Contents

What did Bayes really say and what's the big deal?
Coin Problem and Billiards Problem

Michael A. Kohn, MD, MPP ©2023

1/9/2023

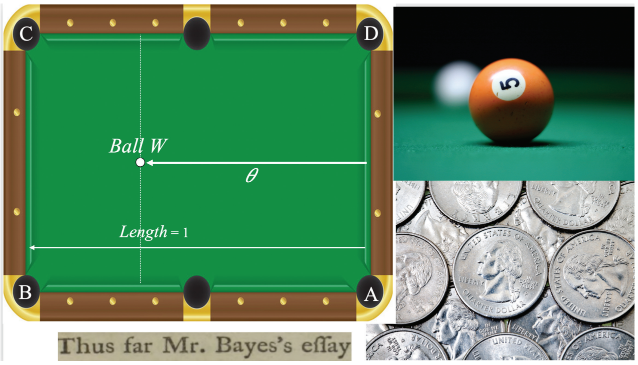


Figure 1: Bayes collage

1 Introduction

You may be confused about the debate between Frequentist and Bayesian statistics over how to use new data to judge a hypothesis. You may also know the formula that updates the probability of a hypothesis based on the likelihood of the observed data under the hypothesis. Since this formula is called Bayes's Rule, you might think Bayes wrote it; he didn't. You might think the disagreement is about the formula's validity; it isn't.

The debate is about whether we can judge a hypothesis based *only* on how likely or unlikely the observation would be if the hypothesis were true.

Frequentisits say yes. Bayesians say no, we must consider not just the likelihood of the observation given the hypothesis but also the prior probability of the hypothesis and of other explanations for the observation. Adding to the confusion is the association of Bayes with the assumption that, prior to making an observation, all potential explanations are equally likely. But this idea, called "equal priors" is *not* the issue. At its core, the disagreement is about the meaning of probability. Is it the long-run frequency of an event or the plausibility of a proposition based on background information?

I believe the best way to understand both the difference between Frequentist and Bayesian viewpoints and the difference between Bayes's Rule and what Bayes actually wrote is to present two related numerical examples of using an observation to judge a hypothesis. The first is a coin problem that was given to me during a job interview many years ago. The second is the "billiards" problem that Bayes posed in the 1763 essay, published posthumously, that made him famous. In the initial version of each problem, the probabilities of the hypothesis and its alternatives are clear. I will show how Bayesians and Frequentists both would solve the problems using the formula known, somewhat inaccurately, as Bayes's Rule. Along the way, I will summarize what Bayes actually said. Then, I will change each problem so that the probabilities of the hypothesis and its alternatives are no longer clear. This will allow me to differentiate between the Bayesian and Frequentist approaches. We will see how it ultimately comes down to whether probability is the plausibility of a proposition or the long run frequency of an event.

2 Coin Problem

Figure 2: This photo is just for a visual related to the coin problem. A cartoon would be better.

Many years ago, I was asked in a job interview to solve the following problem:

> A bag contains three coins: one fair coin, one 2-headed coin, and one 2-tailed coin. One of the three coins is selected and flipped. It shows heads. What is the probability that it is the 2-headed coin?

In this problem, the hypothesis is that the selected coin is 2-headed. The observation is that it comes up heads on a single toss. To solve the problem, we calculate one *unknown* probability from three *known* probabilities.

The *unknown* probability is the probability of having selected the 2-headed coin given that it comes up heads on a single toss:

$$P(\text{2-headed} \mid \text{heads}).$$

The three *known* probabilities are:

1) the probability that it comes up heads given that it is the 2-headed coin,

$$P(\text{heads} \mid \text{2-headed}) = 1,$$

2) the probability of selecting the 2-headed coin,

$$P(\text{2-headed}) = \frac{1}{3},$$

3) the overall probability of heads,

$$P(\text{heads}) = P(\text{heads}\,|\,\text{fair})\times\frac{1}{3} + P(\text{heads}\,|\,\text{2-headed})\times\frac{1}{3} + P(\text{heads}\,|\,\text{2-tailed})\times\frac{1}{3}$$
$$P(\text{heads}) = (\frac{1}{2}\times\frac{1}{3}) + (1\times\frac{1}{3}) + (0\times\frac{1}{3}) = \frac{1}{2}$$

The formula used for this calculation is universally known as "Bayes's Theorem" or "Bayes's Rule". (If you think I should be punctuating the possessive in some way other than "Bayes's", see Endnote #1.) I will give a general version of Bayes's Rule in the next section. For this problem, it is:

$$\begin{aligned} P(\text{2-headed}\,|\,\text{heads}) &= \frac{P(\text{heads}\,|\,\text{2-headed}) \times P(\text{2-headed})}{P(\text{heads})} \\ &= \frac{1\times\frac{1}{3}}{\frac{1}{2}} \end{aligned}$$

In the interview, I didn't simplify the answer to $\frac{2}{3}$, but the interviewer passed me to the next level anyway.

3 Bayes's Rule: History

Bayes's Rule calculates the probability of A given B from

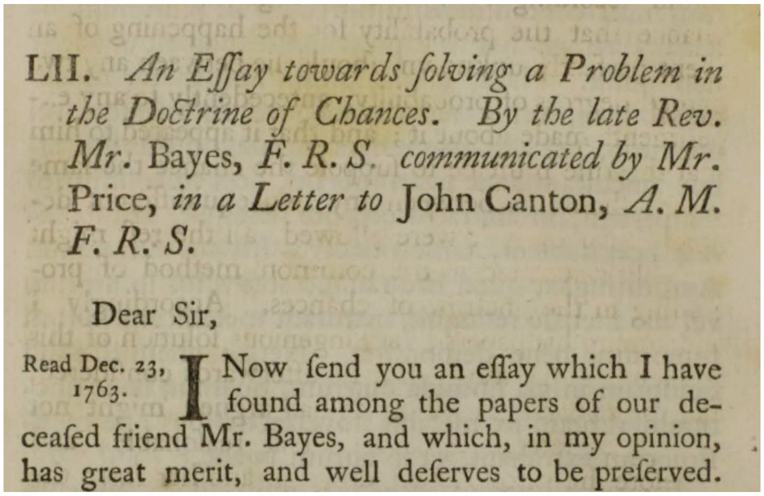
1) the probability of B given A,
2) the probability of A before observing B, and
3) the overall probability of B.

$$P(A|B) = \frac{P(B|A)\times P(A)}{P(B)}$$

But Bayes never wrote it.

