a model of Sequential Analysis.

The ubiquity of frequentist statistical analyses is understandable given its usefulness in modeling in a vast array of methods from linear regression to neural networks. These can predict the presence of some correlated condition or the likelihood of some future event with surprising accuracy. But some events we may want to predict may not have data readily available for advanced modeling. In these situations, experts are often elicited to give their best estimation of the likelihood of some event. This estimation is a probability in the Bayesian sense.

Making predictions using experts' subjective probability tremendously expands the feasible set of events we can forecast. Some have written on the best way to elicit expert judgment on forecasting the likelihood of certain events, evaluating the accuracy of different decision makers (Mellers et. al, 2014; Morgan, 2014).

Another strength of Bayesian Statistics is it can allow personal belief to be incorporated with data analysis. In a technique called Bayesian Inference, subjective belief about a random process is combined with data generated from that random process to generate a new estimate of the random process as follows

(Wasserman, 2013, p. 176-177):

$P(\mathbf{x}|\theta)$: a statistical model representing the likelihood of the data $\mathbf{x}$ given some value of θ

$P(\theta)$: prior distribution giving the modeler's belief of the likelihood of some value of θ

$$P(\theta|\mathbf{x}) = \frac{P(\mathbf{x}|\theta)P(\theta)}{\sum P(\mathbf{x}|\theta)P(\theta)}$$

where $c_n = \sum P(\mathbf{x}|\theta)P(\theta)$ is called the normalizing constant. Since it is a constant, choosing θ to maximize the likelihood of the posterior distribution is the same as choosing θ to maximize the product $P(\mathbf{x}|\theta)P(\theta)$

3 Sequential Analysis

Gold and Shadlen (2007) give a thorough description of the model called Sequential Analysis, demonstrating it using repeated trials of flipping two different coins (Bernoulli random variables) with two different probabilities of success. One coin is a fair coin with 50% probability of heads and the other is a trick coin with 60% probability of heads. Sequential analysis models the situation in which a person is trying to determine which random variable is generating the data they observe, and in this example, whether they are flipping the trick coin or the fair coin based on the number of heads or tails they observe. The analysis is called sequential because each flip is accumulated to help determine which coin they are flipping. When the accumulated evidence passes one of two thresholds, the process ends, and the individual decides which type of coin he or she is flipping.

The common maximum likelihood method would be to choose theta that maximizes the likelihood of $P(\mathbf{x}|\theta)$ where $\mathbf{x}$ is a sequence of flips (e.g. H, T, H, H, T, H). In that case, the decision criterion would be

$$\begin{cases} \text{trick} & \text{if } P(x_1, x_2, ..., x_n|\theta = .6) > P(x_1, x_2, ..., x_n|\theta = .5) \\ \text{fair} & \text{if } P(x_1, x_2, ..., x_n|\theta = .6) \leq P(x_1, x_2, ..., x_n|\theta = .5) \end{cases}$$

This requires the modeler to decide the number of coin flips in advance. In sequential analysis, the relative likelihoods are used as weights in a random-walk diffusion model to move towards one threshold or the other. The number of coin flips are determined by the path of the random walk and when it crosses either threshold. The weight of each step is as follows:

$$w_i = \begin{cases} \log \frac{P(x_i=Heads|\theta=.6)}{P(x_i=Heads|\theta=.5)} & \text{if flip is heads} \\ \\ \log \frac{P(x_i=Tails|\theta=.6)}{P(x_i=Tails|\theta=.5)} & \text{if flip is tails} \end{cases}$$

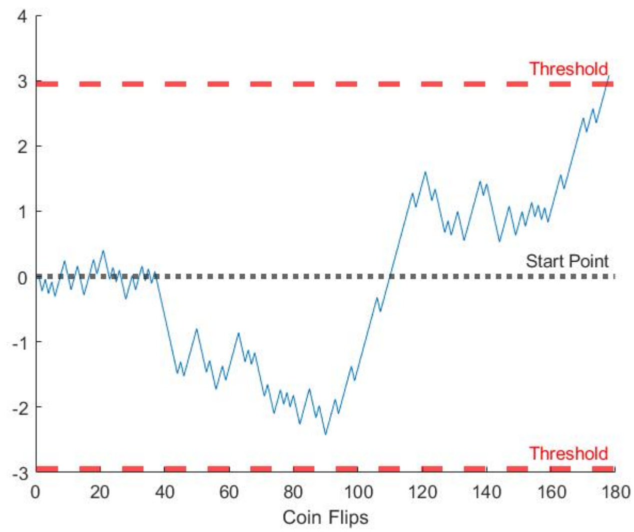
Beginning then at a starting point of zero, a hypothetical person using sequential analysis to determine whether a coin was fair or not is graphed below.

