Bayesian Conditioning and Hypothesis Testing in Bernoulli Experiments
Understanding Popular Paradoxes

Ethan Pronovost
Pl 122 - Probability, Evidence, and Belief

March 21, 2019

1 Introduction

The dichotomoy between Bayesian and Frequentist approaches defines two radically different ways of thinking about statistical inference. Exploring this debate helps us understand the assumptions underlying our attempts to understand the nature of inductive reasoning.

Many have crafted particular examples in which one approach seems to be inferior, often by contradicting intuition. Due to their relative simplicity, Bernoulli experiments - which center around repeated sampling from a random Bernoulli process - are popular choices for such examples. These experiments yield well-defined, uncontroversial probability distributions, and are simple enough for the reader to have intuitions for what such an experiment should yield. Yet they are rich enough to allow for examples that contradict these intuitions, and differentiate between the two types of statistics.

It is important to have an appropriate understanding of the common statistical procedures of both Bayesian and Frequentist approaches. The fact that such examples can yield supposed contradictions signals that we need to enrich our intuitions. By formalizing the class of Bernoulli experiments, we are able to completely explain proposed "paradoxes". This framework allows us to build a better understanding of the Likelihood Principle, Bayesian conditioning, and hypothesis testing in Bernoulli experiments. By developing this more accurate intuition for these statistical procedures, we can gain insight into the nature of these procedures overall.

In Section 2, we will fully characterize the probability space governing Bernoulli experiments. Doing so will allow us to examine the Likelihood Principle in Section 3 and significance tests in Section 4 with sufficient foundation to make sense of the counterintuitive results. With this foundation, we can fully understand the type of paradox described in Lindley and Phillips [1976] as a natural consequence of the topology of the probability space, and the calculations performed on them by the two statistical approaches. In Section 5, we will demonstrate additional ways in which p-values can yield unintuitive results. Finally, in Section 6, we will conclude with a commentary on the correct attitude to hold towards significance tests.

2 Probability Space of Bernoulli Experiments

We wish to efficiently describe Bernoulli experiments in a way that allows us to make general probability computations. This will allow us to investigate the role of experimental design in the results of that experiment. As we will see below, for Bernoulli experiments the experimental setup amounts to specifying when you will stop the trials. This *stopping condition* will entirely determine the results of the experiment. By formalizing this notion, we establish a framework within which we can model various experiments in subsequent sections.

2.1 Experimental Setup

Following a canonical example, we will describe binomial experiments as coin tosses. Assume that the probability of a given coin toss coming up heads is $\Pr[H] = \theta$, where $\theta \in (0, 1)$. This parameter will be the quantity we wish to estimate.

Consider a general experiment involving coin tosses: we will toss a coin until the sequence of measured results satisfies some *stopping condition*, at which point we will return the full sequence we have measured. The choice of stopping rule completely determines the experiment. We can generalize such a rule as a subset S of all possible finite sequences of heads and tails (of arbitrary length). However, not any such subset should suffice. In order to obtain meaningful stopping conditions, we require two conditions.

- Termination: Let r be an infinite random Bernoulli(θ) sequence. Then for all $\theta \in (0, 1)$, there should exist some finite prefix of r in S with probability 1.
- No Preemption: If $s \in S$, then there is no other $s' \in S$ that has s as a prefix.

This set S is also the set of possible observable outcomes of the experiment.

The first condition requires our experiment to be well-defined. Without it, we would have experiments that may never end. Consider the experiment "stop if we obtain a sequence with 3 heads and 3 tails". If after 6 flips we have something other than 3 heads and 3 tails, we will keep flipping forever! This is an unrealistic model of experimentation in the real world.

The second condition is a simplifying minimality assumption. If there were two $s, s' \in S$ with s a prefix of s', then no sequence of flips would ever reach s' under the stopping rule S. Hence, it makes no difference if we consider S or $S \backslash \{s'\}$. The No Preemption condition requires us to make all such removals possible, to obtain the smallest stopping rule that yields the same distribution on outcomes.

Given an observed sequence of coin flips, the total number of heads h and tails t are a sufficient statistic for estimating θ. For simplicity, we will restrict the types of stopping rules considered to only depend on these two summary statistics of a sequence.

Note that this does not imply that our stopping rule S must be closed under permutations. As a clear example, consider the stopping rule "stop after the first heads". We can formally characterize this as the set $S = \{T^*H\}$, where T^* denotes a sequence of zero or more tails. Clearly, this set is not closed under permutations; for example if we permute $s = TTH$ to get THT, this will not be in S, since we would have stopped at TH and never proceeded further.

However, a constraint that is imposed by this restriction is that for any $s \in S$ and any permutation $\sigma(s)$, there must exist some prefix of $\sigma(s)$ (possibly the whole sequence) in S.

$$\forall\, s \in S, \forall\, \sigma \in S_{|s|}, \exists\, s' \text{ prefix of } \sigma(s) \text{ s.t. } s' \in S.$$

This amounts to the requirement that, if we would not have stopped earlier given the sequence $\sigma(s)$, then since s and $\sigma(s)$ have the same number of heads and tails, and we stop at s, we must also stop at $\sigma(s)$.

With this simplification, we can think of a sequence of coin flips as navigating a two dimensional grid which counts the number of heads in one dimension and the number of tails in the other. From each non-terminal node (h, t), we can transition to $(h + 1, t)$ with probability θ, and $(h, t + 1)$ with probability $1 - \theta$.

Viewing the states of the experiment as points on this two dimensional lattice, a stopping rule S is simply a subset of nodes that form the boundary of this experimental graph. For example, the stopping rule "stop after four tosses" corresponds to the set

$$S = \{(4, 0), (3, 1), (2, 2), (1, 3), (0, 4)\}.$$

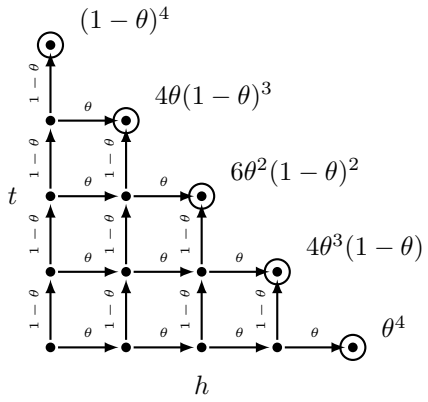

Figure 1: An experiment with stopping rule "stop after 4 tosses".

This modifies the graph of possible experimental states by marking these nodes as terminal, as shown in Figure 1.

In this framework, the experiment, stopping rule, and experimental graph all determine each other, so can be thought of as equivalent.

2.2 Probability Distribution

In this subsection, we will define the probability distribution for an experiment with stopping rule S. This will form the basis of all computations going forward. Both frequentist and Bayesian analyses require a model of the data generating process. It is an open question how to discern the "true" model underlying a data source. In this paper, we defer this question and assume that we have knowledge of the true generating distribution.

We can parametrize the set of possible worlds for an experiment as

$$\Omega = \{(h, t, \theta) : (h, t) \in S,\, \theta \in [0, 1]\}.$$

The probability of a given Bernoulli(θ) sequence with h heads and t tails is

$$\Upsilon(h, t, \theta) = \theta^h (1-\theta)^t.$$

Thus, the probability of reaching a given state (h, t) is $\Upsilon(h, t, \theta)$ times the number of possible sequences that end at (h, t). This later factor is a combinatorial problem in the experimental graph induced by the stopping rule S. The set of terminal nodes form a "cut" in the lattice that restricts the number of paths. For a stopping rule S and each node (h, t), let $\Gamma(h, t; S)$ be the number of paths from the origin to (h, t) in the corresponding experimental graph. For example, in Figure 1, this path number is simply the binomial coefficient $\Gamma(h, t; S) = \binom{h+t}{h}$ for $h + t \leq 4$, and 0 otherwise.