Thomas Bayes (1701-1761) was a Presbyterian minister, amateur mathematician, and member of the Royal Society of London who lived in Tunbridge Wells, England. He is famous for "An Essay Towards Solving A

Problem in the Doctrine of Chances", which was published in the Royal Society's *Philosophical Transactions* on 23 December 1763, more than two and a half years after his death (Bayes 1763; Barnard 1958). His friend Richard Price (1723-1791) found the essay among Bayes's papers and sent it to the Royal Society along with an introductory letter, footnotes, an abridgement of the latter part of the essay, and an appendix containing numerical examples (Stigler 2018).



LII. *An Eſſay towards ſolving a Problem in the Doctrine of Chances. By the late Rev. Mr.* Bayes, *F. R. S. communicated by Mr.* Price, *in a Letter to* John Canton, *A. M. F. R. S.*

Dear Sir,

Read Dec. 23, 1763.

I Now ſend you an eſſay which I have found among the papers of our deceaſed friend Mr. Bayes, and which, in my opinion, has great merit, and well deſerves to be preſerved.

Figure 3: The title of the essay and first sentence of Richard Price's introductory letter. John Canton was secretary of the Royal Society of London.

The essay as originally published is 49 pages – 24 pages written by Bayes and 25 by Price. It is difficult to read today because, to us, the 18th century English seems stilted and the mathematical notation is unfamiliar.

The closest Bayes gets to stating the rule that now bears his name is Proposition 5 of 7 in an introductory section on "the general laws of chance".

> **Original text**
> If there be two subsequent events, the probability of the 2nd b/N and the probability of both together P/N, and it being first discovered that the 2nd event has happened, from hence I guess that the 1st event has also happened, the probability I am in the right is P/b.

Modern equivalent
If A and B are two events , then

$$P(A|B) = \frac{P(B\&A)}{P(B)}$$

In many textbooks, this is presented as the definition of conditional probability (Blitzstein, 2019, page 46). Again, it is just one of seven propositions preparatory to discussing the main problem, which we will get to shortly.

4 Bayes's Rule: Derivation

Whether or not Bayes wrote it, there is nothing controversial about his rule. Everybody accepts that the probability of *both A and B* is the probability of *A given B* times the probability of B:

$$P(A\&B) = P(A|B) \times P(B),$$

that the probability of *both B and A* is the probability of *B given A* times the probability of A,

$$P(B\&A) = P(B|A) \times P(A),$$

and that the probability of *A and B* equals the the probability of *B and A*:

$$P(A\&B) = P(B\&A).$$

So,

$$\begin{aligned} P(A|B) \times P(B) &= P(B|A) \times P(A) \\ P(A|B) &= \frac{P(B|A) \times P(A)}{P(B)} \end{aligned}$$

Nobody can argue with this one-step derivation. The identity was likely known before Bayes and is *not* the focus of his essay.

5 Bayes's Billiards Problem

The billiards problem *is* the focus of Bayes's essay, but it is harder than the coin problem. Both problems start with the probability of "success" in a binary trial. In the coin problem, we designate heads as a "success", and there are three discrete success probabilities: 0% for the 2-tailed coin; 50% for the fair coin; and 100% for the 2-headed coin. As we shall see, in the billiards problem, the success probability ranges continuously from 0 to 1. To move from the coin problem to the billiards problem, I introduce a variable θ equal to the probability of success on a single trial. For the 2-tailed coin, $\theta = 0$; for the fair coin, $\theta = 0.5$, and for the 2-headed coin, $\theta = 1$. Since the probability of selecting each of the three coins is $\frac{1}{3}$, before the coin toss,

$$\begin{aligned} P(\theta{=}0) &= \frac{1}{3} \quad \text{(2-tailed)} \\ P(\theta{=}0.5) &= \frac{1}{3} \quad \text{(fair)} \\ P(\theta{=}1) &= \frac{1}{3} \quad \text{(2-headed)} \end{aligned}$$

This can be confusing because θ is itself a probability which can take on three possible values, so $\frac{1}{3}$ is the probability of a *probability.*

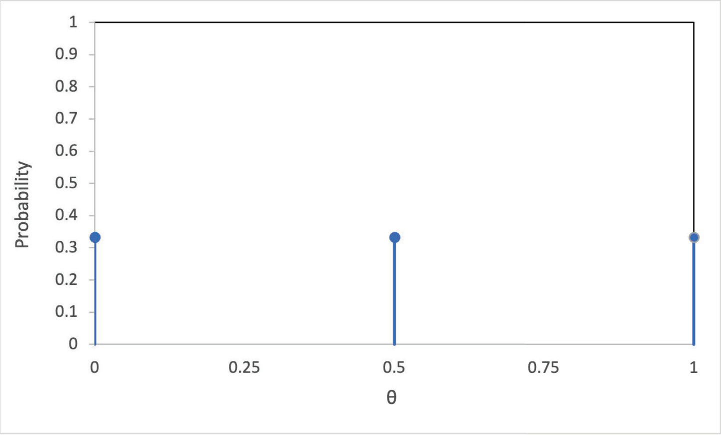


Figure 4: Prior to the coin toss, this is the discrete uniform probability distribution on the variable θ. θ is the single-toss probability of heads, so this is the probability distribution of a *probability*.

What is the distribution of θ *after* the coin has come up heads on a single toss? We already know that, after seeing heads, the probability of the 2-headed coin ($\theta = 1$) is $\frac{2}{3}$. The possibility of a 2-tailed coin has been eliminated, so $P(\theta = 0.0) = 0$. That leaves only the fair coin, so $P(\theta = 0.5) = \frac{1}{3}$.

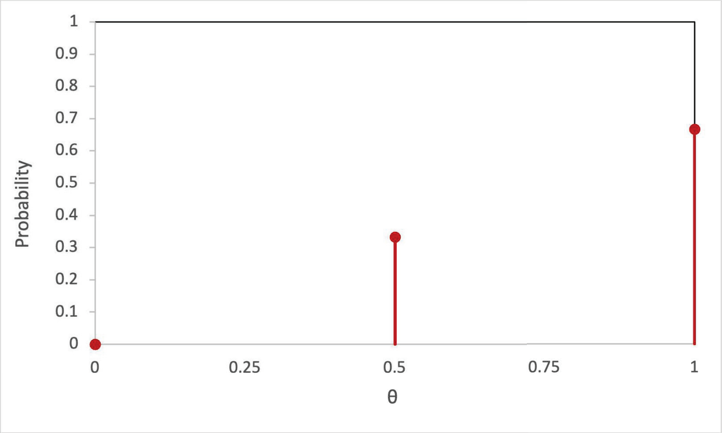


Figure 5: After seeing the coin come up heads, this is the probability distribution on the variable θ.

