Contents

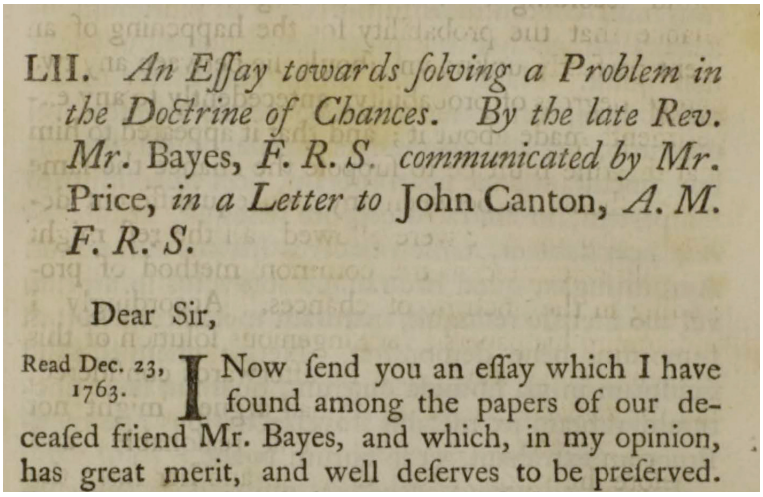
What did Bayes really say?

Michael A. Kohn, MD, MPP ©2022

6/16/2022

1 Introduction

The Reverend Thomas Bayes (1701-1761) is famous for "An Essay Towards Solving A Problem in the Doctrine of Chances", which was published in the Royal Society of London's *Philosophical Transactions* on 23 December 1763, more than two and a half years after Bayes's death. (If you think I should be forming the possessive of Bayes in some way other than "Bayes's", see Endnote #1.) His friend Richard Price 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.



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 1.1: The title of the essay and first sentence of Richard Price's introductory letter. John Canton was secretary of the Royal Society of London. F.R.S means Fellow of the Royal Society.

The essay as originally published is 49 pages – 24 pages written by Bayes and 25 by Price: introductory letter (6 pages), abridged conclusion (4 pages), and appendix (15 pages). It is difficult to read today because, to us, the 18th century English seems stilted and the mathematical notation is unfamiliar. So, I have tried to "translate" it into modern language and mathematical notation.

The essay is *not* focused on what we now call Bayes's Rule:

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

It is about the more specific problem of how to draw inferences about the probability of a binary event by observing the number of times it does and doesn't happen.

Bayes starts with a statement of the problem. Next, he provides 7 definitions, including the definition of "inconsistent" (disjoint) events, "contrary" (complementary) events, and independent events. In Definition 5, he defines the probability of an event as the ratio of its expected value to the value realized if the event occurs.

After the definitions, Bayes presents a set of 6 propositions, including what most modern textbooks (following Kolmogorov) present as axioms, as well as the complement rule, the multiplication rule, and the product rule for independent events, but he does not explicitly present the rule that now bears his name. In *Prop. 7*, he proceeds to the binomial distribution.

This introductory material might be considered the first textbook coverage of the definitions, axioms, and basic rules of probability theory. Richard Price's introductory letter says, "Mr Bayes has thought fit to begin his work with a brief demonstration of the general laws of chance. His reason for doing this... was not merely that his reader might not have the trouble of searching elsewhere for the principles on which he has argued, but because he did not know whither to refer him for a clear demonstration of them." De Moivre's "Doctrine of Chances" (1718, 1738, and 1756) is sometimes called the first probability textbook, but if Bayes thought it provided a clear demonstration of the "general laws of chance", he would have known "whither to refer" the reader.

One minor problem is that the notation $\binom{n}{k}$ for "n choose k" did not exist,

so Bayes relies on the well-known binomial expansion of $(a+b)^n$,

$$a^n + na^{n-1}b + \frac{n(n-1)}{2}a^{n-2}b^2 + \frac{n(n-1)(n-2)}{3\cdot 2}a^{n-3}b^3 + ... + nab^{n-1} + b^n,$$

which we now write as

$$\binom{n}{n}a^n b^0 + \binom{n}{n-1}a^{n-1}b^1 + \binom{n}{n-2}a^{n-2}b^2 + \binom{n}{n-3}a^{n-3}b^3 + ... + \binom{n}{1}a^1 b^{n-1} + \binom{n}{0}a^0 b^n$$

or

$$\sum_{k=0}^{n}\binom{n}{k}a^k b^{n-k}.$$

Bayes refers to $\binom{n}{k}$ as "the coefficient of the term in which occurs $a^k b^{n-k}$ when the binomial $(a+b)^n$ is expanded". In this quoted phrase, I have already substituted k for p and $n-k$ for q, because of another potential source of confusion for modern readers....

Today, we often present the binomial distribution this way:

> If k is the number of successes in n binary trials with success probability p and failure probability $q = 1-p$, then the probability mass function (PMF) of k is given as follows:
>
> $$BinomPMF(k;n,p) = P(k;n,p) = \binom{n}{k}p^k q^{n-k} \quad k = 0,1,2,...,n$$

Unfortunately for those of us accustomed to p as the probablity of a success and $q = 1-p$ as the probability of a failure, Bayes used p as the *number* of successes, where we now often use k, and q as the *number* of failures, where we now often use $n-k$. When I translate Bayes's original text, I use θ for the probability of a success and, when I need it, $\gamma = 1-\theta$ for the probability of a failure. I still use k for the the number of successes and, when I need it, $j = n-k$ for the number of failures. The binomial distribution completes Section 1 of the essay.

In the second section, Bayes describes a hypothetical square table onto which he imagines throwing first ball W and then ball O repeatedly. I will follow many others and call his table a billiards table, although Bayes never mentions billiards. He measures the distance of ball W from the right side of his table, so I will assume that balls are thrown onto the billiards table from the right end, not the left end.

Yet another potential source of confusion is that the horizontal axis in Bayes's figures goes from 0 on the *right* to 1 on the *left*, which is the reverse of the way we usually do it now. For example, he did a free hand drawing of the function $x^k(1-x)^j$ for x from 0 to 1. Here is how it would look with with $k = 4$ and $j = 6$.

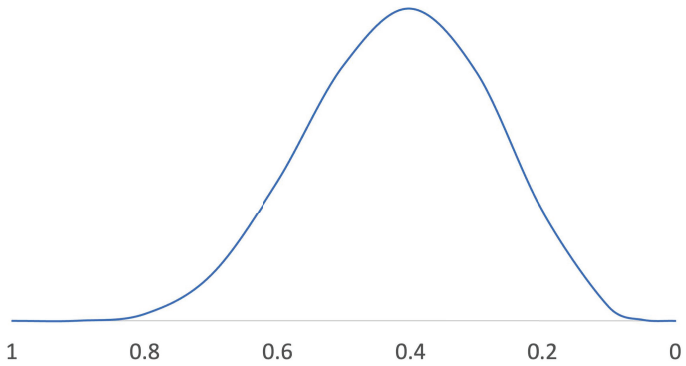


Figure 1.2: In Bayes's figures, the horizontal axis goes from 0 on the *right* to 1 on the *left*. This is $x^4(1-x)^6$. The maximum is at $x = 0.4$, which is to the *right* of midline.

Bayes's key insight was that the area under the curve $x^k(1-x)^j$ for x from 0 to 1 is

$$\frac{1}{\binom{k+j}{k}(k+j+1)}.$$

For example, the area under the curve with $k = 4$ and $j = 6$ in Figure 1.2 is

$$\frac{1}{\binom{10}{4}(11)} = \frac{1}{(210 \times 11)} = \frac{1}{2310} = 0.0004329.$$

As we shall see, Bayes first gives what Harvard statistics professor Joseph Blitzstein calls a"story proof" (Blitzstein page 382) and then a derivation using algebra and calculus, which Bayes calls "fluxions".

Towards the end of the second section, Richard Price wrote, "Thus far Mr. Bayes's essay." The rest of the section is Price's abridgement of what Bayes wrote. Price also added an appendix with numerical examples.

Read on to find out what Thomas Bayes really said and what Richard Price added. I present the essay in small sections of the original text, followed by my translation into the modern equivalent. I also intersperse explanatory comments. If a comment seemed too long, I moved it to an endnote. I finish with annotated references.

2 Problem and Definitions

Problem

Comment
Bayes's statement of the problem is clear: What can we infer about the probability of a binary event by observing the number of times it does and doesn't happen?

Original text
Given the number of times in which an unknown event has happened and failed: *Required* the chance that the probability of its happening in a single trial lies somewhere between any two degrees of probability that can be named.

Modern equivalent
Given: that n independent binary trials with unknown probability of success θ result in k successes and $n - k$ failures.
Required: the probability that θ lies in the interval between θ_1 and θ_2.

Comment
Bayes assumes that the prior distribution of θ is the uniform distribution between 0 and 1; $\theta \sim Unif(0, 1)$.

Definitions

Comment
After presenting the problem, Bayes starts off with 7 definitions.

Today's textbooks, except for Jaynes (see references), use the notation of set theory. $A \cap B$ means that events A and B both occur; $A \cup B$ means that at least one of A and/or B occurs. They introduce a sample space S of

all possible events and a probability function P that takes an event $A \subseteq S$ as input and returns $P(A)$, a real number between 0 and 1, as output.

Bayes preceded the use of set notation for probability definitions by more than 150 years. As we will see in his Definition 5, he defines the probability of an event as the ratio of its expected value to the value realized if it occurs. But he covers the same points about probability as do today's textbooks.

Original text
Definition 1. Several events are inconsistent when, if one of them happens, none of the rest can.

Modern Equivalent
We now use "disjoint" instead of "inconsistent". Saying that events are *disjoint* means that they are mutually exclusive.

Original text
2. Two events are contrary when one, or other of them must; and both together cannot happen.

Modern Equivalent
We now use "complementary" instead of "contrary". The *complement* of event A is $Not(A)$, which I will denote A^c.

Original text
3. An event is said to fail, when it cannot happen; or, which comes to the same thing, when its contrary has happened.
4. An event is said to be determined when it has either happened or failed.

Comment
Definitions 3 and 4 do not require translation or elaboration.

Original text
5. The probability of any event is the ratio between the value at which an expectation depending on the happening of the event ought to be computed, and the value of the thing expected upon its happening.

Modern Equivalent
The probability of an event is the ratio of its expected value to the value

realized if the event occurs.

Comment
Later, Bayes will talk about receiving N if event A occurs. Definition 5 says that, if the expected value of event A is $E(A)$, then $P(A) = E(A)/N$. He will denote $E(A)$ with $\mathbf{a}$ and therefore $P(A) = \mathbf{a}/N$. While Bayes defines probability as the ratio of expected value to amount received, essentially all others define expected value as probability times amount: $E(A) = P(A) \times N$.

When Bayes talks about receiving a value N, he means utility, not monetary value. (See Endnote #2.) Awkwardly, he assumes all events result in receiving N. If events A, B, and C all result in N and have expected values $\mathbf{a}$, $\mathbf{b}$, and $\mathbf{c}$, respectively, then $P(A) = \mathbf{a}/N$, $P(B) = \mathbf{b}/N$, and $P(C) = \mathbf{c}/N$.