The thresholds are set in a such a way that the probability of Type I and Type II errors are each 5%.

After performing sequential analysis on 10,000 repeated trials of Sequential Analysis with a 50% chance of the coin being a trick coin, there are 225 Type I errors, 233 Type II errors, and it takes an average of 135.6 flips to determine the correct coin type.

But what if instead of the coin being equally likely to be trick or fair, it is more often a trick coin? Let's assume now that the coin is a trick coin 70% of the time. The number of Type I errors increases to 313, the number of Type II errors decreases to 134, but the average number of flips required to reach a decision is an almost identical 135.2

Figure 1: Sequential Analysis on Coin Flips



What if knowledge of the likelihood of the coin being a trick coin could be incorporated into the model?

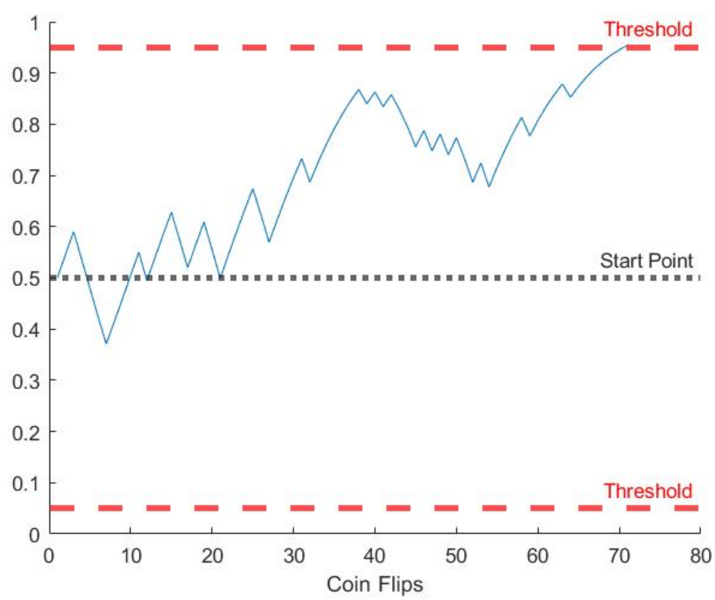
One way to do that is to use Bayesian Inference explicitly in the following way, where "Prior" is the decision maker's belief in the chance he or she will be given a trick coin, and the "Posterior" distribution is the result of updating the prior belief with data from coin flips.

$$P(\theta|x)_0 = P(\theta)_0 = \text{Prior}$$

$$P(\theta|x)_i = \begin{cases} \frac{P(\text{Heads}|\text{Trick})P(\theta)_i}{P(\text{Heads}|\text{Trick})P(\theta)_i + P(\text{Heads}|\text{Fair})(1-P(\theta)_i)} & \text{if Heads} \\ \\ \frac{P(\text{Tails}|\text{Trick})P(\theta)_i}{P(\text{Tails}|\text{Trick})P(\theta)_i + P(\text{Tails}|\text{Fair})(1-P(\theta)_i)} & \text{if Tails} \end{cases}$$