In the billiards problem, the success probability θ isn't limited to being 0, 0.5, or 1. Bayes assumes that θ is equally likely to take on any value between 0 and 1. It's easy to imagine selecting θ from three equally likely alternatives, but how does one imagine selecting θ from any of the possible values between 0 and 1? Bayes describes a hypothetical square table onto which someone else (besides Bayes) throws a ball labelled W from the right end. I follow many others and call the table a billiard table although Bayes never mentions billiards. Picture Bayes sitting with his back to the table because he doesn't know where ball W ends up, but it is equally likely to end up anywhere along the length of the table. To get θ (which is unknown to Bayes), divide the distance of W from the right side of the table by the length of the table.

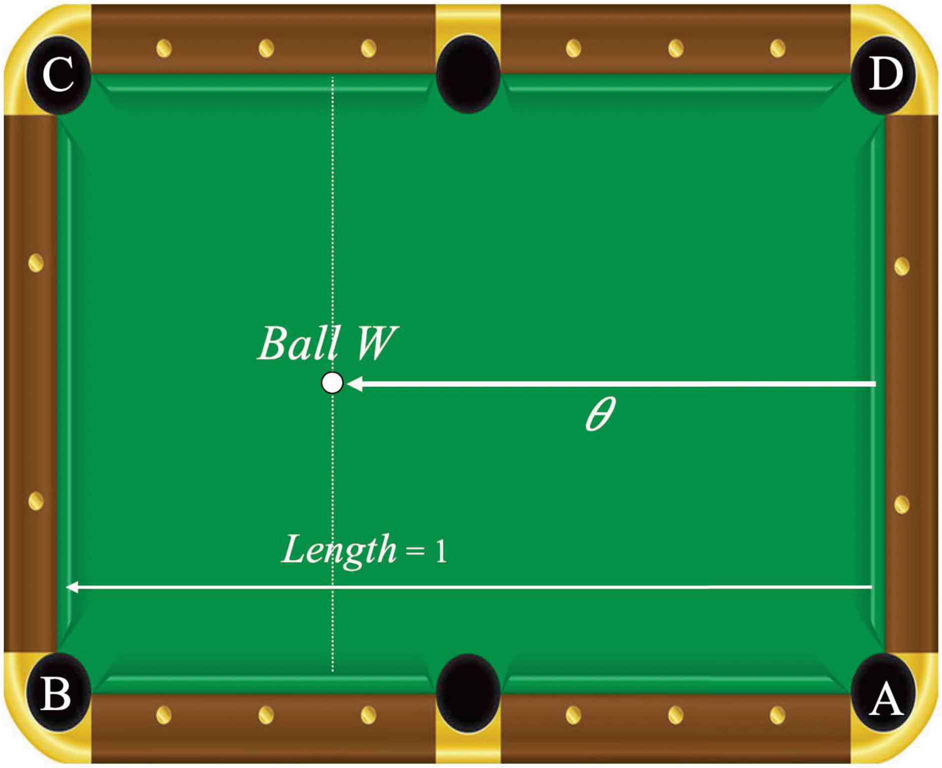


Figure 6: Bayes's "Billiard" Table: Stand on the right end AD and toss ball W onto the table and towards the left end BC. It is equally likely to end up anywhere along the length of the table. Call its unknown distance from the right end θ where $0 \leq \theta \leq 1$

Bayes's binary event is not tossing a coin for heads or tails. Instead, while he still has his back to the table, the same "someone else" tosses a second ball O onto it. The equivalent of "heads" is having O end up to the right of the first ball W. We will call this a "success". After tossing ball O, the "someone else" reports the result, success or failure. Ball O can be tossed repeatedly, but we will start with one round. Based on the result, Bayes calculates the probability that θ is in a particular range, say between 0.5 and 1, which is the probability that the first ball W made it more than halfway across the table.

Figure 7: I like to think of the balls that Bayes imagines throwing as pool balls with ball W as the (white) cue ball, and ball O as the (orange) 5-ball.

So, here is the billiards problem:

> A cue ball $W$is tossed from the right end onto a billiards table and ends up at an unknown distance θ from the right. Then an (orange) 5-ball O is tossed and ends up nearer to the right end than the cue ball W. This is arbitrarily called a "success". Given this one success, what is the probability that the cue ball W made it more than halfway across the table? In other words what is $P(0.5 < \theta < 1)$?

Again, the billiards problem is more difficult than the coin problem. We have moved from a discrete uniform probability distribution $P(\theta)$ to a continuous uniform probability *density* function $p(\theta)$.

I will use upper case $P(\theta)$ for the discrete probability distribution, also called the probability mass function (PMF), and I will use lower case $p(\theta)$ for the continuous probability density function (PDF). If you are new to PDFs, they take some getting used to. Like a probability, a probability *density* is always greater than or equal to 0, $p(\theta) \geq 0$, but it doesn't have to be less than 1. In a figure, probability is no longer represented by the height of a discrete point but by the area under the continuous PDF ($P(\theta_a < \theta < \theta_b) = \int_{\theta_a}^{\theta_b} p(\theta) d\theta$). For the PDF to be valid, the area under it over the entire range of θ must equal 1 ($\int_{-\infty}^{+\infty} p(\theta) d\theta = 1$).

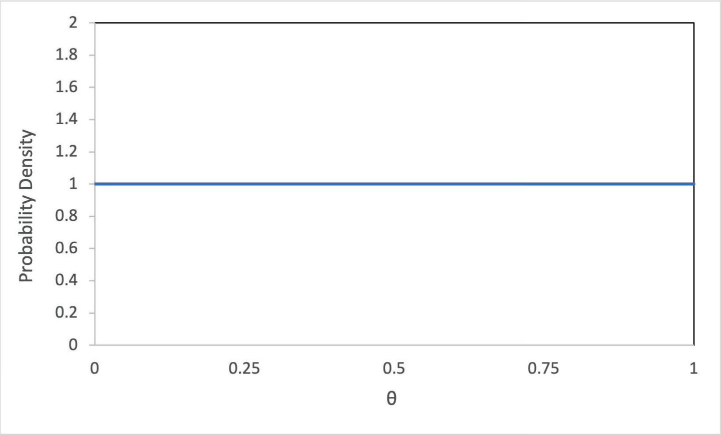


Figure 8: This is the uniform probability density function (PDF) for the variable θ prior to any trials. The probability that θ is between any two values is the area under the PDF between those two values.

6 Bayes's Rule with Different Notation, Prior and Posterior

In the billiards problem, θ can take on any value between 0 and 1, and the discrete probability distribution has been replaced by a continuous probability density function (PDF). Bayes's Rule still applies, but let's change the variables and notation. Our one-step derivation of Bayes's Rule resulted in

$$P(A|B) = \frac{P(B|A) \times P(A)}{P(B)}$$

Let θ be any quantity about which we are interested, such as the single-toss probability of a success, and replace A with θ. Let "*data*" be what we observe (or have reported to us), such as ball O ending up to the right of ball W, and replace B with *data*. Finally, to remind ourselves that we are working with continuous probability density functions, replace upper case $P(\cdot)$ with lower case $p(\cdot)$. Bayes's Rule is then

$$p(\theta|data) = \frac{p(data|\theta) \times p(\theta)}{p(data)}$$

This is how Bayes's Rule is presented in *A Student's Guide to Bayesian Statistics* (Lambert 2018).