Bayes's definition is related to indicator random variables and what Blitzstein (p. 164) calls the "fundamental bridge" between probability and expectation. If indicator random variable $I_A = 1$ when event A occurs and $I_A = 0$ when A fails to occur, then $E(I_A) = P(A)$. For Bayes, this isn't a bridge between probability and expectation, it's the definition of probability. You can see this, without loss of generality, by setting his $N = 1$. In short, Definition **5** *defines* the probability of an event as the expected value of its indicator variable.

Original text
6. By chance I mean the same as probability.
7. Events are independent when the happening of any one of them does neither increase nor abate the probability of the rest.

Comment
Definitions 6 and 7 do not require translation or elaboration. On to the propositions...

3 Propositions 1 - 7

Comment
It is interesting to note that Bayes covers the now standard probability axioms either implicitly or explicitly.

Prop. 1 [Axiom 1: Sum Rule for Disjoint Events]

The first part of Bayes's first proposition, which I will call the "sum rule for disjoint events", is that the probability of one or the other of several mutually exclusive events is the sum of their individual probabilities. He introduces 3 disjoint ("inconsistent") events A (1st event), B (2nd event), and C (3rd event), each of which results in receiving value N, and with expected values **a**, **b**, and **c**, respectively. Per Definition **5**, he defines probability as the ratio of expected value to amount received, $P(A) = \mathbf{a}/N$, $P(B) = \mathbf{b}/N$, and $P(C) = \mathbf{c}/N$.

Original text
When several events are inconsistent the probability of the happening of one or other of them is the sum of the probabilities of each of them.

Suppose there be three such [inconsistent] events, and whichever of them happens I am to receive N, and that the probability of the 1st, 2nd, and 3rd are respectively **a**/N, **b**/N, **c**/N. Then (by the definition of probability) the value of my expectation from the 1st will be **a**, from the 2nd **b**, and from the 3rd **c**. Wherefore the value of my expectations from all three will be **a** + **b** + **c**. But the sum of my expectations from all three is in this case an expectation of receiving N upon the happening of one or other of them. Wherefore (by definition 5) the probability of one or other of them is (**a** + **b** + **c**)/N or **a**/N + **b**/N + **c**/N. The sum of the probabilities of each of them.

Modern equivalent
If A_1, A_2, ..., A_n are disjoint events, then

$$P\left(\bigcup_{j=1}^{n} A_j\right) = \sum_{j=1}^{n} P(A_j)$$

Saying that these events are *disjoint* means that they are mutually *exclusive*: $A_i \cap A_j = \emptyset$. Bayes used "inconsistent" instead of "disjoint".

Comment
I call this the sum rule for disjoint events to distinguish it from the complement rule (below), which at least Jaynes (p. 33) refers to as "the sum rule". Although conventional expositions present the sum rule for

disjoint events as an axiom, Jaynes (p. 38) deduces it from "simple qualitative conditions of consistency".

Prop. 1 (continued) [Axiom 2: P(S) = 1]

The next part of *Prop. 1* is that at least one of all possible events must occur, and therefore the union of all possible events has probability 1.

Original text
Corollary. If it be certain that one or other of the three events must happen, then $\mathbf{a} + \mathbf{b} + \mathbf{c} = \mathrm{N}$. For in this case all the expectations together amounting to a certain expectation of receiving N, their values together must be equal to N.

Modern equivalent
Since S includes all possible events,

$$P(S) = 1$$

Comment
It is awkward that Bayes presents this using expectations instead of probabilities, but as noted, $P(A) = \mathbf{a}/N$, $P(B) = \mathbf{b}/N$, and $P(C) = \mathbf{c}/N$, so $P(A) + P(B) + P(C) = 1$. This proposition identifies certainty with a probability of 1.

Implicit in the above axioms is that the probability of the empty set $\emptyset$ is 0.

$$P(\emptyset) = 0$$

Prop. 2 [Complement Rule]

Comment
Again, Bayes's exposition parallels what we now see in probability textbooks, which introduce the complement rule right after the axioms and then move to the definition of conditional probability and the multiplication rule.

Original text
And from hence it is plain that the probability of an event added to the probability of its failure (or of its contrary) is the ratio of equality. If a

person has an expectation depending on the happening of an event, the probability of the event is to the probability of its failure as his loss if it fails to his gain if it happens.

Modern equivalent
If A and A^c are complementary events (i.e., $A^c = Not(A)$), then

$$P(A^c) = 1 - P(A)$$

***Prop.* 3 [Multiplication Rule]**

Original text
The probability that two subsequent events will both happen is a ratio compounded of the probability of the 1st, and the probability of the 2nd on supposition the 1st happens.

Modern equivalent

$$P(A \cap B) = P(A)P(B|A)$$

Comment
This is the multiplication rule. Bayes presents this first and then the definition of conditional probability as a corollary. Most modern textbooks (again, except for Jaynes) present them the other way around.

Original text
COROLLARY. Hence if of two subsequent events the probability of the 1st be a/N, and the probability of both together be P/N, then the probability of the 2nd on supposition the 1st happens is P/a.

Modern equivalent
This is the definition of conditional probability. If A and B are any two events in the sample space S and $P(A) \neq 0$, then

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

Comment
Bayes presents a temporal sequence with A being "determined" (occurring or failing to occur) *before* B. This is important in *Props* 4 and 5.

***Prop.* 4 [One that doesn't fit in]**

Comment
This cryptic proposition with its even more cryptic footnote covers 2 pages of the published essay. It does *not* have a modern textbook equivalent. Here is the first sentence:

Original text
If there be two subsequent events to be determined every day, and each day the probability of the 2nd is b/N and the probability of both P/N, and I am to receive N if both the events happen the first day on which the 2nd does; I say, according to these conditions, the probability of my obtaining N is P/b.

Modern equivalent
Assume A and B are two events that can occur daily. I am to receive N if, on the first day that B occurs, A also occurs. Let W be the event of receiving N, i.e., the event of A occurring on the first day that B occurs. $P(W) = P(A \cap B)/P(B)$

Comment
Here is one explanation of *Prop.* 4. (Endnote #3 gives an alternative explanation.) Let $E(W)$ be the expected value of this situation, which I refer to as a "game". On each day, there are four possible outcomes: $A \cap B$, $A^c \cap B$, $A \cap B^c$, $A^c \cap B^c$. On Day 1, if both A and B occur ($A \cap B$), I receive N and the game is over. If B occurs but A doesn't ($A^c \cap B$), I receive 0 and the game is over. If B doesn't occur ($A \cap B^c$ or $A^c \cap B^c$), then I'm back to where I started. Bayes refers to this as "being reinstated in my former circumstances". This translates into the following equation for the expected value $E(W)$:

$$E(W) = P(A \cap B) \times N + P(A^c \cap B) \times 0 + (1 - P(B)) \times E(W)$$

Let $P(B) = \mathbf{b}/N$ and $P(A \cap B) = \mathbf{P}/N$.

$$\begin{aligned} E(W) &= \mathbf{P}/N \times N + 0 + (1 - \mathbf{b}/N)E(W) \\ E(W) - (1 - \mathbf{b}/N)E(W) &= \mathbf{P} \\ \frac{\mathbf{b}E(W)}{N} &= \mathbf{P} \\ E(W) &= \frac{\mathbf{P}N}{b} \end{aligned}$$

Bayes defines the probability of receiving N as the ratio of $E(W)$ to N, so if $P(W)$ is the probability of receiving N,

$$P(W) = \frac{\mathbf{P}}{b}$$

***Prop.* 5 [Same as *Prop* 3 switching A and B]**

Comment
Recall that Bayes thought of event A as being "determined" (occurring or failing to occur) before event B. By repeating *Prop 3* and switching A and B, he gives us the probability that the earlier event A occurred based only on knowing whether the later event B occurred. This is one step away from what we now call Bayes's Rule.

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 any two events in the sample space S and $P(B) \neq 0$, then

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

Comment
From *Prop* 3's multiplication rule, we know that $P(B \cap A) = P(B|A)P(A)$, so

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

Although this is what we think of as Bayes's Rule, *Prop* 5 is as close as he comes to saying it. It is just one of 7 propositions in this introductory section on "the general laws of chance" before addressing the problem stated at the beginning: inferring the probability of a binary event by observing the number of times it happened and failed to happen.

Since, $P(B) = P(B|A)P(A) + P(B|A^c)P(A^c)$, we could also write

$$P(A|B) = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|A^c)P(A^c)}.$$

I prefer the odds form of Bayes's rule, which is derived as follows:

$$P(A \cap B) = P(B \cap A)$$
$$P(A|B)P(B) = P(B|A)P(A)$$

Similarly,

$$P(A^c \cap B) = P(B \cap A^c)$$
$$P(A^c|B)P(B) = P(B|A^c)P(A^c)$$

Dividing,

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

Some terminology:

$$\frac{P(A)}{P(A^c)} = Odds(A) = \text{prior odds}$$
$$\frac{P(A|B)}{P(A^c|B)} = Odds(A|B) = \text{posterior odds}$$
$$\frac{P(B|A)}{P(B|A^c)} = LR_A(B) = \text{likelihood ratio for } A \text{ of } B$$

So,

$$\frac{P(A|B)}{P(A^c|B)} = \frac{P(B|A)}{P(B|A^c)}\frac{P(A)}{P(A^c)}$$
$$Odds(A|B) = LR_A(B) \times Odds(A)$$
$$\text{posterior odds} = \text{likelihood ratio} \times \text{prior odds}$$

Comment

According to Dale (see references), John Maynard Keynes presented this as Bayes's Rule in his *Treatise on Probability* (1921). It nicely displays "what we think now" (posterior odds) as the product of "what we thought before" (prior odds) and "what we learned" (likelihood ratio). (For more on the odds form of Bayes's Rule, see Endnote #4.)

***Prop.* 6 [Product Rule for Independent Events]**

Comment
Going from axioms to the complement rule to the multiplication rule to Bayes's Rule to the product rule for independent events (below) is the way we present it today. Bayes differs by including *Prop.* 4 and *by never stating the rule that bears his name.*

Original text
The probability that several independent events shall all happen is a ratio compounded of the probabilities of each.

Modern equivalent
If events A and B are independent,

$$P(A \cap B) = P(A)P(B)$$

Original text
If there be several independent events, and the probability of each one be a, and that of its failing be b, the probability that the 1st happens and the 2nd fails, and the 3rd fails and the 4th happens, etc. will be *abba*, etc. For, according to the algebraic way of notation, if a denote any ratio and b another, *abba* denotes the ratio compounded of the ratios a, b, b, a. This corollary therefore is only a particular case of the foregoing.

Modern equivalent
In a sequence of independent binary trials with success probability p and failure probability $q = 1 - p$, the probability of a sequence of successes and failures is given by multiplying all the individual probabilities. If the 1st succeeds, the 2nd fails, the 3rd fails, and the 4th succeeds, the probability of the sequence will be *pqqp*. This is a specific example of the general definition of independence.

Comment
By introducing an independent event that can either occur or fail with a

given probability, Bayes has covered what we now call the Bernoulli distribution, so it is natural for him to proceed to the binomial distribution.

Prop. 7 [Binomial Distribution]

Original text
If the probability of an event be a, and that of its failure be b in each single trial, the probability of its happening p times, and failing q times in $p+q$ trials is Ea^pb^q if E be the coefficient of the term in which occurs a^pb^q when the binomial $(a+b)^{p+q}$ is expanded.