Given this, we can define our likelihood function

$$L(h, t, \theta; S) = \Gamma(h, t; S)\ \Upsilon(h, t, \theta). \tag{2.1}$$

Hence, the probability of measuring (h, t) given an experiment S and true parameter θ is

$$\Pr[(h, t) \mid S, \theta] = L(h, t, \theta; S).$$

This equation will be central to later computations. In the frequentist approach, this equation gives us the likelihood of a given measurement for some fixed value of θ. In the Bayesian approach, we can use this conditional probability via Bayes' Theorem to compute the posterior distribution of θ given S, and an observed datum (h, t).

2.3 Bayesian Updating

With the likelihood defined in Equation (2.1), the Bayesian update rule for an experiment is clear. Given an experiment with stopping rule S and $P(\theta)$ over the interval $[0, 1]$, we measure a sequence with h heads and t tails. Then the posterior over θ given this measurement is

$$\Pr[\theta \mid S, (h,t)] = \frac{P(\theta)\ L(h,t,\theta;S)}{\Pr[(h,t) \mid S]} = \frac{P(\theta)\ L(h,t,\theta;S)}{\int_0^1 P(\theta')\, L(h,t,\theta';S)\ d\theta'} = \frac{P(\theta)\ \Upsilon(h,t;\theta)}{\int_0^1 P(\theta')\ \Upsilon(h,t;\theta')\ d\theta'}. \tag{2.2}$$

This equation describes how a Bayesian's beliefs about the true value of θ should change after conducting an experiment. We can interpret this update as reweighting our beliefs of θ to make the observed result seem least surprising. How "surprising" an outcome (h, t) is if the true parameter is θ is exactly captured in the likelihood function $L(h, t, \theta; S)$; the lower the likelihood is, the more surprising the result. For example, if $\theta = 0.1$, we would be very surprised to measure a sequence of 9 heads and only 1 tail. By reweighting our beliefs by this likelihood, we give more weight to values of θ that are not very surprised by measuring (h, t).

A second important observation of Equation (2.2) is the last equality; we can replace the full likelihood $L(h, t, \theta; S)$ with just the binomial probability $\Upsilon(h, t, \theta)$; the path count and all influence of S drops out. This follows from the clean factorization of the likelihood given in Equation (2.1). The likelihood of measuring (h, t) given a true value of θ in experiment S has two components: a component Υ coming from the ratio between h and t in relation to θ, and a component Γ coming from how possible it is to achieve such a sequence (h, t) given the stopping rule S. Since a Bayesian update only considers the likelihood of the observed measurement, this second component cancels out in the numerator and denominator.

In this work we will not take Bayesian updating further. Some have sought to compare Bayesian updating to frequentist approaches by using the posterior for hypothesis testing (Steele [2010]). Such approaches conscript Bayesian statistics to support a fundamentally incompatible world view. While comparisons between these "Bayesian-influenced significance tests" and traditional frequentist significance tests are possible, such comparisons do not shed light on the fundamental perspective of the Bayesian approach to inference.

3 The Likelihood Principle

The likelihood principle is discussed in [Hill, 1987]. In that paper, Hill examines the Formal Likelihood Principle, and argues why the statement in full generality is not justifiable. We can consider this argument in the context of coin flip experiments described above. In this setting, the rationality of Hill's argument is obvious.

3.1 Hill's Argument

[Hill, 1987] begins by stating a principle that, at first glance, seems reasonable. In the context described in Section 2.1, this Formal Likelihood Principle is realized in the following statement.

Definition 3.1 (Formal Likelihood Principle)**.** Consider two experiments S_1 and S_2. Suppose that for some measurements $(h_1, t_1) \in S_1$ and $(h_2, t_2) \in S_2$, the likelihood functions (as functions of θ) are proportional for all $\theta \in [0, 1]$; i.e.

$$L(h_1, t_1, \theta; S_1) \propto L(h_2, t_2, \theta; S_2).$$

Then any conclusion that can be drawn from the measurement of (h_1, t_1) in the first experiment can also be drawn from the measurement of (h_2, t_2) in the second experiment.

On first pass, this seems reasonable. Most readers will take the *spirit* of this principle, instead of the explicit definition. Indeed, the intention behind this principle is valid. However, the Formal Likelihood

Principle as stated above over-generalizes the conclusions that can be drawn from the proportionality of the likelihoods. As Hill points out, it is possible that there are implications from either experiment that extend beyond the variables described by the likelihood functions. While the impacts of these measurements should yield equivalent conclusions about θ (given several other mundane assumptions), the fact that the likelihoods are proportional poses no constraint on implications beyond the variables that are proportional.

Hill offers a framework for a counterexample to the Formal Likelihood Principle. However, without a formal experimental context such as is defined in Section 2.1, he is unable to give a concrete, tangible example. Here, we will be able to do so.

Having highlighted flaws in the Formal Likelihood Principle, Hill then proposes an alternative Restricted Likelihood Principle. In the context of Section 2.1, this takes the following form. The additions to Definition 3.1 are shown in blue.

Definition 3.2 (Restricted Likelihood Principle)**.** Consider two experiments S_1 and S_2. Suppose that for some measurements $(h_1, t_1) \in S_1$ and $(h_2, t_2) \in S_2$, the likelihood functions (as functions of θ) are proportional for all $\theta \in [0, 1]$; i.e.

$$L(h_1, t_1, \theta; S_1) \propto L(h_2, t_2, \theta; S_2).$$

Furthermore, suppose that the choice of experiment is independent of θ (i.e. our beliefs about θ before conducting the chosen experiment are identical whether we chose Experiment 1 or Experiment 2). Then any conclusion involving θ and nothing else that can be drawn from the measurement of (h_1, t_1) in the first experiment can also be drawn from the measurement of (h_2, t_2) in the second experiment.

3.2 Formal Likelihood Principle

Given the definition of the likelihood from Equation (2.1), the experimental setup only impacts this likelihood through Γ, the path count in the experiment graph. Crucially, this path count is independent of θ. Taking (h_1, t_1) as a fixed observed datum, this path count will be a constant, so that $L(h_1, t_1, \theta; S_1) \propto \Upsilon(h_1, t_1, \theta)$. Thus, for any two experiments respective measurements that yield proportional likelihoods,

$$L(h_1, t_1, \theta; S_1) \propto L(h_2, t_2, \theta; S_2) \iff \Upsilon(h_1, t_1, \theta) \propto \Upsilon(h_2, t_2, \theta) \iff (h_1, t_1) = (h_2, t_2).$$

Thus, we can simplify the Formal Likelihood Principle to state that if two experiments S_1 and S_2 measure the same outcome, they should draw exactly the same conclusions.

As a counterexample to show how this principle can fail if the experiments shed light on information beyond the scope of the likelihood function, consider the following scenario. An election is taking place, with nine voters each either voting "Yes" or "No". The person counting the votes does not wish to announce the results explicitly. Instead; they will play a coin flip game. They will give you a coin of unknown bias, and ask you to repeatedly flip it. In Experiment 1, they agree to tell you to stop when the number of heads equals the number of "Yes" votes; in Experiment 2, they agree to tell you to stop when the number of heads equals the number of "No" votes.

This setup is a slight modification of the context described in Equation (2.1). Since the votes are cast, the stopping rules in both experiments are well-defined beforehand; the experimenter simply doesn't know what they are going into the experiment. However, after making a measurement, they will fully know the stopping rule, and can compute any relevant statistic that someone with pre-knowledge of the stopping rule could.