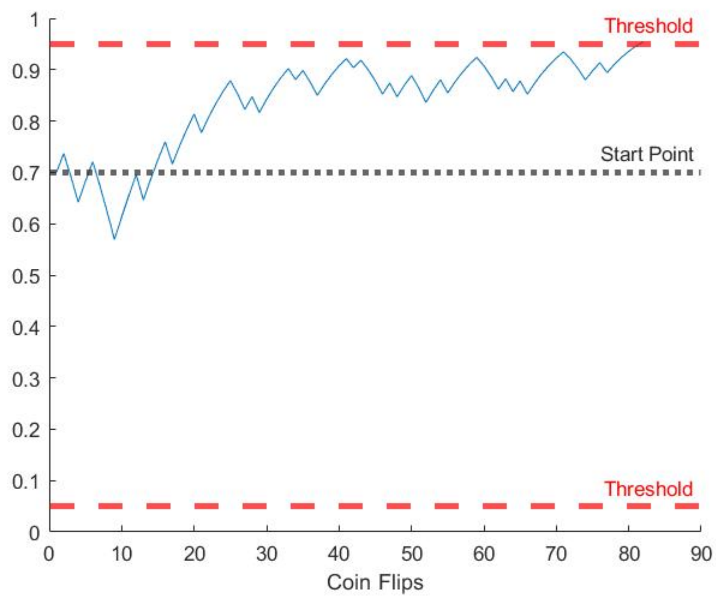
$$P(\theta)_i = P(\theta|x)_{i-1}$$

With this method we must apply different thresholds. Perhaps a natural choice is an upper bound of 95% and a lower bound of 5%. Once the posterior surpasses 95% or drops below 5%, the decision is made. In this case we are no longer using weights to make sequential steps, but instead plotting the diffusion model as the sequence of posterior distributions after each coin flip. In our first demonstration, I will use the prior belief that receiving a fair or trick coin is equally likely even though the coin has a 70% probability of being trick coin. The resulting plot appears quite similar to Gold and Shadlen's model, but uses a different scale for the y-axis (each trial will appear different in terms of the overall trend because the data generating process is random).



Also similar to Gold and Shadlen's model, with a prior of equal likelihood for fair and trick coins on 10,000 repeated trials, using Bayesian inference results in 317 Type I errors, 139 Type II errors, and an average of 136.9 coin tosses to make the decision.

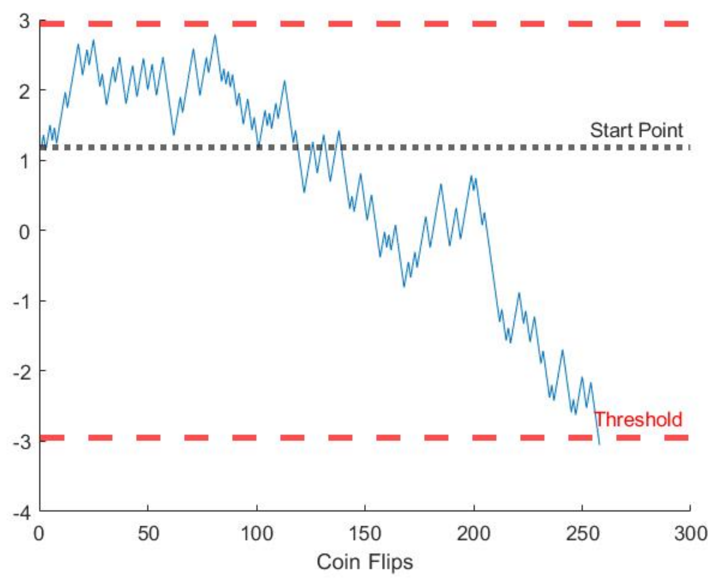
But, using a prior of 50% is not using all the information available to us! When setting the prior to 70%, the result on 10,000 repeated trials becomes 137 Type I errors, 330 Type II errors, and an average number of coin tosses of 119.3. This method improved on the time required to decide on the type of coin without decreasing overall accuracy. A plot of one of the 10,000 trials is shown below.