Comment
Here, Bayes uses a for the probability of success and $b = 1 - a$ for the probability of failure. Later, he will use x and $r = 1 - x$. Today, we commonly use p and $q = 1 - p$, but as I mentioned in the introduction, Bayes uses p for the *number* of successes and q for the *number* of failures. Instead of Bayes's a and b, his later x and r, or our common p and q, I will use θ and $\gamma = 1 - \theta$ for the probabilities of success and failure, and I will use k and $n - k$ for the number of successes and failures. Also as mentioned in the introduction, I use $\binom{n}{k}$ as "the coefficient of the term in which occurs a^kb^{n-k} when the binomial $(a+b)^n$ is expanded."

Modern equivalent
In a sequence of n independent binary trials with success probability θ and failure probability $\gamma = 1 - \theta$, the probability of k successes and $n - k$ failures is given as follows:

$$P(k; n, \theta) = \binom{n}{k} \theta^k \gamma^{n-k}$$

Comment
This is the probability mass function (PMF) of the binomial distribution.

$$BinomPMF(k; n, \theta) = \binom{n}{k} \theta^k (1-\theta)^{n-k}$$

Bayes does not discuss the cumulative distribution function (CDF) for the

binomial distribution, but it's just the sum of all the probabilities up to k.

$$P(K \leq k; n, \theta) = \sum_{i=0}^{k} \binom{n}{i} \theta^i (1-\theta)^{n-i}$$

$$BinomCDF(k; n, \theta) = \sum_{i=0}^{k} \binom{n}{i} \theta^i (1-\theta)^{n-i}$$

We will need this further on.
The first section of the essay ends with the binomial distribution. The language seems stilted to modern readers. However, except for *Prop* 4, this first section is reasonably clear as what Price calls "a brief demonstration of the general laws of chance". According to Price, this "brief demonstration" may not have been available elsewhere, although clearly the material was already known and not being presented for the first time. The second section starts with the "billiards" table.

4 Bayes's "Billiards"

Original text
POSTULATE. 1. I suppose the square table or plane $ABCD$ to be so made and levelled, that if either of the balls O or W be thrown upon it, there shall be the same probability that it rests upon any one equal part of the plane as another, and that it must necessarily rest somewhere upon it.

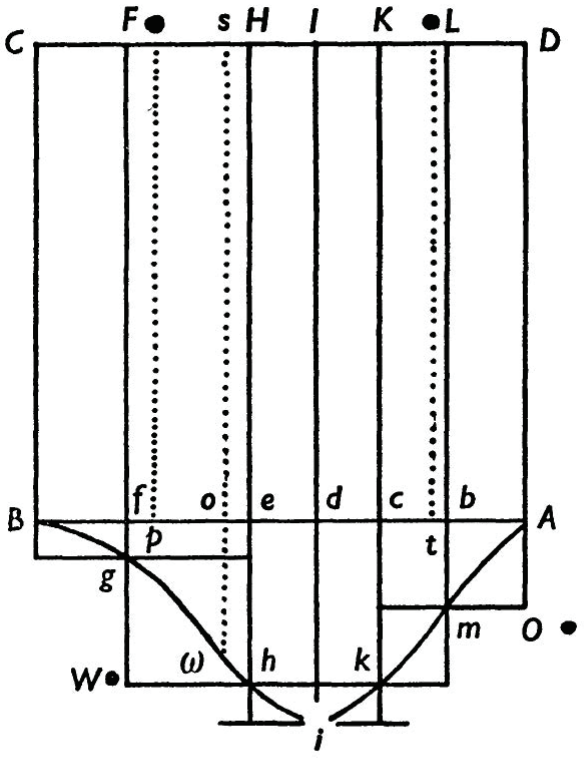


Figure 4.1: I suppose that the ball W shall be first thrown, and through the point where it rests a line os shall be drawn parallel to AD, and meeting CD and AB in s and o; and that afterwards the ball O shall be thrown $p+q$ or n times, and that its resting between AD and os after a single throw be called the happening of the event M in a single trial.

[skipping down] LEM. 2. The ball W having been thrown, and the line os drawn, the probability of the event M in a single trial is the ratio of Ao to AB.

Comment
It's easier to imagine a rectangular billiards table $ABCD$. Let the length of the table $AB = CD = 1$.

Modern equivalent

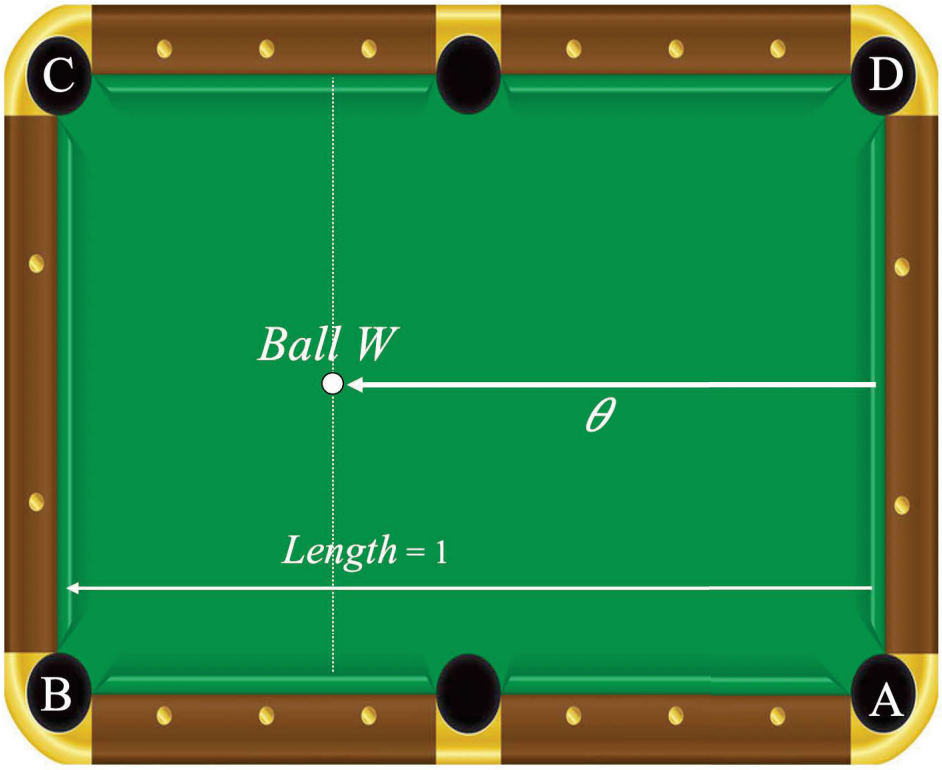


Figure 4.2: Bayes's "Billiard" Table: Stand on the right end AD and toss ball W onto the table and towards the left end BC. It is equally likely to end up anywhere along the length of the table. Call its unknown distance from the right end θ where $0 \leq \theta \leq 1$

Now suppose O is thrown. If it ends up nearer than W did, we will call that a success (Bayes's "event M"). If it ends up farther away, we will call that a failure. The probability of a success is θ, which is the unknown distance of W from the near end. Throw O a total of n times and count the number of successes k.

Figure 4.3: 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.

5 Propositions 8 - 10

***Prop.* 8 [Joint Probability Density]**

Comment
This is where Bayes calculated the joint probability density of a value θ and k successes in n trials. In modern notation, it is $P(\theta, k; n)$. This is a hybrid probability density function because k is a discrete random variable with possible values $0, 1, ..., n$, and θ is a continuous random variable between 0 and 1. By saying above that "there shall be the same probability that [the ball] rests upon any one equal part of the plane as another", Bayes assumed a uniform prior on θ, so $\theta \sim Unif(0, 1)$, and the prior probability density function $f(\theta) = 1$ for $0 \leq \theta \leq 1$.

In the following, Bayes uses y for this joint probability density $P(\theta, k; n)$. He also uses x and $r = 1 - x$ for what we have been calling θ and $1 - \theta$ and uses (sorry) p and q for what we have been calling k and $n - k$.

Original text
[With] E being the coefficient of the term in which occurs $a^p b^q$ when the binomial $(a + b)^{p+q}$ is expanded, $y = Ex^p r^q$.

Modern equivalent

$$P(k, \theta; n) = \binom{n}{k} \theta^k (1-\theta)^{n-k} f(\theta)$$

But since $f(\theta) = 1$,

$$P(k, \theta; n) = \binom{n}{k} \theta^k (1-\theta)^{n-k}$$

***Prop.* 9 [Bayes's Key Insight?]**

Comment
In order to understand what follows, note that Bayes drew an upside-down, somewhat bell-shaped curve AiB in Figure 5.1 that represents the joint probability density function $P(k, \theta; n)$ for particular values of k and n. For example, he would get a curve like that if he observed 4 successes out of 10 trials, $P(k = 4, \theta; n = 10)$. The equation for the upside-down curve would be $\binom{10}{4}\theta^4(1-\theta)^6$.

Original text
If before anything is discovered concerning the place of the point o [where Ball W ended up], it should appear that the event M had happened p times and failed q in $p + q$ trials, and from hence I guess that the point o lies between any two points in the line AB, as f and b, and consequently that the probability of the event M in a single trial was somewhere between the ratio of Ab to AB and that of Af to AB: the probability I am in the right is the ratio of that part of the figure AiB described as before which is intercepted between perpendiculars erected upon AB at the points f and b, to the whole figure AiB.

Modern equivalent
The probability that Ball W is between $\theta = b$ and $\theta = f$ is given by the area under the upside-down curve between b and f divided by the area under the entire curve.

$$P(b < \theta < f | k; n) = \frac{\int_b^f \binom{n}{k} \theta^k (1-\theta)^{n-k} d\theta}{\int_0^1 \binom{n}{k} \theta^k (1-\theta)^{n-k} d\theta}$$

Original text

COR. The same things supposed, if I guess that the probability of the event M lies somewhere between 0 and the ratio of Ab to AB, my chance to be in the right is the ratio of Abm to AiB.

Modern equivalent
Instead of looking at the area under the curve between b and f, we are looking at the area between 0 and b.

$$P(0 < \theta < b|k; n) = \frac{\int_0^b \binom{n}{k}\theta^k(1-\theta)^{n-k}d\theta}{\int_0^1 \binom{n}{k}\theta^k(1-\theta)^{n-k}d\theta}$$

Scholium

Comment
Before we get to the original text of the *Scholium*, we should do some preparatory discussion.