Prior to obtaining the data, the PDF of θ is $p(\theta)$, so this is called the "prior".

$$\text{prior} = p(\theta)$$

Remember, this is a continuous PDF, not a discrete distribution.

After obtaining the data, the PDF of θ is $p(\theta|data)$, so this is called the "posterior".

$$\text{posterior} = p(\theta|data)$$

Again, this is a continuous PDF, not a discrete distribution.

Besides the prior, the other term in the numerator of the formula is $p(data|\theta)$, which is called the "likelihood".

$$\text{likelihood} = p(data|\theta)$$

The likelihood function is uncontroversial. Bayesians and Frequentists generally agree on $p(data|\theta)$, but it also takes some getting used to. The *data* is fixed but θ varies over its range of possible values (e.g., between 0 and 1). Like a probability, $p(data|\theta)$ is greater than 0, but unlike the probabilities in a probability distribution, the likelihoods in a likelihood function do not sum to 1. (See EndNote #2.)

The weighted sum of $p(data|\theta) \times p(\theta)$ over all possible θ is the denominator in the formula known as Bayes's Rule. We will just call it "the denominator".

$$\text{denominator} = p(data)$$

This represents the probability of the observed data given each hypothesis averaged over all possible hypotheses. Calculating the denominator can be challenging. Ironically, Bayes's essay does not include or focus on the formula that ended up bearing his name, but it does include and focus on the formula for the denominator in the billiards problem. For interested readers:

> The formula for the denominator in a game of Bayes's billiards with s successes and f failures in $s + f = n$ trials is
>
> $$p(data = s, f) = \frac{1}{s+f+1} = \frac{1}{n+1} \quad .$$
>
> For more on this formula, see Endnote #3.

Perhaps this formula for the denominator should be called Bayes's Rule, but it's not. Again, the formula widely known as Bayes's Rule is

$$p(\theta|data) = \frac{p(data|\theta) \times p(\theta)}{p(data)}$$

Bayes's Rule rule simply says that what we think now (the posterior) depends on what we thought before (the prior) and what we learned (the data).

7 Solving Bayes's Billiards Problem

Since ball W could end up anywhere on the billiard table, before getting the *data*, the prior is

$$p(\theta) = 1 \quad \text{for } 0 < \theta < 1$$

Remember, this is a continuous PDF, not a discrete probability.

The *data* = success×1, so the the likelihood is

$$p(data|\theta) = p(\text{success}\times 1\,|\,\theta) = \theta$$

The *data* is fixed at success×1 but θ is variable. Say that ball W ends up seven-tenths of the way across the table. Then, $\theta = 0.7$ and $P(\text{success}\times 1|\theta) = \frac{7}{10}$.

Since θ is equally likely to be any value between 0 and 1, symmetry requires that the overall probability of a success must be $\frac{1}{2}$.

$$p(data) = p(\text{success}\times 1) = \frac{1}{2}$$

(See Endnote #4.)

We have the prior, the likelihood, and the denominator, so we use Bayes's Rule to get the posterior:

$$\begin{aligned} p(\theta|data) &= \frac{p(data|\theta) \times p(\theta)}{p(data)} \\ &= \frac{\theta \times 1}{\frac{1}{2}} \\ &= 2\theta \quad \text{for } 0 < \theta < 1 \end{aligned}$$

In summary, after one success, the posterior PDF for θ is

$$p(\theta|\text{success}\times 1) = 2\theta \quad \text{for } 0 < \theta < 1$$

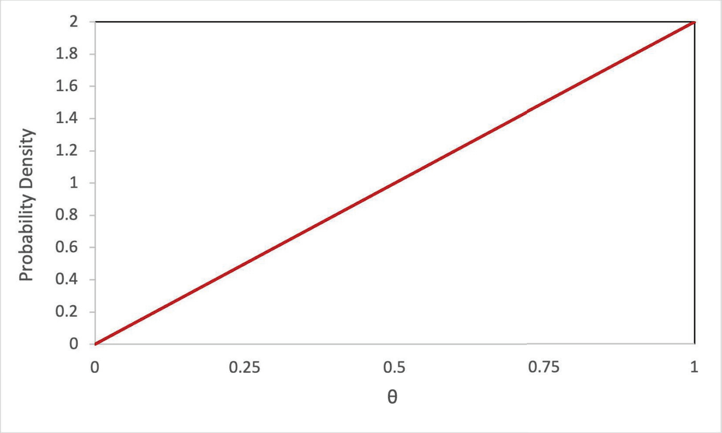


Figure 9: This is the posterior probability density function (PDF) for the variable θ after seeing one success. The probability that θ is between any two values is the area under the PDF between those two values. The area under the curve from $\theta = 0.5$ to $\theta = 1$ is $\frac{3}{4}$.

We are to find the probability, after one success, that ball W made it more than halfway across the table, that is $P(0.5 < \theta < 1.0)$. This is the area under the posterior PDF $p(\theta|data)$ between $\theta = 0.5$ and $\theta = 1$. We can see that this probability is $\frac{3}{4}$.

$$P(0.5 < \theta < 1 \,|\, \text{success} \times 1) = \frac{3}{4}$$

One way to see this is to subtract from 1 the area of the triangle from $[\theta = 0, p(\theta) = 0]$ to $[\theta = 0.5, p(\theta) = 1]$). The area of the triangle is $\frac{1}{4}$, so our answer is $1 - \frac{1}{4} = \frac{3}{4}$. For the record, I wouldn't have gotten the billiards problem right if it had been given to me in the job interview.

Pierre Simone Laplace (1774), writing 11 years after the Bayes/Price essay and apparently unaware of it, was more interested in the expected value or mean, $E(\theta \,|\, \text{heads})$, which is $\frac{2}{3}$. (Stigler, 1986a, 1986b) By the way, in his 1774 memoir, Laplace went farther in analysing the problem than Bayes and Price, but he didn't present the formula now known as Bayes's Rule either.