The contradiction here seems both obvious and unavoidable. Consider being told to stop after having flipped 6 heads and 2 tails. Clearly the likelihoods will be proportional; given an uniform prior there is weak evidence that the coin is biased towards heads. However, in Experiment 1, we conclude that there are 6 "Yes" votes, and the resolution passes; in Experiment 2, we conclude that there are 6 "No" votes, and the resolution fails.

This counterexample seems silly; clearly nobody would claim that the same measurement should yield identical conclusions in these two experiments. In this simple context, ways in which the Formal Likelihood Principle can overstate conclusions are obvious. However, in the real world of far more complicated experiments, the cases may not be this clear and yet the issue is just as present. We should demand that any governing principle should be able to be blindly followed, and never mislead. It is with this motivation in mind that Hill proposes his Restricted Likelihood Principle.

3.3 Restricted Likelihood Principle

In the concrete context of Section 2.1, Hill's Restricted Likelihood Principle becomes a trivial corollary of Bayesian updating. Since the choice between S_1 and S_2 must make no difference to our beliefs about θ, the prior $P(\theta)$ must be identical in both cases.

With this restriction, then since $L(h_1, t_1, \theta; S_1) = c \cdot L(h_2, t_2, \theta; S_2)$ for some positive constant c, by Equation (2.2), the posterior distributions

$$\begin{aligned}
\Pr[\theta \mid S_1, (h_1, t_1)] &= \frac{P(\theta)L(h_1, t_1, \theta; S_1)}{\int_0^1 P(\theta')L(h_1, t_1, \theta'; S_1)\, d\theta'} \\
&= \frac{c \cdot P(\theta)L(h_2, t_2, \theta; S_2)}{c \cdot \int_0^1 P(\theta')L(h_2, t_2, \theta'; S_2)\, d\theta'} \\
&= \frac{P(\theta)L(h_2, t_2, \theta; S_2)}{\int_0^1 P(\theta')L(h_2, t_2, \theta'; S_2)\, d\theta'} \\
&= \Pr[\theta \mid S_2, (h_2, t_2)].
\end{aligned}$$

In this setting, we have seen that proportionality of likelihoods implies that $(h_1, t_1) = (h_2, t_2)$. But then by Equation (2.2), the posterior distribution depends only on the prior $P(\theta)$ and the observed measurement. Thus, the Restricted Likelihood Principle can be paraphrased as saying that if the two things that impact the posterior of θ are identical, then any conclusion based on that posterior is valid in either experiment.

Again, this specificity may seem silly. Indeed, in this simple context the results are not revolutionary. However, it is important to establish a solid foundation of principles that can always be trusted as we venture into the real world of extremely complicated experiments.

4 Significance Tests

A major debate between frequentist and Bayesian approaches to statistics lies over the importance of experimental design in interpreting the outputs. As we have seen above, for coin flip experiments, the experimental design (i.e. the stopping rule) will not affect the posterior distribution of θ beyond determining the possible measured outcomes. In other words, the posterior of θ is independent of the experimental design, conditional on the observed datum; we may write

$$\Pr[\theta \mid S, (h, t)] = \Pr[\theta \mid (h, t)].$$

In Bayesian statistics, the posterior is used to obtain information about θ. For example, we can compute the probability that $\theta > \frac{1}{2}$ by integrating the posterior from $\frac{1}{2}$ to 1. In contrast, frequentist statistics commonly utilize tests of significance to evaluate a null hypothesis. While the posterior is independent of experimental design conditional on the observed datapoint, such significance tests are not. As we will see, p-values reflect more of the experimental design then the actual information about θ.

Let us consider the example from [Lindley and Phillips, 1976]. As above, we perform two experiments involving flipping a coin. In the first experiment E_1, we flip the coin 12 times. In the second experiment E_2, we flip until we get 3 tails. In both, we measure a sequence with $h = 9$ heads and $t = 3$ tails.

4.1 Formalizing the Experiments

First consider Experiment 1. This is a similar situation to that described in Figure 1. Concretely, the stopping condition $S_1 = \{(h,t) : h+t = 12\}$. The experimental graph for this case looks similar to that of Figure 1, and the path count is given by the simple combinatorial combination function

$$\Gamma(h,t;S_1) = \begin{cases} \binom{h+t}{h} & h+t \leq 12 \\ 0 & \text{otherwise} \end{cases}.$$

The second experiment is a different situation to analyze. The stopping rule $S_2 = \{(h,3) : h \in \mathbb{N}\}$. Graphically, this yields a very different looking graph.

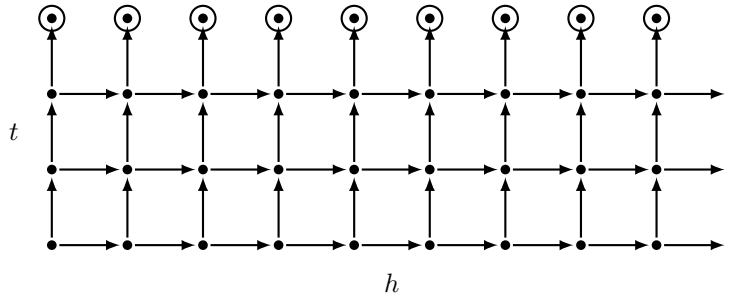

Figure 2: An experiment with stopping rule "stop after 3 tails".

Unlike the previous graph, this graph is infinite. However, unless $\theta = 1$ any sequence of coin tosses will eventually terminate with probability 1, so we still have a well-defined stopping rule. Figure 2 shows that the path count

$$\Gamma(h,t;S_2) = \begin{cases} \binom{h+t}{h} & t \leq 2 \\ \binom{h+2}{h} & t = 3 \\ 0 & \text{otherwise} \end{cases}.$$

Note the different case when $t = 3$; because $(h, t-1)$ is terminal, the only paths that reach (h,t) must go through $(h-1,t)$.

Our observed measurement $(9,3)$ is the only point in $S_1 \cap S_2$.

4.2 Bayesian Posterior

The Bayesian posterior for these experiments is rather straight forward. Using Equation (2.2),

$$\Pr[\theta \mid (9,3)] = \frac{P(\theta)\cdot\theta^9(1-\theta)^3}{\int_0^1 P(\theta)\cdot\theta^9(1-\theta)^3}.$$

If we take P to be the uniform prior, this yields

$$\Pr[\theta \mid (9,3)] = 2860\,\theta^9(1-\theta)^3,$$

which appears as we would expect, with a concentration of mass around $\theta = \frac{3}{4}$ (see Figure 3).

This posterior is identical from either experiment. With this result, we would conclude that the probability of θ being larger than $\frac{1}{2}$ is

$$\Pr\left[\theta > \frac{1}{2} \mid (9,3)\right] = \int_{\frac{1}{2}}^1 \Pr[\theta \mid (9,3)] \approx 0.954.$$

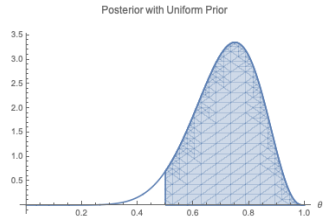

Figure 3: Posterior $\Pr[\theta \mid (9,3)]$ under the uniform prior.

4.3 Null Hypothesis Testing

To offer a frequentist analysis of whether the coin is biased, we can perform a p-value test on the null hypothesis that $\theta_0 = \frac{1}{2}$. To do so, we compute the probability, given $\theta = \frac{1}{2}$, of making a measurement of $(9,3)$ or something more extreme. Commonly, if this p-value is less than 5%, the result is deemed statistically significant. For simplicity, we will use this threshold in this work. Similar findings can be obtained for any other choice of threshold.