The area under the entire upside-down curve in Figure 5.1 is

$$P(k; n) = \int_0^1 \binom{n}{k}\theta^k(1-\theta)^{n-k}d\theta$$

This is the the probability of seeing k successes in n trials before starting the experiment. For example, if I know that (after throwing ball W) I am going to throw ball O 10 times, this could be the probability of seeing $k = 4$ successes.

$$P(k = 4; n = 10) = \int_0^1 \binom{10}{4}\theta^4(1-\theta)^6d\theta$$

Here is the key insight in this essay. Bayes is about to argue that

$$P(k; n) = \frac{1}{n+1}, \quad \textit{regardless of } k.$$

In other words, it is generally true that

$$\int_0^1 \binom{n}{k}\theta^k(1-\theta)^{n-k}d\theta = \frac{1}{n+1}$$

Note that we are no longer talking about the probability of θ being between two numbers such as $b = 0.25$ and $f = 0.35$. This is a marginal probability

calculated by "integrating out" θ, so we are talking about the probability of k being 0, 1, ..., n. Above, I gave the example of $n = 10$, meaning that, after throwing ball W, we will throw ball O 10 times. We could see $k = 0$, 1, .., 10 successes and the probability for each of those values of k is 1/11.

$$\int_0^1 \binom{10}{k} \theta^k (1-\theta)^{10-k} d\theta = \frac{1}{11} \quad , \quad k = 0, 1, 2...10.$$

For $k = 0$ and $k = 10$, this is a simple integral, and it evaluates to $\frac{1^{11}}{11}$. But if $k = 4$, the equation for the curve is $\binom{10}{4}\theta^4(1-\theta)^6$, and the area under it is still $\frac{1}{11}$.

$$\int_0^1 \binom{10}{4} \theta^4 (1-\theta)^6 d\theta = \frac{1}{10+1} = \frac{1}{11}.$$

If $k = 9$, the equation for the curve is $\binom{10}{9}\theta^9(1-\theta)^1 = 10\theta^9(1-\theta)$, and the area under it? Still $\frac{1}{11}$.

$$\int_0^1 \binom{10}{9} \theta^9 (1-\theta)^1 d\theta = \frac{1}{11}.$$

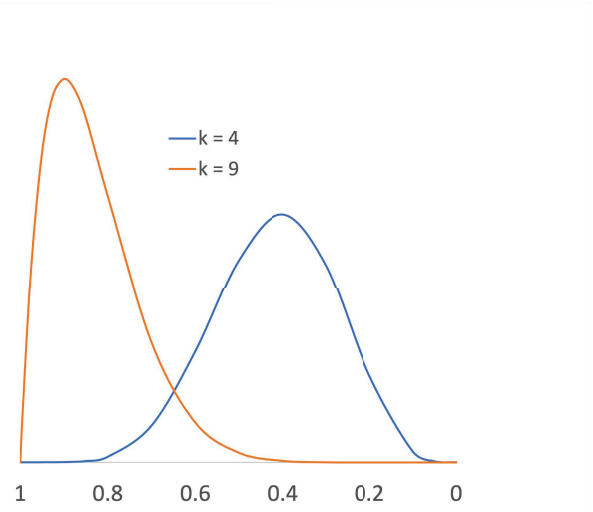


Figure 5.1: For the blue curve on the right, the equation is $\binom{10}{4}\theta^4(1-\theta)^6$, and for the orange curve on the left it is $\binom{10}{9}\theta^9(1-\theta)^1 = 10\theta^9(1-\theta)$. The area under both curves is the same $(\frac{1}{11})$.

To show that

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

we could integrate by parts repeatedly until we reached

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

which is

$$\frac{1}{n+1}.$$

But this bypasses Bayes's key insight. Instead, let's follow Blitztein (p. 382) and assume that instead of starting with two balls: W (which you toss once) and O (which you toss n times), you start with $n+1$ balls and tell two different stories for the left-hand and right-hand side of

$$\int_0^1 \binom{n}{k}\theta^k(1-\theta)^{n-k}d\theta = \frac{1}{n+1}.$$

Story 1: This is essentially the same story we have heard so far. You mark one of the $n+1$ balls as W and toss it onto the table from the right-hand side. Then you toss the other n balls and see that k end up to the right of W. The position of W is θ. Conditional on θ, the probability of k "successes" in n trials is

$$P(k|\theta;n) = \binom{n}{k}\theta^k(1-\theta)^{n-k}$$

Now, to get the unconditional (or marginal) probability of k successes, we have to integrate over all θ and multiply by the probability density $f(\theta)$,

$$P(k;n) = \int_0^1 \binom{n}{k}\theta^k(1-\theta)^{n-k}f(\theta)d\theta$$

But again, $\theta \sim$ Unif(0,1), so $f(\theta) = 1$, and

$$P(k;n) = \int_0^1 \binom{n}{k}\theta^k(1-\theta)^{n-k}d\theta$$

Story 2: This is the new story. Start with $n+1$ balls, all unmarked. Throw them onto the table. Pick one at random and mark it W. Since each of the

$n+1$ balls has the same probability of being marked, the probability of there being $k = 0, 1, 2, ..., n$ balls to the right of the marked ball is

$$P(k; n) = \frac{1}{n+1}, \quad \text{regardless of k.}$$

The following is Bayes (sort of) telling *Story 2*:

Original text
Notes: "Event M" is Ball O landing to the right of Ball W. The *italics* are mine.

... in the case of an event concerning the probability of which we absolutely know nothing antecedently to any trials made concerning it, ...I have no reason to think that, in a certain number of trials, it should rather happen any one possible number of times than another. For, on this account, I may justly reason concerning it as if its probability had been at first unfixed, ... But this is exactly the case of the event M. For before the ball W is thrown, which determines it's probability in a single trial ..., the probability it has to happen p times and fail q in p + q or n trials is the ratio of AiB to CA, which ratio is the same when p + q or n is given, *whatever number p is*; as will appear [later in this essay] by computing the magnitude of AiB by the method of fluxions.* And consequently before the place of the point o is discovered or the number of times the event M has happened in n trials, I can have no reason to think it should happen one possible number of times than another.

*It will be proved presently in art. 4 by [fluxions] that AiB contracted in the ratio of E to 1 is to CA as 1 to (n+1)E: from whence it plainly follows that, antecedently to this contraction, AiB must be to CA in the ratio 1 to n + 1, which is a constant ratio when n is given, whatever p is.

Modern equivalent
If all we know is that, after Ball W, Ball O will thrown n times, the marginal distribution of the number of successes k must be discrete uniform with $P(k) = \frac{1}{n+1}$ for $k = 0, 1, 2, ..., n$.

Original text
In what follows therefore I shall take for granted that the rule given concerning the event M in prop. 9 is also the rule to be used in relation to any event concerning the probability of which nothing at all is known

antecedently to any trials made or observed concerning it. And such an event I shall call an unknown event.

Modern equivalent
In independent trials of an unknown binary event, all possible success counts are equally likely.

Comment
Bayes says an "unknown event" is a binary event in which the number of successes k in n trials, has the discrete uniform distribution: $P(k) = \frac{1}{n+1}$ for $k = 0, 1, 2, ...n$. Note that the definition refers to the distribution of the discrete variable k, not the continuous variable θ. It is true that, if $P(k) = \frac{1}{n+1}$ for *any value* of $n \geq 1$, then $\theta \sim$ Unif(0,1). Bayes didn't know that other prior distributions for θ than Unif(0,1) can lead to $P(k) = \frac{1}{n+1}$ for specific values of n (Stigler p. 129). So, he is saying that the solution to the "billiards" problem (which does assume a uniform prior distribution on θ) applies generally to an "unknown event".

In summary, this *Scholium* says:

> In n independent trials of an unknown binary event, all possible success counts k are equally likely:
> $$P(k; n) = \frac{1}{n+1}, \quad \text{regardless of k.}$$

This is the "rule" that Bayes highlights in this essay, "the rule to be used in relation to any event concerning the probability of which nothing at all is known antecedently to any trials made or observed concerning it."

In the footnote, Bayes says that he will prove by calculus ("the method of fluxions") that

$$\int_0^1 \theta^k (1-\theta)^{n-k} d\theta = \frac{1}{(n+1)\binom{n}{k}}.$$

Quite reasonably, Bayes treats as equivalent the equation with $\binom{n}{k}$ in the integral on the left and the equation with it in the denominator on the right.

The quantity $(n+1)\binom{n}{k}$ can be large. For $n = 10$ and $k = 4$, it is 2,310; for $n = 30$ and $k = 12$, it is 2,681,289,975, so its inverse can be tiny. ($1/2310 = 0.0004329$ and $1/2681289975 = 0.000000000373$)

In modern terms, this integral is the value of the (complete) Beta function with $a = k + 1$ and $b = n - k + 1$.

Here is the (complete) Beta function.

$$B(a, b) = \int_0^1 \theta^{a-1}(1 - \theta)^{b-1} d\theta$$

Substituting and simplifying we see that, for positive integer values of a and b,

$$B(a, b) = \frac{(a - 1)!(b - 1)!}{(a + b - 1)!}$$

Again, the value of this function can be tiny.

***Prop.* 10 [Evaluating the Incomplete Beta Function]**

Comment
In what follows, Bayes will let x (instead of our θ) be the probability of success and $r = 1 - x$ (instead of our $\gamma = 1 - \theta$) be the probability of a failure in n independent binary trials. Again, he uses p (instead of k) for the number of successes and q (instead of j) for the number of failures. As mentioned in the Introduction, Bayes drew a rough plot of the function $y = x^p r^q$ from $x = 0$ on the *right* to $x = 1$ on the *left*. Besides being right side up, the only difference between this curve and the upside down curve in Bayes's first figure (our Figure 5.1), is that there is no factor of $\binom{n}{k}$.

Original text

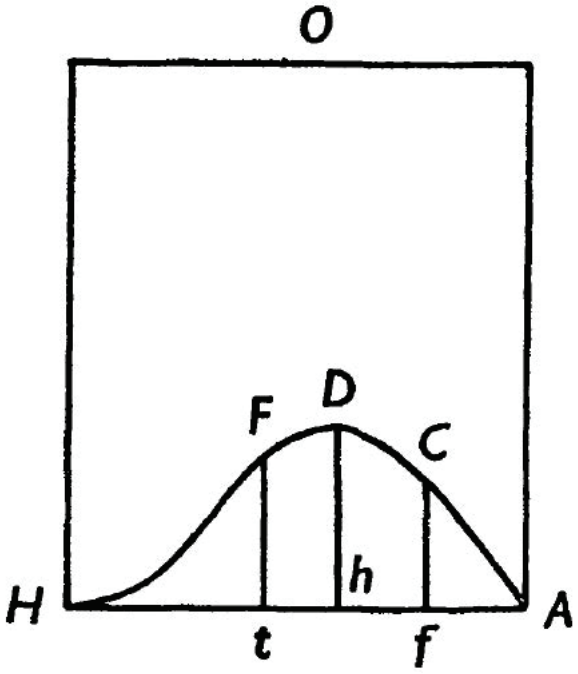


Figure 5.2: If a figure be described upon any base AH (Vid. Fig.) having for its equation $y = x^p r^q$; where y, x, r are respectively the ratios of an ordinate of the figure insisting on the base at right angles, of the segment of the base intercepted between the ordinate and A the beginning of the base, and of the other segment of the base lying between the ordinate and the point H, to the base as their common consequent.

I say then that if an unknown event has happened p times and failed q in $p + q$ trials, and in the base AH taking any two points as f and t you erect the ordinates fC, tF at right angles with it, the chance that the probability of the event lies somewhere between the ratio of Af to AH and that of At to AH, is the ratio of $tFCf$, that part of the before-described figure which is intercepted between the two ordinates, to $ACFH$ the whole figure insisting on the base AH.

Modern equivalent
Before observing the trials, $\theta \sim Unif(0, 1)$. After observing k successes

and j failures in $k + j = n$ trials,

$$P(f \le \theta \le t) = \frac{\int_f^t \theta^k (1-\theta)^j d\theta}{\int_0^1 \theta^k (1-\theta)^j d\theta}.$$

Comment
Of course, we have everything we need if we can evaluate

$$\int_0^g \theta^k (1-\theta)^j d\theta \qquad 0 < g < 1$$

In modern terms, this is the incomplete Beta function with parameters $a = k + 1$ and $b = j + 1$.