8 Unclear Priors

Nothing so far in the coin problem or the billiards problem would cause any debate between a Frequentist and a Bayesian. They would both use Bayes's Rule to get to the same answer. In the coin problem, after seeing one heads, the probability of a 2-headed coin is $\frac{2}{3}$. In the billiards problem, after getting one success, the probability that ball W made it more than halfway across the table is $\frac{3}{4}$. The disagreement between Frequentist and Bayesian statistics can't be about the validity of the formula known as Bayes's Rule. It isn't about about the likelihood either. Under explicit assumptions about how the data is generated, $P(data|\theta)$ is clear. In the coin problem, the probability of heads given that the coin is fair is $\frac{1}{2}$. In the billiards problem, if ball W ended up seven-tenths of the way across the table, the probability that Ball O ends up to the right of W must be $\frac{7}{10}$. Frequentists use the same likelihoods as Bayesians do. The debate focuses on the prior.

In the coin problem, the bag contains 3 coins, one fair, one 2-headed, and one 2-tailed, but what if the bag contains an unspecified number of fair, 2-headed, and 2-tailed coins? Now, after observing one flip that comes up heads, what is the probability that the coin is 2-headed?

I can think of three options:
1) Say something low, like 0.0002, because I know 2-headed coins are rare.
2) Say two-thirds, because that's what I get from a prior in which fair, 2-headed, and 2-tailed coins are equally likely.
3) Dodge the question.

Option 1 uses information that wasn't given and is sometimes referred to as using a "subjective prior". I have heard of 2-headed coins and decided that, before it was flipped, this coin had a 1-in-10,000 chance of being 2-headed.

Option 2 counts three hypotheses (fair, 2-headed, and 2-tailed) and uses a prior in which they are equally likely. I will call this the "equal priors" approach, although it has been called *the principle of insufficient reason* or the *principle of indifference* (Keynes 1921 page 44, Jaynes 2003 page 40)

Option 3 (dodging the question) is the Frequentist approach.

In the billiards problem, the probability θ is the distance of ball W from the right end of the billiard table and is thus equally likely to be any number between 0 and 1, but what if θ is an unspecified function of the distance? Now, after a success, I'm asked the probability that θ is between 0.5 and 1. (See Endnote #5.) Since I can't think of any better place to start, I would still assume a uniform prior, and say $P(0.5 < \theta < 1) = \frac{3}{4}$.

The section of Bayes's essay called the *scholium* (i.e., explanation) has been taken by many, including the brilliant, truculent Frequentist, R.A. Fisher (1922 p.324), to say that a uniform prior is appropriate when one knows nothing at all about the success probability of a binary event (Stigler, 1982). Certainly, Bayes's billiards problem starts with a uniform prior. So, the term "Bayesian" has been identified with the use of equal priors (McGrayne, 2011 page 87). Nowadays, "Bayesian" does not mean "equal priors" but refusal to ignore the prior. For modern Bayesians, the prior is not necessarily uniform, but it must be specified. As we shall see, forcing oneself to specify a prior comes with a different understanding of what probability means. Meanwhile, the Frequentist approach is to avoid the prior completely.

9 Frequentist Approach

Frequentists don't like the idea of going outside the boundaries of a problem as in Option 1 above, because that option isn't always available when the question isn't about a coin. They also don't like the equal priors approach, as in Option 2. So here is what I consider to be the Frequentist approach to the coin problem.

I know the likelihood function:

$$P(\text{heads} \,|\, \text{fair}) = P(data \,|\, \theta{=}0.5) = \frac{1}{2}$$

$$P(\text{heads} \,|\, \text{2-headed}) = P(data \,|\, \theta{=}1.0) = 1$$

$$P(\text{heads} \,|\, \text{2-tailed}) = P(data \,|\, \theta{=}0.0) = 0$$

I could hypothesize that the coin is fair. Call this hypothesis H_0. Then I could compare $P(data|H_0)$ to an arbitrary threshold for rejecting H_0. For example, we might choose a threshold of 0.05, so one rejects H_0 if $P(data|H_0) < 0.05$. Since $P(\text{heads} \,|\, \text{fair}) = 0.5$, and $0.5 > 0.05$, I can't reject the hypothesis of a fair coin.

I could also list all of the hypotheses that I cannot reject and call this list a "confidence interval". The list would include "fair" ($\theta = 0.5$) and "2-headed" ($\theta = 1$), but not "2-tailed" ($\theta = 0$). After seeing heads come up, both common sense and Bayes's Rule tell me that the probability of a 2-tailed coin is now 0. The probability that the coin is 2-headed must be at least as great as it was before I saw heads come up, and unless the prior probability of a 2-headed coin was 0, it had to increase by twice as much as the probability of a fair coin.

In the billiards problem, θ can have any value between 0 and 1. After one success, as a Frequentist, I could generate a confidence interval on θ that includes values of θ that I cannot reject. By a common criterion, I can reject $\theta < 0.05$, so I could report a confidence interval for θ of 0.05 to 1.

The coin problem asks for the probability of the 2-headed coin. I responded by rejecting the hypothesis that the coin is 2-tailed, but not rejecting the hypothesis that it was fair or that it was 2-headed. The billiards problem asks for the probability that ball W is more than halfway across the table. I responded with a 95% confidence interval from 0.05 to 1.0. Do you see why I say that the Frequentist approach is to dodge the question?

10 More Data

As long as we have a clear prior, we can use Bayes's Rule with more data. In the coin problem, perhaps the selected coin is flipped twice and comes up heads both times. Since $P(\text{2-tailed}) = 0$, we can ignore the two-tailed coin. Here are the steps:

$$
\begin{aligned}
&P(\text{2-headed}\,|\,\text{heads}\times 2) = \\
&\quad = \frac{P(\text{heads}\times 2\,|\,\text{2-headed}) \times P(\text{2-headed})}{P(\text{heads}\times 2)} \\
&\quad = \frac{P(\text{heads}\times 2|\text{2-headed}) \times P(\text{2-headed})}{P(\text{heads}\times 2|\text{fair}) \times P(\text{fair}) + P(\text{heads}\times 2|\text{2-headed}) \times P(\text{2-headed})} \\
&\quad = \frac{1 \times \frac{1}{3}}{(\frac{1}{2^2} \times \frac{1}{3}) + (1 \times \frac{1}{3})} \\
&\quad = \frac{\frac{1}{3}}{(\frac{1}{2^2} \times \frac{1}{3}) + \frac{1}{3}}
\end{aligned}
$$

Cancel the $\frac{1}{3}$ and we get

$$\frac{1}{\frac{1}{4}+1} = \frac{4}{5}.$$

What if it comes up heads three times?

$$\frac{1}{\frac{1}{8}+1} = \frac{8}{9}$$

s times?

$$\frac{1}{\frac{1}{2^s}+1} = \frac{2^s}{1+2^s}$$

In the billiards problem, θ can be any number between 0 and 1. We are asked the probability, after s successes, that ball W is more than halfway across the table, that is, the probability that θ is in the interval between 0.5 and 1. Richard Price covers this in the first part of his appendix to Bayes's essay and gives the following formula:

$$P(0.5 < \theta \leq 1\,|\,s) = \frac{2^{s+1}-1}{2^{s+1}}$$

(See Endnote #6)

So, if we have seen one success ($s = 1$), $P(0.5 < \theta < 1) = \frac{3}{4}$; for two successes ($s = 2$), it's $\frac{7}{8}$; for $s = 3$, $\frac{15}{16}$, and so on.