In Experiment 1, the more extreme measurements would be $(9,3)$, $(10,2)$, $(11,1)$, and $(12,0)$. Thus, we get a p-value of

$$\begin{aligned} p_1 &= \sum_{h=9}^{12} \Pr\left[(h, 12-h) \mid S_1,\, \theta = \frac{1}{2}\right] \\ &= \sum_{h=9}^{12} \Gamma(h, 12-h; S_1)\ \Upsilon\left(h, 12-h, 1/2\right) \\ &= \frac{1}{2^{12}} \sum_{h=9}^{12} \binom{12}{h} \\ &= \frac{299}{4096} \\ &\approx 0.0730. \end{aligned}$$

Crucially, this is above the 5% threshold commonly used in statistics, so would not be deemed significant.

For Experiment 2, the more extreme measurements would be $(9,3)$, $(10,3)$, $(11,3)$, etc... on to infinity. Thus, we get a p-value of

$$\begin{aligned} p_2 &= \sum_{h=9}^{\infty} \Pr\left[(h,3) \mid S_2,\, \theta = \frac{1}{2}\right] \\ &= \sum_{h=9}^{\infty} \Gamma(h, 3; S_2)\ \Upsilon\left(h, 3, 1/2\right) \\ &= \sum_{h=9}^{\infty} \binom{h+2}{2} 2^{-(h+3)} \\ &= \frac{134}{4096} \\ &\approx 0.0327. \end{aligned}$$

This is less than half of p_1, and below the 5% threshold. Thus, the same result is deemed insignificant in the first experiment, but significant in the second experiment.

4.4 Making Sense of the Paradox

[Lindley and Phillips, 1976] provides this example as evidence of the inherent problem with significance tests, without digging into the details of what causes this discrepancy. With our formalization of stopping rules and graph path counts, we are equipped to do so.

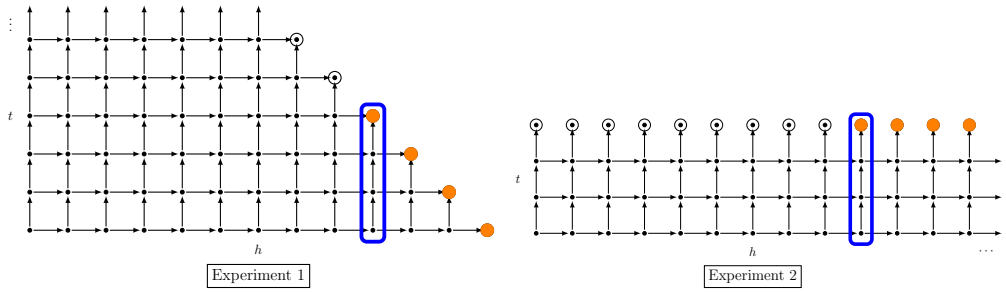

Figure 4: The "more extreme values" of these two experiments. In both cases, a sequence of coin flips will yield one of the extreme values (shown in orange) if and only if it passes through one of the states in the blue box.

Refer back to the experimental graphs defined by these two stopping rules. As shown in Figure 4, in both experiments we will measure one of our "more extreme values" if and only if the sequence passes through one of the four states indicated by the blue box at some point in the experiment. Thus, the p-value is exactly the probability of reaching one of these blue states given each of the two experimental setups.

$$p_i = \sum_{t=0}^{3} L(9, t, \theta_0; S_i) = \sum_{t=0}^{3} \Gamma(9, t; S_i)\ \Upsilon(9, t, \theta_0).$$

Examining this equation, we see that p_1 and p_2 will only differ in the term $\Gamma(9, 3; S_i)$. This difference amounts to the fact that we can reach $(9, 3)$ via the edge $(8, 3) \to (9, 3)$ in Experiment 1, but not Experiment 2. The probability of taking this edge in the first experiment is

$$\Gamma(8, 3; S_1)\ \Upsilon(8, 3; \theta_0) \cdot \theta_0 = \binom{11}{8} \theta_0^9\ (1 - \theta_0)^3.$$

With our null hypothesis $\theta_0 = \frac{1}{2}$, this simplifies to

$$\binom{11}{8} \cdot \frac{1}{2^{12}} = \frac{165}{4096} = p_1 - p_2.$$

The difference in p-values between the two experiments is entirely explained by the different layouts of the stopping rules. In Experiment 1, there were more ways to achieve a "more extreme value" than in Experiment 2. Concretely, if we were to take $S_3 = S_1 \cup \{(8, 3)\}$, so that the route $(8, 3) \to (9, 3)$ is disallowed, then the p-value of $(9, 3)$ under this third experiment will match p_2.

What about for other null hypotheses? We can consider p_1 and p_2 as functions of the null hypothesis θ_0. Examining the above results, we see that

$$p_1 - p_2 = \Gamma_1(8, 3)\ \Upsilon(8, 3; \theta_0) \cdot \theta_0$$

for all $\theta_0 \in [0, 1]$. Thus, $p_1 > p_2$ for all $\theta_0 \in (0, 1)$. Not surprisingly, the difference in these p-values is greatest around $\theta = \frac{3}{4}$. As shown in the black dotted lines of Figure 5, there is an interval of approximately $\theta_0 \in (0.48, 0.53)$ in which a null hypothesis is refuted with a 5% p-value threshold in Experiment 2, but not Experiment 1.

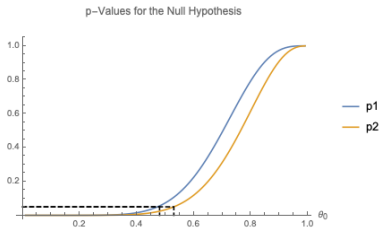

Figure 5: Significance values $p_1(\theta)$ and $p_2(\theta)$ of $(9, 3)$ for various null hypotheses θ.

4.5 Two Different Dimensions of Analysis

Comparing the Bayesian posterior to the frequentist p-value testing, we see that these are two orthogonal perspectives on how to evaluate θ. Recall from Section 2.1 the description of all possible worlds

$$\Omega = \{(h, t, \theta) : (h, t) \in S,\, \theta \in [0, 1]\}.$$

In this set of worlds, the experiment's stopping rule determines the horizontal $h - t$ plane; the vertical θ axis at each point (h, t) is determined by $\Upsilon(h, t; \theta)$, which is independent of the stopping rule.

After measuring a value (h, t), we can use Equation (2.2) to update a Bayesian posterior. This amounts to restricting to the vertical line in Ω corresponding to the observed pair (h, t) (see Figure 6 left). This restricted set is in the dimension "independent of the experimental design"; the only influence of the experiment is in making this value (h, t) possible to observe.

In contrast, evaluating a null-hypothesis in the frequentist approach amounts to examining a horizontal slice of this space Ω at the value θ_0 of the null hypothesis (see Figure 6 right). Clearly this slice will depend heavily on the experimental design.

These two approaches examine orthogonal dimensions of the total world space Ω. We can consider the horizontal $h - t$ plane as "data dimensions", and the vertical θ axis as the "parameter dimension". In the Bayesian approach, the observed data is fixed, and we consider the resulting distribution in the unknown parameter. In contrast, a null hypothesis test fixes a parameter value θ_0, and considers what the experimental data distribution would look like in this context. These two different ways of "slicing" Ω for the set of worlds Ω_1 of Experiment 1 are visualized in Figure 6.