Here is the incomplete Beta function.

$$B(g; a, b) = \int_0^g \theta^{a-1} (1-\theta)^{b-1} d\theta \qquad 0 < g < 1$$

But let's continue to use k and j:

$$\int_0^g \theta^k (1-\theta)^j d\theta$$

Bayes's inability to find a good approximation for this integral for k and j large (> 15) may be why he never submitted this essay for publication. But he was able to express the integral as a series that can be evaluated for either small k or small j.

Original text
Now, in order to reduce the foregoing rule to practice, we must find the area of the figure described and the several parts of it separated by ordinates perpendicular to its base. For which purpose, suppose $AH = 1$ and HO the square upon AH likewise $= 1$, and Cf will be $= y$, and $Af = x$, and $Hf = r$, because y, x and r denote the ratios of Cf, Af, and Hf respectively to AH. And by the equation of the curve $y = x^p r^q$ and (because $Af + fH = AH$) $r + x = 1$. Wherefore

$$\begin{aligned} y &= x^p (1-x)^q \\ &= x^p - q x^{p+1} + \frac{q(q-1)x^{p+2}}{2} - \frac{q(q-1)(q-2)x^{p+3}}{2 \cdot 3} + \text{etc.} \end{aligned}$$

Modern equivalent
We need to evaluate the integral

$$\int \theta^k (1-\theta)^j d\theta.$$

We can expand $(1-\theta)^j$ using the binomial formula.

$$(1-\theta)^j = 1 - j\theta + \binom{j}{2}\theta^2 - \binom{j}{3}\theta^3 + \text{etc.}$$

Now mutliply by θ^k.

$$\theta^k(1-\theta)^j = \theta^k - j\theta^{k+1} + \binom{j}{2}\theta^{k+2} - \binom{j}{3}\theta^{k+3} + \text{etc.}$$

Comment
This is where Bayes uses the "method of fluxions". He integrates term by term. Rather than provide his original text, we will continue with the modern equivalent.

$$\int \theta^k (1-\theta)^j = \frac{\theta^{k+1}}{k+1} - \frac{j\theta^{k+2}}{k+2} + \binom{j}{2}\frac{\theta^{k+3}}{k+3} - \binom{j}{3}\frac{\theta^{k+4}}{k+4} + \text{etc.}$$

The above expression can be written as follows:

$$\int \theta^k (1-\theta)^j = \sum_{i=0}^{j} (-1)^i \binom{j}{i} \frac{\theta^{k+i+1}}{k+i+1}$$

Bayes points out that this can be evaluated if j is small. Also, if k is small, let $\gamma = 1-\theta$ (the probability of a failure).

$$\int (1-\gamma)^k (\gamma)^j = \sum_{i=0}^{k} (-1)^i \binom{k}{i} \frac{\gamma^{j+i+1}}{j+i+1}$$

When Bayes's friend Richard Price, in his numerical appendix, examines cases where k is small, he uses this version of the series.

But let's return to the version of the series that works when j is small. For i between 0 and j, the i^{th} term in the summation is

$$(-1)^i \binom{j}{i} \frac{\theta^{k+i+1}}{k+i+1}.$$

Bayes next did something algebraically clever and hard to follow. He multiplied each of these terms by a new term in γ:

$$(-1)^i \frac{\gamma^{j-i}}{\binom{k+i}{i}}$$

$(-1)^i$ in the new term cancels $(-1)^i$ in the original term; the exponents of θ and γ now add to $k+j+1$; and the other terms combine to become

$$\frac{1}{\binom{k+j}{k}} \binom{k+j}{k+i} \frac{1}{k+i+1}.$$

The new series is

$$\begin{aligned} \int \theta^k (1-\theta)^j &= \sum_{i=0}^{j} (-1)^i \binom{j}{i} \left(\frac{\theta^{k+i+1}}{k+i+1} \right) \left((-1)^i \frac{\gamma^{j-i}}{\binom{k+i}{i}} \right) \\ &= \frac{1}{\binom{k+j}{k}} \sum_{i=0}^{j} \binom{k+j}{k+i} \left(\frac{\theta^{k+i+1} \gamma^{j-i}}{k+i+1} \right) \end{aligned}$$

For $i = 3$, the element in the new series would be

$$\frac{1}{\binom{k+j}{k}} \binom{k+j}{k+3} \left(\frac{\theta^{k+4} \gamma^{j-3}}{k+4} \right)$$

Bayes would write it out this way:

$$\frac{j(j-1)(j-2)\theta^{k+4}\gamma^{j-3}}{(k+4)(k+3)(k+2)(k+1)}$$

For $i = j$, the term would be

$$\frac{1}{\binom{k+j}{k}} \binom{k+j}{k+j} \left(\frac{\theta^{k+j} \gamma^{j-j}}{k+j+1} \right)$$

$$\frac{1}{\binom{k+j}{k}} \left(\frac{\theta^{k+j+1}}{k+j+1} \right)$$

Bayes says that this new series with terms involving γ is the same as the one with just θ, "as will easily be seen" by replacing γ with $1-\theta$, expanding the terms, and ordering them according to powers of θ. "Or more readily", by comparing the derivatives of the two series and, in the

second series, substituting $-\dot{\theta}$ for $\dot{\gamma}$. Price added a footnote to show how the derivatives of the two series are equal. See Endnote #5.

Now let's make it a definite integral from $\theta = 0$ ($\gamma = 1 - \theta = 1$) to $\theta = g$ ($\gamma = 1 - g$).

$$\int_0^g \theta^k (1-\theta)^j d\theta = \frac{1}{\binom{k+j}{k}} \sum_{i=0}^{j} \binom{k+j}{k+i} \left(\frac{\theta^{k+i+1} \gamma^{j-i}}{k+i+1} \right) \Bigg|_{\theta=0}^{g}$$

When $\theta = 0$ (and $\gamma = 1$), all the terms are 0, so we are left with the following:

$$\int_0^g \theta^k (1-\theta)^j d\theta = \frac{1}{\binom{k+j}{k}} \sum_{i=0}^{j} \binom{k+j}{k+i} \left(\frac{g^{k+i+1} (1-g)^{j-i}}{k+i+1} \right)$$

This is Bayes's new series to calculate the *incomplete* Beta function.

With some effort, one can substitute $s = k + i + 1$ into Bayes's expression and see that it is equivalent to

$$\int_0^g \theta^k (1-\theta)^j d\theta = \frac{1}{(k+j+1)\binom{k+j}{k}} \sum_{s=k+1}^{k+j+1} \binom{k+j+1}{s} g^s (1-g)^{k+j+1-s}$$

Note that the summation above is the cumulative binomial probability of *more than* k successes in $k + j + 1 = n + 1$ trials: $P(K > k; n+1, g)$. This shows the relationship between the Beta and Binomial distributions. See Endnote #6.

For now, Bayes is confirming via calculus and algebra what he said in the Scholium, essentially that

$$\int_0^1 \theta^k (1-\theta)^j d\theta = \frac{1}{(n+1)\binom{n}{k}}.$$

Here again is Bayes's new series to calculate the incomplete Beta function:

$$\int_0^g \theta^k (1-\theta)^j d\theta = \frac{1}{\binom{k+j}{k}} \sum_{i=0}^{j} \binom{k+j}{k+i} \left(\frac{g^{k+i+1} (1-g)^{j-i}}{k+i+1} \right)$$

If we evaluate the *complete* Beta function by letting $g = 1$ and $1 - g = 0$, all the terms are 0 except the last $i = j$ term.

$$\int_0^1 \theta^k (1-\theta)^j d\theta = \frac{1}{\binom{k+j}{k}} \binom{k+j}{k+j} \left(\frac{1^{k+j+1}}{k+j+1} \right)$$
$$= \frac{1}{(n+1)\binom{n}{k}}$$

Now we can return to the original text.

Original text
If E be the coefficient of that term of the binomial $(a+b)^{p+q}$ expanded in which occurs $a^p b^q$, the ratio of the whole figure $ACFH$ to HO is $[(n+1)E]^{-1}$, n being $= p+q$.

Modern equivalent

$$\int_0^1 \theta^k (1-\theta)^{n-k} d\theta = \frac{1}{(n+1)\binom{n}{k}}$$

Comment
Bayes shows this to support his argument in the *Scholium* that, *knowing nothing* about the probability θ of success in a single trial, the number of successes in n trials has the discrete uniform distribution from 0 to n, meaning that $P(k; n) = \frac{1}{n+1}$ for $k = 0, 1, 2, ..., n$. At least in this example, *knowing nothing* means that $\theta \sim Unif(0, 1)$.

He also needs it to provide the normalizing constant $(n+1)E = (n+1)\binom{n}{k}$ for Rule 1.

6 *Rule 1* [Solution to the Problem]

Original text
[Recall that Bayes's p is our k; his q is our j.]
If nothing is known concerning an event but that it has happened p times and failed q in $p+q$ or n trials, and from hence I guess that the probability of its happening in a single trial lies somewhere between any two degrees of probability as X and x, the chance I am in the right in my guess is $(n+1)E$ multiplied into the difference between the series

$$\frac{X^{p+1}}{p+1} - \frac{qX^{p+2}}{p+2} + \frac{q(q-1)X^{p+3}}{2(p+3)} - etc.$$

and the series

$$\frac{x^{p+1}}{p+1} - \frac{qx^{p+2}}{p+2} + \frac{q(q-1)x^{p+3}}{2(p+3)} - etc.$$

E being the coefficient of $a^p b^q$ when $(a+b)^n$ is expanded.

This is the proper rule to be used when q is a small number; but if q is large and p small, change everywhere in the series here set down p into q and q into p and x into r or (1 -x), and X into R = (1 - X); which will not make any alteration in the difference between the two series.

Modern equivalent

In n independent binary trials with unknown success probability θ, where $\theta \sim Unif(0,1)$, with k successes and $n-k=j$ failures, the posterior probability of θ is given as follows:

$$P(x < \theta < X|k;n) = (n+1)\binom{n}{k}\left(\int_x^X \theta^k(1-\theta)^j d\theta\right).$$

Since

$$\int \theta^k(1-\theta)^j d\theta = \sum_{i=0}^{j}(-1)^i\binom{j}{i}\frac{\theta^{k+i+1}}{k+i+1},$$

the integral in the expression is

$$\sum_{i=0}^{j}(-1)^i\binom{j}{i}\frac{X^{k+i+1}}{k+i+1} - \sum_{i=0}^{j}(-1)^i\binom{j}{i}\frac{x^{k+i+1}}{k+i+1},$$

which must be multiplied by

$$(n+1)\binom{n}{k}$$

Comment

Substituting $\frac{1}{B(k+1,n-k+1)}$ for $(n+1)\binom{n}{k}$,

$$P(x < \theta < X|k;n) = \frac{1}{B(k+1,n-k+1)}\left(\int_x^X \theta^k(1-\theta)^j d\theta\right)$$

This is the modern Beta CDF (cumulative distribution function) evaluated between x and X. Endnote #7 provides definitions and properties for the modern complete and incomplete Beta *functions* and the PDF and CDF of the Beta *distribution*.