But what if the prior isn't clear? We are back to 1) using a subjective prior based on information that wasn't given in the problem, 2) assuming a uniform prior, or 3) dodging the question with a Frequentist approach.

11 Frequentist Approach to More Data

In the frequentist approach to the coin problem (using a significance level of 0.05), we can reject the hypothesis that the coin is fair if we see 5 heads in a row, because

$$P(\text{Heads}\times 5 \mid \text{Fair}) = \frac{1}{2^5} = \frac{1}{32} = 0.03125 < 0.05.$$

In the billiards problem, θ can be any number between 0 and 1. We need the range of hypothesized values for θ that we would reject if we see 5 successes in a row. This turns out to be $\theta < 0.55$. (See Endnote #7)

So, the Frequentist confidence interval for θ after seeing 5 successes in a row is 0.55 to 1. This is a one-sided 95% confidence interval, which is appropriate in this situation. Frequentists reject a hypothesis (such as $\theta = 0.5$) when, under this hypothesis, the probability of the observed data *or more extreme results* is less than a critical value such as 0.05. I have not had to worry about this because there is no more extreme result than all successes and no failures. Note that the confidence interval on θ (0.55 to 1) does not include $\theta = 0.50$, meaning that, after 5 successes in a row, we can reject the hypothesis that $\theta = 0.5$ at the 0.05 significance level. (See Endnote #8.)

If you saw a coin come up heads 5 times in a row, would you conclude that it was not a fair coin but rather a 2-headed coin? You probably would if you knew the coin was drawn from a bag containing 3 coins: one fair, one 2-headed, and one 2-tailed. The probability that it is a 2-headed coin is then

$$\frac{2^5}{2^5+1} = 32/33 = 0.97$$

On the other hand, if you thought the coin was taken out of normal circulation, you would likely insist on seeing more than 5 heads in a row before concluding that it was biased. Even though the observed data (5 heads) is 32 times more likely with a 2-headed coin than a fair coin, you start out with such a low prior probability of 2-headed that increasing it substantially (by approximately a factor of 32) leaves you with a posterior probability that is still low. If you specify the prior probability that the coin is fair versus 2-headed, you can calculate the posterior probability exactly using Bayes's Rule. For example, starting with a 1-in-10,000 chance of a 2-headed coin, the posterior probability is 32 in 10,000. But specifying that prior probability seems subjective and makes us uncomfortable.

The Frequentist response would be to say something like “extreme claims require extreme evidence” and, on an ad hoc basis, set the significance level to 0.001, thereby requiring 10 heads in a row (because $1/2^{10} = 1/1024 < 0.001$. For the Bayesian starting with a 1-in-10,000 chance of a 2-headed coin, 10 heads in a row would raise the chance to 10%. Whether I set the significance level to 0.001 or start with a 1-in-10,000 chance of a 2-headed coin, I have made a subjective judgement.

12 Frequentist v. Bayesian

Frequentist statistics avoids priors and focuses on the uncontroversial likelihood function. A hypothesis is rejected if the probability of the observed data or more extreme results is less than some arbitrary threshold. In Bayesian statistics, you can’t ignore the prior and must consider the probability of all competing hypotheses. This leads to a fundamental disagreement about the nature of probability. Is it an objectively measurable long-term frequency or the plausibility of a proposition based on available information?

For example, in the coin problem, if I don’t have to specify a prior, all I need is the probability that a specific coin will come up heads. This might be something I could determine by experiment, flipping the coin many times and counting how often it comes up heads. But if I have to specify the probability that the coin is of a specific type, my answer depends on my state of information. If I know that it was selected from 3 coins in a bag, then the probability of selecting a specific coin is one-third. If I don’t know where the coin came from, but I know that 2-headed coins exist, I might place the probability of a 2-headed coin at 1 in 10,000. That isn’t based on an experiment.

In the billiards problem, ball W is equally likely to stop anywhere along the length of the table, so the prior is uniform. But for a different binary event, determining the prior will be difficult. Still, the Bayesian does it, using all available information, while the Frequentist avoids it. Again, for the Bayesian, probability is a state of knowledge, while the Frequentist tries to equate it with a long-run frequency.

I probably haven't been fair to the Frequentists, because I agree with the Bayesians that probability "describes only a state of knowledge, and not anything that could be measured in a physical experiment" (Jaynes 2003 p. 44). Sometimes, in determining plausibility, it helps to do a thought experiment and imagine a long series of experiments (coin flips) or an infinite population (all the coins in circulation). But in the end, probability is still a rational assessment of a proposition's plausibility based on all available information.

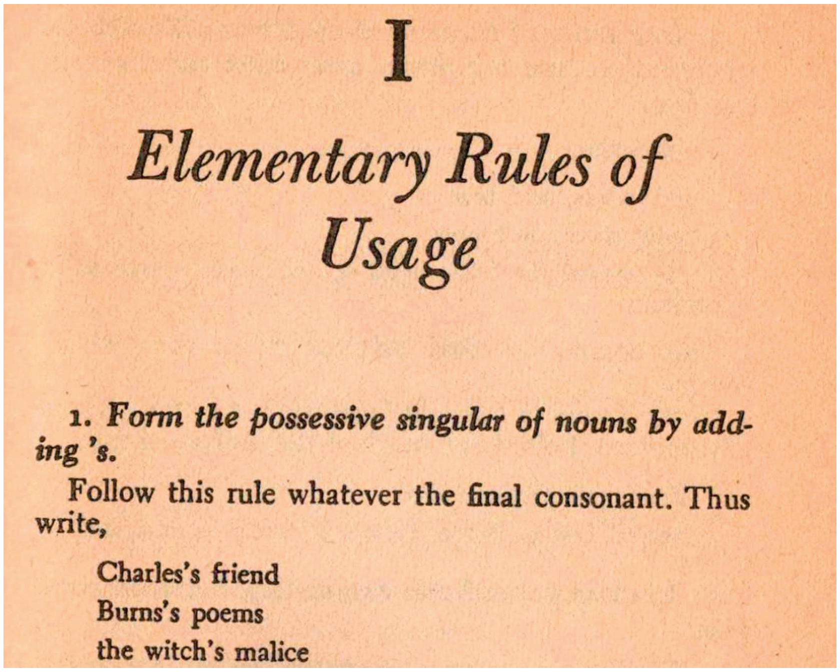
Bayesians are correct when they say that one can't evaluate a hypothesis based only on how likely the observed data would be if the hypothesis were true. Calculating this likelihood is helpful, but we still must consider the prior probability of the hypothesis and of other possible explanations for the observed data.