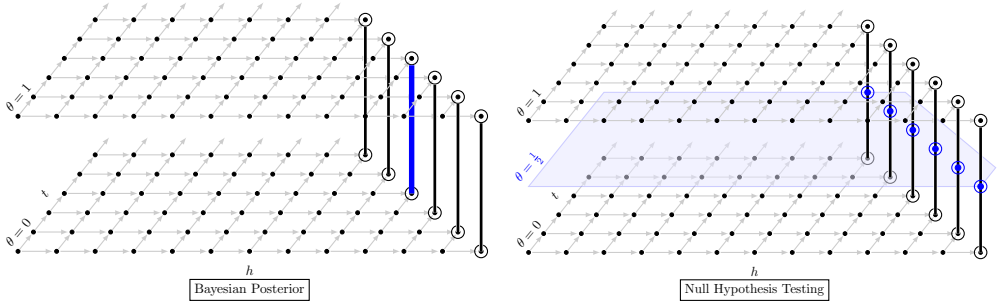

Figure 6: The two different modalities for examining a subset of possible worlds $\Omega_1 = \{(h, t, \theta) \mid (h, t) \in S_1, \theta \in [0, 1]\}$ from Experiment 1.

This difference lies at the heart of the debate between Bayesian and Frequentist Statistics. The Bayesians would argue that once the data is fixed, we should restrict to the still-viable world states (i.e. those on the vertical string at (h, t) of the observation), and consider the probability distribution over this string via Bayesian conditining. The frequentist perspective considers some fixed θ_0, and asks how likely it is that that θ_0 would lead to the results of the given experiment. The former can give us predictive power not only on

the true value of θ, but on our uncertainty of this measurement. The later, on the other hand, is useful when considering whether or not θ takes on a specific value.

5 Experimental Gerrymandering

Because of the dependence on experimental setup, comparing p-values directly between experiments can lead to questionable results (as demonstrated above in Section 4.3). This problem becomes all the more serious when we consider the selection bias of researchers and journals to only report findings that achieve low p-values.

Expanding on this example, we can consider asking the following question: given a stopping rule S, a null hypothesis $\theta_0 = \frac{1}{2}$, and a true parameter value θ, what is the probability of obtaining significant results at the 5% p-value threshold? How can we modify S to improve this quantity?

A Jupyter notebook computing these examples can be found on my Github.

5.1 Formalizing p-Values

First, we must concretely specify what the p-value is for an arbitrary experiment S. Given a measurement (h, t), let $\phi = \frac{h}{h+t} \in [0, 1]$ be the fraction of observed flips that were heads. This is the quantity we would first think to care about, the maximum likelihood estimate for the unknown parameter θ. For a given stopping rule S and true parameter value θ, we can define the conditional probability on ϕ by

$$\Pr[\phi \mid S, \theta] = \sum_{\substack{(h,t) \in S \\ \frac{h}{h+t} = \phi}} \Pr[(h, t) \mid S, \theta].$$

Using this, we can describe the cumulative density function of ϕ as

$$\text{CDF}\,[\phi \mid S, \theta] = \sum_{\substack{(h,t) \in S \\ \frac{h}{h+t} \leq \phi}} \Pr[(h, t) \mid S, \theta].$$

We will define "more extreme values" than some ϕ to be all states $(h, t) \in S$ such that $\frac{h}{h+t}$ lies further towards the closest endpoint of $[0, 1]$ than ϕ. In terms of the experimental graph, this amounts to all terminal states lying at an angle closer to either the horizontal or vertical axis. Note that this definition agrees with our intuitive choice for "extreme values" in Section 4.3 for Experiments 1 and 2 (see Figure 7).

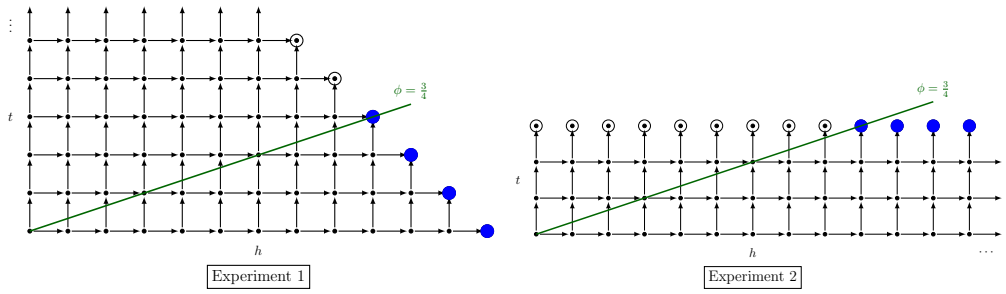

Figure 7: The "more extreme values" can be viewed in terms of the fraction ϕ of heads. This definition agrees with intuition for conventional stopping rules like those described in Section 4.1.

Concretely, for a null hypothesis θ_0, the p-value of ϕ will be[1]

$$\text{pVal}(\phi; S, \theta_0) = \begin{cases} 1 - \text{CDF}[\phi \mid S, \theta_0] & \text{CDF}[\phi \mid S, \theta_0] \geq \frac{1}{2} \\ \text{CDF}[\phi \mid S, \theta_0] & \text{CDF}[\phi \mid S, \theta_0] \leq \frac{1}{2} \end{cases}.$$

Thus, for a true parameter value θ, and the standard 5% threshold, the probability of obtaining a "significant" result with a threshold of 0.05 is

$$\text{ProbSig}(\theta; S, \theta_0) = \Pr_{\phi \sim \Pr[\phi \mid S, \theta]} [\text{pVal}(\phi; S, \theta_0) < 0.05]. \tag{5.1}$$

For $\text{ProbSig}(\theta; S, \theta_0)$ to be high for some value of θ, we require that $\text{pVal}(\phi; S, \theta_0)$ be below 0.05 for the majority of the probability mass of $\Pr[\phi \mid S, \theta]$. That is, we want $\Pr[\phi \mid S, \theta_0]$ to decay rapidly to zero for ϕ far from θ_0, and for the majority of the probability of mass of $\Pr[\phi \mid S, \theta]$ to be centered around θ. In other words, we want the variance of both $\Pr[\phi \mid S, \theta]$ and $\Pr[\phi \mid S, \theta_0]$ to be small. However, there are also subtler ways to modify this quantity $\text{ProbSig}(\theta; S, \theta_0)$. We will see this play out in various experimental designs below.

For any stopping rule S, by definition

$$\begin{aligned} \text{ProbSig}(\theta_0; S, \theta_0) &= \Pr_{\phi \sim \Pr[\phi \mid S, \theta_0]} [\text{pVal}(\phi; S, \theta_0) < 0.05] \\ &= \Pr_{\phi \sim \Pr[\phi \mid S, \theta_0]} [\text{CDF}(\phi \mid S, \theta_0) < 0.05 \vee \text{CDF}(\phi \mid S, \theta_0) > 0.95] \\ &\leq 0.1 \end{aligned}$$

For a continuous distribution, this value would be exactly 0.1. However, for discrete distributions, this value can be less than 0.1.

Using the null hypothesis $\theta_0 = \frac{1}{2}$, we thus cannot hope to achieve a high probability of significant results if $\theta = \theta_0$ (the definition of "significant results"). However, we can consider what experimental designs will yield a high probability of obtaining significant results for other values of $\theta \neq \theta_0$.