Using the modern Beta CDF, we see that Bayes wants to evaluate

$$P(x < \theta < X|k;n) = BetaCDF(X;k{+}1,n{-}k{+}1) - BetaCDF(x;k{+}1,n{-}k{+}1)$$

Rule 1 provides a series expression for BetaCDF that works exactly for k small or $j = n - k$ small, but Bayes needs a good approximation when k and j are both large.

After Rule 1 is presented, Richard Price takes over.

Original text
Thus far Mr Bayes's essay.

With respect to the rule here given, it is further to be observed, that when both p and q [k and j] are very large numbers, it will not be possible to apply it to practice on account of the multitude of terms which the series in it will contain.

Comment
Bayes and Price don't say what they mean by "large numbers" for k and j, but the largest values for which Price attempted to use Rule 1 were $k = 10$, and $j = 100$ and he didn't get the correct answer, so it's safe to say that Rule 1 was impractical for k and j both > 15.

Price next presents an abridgement of the rest of Bayes's essay. He presents *Rule 2* and *Rule 3* for approximating the probability that θ lies within a given distance of k/n. Price and Bayes defined additional variables that I will call $\hat{\theta}$ and σ:

$$\hat{\theta} = k/n$$

$$\sigma = \sqrt{\frac{k(n-k)}{n^3}} = \sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}}$$

Using these new variables, *Rule 2* and *Rule 3* are for approximating the probability that θ lies within a given distance $w\sigma$ of $\hat{\theta}$, i.e, $P(\hat{\theta} - w\sigma \leq \theta \leq \hat{\theta} + w\sigma)$. In other words, *Rule 2* and *Rule 3* are approximations for the following quantity:

$$\frac{\int_{\hat{\theta}-w\sigma}^{\hat{\theta}+w\sigma} \theta^k(1-\theta)^{n-k}d\theta}{\int_0^1 \theta^k(1-\theta)^{n-k}d\theta}$$

$$w > 0$$

This is the Beta CDF with $a = k + 1$ and $b = n - k + 1$ between $\hat{\theta} - w\sigma$ and $\hat{\theta} + w\sigma$.

Rules 2 and 3 do not appear to provide good approximations, so I will not present them, but I will present the normal approximation to

$$f(\theta) = \frac{\theta^k (1-\theta)^{n-k}}{\int_0^1 \theta^k (1-\theta)^{n-k} d\theta},$$

which is the Beta PDF (again with $a = k + 1$ and $b = n - k + 1$) . Endnote #8 shows that

$$f(\theta) \approx \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(\theta - \hat{\theta})^2}{2\sigma^2}\right)$$

Let

$$u = \frac{(\theta - \hat{\theta})}{\sigma}.$$

Then, the normal approximation to the Beta PDF $f(\theta)$ is given by

$$f(\theta) \approx \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-u^2}{2}\right) = \phi(u),$$

where

$$u = (\theta - \hat{\theta})/\sigma$$

$$\phi(u) = \text{standard normal PDF .}$$

Recall that our objective is to approximate the Beta CDF $F(\theta)$ between $\hat{\theta} - w\sigma$ and $\hat{\theta} + w\sigma$:

$$F(\hat{\theta} + w\sigma)) - F(\hat{\theta} - w\sigma) = \frac{\int_{k/n - w\sigma}^{k/n + w\sigma} \theta^k (1-\theta)^{n-k} d\theta}{\int_0^1 \theta^k (1-\theta)^{n-k} d\theta}$$

The normal approximation is given by

$$\begin{aligned} F(\hat{\theta} + w\sigma)) - F(\hat{\theta} - w\sigma) &\approx \frac{1}{\sqrt{2\pi}} \int_{-w}^{w} \exp\left(\frac{-u^2}{2}\right) du \\ &\approx \Phi(w) - \Phi(-w) \\ &\approx 2\Phi(w) - 1 \end{aligned}$$

where $\Phi(w)$ = standard normal CDF. I will use this in the discussion of Price's appendix of numerical examples.

The normal approximation to the Beta CDF was derived by Laplace in the 1780s, two decades after Bayes's essay was published.

Although Bayes and Price were trying to approximate the area under the Beta PDF for an interval around $\hat{\theta} = k/n$, $\hat{\theta}$ is not the mean of the distribution but the mode, i.e., the value of θ where $f(\theta)$ is maximum. You can see this by setting the derivative equal to 0. Endnote #9 shows that the mean of Beta PDF is

$$\frac{k+1}{n+2}$$

Like the normal approximation to the Beta CDF, this expression was derived by Laplace about two decades after publication of the Bayes/Price essay. For the special case of $k = n$ (all n successes and no failures), it is called *Laplace's Rule of Succession*:

$$\frac{n+1}{n+2} = E(\theta|k = n).$$

Laplace famously applied this to the probability that the sun would rise on day $n + 1$ after having been observed to rise for n days. As we shall see momentarily, well before Laplace, Price addressed the probability of sunrise in his appendix to this essay. Bayes and Price aren't interested in the mean of what we now call the Beta posterior distribution. In the appendix, Price focuses on the probability that θ is greater than 0.5 given k successes and j failures in $k + j = n$ trials, i.e., $P(\theta > 0.5; k, j)$.

7 Richard Price's Appendix of Numerical Examples

AN APPENDIX

Containing an application of the foregoing Rules to some particular Cases

Comment

Price starts the Appendix by applying *Rule 1*, which is

$$P(x<\theta<X \mid k, n) = (n+1)\binom{n}{k}\left(\sum_{i=0}^{j}(-1)^i\binom{j}{i}\frac{X^{k+i+1}}{k+i+1} - \sum_{i=0}^{j}(-1)^i\binom{j}{i}\frac{x^{k+i+1}}{k+i+1}\right).$$

But, he starts with $j = 0$ (no failures), so $k = n$.

$$\begin{aligned} P(x < \theta < X \,|\, k{=}n) &= (n+1)\left(\frac{X^{n+1}}{n+1} - \frac{x^{n+1}}{n+1}\right) \\ &= X^{n+1} - x^{n+1} \end{aligned}$$

He focuses on the probability of "more than an even chance" of success, i.e., $P(\theta > \frac{1}{2})$.

$$\begin{aligned} P(1/2 < \theta \leq 1 \,|\, k{=}n) &= 1^{n+1} - (1/2)^{n+1} \\ &= 1 - \frac{1}{2^{n+1}} \\ &= \frac{2^{n+1} - 1}{2^{n+1}} \end{aligned}$$

So, if we have seen one success ($k{=}n{=}1$), $P(\theta > \frac{1}{2}) = \frac{3}{4}$; for $k{=}2$, it's $\frac{7}{8}$; for $k{=}3$, $\frac{15}{16}$, and so on. Which leads him to his sunrise example.

Original Text

Figure 7.1: Let us imagine to ourselves the case of a person just brought forth into this world, and left to collect from his observation of the order and course of events what powers and causes take place in it. The Sun would, probably, be the first object that would engage his attention; but after losing it the first night he would be entirely ignorant whether he should ever see it again. He would therefore be in the condition of a person making a first experiment about an event entirely unknown to him.

But let him see a second appearance or one return of the Sun, and an expectation would be raised in him of a second return, and he might know

that there was an odds of 3 to 1 for some probability of this. This odds would increase, as before represented, with the number of returns to which he was witness. But no finite number of returns would be sufficient to produce absolute or physical certainty. For let it be supposed that he has seen it return at regular and stated intervals a million of times. The conclusions this would warrant would be such as follow. There would be the odds of the millionth power of 2, to one, that it was likely that it would return again at the end of the usual interval. There would be the probability expressed by 0.5352, that the odds for this was not greater than 1,600,000 to 1; and the probability expressed by 0-5105, that it was not less than 1,400,000 to 1 .

Modern Equivalent

After 1,000,000 successes in a row, the distribution of θ is such that

$$P(\theta > 0.5) = \frac{2^{1,000,000} - 1}{2^{1,000,000}} \approx 1.$$

Of more interest, are the following:

$$P\left(\theta < \frac{1,600,000}{1,600,001}\right) = 0.5352$$

$$P\left(\theta > \frac{1,400,000}{1,400,001}\right) = 0.5105$$

Comment

Note that

$$P\left(\theta < \frac{1,400,000}{1,400,001}\right) = 1 - 0.5105 = 0.4895.$$

By definition, the *median* of the posterior distribution on θ is such that

$$P(\theta < median) = 0.5.$$

So, since

$$P\left(\theta < \frac{1,400,000}{1,400,001}\right) = 0.4895$$

$$P\left(\theta < \frac{1,600,000}{1,600,001}\right) = 0.5352,$$

we have bounds for the median:

$$\frac{1,400,000}{1,400,001} < median < \frac{1,600,000}{1,600,001}.$$

Using Excel, I get the value of the median to be

$$\frac{1,442,695}{1,442,696}.$$

Laplace, writing 20 years later, was more interested in the mean or $E(\theta)$, which is given by

$$E(\theta) = \frac{1,000,001}{1,000,002}$$

and thus is lower than the median.

Laplace, like Price, chose to use sunrises as an example of a long string of successes without any failures. After calculating the probability of another sunrise after a series of 1,826,213 as

$$\frac{1,826,214}{1,826,215},$$

he added, "But this number is incomparably greater for him who, recognizing in the totality of phenomena the principal regulator of days and seasons, sees that nothing at the present moment can arrest the course of it." (Pierre Simon Laplace. A Philosophical Essay on Probabilities . 1795. Blackmore Dennett. Kindle Edition. Chapter III.)

Price made a similar disclaimer...

Original Text
It should be carefully remembered that these deductions suppose a previous total ignorance of nature. After having observed for some time the course of events it would be found that the operations of nature are in general regular, and that the powers and laws which prevail in it are stable and permanent.

Comment
After discussing the the sunrise problem, Price turns to runs of binary trials that include failures as well as successes. His example of a binary trial is a lottery draw that can result in either a "prize" or a "blank". The success probability θ is the unknown proportion of prizes overall. This isn't as good an example as Bayes's "billiards" table. One wonders whether successive draws are made without replacement, in which case they wouldn't be independent. In presenting Price's results, I will generalize to n independent binary trials with k successes and $j = n - k$ failures.

Price likes a series of trials with 1 success for every 10 failures, so $k{:}j = 1{:}10$ and $k/(k+j) = k/n = \frac{1}{11} = 0.0909....$ He starts with $n = 11$ trials, $k = 1$ success, and $j = 10$ failures. Then, he moves to 22, 44, and 110 trials with, respectively, 2, 4, and 10 successes. He uses *Rule 1* to calculate the probability that θ is between $\frac{1}{12}$ and $\frac{1}{10}$, $P(\frac{1}{12} < \theta < \frac{1}{10})$. (Note: Price presents this as the (equivalent) probability that γ, the proportion of "blanks", is between $\frac{9}{10}$ and $\frac{11}{12}$, but I prefer to present the probability that θ is between $\frac{1}{12}$ and $\frac{1}{10}$.)

$$P(\frac{1}{12}<\theta<\frac{1}{10} \mid k,n) = (n+1)\binom{n}{k}\left(\sum_{i=0}^{k}(-1)^i\binom{k}{i}\frac{\frac{11}{12}^{j+i+1}}{j+i+1} - \sum_{i=0}^{k}(-1)^i\binom{k}{i}\frac{\frac{9}{10}^{j+i+1}}{j+i+1}\right).$$

Table 1 shows Price's results. For comparison, it also provides the actual values (obtained using Excel's Beta distribution function).