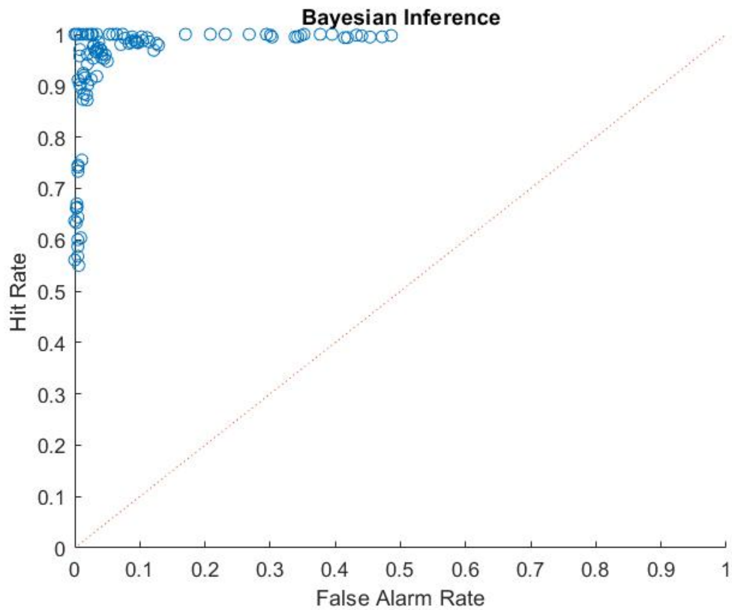
Another way to incorporate Bayesian Statistics to this decision process is to use the prior to inform our choice of the Starting Point in Gold and Shadlen's method. This is possible while keeping their scale and performing the following adjustment:

$$\text{Starting Point} = P(\theta)(\text{Upper Bound} - \text{Lower Bound}) + \text{Lower Bound}$$

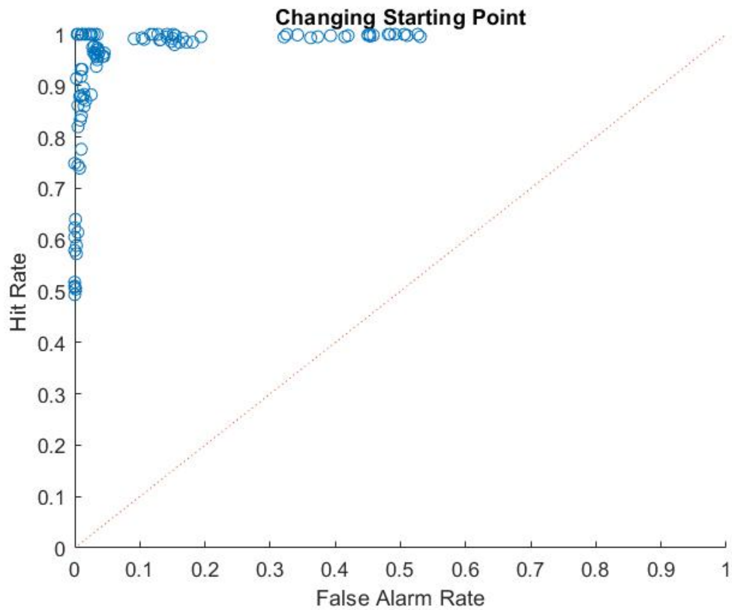
Using the above adjustment with a prior of 70% on 10,000 repeated trials results in 71 Type I errors, 467 Type II errors, and an average of 109.4 tosses to make a decision. It is even faster than the bayesian inference method, with fewer Type I errors and 50% more Type II errors. One of the 10,000 trials is shown below.