5.2 Asymmetric Stopping Rules

Let us first examine what these functions look like for the two experiments from above. We can visualize the distributions $\Pr[\phi \mid S, \theta]$ for various values of θ. Note that there is some ambiguity in how to display these probability distributions. To standardize between experiments, we plot ϕ in $[0, 1]$. However, in Experiment 2, the possible values of ϕ are not uniformly distributed. Thus, by plotting the probabilities themselves, the "area under the curve" will not be one (as evident in the distribution with $\theta = 0.9$ for Experiment 2 in Figure 8). However, if we examined the heights of each point on any of the probability distributions, they would indeed sum to one.

Experiment 1 is symmetric about $\theta = \frac{1}{2}$; however, Experiment 2 is not. Looking at the right hand plot of Figure 8, it appears as if the distribution $\Pr[\phi \mid S_2, \theta = 0.1]$ falls almost entirely within regions of high probability for the null hypothesis $\Pr[\phi \mid S_2, \theta = 0.5]$.

This asymmetry is reflected in the probability of obtaining a significant result. For values of θ less than a half, Experiment 2 is extremely unlikely to yield significant results. Yet for values of θ greater than a half, this second experiment is more likely than the first to yield significant results.

It seems like Experiment 2 would only be advisable if one were to believe that the true value of θ is greater than one half. This observation serves as an introduction to the world of experimental design. We can design stopping rules to be more sensitive to situations we believe more probable, or care more about.

[1] If we treat ϕ as a discrete quantity, we should replace $1 - \text{CDF}[\phi \mid S, \theta_0]$ with $1 - \text{CDF}[\phi \mid S, \theta_0] + \Pr[\phi \mid S, \theta]$ to get the probability of obtaining a measurement greater than *or equal to* ϕ.

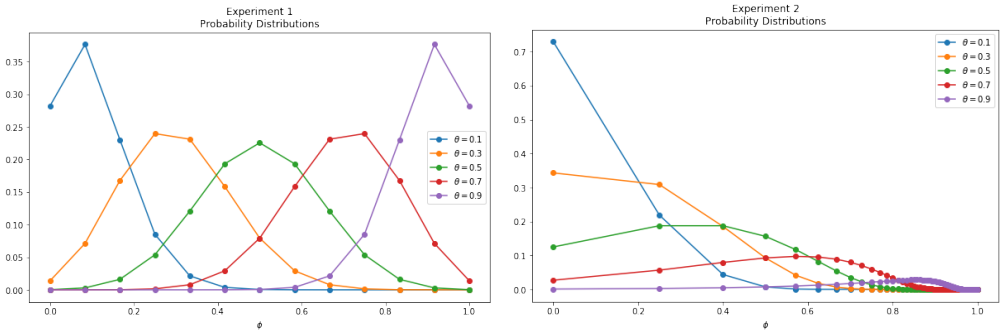

Figure 8: The distributions $\Pr[\phi \mid S, \theta]$ for various values of θ in both experiments.

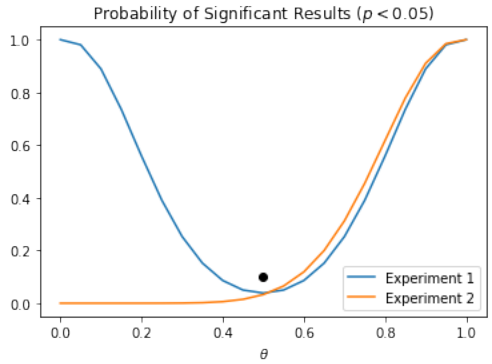

Figure 9: The probability of significant ($p < 0.05$) results in the two experiments, for various values of the true parameter θ. The black dot is plotted at $(\theta_0 = 0.5, 0.1)$.

5.3 Early Stopping for Unpromising Results

Consider first modifying Experiment 1 by stopping early if the observed sequence is close to evenly split. This yields a significantly different distribution from that of Experiment 1 (see Figure 10, right), with a much more irregular pattern for the middle interval of ϕ.

When computing $\text{ProbSig}(\theta; S, \theta_0)$, the only thing that matters is the set of values that would be significant to observe,

$$\Psi(S, \theta_0) = \{\phi \in S : \text{pVal}(\phi; S, \theta_0) < 0.05\}$$

and the probability

$$\Pr_{\phi \sim \Pr[\phi|S,\theta]}[\phi \in \Psi(S, \theta_0)] = \text{ProbSig}(\theta; S, \theta_0).$$

Thus, if we are given two stopping rules S and S' such that $\Psi(S, \theta_0) = \Psi(S', \theta_0)$, and $\Pr[\phi \mid S', \theta] = \Pr[\phi \mid S, \theta]$ for all $\theta \in [0, 1]$ and $\phi \in \Psi(S, \theta_0)$, then $\text{ProbSig}(\theta; S, \theta_0) = \text{ProbSig}(\theta; S', \theta_0)$ for all $\theta \in [0, 1]$.

Applying this to the proposed Experiment 3, one can check that

$$\Psi(S_1, \theta_0 = 1/2) = \{(12, 0), (11, 1), (10, 2), (2, 10), (1, 11), (0, 12)\} = \Psi(S_3, \theta_0 = 1/2).$$

Since the path count to a node (h, t) only depends on nodes to the left and below it in the experimental graph (i.e. nodes (h', t') where $h' \leq h$ and $t' \leq t$),

$$\Gamma(h, t; S_3) = \Gamma(h, t; S_1), \qquad \forall\, (h, t) \in \Psi(S_1, \theta_0 = 1/2) = \Psi(S_3, \theta_0 = 1/2).$$

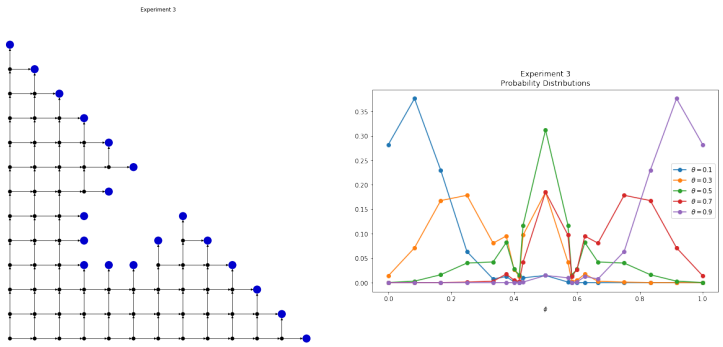

Figure 10: Left: Experiment 3 graph. Right: The probability distribution of Experiment 3.

Hence, the modification we made to Experiment 1 to obtain Experiment 3 will not impact the probability of significant results. This can be empirically verified in Figure 11 (the red line is superimposed on the blue).

One might hope to drop the first condition that $\Psi(S, \theta_0) = \Psi(S', \theta_0)$; it seems that if the second condition holds, this first condition would be implied. While this argument might hold in continuous distributions, it fails for discrete distributions such as Bernoulli experiments. To see this, consider a stopping rule $S_4 = S_3 \cup \{(7, 3)\}$ (see left, Figure 11). We have added a single additional stopping condition. Still, we have that $\Gamma(h, t; S_4) = \Gamma(h, t; S_3) = \Gamma(h, t; S_1)$ for all $(h, t) \in \Psi(S_1, \theta_0 = 1/2) = \Psi(S_3, \theta_0 = 1/2)$. However, with the addition of this extra stopping condition, we decrease $\Gamma(9, 3; S_4)$, such that now $\Psi(S_4, \theta_0 = 1/2) = \Psi(S_3, \theta_0 = 1/2) \cup \{(9, 3)\}$.