Table 1: Richard Price's probability calculations.

n	**k**	**R. Price's P**	$P(\frac{1}{12} < \theta < \frac{1}{10})$
11	1	0.0770	0.0770
22	2	0.1084	0.1084
44	4	0.1525	0.1527
110	10	0.2506	0.2390

The discrepancy in the 4th row means Price was off somewhat when he did the following 11-term summation:

$$\sum_{i=0}^{10}(-1)^i\binom{10}{i}\frac{\frac{11}{12}^{100+i+1}}{100+i+1}$$

I attempted to duplicate the calculation using the series formula instead of Excel's Beta distribution function and was off by 0.0002. It's impressive that Price did so well presumably using a feather quill and a table of logarithms.

Price then moves on to $n = 1100$ and $k = 100$. Now, he must use Rule 2, which requires a symmetrical interval around $\theta = \frac{k}{n} = \frac{1}{11}$, so he changes

the lower bound of the interval from $\frac{1}{12}$ to $\frac{1}{11} - \frac{1}{110} = \frac{9}{110}$, which is slightly less than $\frac{1}{12}$ (0.081818 instead of 0.083333). The upper bound remains $\frac{1}{11} + \frac{1}{110} = \frac{11}{110} = \frac{1}{10}$. For $P(\frac{9}{110} < \theta < \frac{1}{10})$, he gives a range of 0.7953 to 0.9405, which is quite wide and does not even contain the true probability of 0.7055.

In subsequent work, Price got close to deriving Laplace's normal approximation to the Beta distribution (Dale p. 36). Here is the normal approximation for $P(\frac{9}{110} < \theta < \frac{1}{10})$ given $k = 100$ successes and $j = 1000$ failures in $n = 1100$ trials:

$$2\Phi(\frac{\frac{1}{10} - \frac{1}{11}}{\sigma}) - 1$$

$$\sigma = \sqrt{\frac{(1/11)(10/11)}{1100}} = 0.0087$$

$$2\Phi(\frac{0.1000 - 0.0909}{0.0087}) - 1$$

$$2\Phi(1.049) - 1$$

$$2(0.853) - 1 = 0.706$$

Both the calculation with $n = 1100$ and another with $n = 11000$, which I will omit, reveal the inadequacy of Rules 2 and 3. In both his cover letter and this appendix, Price expresses his hope that "some person shall discover a better approximation to the value of the two series in the first rule". As we have seen, Laplace did, possibly within Price's lifetime. Price died in 1791.

8 My Addendum on Bayes's Real Rule

In modern terminology, the posterior distribution of success probability θ after observing k successes and j failures in $k + j = n$ trials is

$$BetaPDF(\theta; k + 1, j + 1)$$

if the prior distribution of θ was $Unif(0, 1)$.

Remember that a uniform prior is also a Beta prior with parameters $a = 1$ and $b = 1$, $BetaPDF(1, 1)$.

So, if you start with

$$\theta \sim BetaPDF(1,1),$$

and you observe k successes and j failures, you finish with

$$\theta \sim BetaPDF(1+k,1+j).$$

More generally, if you start with

$$\theta \sim BetaPDF(a,b),$$

and you observe k successes and j failures, you finish with

$$\theta \sim BetaPDF(a+k,b+j).$$

Nowadays, anyone using Microsoft Excel, which has a Beta distribution function, can solve Bayes's problem, not just for a uniform prior distribution, but for a wide variety of prior (Beta) distributions.

Bayes's key insight, presented in the *Scholium*, was to define an "unknown event" as one for which the number of occurrences k in n trials has the discrete uniform distribution

$$P(k;n) = \frac{1}{n+1}, \quad k = 0,1,...,n.$$

He used his "billiards" example to also show that, for any allowable k,

$$P(k;n) = \int_0^1 \binom{n}{k} \theta^k (1-\theta)^{n-k} d\theta.$$

Setting these two expressions for $P(k;n)$ equal, we get

$$\int_0^1 \binom{n}{k} \theta^k (1-\theta)^{n-k} d\theta = \frac{1}{(n+1)}$$

or dividing through by $\binom{n}{k}$,

$$\int_0^1 \theta^k (1-\theta)^{n-k} d\theta = \frac{1}{(n+1)\binom{n}{k}}$$

He then confirmed this using calculus and algebra in ***Prop. 10***.

We all use "Bayes's Rule" to refer to

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

but Bayes never wrote it.

The mathematical rule that he highlighted was, in modern notation,

$$\int_0^1 \binom{n}{k} \theta^k (1-\theta)^{n-k} d\theta = \frac{1}{(n+1)} \quad k = 0, 1, 2, ...n.$$

Here is how he said it in words:

> ... in the case of an event concerning the probability of which we absolutely know nothing antecedently, ...I have no reason to think that, in a certain number of trials, it should rather happen any one possible number of times than another.

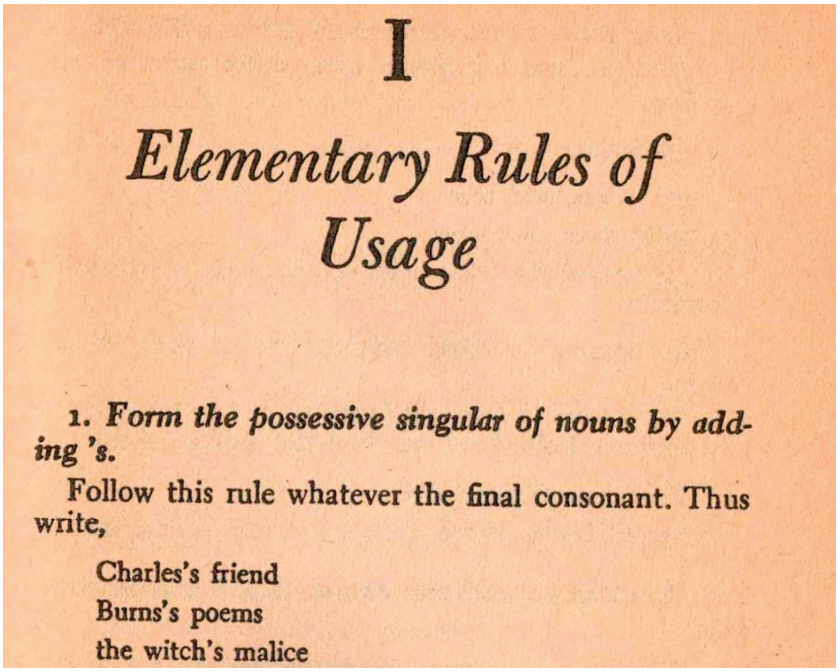
And here is my "translation":

> In independent trials of an unknown binary event, all possible success counts are equally likely.

According to Stigler (p. 98), this essay by Bayes "was ignored by his contemporaries (save Richard Price) and seems to have had little or no impact upon the early development of statistics" until rediscovery and development by Laplace. It seems to have come to European notice around 1780, almost 20 years after Bayes's death and 17 years after publication. Apparently, the term "Bayesian" only goes back to about 1950. (See the conclusion of Stigler's paper on Price.)

9 Endnotes

#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,

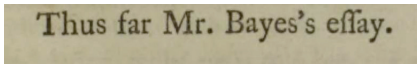
Charles's friend
Burns's poems
the witch's malice

Figure 9.1: 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.

#2 Utility vs. monetary value

Bayes may not have known that, in a 1738 essay, Daniel Bernoulli distinguished between price and utility:

> . . . the *value* of an item must not be based on its *price*, but rather on the *utility* it yields. The price of the item is dependent only on the thing itself and is equal for everyone; the utility, however, is dependent on the particular circumstances of the person making the estimate. Thus there is no doubt that a gain of one thousand ducats is more significant to a pauper than to a rich man though both gain the same amount.
> Daniel Bernoulli, *Exposition of a New Theory on the Measurement of Risk*, Papers of the Imperial Academy of Sciences in St. Petersburg, 1738.

Bayes doesn't bother with this distinction between monetary values and utility. If he did, he would equate "value" with utility.

#3 Alternative version of *Prop.* 4
Dispense with the awkward N by setting it equal to 1. Now, $P(B) = b/N = b$ and $P(B \cap A) = \mathbf{P}/N = \mathbf{P}$, also let d_B be the first day that B occurs.

$$\begin{aligned} P(d_B = 1) &= b \\ P(d_B = 2) &= (1-b)b \\ P(d_B = 3) &= (1-b)^2 b \\ P(d_B = i) &= (1-b)^{i-1} b \end{aligned}$$

This is the "First Success" distribution.

The probability that i is the first day that B occurs *and* A also occurs on that day is

$$P(A \cap d_B = i) = (1-b)^{i-1}\mathbf{P}$$

To get $P(W)$, the probability of receiving N, sum over all i.

$$P(W) = \sum_{i=1}^{\infty} (1-b)^{i-1}\mathbf{P}$$

$$P(W) = \frac{\mathbf{P}}{b}$$

#4 More on the Odds Form of Bayes's Rule
Here, again, is the odds form of Bayes's Rule:

$$\frac{P(A|B)}{P(A^c|B)} = \frac{P(B|A)}{P(B|A^c)}\frac{P(A)}{P(A^c)}$$
$$Odds(A|B) = LR_A(B) \times Odds(A)$$

We can convert the multiplication into addition by taking the (base 10) logarithm:

$$\log Odds(A|B) = \log LR_A(B) + \log Odds(A)$$

Jaynes (page 91) multiplies through by 10 to get

$$10 \log Odds(A|B) = 10 \log LR_A(B) + 10 \log Odds(A)$$

He denotes $10 \log Odds(A)$ as $e(A)$ and $10 \log Odds(A|B)$ as $e(A|B)$, so

$$e(A|B) = 10 \log LR_A(B) + e(A)$$

Jaynes's units for $e(A)$ are *decibels*, abbreviated db. A 1 db increase in $e(A)$ corresponds to multiplying $Odds(A)$ by $10^{0.1} \approx 1.26$. If $Odds(A)$ is a small number (e.g., < 0.1), then a 1 db increase in $e(A)$ also corresponds to multiplying $P(A)$ by 1.26. When $Odds(A)$ is a large number, a 1 db increase in $e(A)$ corresponds to a much smaller multiple of $P(A)$. According to Jaynes (p. 93), "a 1 db change in [e(A)] is about the smallest increment in plausability that is perceptible to our intuition."

During his code breaking work in World War II, Alan Turing expressed plausibility using the same quantity, $10 \log Odds$, and called the units "decibans" instead of decibels. (Jaynes, p. 116)

The likelihood ratio for A of B is the ratio of posterior odds to prior odds,

$$LR_A(B) = \frac{Odds(A|B)}{Odds(A)}.$$

But it is not the "Odds Ratio". The Odds Ratio for A with respect to B, $OR_A(B)$, is the ratio of posterior odds of A given B to posterior odds of A given B^c:

$$OR_A(B) = \frac{Odds(A|B)}{Odds(A|B^c)}.$$

$LR_A(B)$ and $OR_A(B)$ have the same numerator but different denominators.