13 Conclusion

Bayes didn't write Bayes's Rule, but nobody disputes that it is mathematically correct and appropriate when priors are clear. When priors are unclear, experience-based, subjective priors and equal priors are problematic, but they can't be avoided. Historically, "Bayesian" was interpreted as favoring equal priors. Nowadays, "Bayesian" means refusal to ignore the prior and acceptance of probability as a numerical representation of a proposition's plausibility.

14 Endnotes

Endnote #1 About forming the possessive of singular nouns
If you think I should be forming the possessive of our author's surname in some way other than "Bayes's", read Strunk and White, *Page 1, Rule 1.*



I

Elementary Rules of Usage

1. ***Form the possessive singular of nouns by adding 's.***

Follow this rule whatever the final consonant. Thus write,

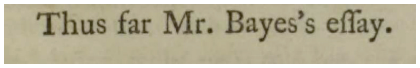
Charles's friend
Burns's poems
the witch's malice

Figure 10: Strunk W, White EB. *The elements of style.* 3d ed. New York: Macmillan; 1979. *Page 1.*

Have things changed? Not according to Benjamin Dreyer, Copy Chief at Random House, who wrote this in 2019:

> . . . you'll save yourself a lot of thinking time by not thinking about those s's and just applying them. I'd even urge you to set aside the Traditional Exceptions for Antiquity and/or being the Son of God and go with: Socrates's, Aeschylus's, Jesus's .
>
> Dreyer B. *Dreyer's English: an utterly correct guide to clarity and style.* First edition. New York: Random House; 2019. p. 39.

And what did Richard Price write in 1763, when he wanted to take over with his abridgement and end the part of the essay written by Bayes?



Thus far Mr. Bayes's eſſay.

Endnote #2 Fine Points about the Likelihood Function

The likelihood function $p(data|\theta)$ is always ≥ 0. In Bayes's billiards problem, it is also ≤ 1 because, for a given value of θ, it is a discrete binomial probability. If, for a given value of θ, the *data* is distributed according to a PDF instead of a discrete probability distribution, then $p(data|\theta)$ can be > 1. Also, in the billiards problem, θ is a continuous variable, so the likelihood function is continuous. Instead of summing the likelihoods over all possible discrete values of θ, we would integrate over θ between 0 and 1. The integral still wouldn't be 1, as would be the case if $p(data|\theta)$ were a valid PDF instead of a likelihood function. If $data =$ heads $\times 1$, then $p(data|\theta) = \theta$, which does not integrate to 1. See Endnote #4 .

Endnote #3 Formula for the denominator with *s* successes and *f* failures

Here again is the formula for the denominator $p(data)$ after seeing s successes and f failures in $n = s + f$ trials.

$$p(data = s, f) = \frac{1}{s + f + 1} = \frac{1}{n + 1}$$

This formula appears in a footnote to a section of the essay called the *scholium* (explanation), but although the formula appears in a footnote, the *idea* is central to the essay: prior to doing $s + f = n$ trials, the probability of s successes is $\frac{1}{n+1}$ regardless whether s is 0, 1, 2, ..., or n.

By definition, the denominator is $p(data|\theta)p(\theta)$ integrated over all possible θ:

$$p(data = s, f) = \int_0^1 \binom{s+f}{s} \theta^s (1-\theta)^f p(\theta) d\theta.$$

Since $p(\theta) = 1$,

$$p(data = s, f) = \int_0^1 \binom{s+f}{s} \theta^s (1-\theta)^f d\theta.$$

To show that

$$\int_0^1 \binom{s+f}{s} \theta^s (1-\theta)^f d\theta = \frac{1}{s+f+1},$$

we could integrate by parts repeatedly until we reached

$$\binom{s+f}{s} \frac{(f)(f-1)...(2)(1)}{(s+1)(s+2)...(s+f-1)(s+f)} \int_0^1 \theta^{s+f} (1-\theta)^0 d\theta,$$

which is

$$\frac{1}{s+f+1}.$$

Bayes didn't use integration by parts, but Laplace did in his 1774 memoir. (See references).

Endnote #4 The denominator $p(data)$ when $data = \text{success } \times 1$

$$\begin{aligned} p(data) &= \int_0^1 p(data|\theta) \times p(\theta)\, d\theta \\ &= \int_0^1 \theta \times 1\, d\theta \\ &= \frac{\theta^2}{2}\bigg|_{\theta=0}^{1} \\ &= \frac{1}{2} \end{aligned}$$

Or, we could use the formula for the denominator as a function of $s = 1$ success and $f = 0$ failures:

$$\begin{aligned} p(data : s = 1, f = 0) &= \frac{1}{s+f+1} \\ &= \frac{1}{1+0+1} \\ &= \frac{1}{2} \end{aligned}$$

Endnote #5 Another potential function of θ to use as a prior
R.A. Fisher (1922, p. 325), made an argument something like this. In the billiards problem, if d is the distance of ball W from the right end of the billiard table, we know $\theta = d$. But in another situation, it might be $\theta = \frac{cos^{-1}(1-2d)}{\pi}$. (See also, Stigler 1982 page 252). Then, after one success, $P(0.5 < \theta < 1) = 0.68$, not 0.75.

Endnote #6 Derivation of Price's formula

$$\begin{aligned}
P(0.5\!<\!\theta\!\leq\!1 \mid s) &= \frac{\int_{0.5}^{1} \theta^s d\theta}{\int_0^1 \theta^s d\theta} \\
&= \frac{\theta^{s+1}/(s+1)\big|_{0.5}^{1}}{\theta^{s+1}/(s+1)\big|_{0}^{1}} \\
&= 1 - 0.5^{s+1} \\
&= \frac{2^{s+1}-1}{2^{s+1}}
\end{aligned}$$

Endnote #7 Confidence interval on θ after 5 successes and no failures

$$\begin{aligned}
\theta^5 &< 0.05 \\
5\ln\theta &< \ln(0.05) \\
\ln\theta &< \frac{\ln(0.05)}{5}
\end{aligned}$$

A fun fact is that $\ln(20) = 3.00$ and $\ln(\frac{1}{20}) = \ln(0.05) = -3.00$, so

$$\begin{aligned}
\ln\theta &< \frac{-3}{5} \\
\theta &< e^{-\frac{3}{5}} \\
\theta &< 0.55
\end{aligned}$$

Endnote #8 Confidence interval on θ after s successes and no failures

If the number of successes k is substantially greater than 5, then we can approximate $e^{-\frac{3}{k}}$ with $1 - \frac{3}{k}$. For example, if $k = 10$, the lower bound of the confidence interval is approximately $1 - \frac{3}{10} = 1 - 0.30 = 0.70$. Thus, after 10 successes in a row, the confidence interval on θ is roughly 0.7 to 1. (With $k = 10$, the approximation is still rough. $e^{-3/10} = 0.74$, not 0.70.)