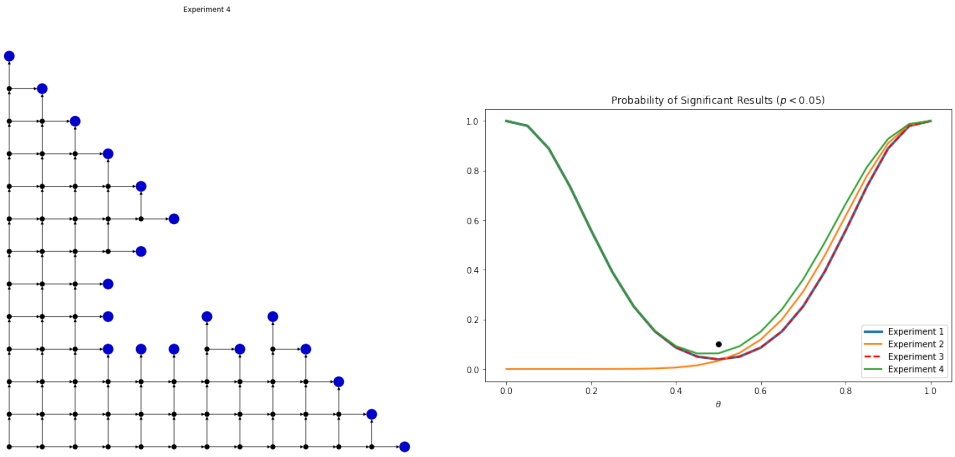

Figure 11: Left: Experiment 4 graph. Right: The probability of significant results from Experiment 3 matches that of Experiment 1, despite the difference in stopping rule (see red line superimposed on blue line). Experiment 4 yields a higher probability of significant results, with the largest difference occurring at $\theta = \frac{3}{4}$.

This difference will mean that Experiment 4 is more likely to yield "statistically significant" results than Experiments 1 and 3. This difference is exactly

$$\text{ProbSig}(\theta; S_4, \theta_0 = 1/2) - \text{ProbSig}(\theta; S_3, \theta_0 = 1/2) = \Pr[(9, 3) \mid S_4, \theta],$$

and so will be greatest for $\theta \approx \frac{3}{4}$. For $\theta < \frac{1}{2}$, this difference will quickly approach 0.

Two reasonable quantities for an experimenter to keep in mind when designing an experiment are the probability of obtaining significant results, and the cost of performing the experiment. In this context, the cost of an experiment is the number of flips necessary. There are many ways an experimenter could measure cost (e.g. the expected value, upper bound, 90^{th} percentile, etc.).

Overall, it seems reasonable that the experimenter should wish to maximize the probability of obtaining significant results, all else being equal. The cost of the experiment may or may not matter; if it does, it seems reasonable to assume the experimenter would like to reduce costs, all else being equal.

The example provided in this section offers a stopping rule S_4 with a probability of significant results that strictly dominates that of S_1 for all $\theta \in (0, 1)$, and with an experimental cost equal or less than that of S_1 (depending on what metric is used). Thus, no matter how the experimenter measures experiment cost or trades off these two considerations, it seems that Experiment 4 would be strictly preferable to perform than Experiment 1. Furthermore, the symmetrization of Experiment 4, $S_5 = S_4 \cup \{(4, 7)\}$, will strictly dominate S_4 in the same way, offering improvements for $\theta < \frac{1}{2}$.

Assuming experimenters are rational and understand the laws of probability, why would one choose to use S_1 over S_5? One possible answer is that such a contrived stopping rule would be looked down on by peer reviewers. Formalizing a notion of what a "contrived" stopping rule is is difficult at best. The stopping rule S_5 can be conducted in a pre-registered, "scientific" manner. If this rule is "too contrived", what stopping rules are admissible?

Any attempt to define a subset of acceptable stopping rules is just as arbitrary as the choice of stopping rule to use in the first place. Are only rules of the form "stop after n tosses" acceptable? Then what about "stop after n tails"? If we allow that, we must also allow "stop after n heads". But then what about the union of these two rules, "stop after n heads or n tails"? Where to stop this chain of reasoning requires drawing a line at some arbitrary point. For any such line, there is likely some example of an "unacceptable" stopping rule that would seem intuitively reasonable in certain contexts.

The Bayesian answer for why one might perfer S_1 over S_5 is clear. The posterior $\Pr[\theta \mid S, (h, t)]$ becomes more concentrated (i.e. with lower uncertainty) as $h + t$ increases. Thus, for the Bayesian, the expected variance of the posterior will be lower in Experiment 1.

5.4 *p* for Patience

The results in Section 5.3 can be described as arising from an edge effect of the discrete nature of our experiments; namely that $\Pr[\text{CDF}[\phi \mid S, \theta] < 0.05]$ can be less than 0.05. One might hope that all issues with p-values fade away as we perform larger and larger experiments, with distributions that appear closer and closer to being continuous. However, the law of large numbers presents its own problem to the use of p-values.

Consider the simple "diagonal" class of stopping rules "stop after n flips"; concretely let $D_n = \{(h, t) : h + t = n\}$. Extending the analysis from Section 4.1, the path count of such graphs will be

$$\Gamma(h, t; D_n) = \begin{cases} \binom{h+t}{h} & h + t \leq n \\ 0 & \text{otherwise} \end{cases},$$

and the distribution of h over D_n is simply the binomial distribution $h \sim \text{Binom}(n, \theta)$.

By the central limit theorem, as $n \to \infty$,[2]

$$h \sim \text{Binom}(n, \theta) \sim \mathcal{N}\left(n\theta, n\theta(1 - \theta)\right).$$

[2]Here, we use the notation of mean and variance for the normal distribution $\mathcal{N}(\mu, \sigma^2)$.

Since we care about the value of $\phi = \frac{h}{h+t} = \frac{h}{n}$, we get that in this limit

$$\phi \sim \mathcal{N}\left(\theta, \frac{\theta(1-\theta)}{n}\right).$$

Hence, as $n \to \infty$, the distribution of ϕ will converge tighter and tighter around θ, with the variance tending towards zero.

Recall from the discussion in Section 5.1 that this is exactly what we would like to get a high probability of obtaining significant results. When n is sufficiently large, the distribution of the null hypothesis $\Pr[\phi \mid D_n, \theta_0]$ will quickly decay to zero for ϕ not close to θ_0, and the set $\Psi(D_n, \theta_0)$ of results that would be deemed significant will be large. Similarly, if the true parameter value θ is sufficiently far away from θ_0, then the true distribution $\Pr[\phi \mid D_n, \theta]$ will have little overlap with the null hypothesis. Thus, taking n sufficiently large, we can get $\text{ProbSig}(\theta; D_n, \theta_0)$ arbitrarily close to 1 for any $\theta \neq \theta_0$. Indeed, in the limit as $n \to \infty$, this probability of signifiant results goes to 1 for all $\theta \neq \theta_0$, with a single point discontinuity of 0.1 at θ_0. This phenomenon can be seen in Figure 12.

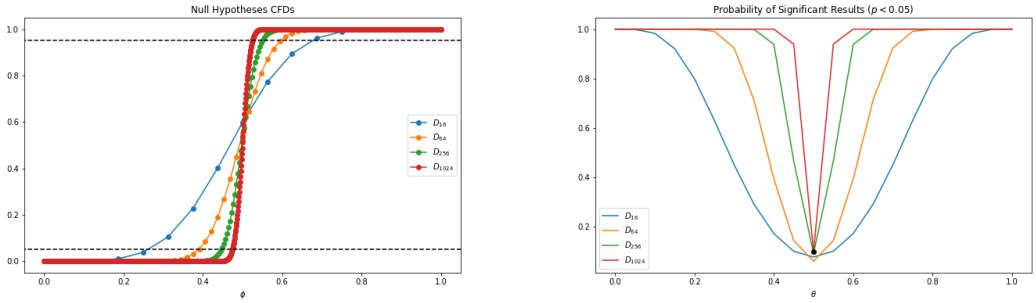

Figure 12: Left: The cumulative density functions of ϕ for diagonal stopping rules. Right: The probability of obtaining significant results under null hypothesis $\theta_0 = 1/2$.