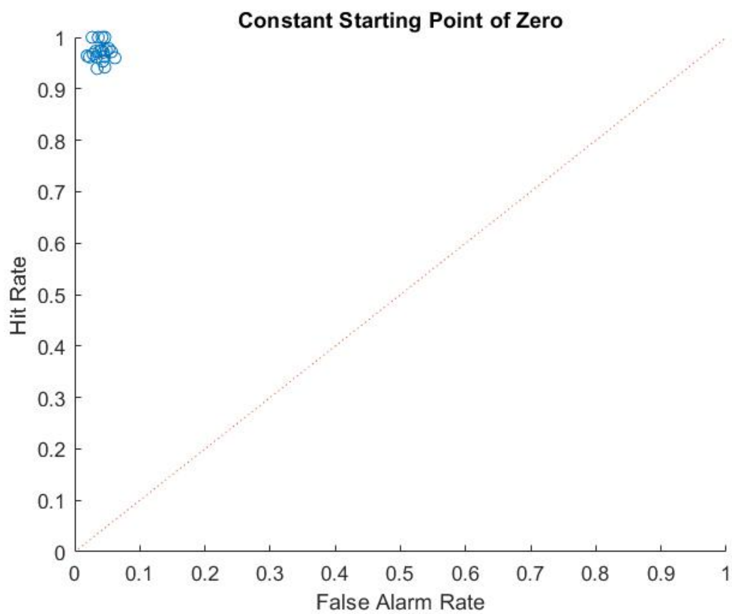


The accuracy of these methods can be shown in ROC curves. This was done by keeping track of hits and false alarms while independently varying the prior, proportion of trick coins, and the size of the trick coin bias. Because the parameters are independently varied, some points represent a mismatch of the belief of the likelihood of the trick coin and the actual likelihood of the trick coin (e.g. a prior of 10% when the actual percentage of trick coins is 80%). Such a feature results in a penalty shown in the increased False Alarm Rates for some trials and decreasing Hit Rates for others. Gold and Shadlen's original method of a constant starting point is not penalized in this way, which explains its more consistent accuracy. However, as seen in the clusters of points near the upper left hand corner, when the prior distributions are correct in the Bayesian methods, the resulting models are more accurate than the constant starting point method.



Of the Bayesian methods, Bayesian Inference has higher discriminatory power than the changing starting point method, as seen by the grouping of points farther in the upper left corner and the higher average D-prime value of 3.68 vs. 3.63 for changing the starting point. But this comes at the trade-off of being slower than the Gold and Shadlen method with changing starting point. The average number of flips over all the parameters was 31.33 for the changing starting point method and 32.94 for the Bayesian Inference method. The constant starting point method continues to require a higher average number of coint flips at 39.2, but its average D-prime is 3.84, another consequence of the penalty for mismatch in Bayesian methods when prior beliefs are inaccurate.


Changing Starting Point
Hit Rate
False Alarm Rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1


Constant Starting Point of Zero
Hit Rate
False Alarm Rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

4 Conclusion

In this work I showed that Bayesian Inference improves on Gold and Shadlen's model of Sequential Analysis by making decisions more quickly. Using priors to update the starting point used in Gold and Shadlen's model improved decision time even more, although with more False Alarms than in the Bayesian Inference method. These results show the efficacy of incorporating subjective probabilities into analytical models so long as those probabilities are accurate. Although the original Sequential Analysis model had more consistent accuracy than the models using Bayesian Inference, analysis of the ROC curves shows when prior beliefs are accurate, resulting joint decision models are also *more* accurate than models that do not incorporate prior beliefs. Given the work done to identify experts who consistently deliver more accurate judgments than others and the work done to efficiently aggregate expert judgment, these results suggest benefits to using Bayesian Inference in decision making.

5 References

Etz, A., & Vandekerckhove, J. (2018). Introduction to Bayesian Inference for Psychology. Psychonomic Bulletin & Review, 25(1), 5–34. https://doi.org/10.3758/s13423-017-1262-3

Gold, J. I., & Shadlen, M. N. (2007). The Neural Basis of Decision Making. Annual Review of Neuroscience, 30(1), 535–574. https://doi.org/10.1146/annurev.neuro.29.051605.113038

Hoo, Z. H., Candlish, J., & Teare, D. (2017). What is an ROC curve?

Kersten, D., Mamassian, P., & Yuille, A. (2004). Object perception as Bayesian inference. Annu. Rev. Psychol., 55, 271-304.

Mellers, B., Ungar, L., Baron, J., Ramos, J., Gurcay, B., Fincher, K., ... &

Tetlock, P. E. (2014). Psychological strategies for winning a geopolitical forecasting tournament. Psychological science, 25(5), 1106-1115.

Morgan, M. G. (2014). Use (and abuse) of expert elicitation in support of decision making for public policy. Proceedings of the National academy of Sciences, 111(20), 7176-7184.

Wagenmakers, E.-J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., Love, J., Selker, R., Gronau, Q. F., Šmíra, M., Epskamp, S., Matzke, D., Rouder, J. N., & Morey, R. D. (2018). Bayesian inference for psychology. Part I: Theoretical advantages and practical ramifications. Psychonomic Bulletin & Review, 25(1), 35–57. https://doi.org/10.3758/s13423-017-1343-3

Wasserman, L. (2013). All of statistics: a concise course in statistical inference. Springer Science & Business Media.