15 References

Bayes, T. (1763). An essay towards solving a problem in the doctrine of chances. Phil Trans R Soc. 1763 Dec 31;53:370418. *The essay as originally published. Available at https://royalsocietypublishing.org/doi/pdf/10.1098/rstl.1763.0053 But don't try to read this; read the next version instead.*

Barnard GA. (1958). Studies in the History of Probability and Statistics: IX. Thomas Bayes's Essay Towards Solving a Problem in the Doctrine of Chances: Reproduced with the permission of the Council of the Royal Society from The Philosophical Transactions (1763), 53, 370-418. Biometrika. 1958;45(34):2935.
This is better than the original essay as published because "the notation has been modernized, some of the archaisms have been removed, and what seem to be obvious printer's errors have been corrected."

Blitzstein JK, Hwang J. (2019) Introduction to probability. Second edition. Boca Raton: CRC Press; 2019.
My favorite probability textbook. Look for Joseph Blitzstein's Stat 110 lectures on YouTube. This book and the other probability textbooks on my shelf present the probability axioms, basic rules, and definitions – what Price calls "the general laws of chance" – in roughly the same order that Bayes does in Section 1 of his essay. Except for Jaynes (see below), they all use set theory notation and define conditional probability as $P(A|B) = \frac{P(A\cap B)}{P(B)}$.

Clayton, Aubrey. (2021). Bernoullis Fallacy: Statistical Illogic and the Crisis of Modern Science. New York: Columbia University Press.
The"Fallacy" in the title is that the validity of a hypothesis may be judged solely on how likely or unlikely the observed data would be if the hypothesis were true. Clayton calls it Bernoulli's Fallacy because Jacob Bernoulli's Ars Conjectandi *is devoted to determining how likely or unlikely an observation is, given that a hypothesis is true, when what we need is, not the probability of the data given the hypothesis, but the probability of the hypothesis given the data. On page 86, Clayton gives a coin example, which by coincidence is similar to mine.*

Fisher, RA. (1922) On the Mathematical Foundations of Theoretical Statistics. 1922. Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character 222 (594604): 30968. https://doi.org/10.1098/rsta.1922.0009.

I refer to Fisher as “the brilliant, truculent, Frequentist” based on many accounts, including Clayton and McGrayne.

Jaynes ET, Bretthorst GL (ed). (2003),. Probability theory: the logic of science. Cambridge, UK; New York, NY: Cambridge University Press; 727 p.
Jaynes extends reasoning about the truth of a proposition to reasoning about its plausibility. His approach is more general than the Kolmogorov system of probability with its use of set notation. Jaynes refers to propositions that are true or false; the other books (and Bayes) refer to events that either occur or fail to occur.

Keynes, J. M. (1921). A Treatise on Probability. Macmillan & Co., London.
Before his “General Theory of Employment, Interest and Money”, Keynes published this book on probability theory, which according to Jaynes (see prior reference) first presents the Principle of Indifference. Both Keynes and Jaynes would likely have problems with my use of the principle in this example.

Lambert B. (2018). A students guide to Bayesian statistics. Los Angeles: SAGE; 2018. 498 p.
Look for Ben Lambert’s Bayesian Statistics videos on YouTube. I chose to use the same notation and terms that Lambert does, including referring to $P(data)$ *as “the denominator”.*

Laplace, Pierre Simon. (1814). “A Philosophical Essay on Probabilities” Blackmore Dennett. Kindle Edition. Published 2019.
This is Laplace’s write-up of lectures that he delivered in 1795 to “the normal schools” where he had been called by the national convention as a professor of mathematics. It covers the same material as his “Analytical Theory of Probabilities”, but without the equations. This is the 1902 English translation by F.W. Truscott and F.L.Emory.

Laplace PS. (1774). Memoir on the Probability of the Causes of Events. Statist Sci [Internet]. 1986 1(3). Available from: https://projecteuclid.org/journals/statistical-science/volume-1/issue-3/Memoir-on-the-Probability-of-the-Causes-of-Events/10.1214/ss/1177013621.full
This is Stephen Stigler’s translation of Laplace’s first major article on

mathematical statistics. It appeared in 1774, when Laplace was 25 years old and before he was aware of Bayes's essay. Section VI presents a coin problem.

McGrayne SB. (2011). The theory that would not die: how Bayes rule cracked the enigma code, hunted down Russian submarines, & emerged triumphant from two centuries of controversy. New Haven: Yale university press; 2011.
This is an entertaining book with fun historical anecdotes. McGrayne knows better, but in some places in the book, one gets the impression that Frequentists dispute the correctness of the formula commonly called Bayes's Rule in the same way that mid-twentieth century geologists (including prominent Bayesian Harold Jeffreys) disputed the existence of continental drift. Of course, Frequentists can't and don't dispute the correctness of the formula itself. They question the use of prior distributions, whether uniform or not.

Stigler, Stephen M. 1982. Thomas Bayess Bayesian Inference. Journal of the Royal Statistical Society. Series A (General) 145 (2): 250. https://doi.org/10.2307/2981538.
A concise summary of Bayes's essay, but not Price's appendix. Stigler differs from some others on the nature of Bayes's key argument. In n trials of an unknown binary event, the number of successes has a discrete uniform distribution. The focus is on the distribution of the discrete number of successes between 0 and n, not on the distribution of the continuous success probability θ.

Stigler SM. (1986a). Laplace's 1774 Memoir on Inverse Probability. Statist Sci [Internet]. 1986 Aug 1. Available from: https://projecteuclid.org/journals/statistical-science/volume-1/issue-3/Laplaces-1774-Memoir-on-Inverse-Probability/10.1214/ss/1177013620.full
Stigler's introduction to his translation of Laplace's 1774 Memoir on the Probability of the Causes of Events.

Stigler SM. (1986b). The history of statistics: the measurement of uncertainty before 1900. Cambridge, Mass: Belknap Press of Harvard University Press; 1986.
See Chapter 3: Inverse Probability. As mentioned above, Stigler differs from some others on the nature of Bayes's key insight.

Stigler SM. (2018). Richard Price, the first Bayesian. Statistical Science. 2018 Feb;33(1):117-25.
An interesting article about Richard Price and especially his appendix to Bayes's essay.