It is because of this finding that one could claim that the "p" in "p-value" stands not for "probability", but "patience"; if we simply let our experiments run long enough, we will almost surely find significant evidence to disprove the null hypothesis.

In the falsificationist viewpoint of scientific discovery put forward by Popper, this is exactly what should happen [Goodman, 1999]. According to his philosophy, science advances only by prediction using hypotheses and falsification of those hypothesis. In principle this is a valid idea; the problem in practice is the observation made by [Gelman and Shalizi, 2013]. Except for the most trivial of examples, any model we put forward is only an approximation of reality. As such, given enough measurement precision, any model can likely be found false. The relevent question then is not whether a given model is true, but whether it is useful in approximating reality.

This issue highlights a fundamental difference in perspective between frequentist and Bayesian statistics. In the frequentist paradigm, there is some true value θ. As such, it makes sense to ask whether a null hypothesis θ_0 is correct: "does $\theta = \theta_0$"?

In the Bayesian viewpoint, however, θ is a random variable following some distribution $P^*(\theta)$. If we assume $P^*(\theta)$ is continuous over $[0, 1]$, then the question "does $\theta = \theta_0$?" amounts to the integral $\int_{\theta_0}^{\theta_0} P^*(\theta)\, d\theta = 0$ for any θ_0. While there may be cases where P^* is discontinuous, so that this integral could be non-zero, there are surely situations in which the underlying distribution P^* is in fact continuous. In these cases the question "does $\theta = \theta_0$?" becomes vacuously false. In its place, we can ask "what is the probability that θ is within $\epsilon > 0$ of θ_0?". But that is a fundamentally Bayesian question.

6 A False Objectivity

In the 18th century, Bayes solved the problem of inductive inference [Goodman, 1999]. In the following two centuries, philosophers struggled with the inherent subjectivity introduced by the prior in Bayes theorem. Recent works have tried to remedy this by making the prior objective; for example using the Maximum Entropy Principle [Jaynes, 2003] or an "ur-prior" [Meacham, 2016].

Long before that, however, philosphers and statisticians in the 20th century developed frequentist statistics to try to remove the subjectivity of the prior, and create a purely objective and deductive model of science. However, this attempt was unsuccessful. The modern notion of p-values arose from a combination of the works of Fisher with those of Neyman and Pearson, and offers the illusion of objectivity. To see the fallacy behind this idea, we need to consider the causal Markov graphs underlying the Bayesian update and p-value computations.

In the Bayesian update, we begin with an experiment S and a prior $P(\theta)$, both subjective choices of the experimenter. We then run the experiment, obtaining a measurement M. This measurement is dependent on the choice of experiment; S determines what values can be measured and the distribution governing these possibilities. For the case of Bernoulli experiments considered in this paper, this is exactly the stopping rule. However, given M, the updated posterior $P(\theta \mid M)$ is independent of S; the experimental design plays no further role in influencing the posterior. This idea was described above in Section 4.5, as the Bayesian posterior considers only a single "vertical string" in the set of all possible worlds Ω. This yields the causal Markov graph shown on the left in in Figure 13.

The issue people take with this approach is the direct connection between $P(\theta)$ and $P(\theta \mid M)$. It seems problematic that the outcome should be directly tied to something subjective. It would be preferable if all subjective aspects could be causally mediated by objective components outside of the experimenter's control, similar to the causal separation of S and $P(\theta \mid M)$ by M.

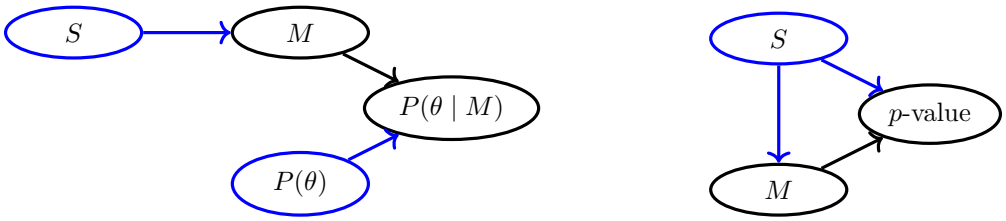

Figure 13: Left: The causal Markov graph of a Bayesian update computation. Right: The causal Markov graph of a p-value computation. Subjective elements are shown in blue.

The p-value attempts to address this issue by offering another procedure that does not depend on a subjective prior. As in the Bayesian case, the experiment S will determine the set of possible values of the observed measurement M. However, in the p-value procedure, this M does not causally separate S from the p-value. As highlighed in Section 4.5, the p-value does not depend on M, but on M *in the context of* S; as such, it is inseparable from the experimental design. The causal Markov graph for this procedure is shown on the right of Figure 13.

In order to remove the subjective influence of the prior, the p-value introduces a direct subjective influence from the experimental design. Unfortunately, this new subjective influence is often overlooked, giving the false impression of an objective procedure.

As scientists, we would love to remove all direct subjective influences on the outcome of our scientific procedure. However, in the context of scientific discovery, this is an unachievable hope. In the real world, no meaningful theory can be definitively proven from the result of an experiment; the only statements that

can be proven in this way are trivial observations and tautologies. To go beyond this requires the scientist, their reasoning, and with it their subjectivity.

In Bayesian updating, the direct subjective influence on the outcome is clear: it is specified by the prior. This can be changed and tested after the fact, to see how the resulting posterior is affected. In this way, one can get a sense of how strongly the subjective influence of a particular prior effects the final posterior distribution.

On the other hand, the subjective influence on a p-value computation cannot be readily tested. Modifying S would require redoing the experiment to obtain a new measurement M as well. Trying to change S without refreshing M as well can lead to the paradoxes described in Section 4.3; one cannot seamlessly compare p-values for the same measurement in different experiments. As such, it is difficult - if not impossible - to determine how much the subjective influence of experimental design is impacting the final p-value in a rigorous way. This unmeasurability may contribute to the false notion that there is no subjective link from S to the p-value. Unfortunately, the link is there, it simply can't be studied in a meaningful way.

Neither option can offer a completely objective method for inductive inference. However, it seems like the Bayesian approach, which is very explicit in the role of subjectivity, offers a more testable, epistemically conservative, and scientific way of handling this necessary weak point for any scientific result.

References

A. Gelman and C. R. Shalizi. Philosophy and the Practice of Bayesian Statistics. *British Journal of Mathematical and Statistical Psychology*, 2013.

S. N. Goodman. Towards Evidence-Based Medical Statistics. 1: The P Value Fallacy. *Annals of Internal Medicine*, 1999.

B. M. Hill. The Validity of the Likelihood Principle. *The American Statistician*, 1987.

E. Jaynes. *Probability Theory*, chapter 11, pages 343 – 371. Cambridge University Press, 2003.

D. Lindley and L. Phillips. Inference For a Bernoulli Process (a Bayesian View). *The American Statistician*, 1976.

C. Meacham. Ur-Priors, Conditionalizaiton, and ur-Prior Conditionalizaiton. *Ergo*, 2016.

K. Steele. Stopping Rules Matter to Bayesians Too. *Statistical Science and Philosophy of Science: Where do/should they meet in 2010 (and beyond)?*, 2010. URL https://www.phil.vt.edu/dmayo/conference_2010/archive.htm.