#5 Show the derivatives of Bayes's two expressions are equal
Recall that the first expression, which contains only θ, not γ, is

$$\int \theta^k (1-\theta)^j = \sum_{i=0}^{j} (-1)^i \binom{j}{i} \frac{\theta^{k+i+1}}{k+i+1}$$

Bayes got this by expanding $(1-\theta)^j$ using the binomial theorem, multiplying by θ^k, and integrating term by term. We can reverse these steps and differentiate term by term, factor out θ^k, and recognize what remains as the binomial expansion of $(1-\theta)^j$. Or we could just remember the fundamental theorem of calculus. Either way, the derivative of the right-hand side must be

$$\theta^k (1-\theta)^j$$

The second expression with the new term in γ is

$$\frac{1}{\binom{k+j}{k}} \sum_{i=0}^{j} \binom{k+j}{k+i} \left(\frac{\theta^{k+i+1} \gamma^{j-i}}{k+i+1} \right).$$

We just need to verify that the derivative of this second expression is also $\theta^k (1-\theta)^j$.

Leaving aside the factor out front of $\frac{1}{\binom{k+j}{k}}$, using the rule for taking the derivative of a product, and remembering that $\gamma = 1-\theta$ and $d\gamma/d\theta = -1$, we get that the derivative is

$$\sum_{i=0}^{j} \left(\binom{k+j}{k+i} \theta^{k+i} \gamma^{j-i} - \binom{k+j}{k+i} \frac{j-i}{k+i+1} \theta^{k+i+1} \gamma^{j-i-1} \right)$$

Simplifying the second (negative) part of the derivative, this is

$$\sum_{i=0}^{j} \left(\binom{k+j}{k+i} \theta^{k+i} \gamma^{j-i} - \binom{k+j}{k+i+1} \theta^{k+i+1} \gamma^{j-i-1} \right)$$

The second (negative) part of element i cancels the first (positive) part of element $i+1$. This is a "telescoping sum". The second part of the last or jth element is 0. So we are left with only the first part of the first element, which is

$$\binom{k+j}{k+0} \theta^{k+0} \gamma^{j-0}$$

$$\binom{k+j}{k}\theta^k\gamma^j$$

Multiplying by the factor of $\frac{1}{\binom{k+j}{k}}$ that we left aside earlier, we have

$$\theta^k(1-\theta)^j.$$

Thus, we have verified

$$\int \theta^k(1-\theta)^j = \frac{1}{\binom{k+j}{k}}\sum_{i=0}^{j}\binom{k+j}{k+i}\left(\frac{\theta^{k+i+1}\gamma^{j-i}}{k+i+1}\right)$$

#6 Relationship between the Beta and Binomial CDFs
Bayes claimed and we verified in Endnote #5 that

$$\int_0^g \theta^k(1-\theta)^{n-k}d\theta = \frac{1}{\binom{k+j}{k}}\sum_{i=0}^{j}\binom{k+j}{k+i}\left(\frac{g^{k+i+1}(1-g)^{j-i}}{k+i+1}\right).$$

I said that with some effort, one can substitute $s = i+k+1$ into Bayes's expression and see that it is equivalent to

$$\int_0^g \theta^k(1-\theta)^j d\theta = \frac{1}{(k+j+1)\binom{k+j}{k}}\sum_{s=k+1}^{k+j+1}\binom{k+j+1}{s}g^s(1-g)^{k+j+1-s}$$

If we multiply through by $(k+j+1)\binom{k+j}{k}$, we get

$$(k+j+1)\binom{k+j}{k}\int_0^g \theta^k(1-\theta)^j d\theta = \sum_{s=k+1}^{k+j+1}\binom{k+j+1}{s}g^s(1-g)^{k+j+1-s}.$$

On the left we have $BetaCDF(g; k+1, j+1)$ and on the right we have the cumulative binomial probability of *more than* k successes in $n+1$ trials: $P(K > k; n+1, g)$, so this expression relates the Beta and Binomial CDFs.

Here is another way to get to that expression. Returning to Bayes's "billiards", throw $n+1$ balls and specifically mark the ball with k balls to its right, i.e., ball $k+1$. The position of this ball is θ_{k+1}. What is the probability that $\theta_{k+1} < g$? It's the probability that ***at least*** $k+1$ balls are to the right of g, which is the probability that ball $k+1$ is to the right of g and balls $k+2, ..., n+1$ are to its left ***plus*** the probability that ball $k+2$

is to the right of g and balls $k+3, ...n+1$ are to its left ***plus*** the probability that ball $k+3$ is to the right of g....

$$F(\theta_{k+1} \leq g) = \sum_{s=k+1}^{n+1} \binom{n+1}{s} g^i (1-g)^{n+1-s} \quad g \in [0,1]$$

This is the CDF for the position of ball $k+1$. It is also the probability of *greater than* k successes in $n+1$ binary trials with success probability g, $P(K > k; n+1, g)$.

To get the PDF $f(\theta_{k+1})$, we could take the derivative of the CDF, $F(\theta_{k+1} \leq g)$, with respect to g, but according to Blitzstein, "the resulting expression is ugly" (Blitzstein p. 401). We know from Bayes's *Prop.* 10 that it's going to be $BetaPDF(k+1, n-k+1)$. Here is my adaptation of Blitztein's explanation.

What is the probability $f(\theta_{k+1})d\theta$ that you mark a ball $k+1$ and it falls in the infinitesimal interval of width $d\theta$ around some arbitrary θ? First, we choose from the $n+1$ balls to get the one that we mark. That gets us a factor of $n+1$. Because of the uniform distribution, the probability that the chosen ball is in the interval $d\theta$ is just $d\theta$, so now we have $(n+1)d\theta$. From the remaining n balls, we choose exactly k to be to right of the marked ball, each with probability θ, leaving $n-k=j$ to its left, each with probability $(1-\theta)$. That gets us a factor of $\binom{n}{k}\theta^k(1-\theta)^j$. We multiply the factors to get

$$f(\theta_{k+1})d\theta = (n+1)d\theta\binom{n}{k}\theta^k(1-\theta)^j$$

Dropping the $d\theta$s from both sides, gets us

$$f(\theta_{k+1}) = (n+1)\binom{n}{k}\theta^k(1-\theta)^j$$

Another expression for the CDF is the integral of the PDF:

$$F(\theta_{k+1} \leq g) = \int_0^g (n+1)\binom{n}{k}\theta^k(1-\theta)^j d\theta \quad g \in [0,1]$$

Setting our two expressions for the CDF $F(\theta_{k+1} \leq g)$ equal, we get

$$\int_0^g (n+1)\binom{n}{k}\theta^k(1-\theta)^{n-k} d\theta = \sum_{s=k+1}^{n+1} \binom{n+1}{s} g^s (1-g)^{n+1-s}$$

This is what we set out to obtain.

Instead of summing from $k+1$ to $n+1$, we could sum from 0 to k and subtract from 1.

$$\begin{aligned} F(\theta_{k+1} \leq g) &= 1 - \sum_{i=0}^{k} \binom{n+1}{i} g^i (1-g)^{n+1-i} \\ &= 1 - BinomCDF(k; n+1, g) \end{aligned}$$

We now have the relationship between the Beta CDF and the Binomial CDF.

$$\int_0^g (n+1)\binom{n}{k}\theta^k (1-\theta)^{n-k} d\theta = 1 - \sum_{i=0}^{k} \binom{n+1}{i} g^i (1-g)^{n+1-i}$$

$$BetaCDF(g; k+1, n-k+1) = 1 - BinomCDF(k; n+1, g)$$

Or let $m = n+1$ and move things around

$$1 - BetaCDF(g; k+1, m-k) = BinomCDF(k; m, g)$$

I digress to point out how the relationship between the $BetaCDF$ and the $BinomCDF$ can be used today when we have a function for the inverse BetaCDF in Microsoft Excel. We can calculate the exact binomial confidence interval about the probability of success g when we observe k successes in m trials. The point estimate is $g = k/m$ and the $1-\alpha$ confidence interval is given by

$$InvBetaCDF(\alpha/2, k, m-k+1) \text{ to } InvBetaCDF(1-\alpha/2, k+1, m-k)$$

For example, if $k = 4$ and $m = 10$, the point estimate for success probability g is $\frac{4}{10} = 0.4$, and then the 95% confidence interval is given by

$$InvBetaCDF(0.025, 4, 7) = 0.1216 \text{ to } InvBetaCDF(0.975, 5, 6) = 0.7376$$

#7 Definitions and properties of the complete and incomplete Beta functions and the Beta distribution's PDF and CDF
Complete Beta Function:

$$B(a, b) = \int_0^1 \theta^{a-1}(1-\theta)^{b-1} d\theta$$

Properties of the Complete Beta Function:

$$B(a,b) = B(b,a),$$

For a and b positive integers,

$$B(a,b) = \frac{1}{(a+b-1)\binom{a+b-2}{a-1}} = \frac{(a-1)!(b-1)!}{(a+b-1)!}.$$

Incomplete Beta Function:

$$B(g;a,b) = \int_0^g \theta^{a-1}(1-\theta)^{b-1}d\theta \qquad 0 < g < 1$$

Properties of the Incomplete Beta Function:

$$\begin{aligned}\int_g^1 \theta^{a-1}(1-\theta)^{b-1}d\theta &= \int_0^{1-g} \theta^{b-1}(1-\theta)^{a-1}d\theta \\ &= B(1-g;b,a)\end{aligned}$$

Since

$$\int_0^g \theta^{a-1}(1-\theta)^{b-1}d\theta + \int_g^1 \theta^{a-1}(1-\theta)^{b-1}d\theta = \int_0^1 \theta^{a-1}(1-\theta)^{b-1}d\theta,$$

$$B(g;a,b) + B(1-g;b,a) = B(a,b) = B(b,a).$$

Beta Distribution PDF:

$$BetaPDF(\theta;a,b) = \frac{1}{B(a,b)}\theta^{a-1}(1-\theta)^{b-1} \qquad 0 < \theta < 1$$

Properties of the Beta PDF:

$BetaPDF(\theta;1,1) = 1$ for $0 < \theta < 1$, so $BetaPDF(1,1)$ and $Unif(0,1)$ are the same distribution.

Beta Distribution CDF:

$$BetaCDF(g;a,b) = \frac{1}{B(a,b)}\int_0^g \theta^{a-1}(1-\theta)^{b-1}d\theta \qquad 0 < g < 1$$

$$= \frac{B(g;a,b)}{B(a,b)}$$

Properties of the Beta CDF:

$$\frac{B(g;a,b)}{B(a,b)} + \frac{B(1-g;b,a)}{B(b,a)} = 1$$
$$BetaCDF(g;a,b) + BetaCDF(1-g;b,a) = 1$$

#8 Normal approximation to the Beta PDF
The Beta PDF is given by

$$f(\theta) = \frac{\theta^k(1-\theta)^{n-k}}{\int_0^1 \theta^k(1-\theta)^{n-k}d\theta}.$$

Let

$$\begin{aligned}
\gamma &= 1-\theta \\
j &= n-k \\
\hat{\theta} &= k/n \\
\hat{\gamma} &= j/n \\
\sigma^2 &= \frac{kj}{n^3} = \frac{\hat{\theta}\hat{\gamma}}{n} \\
B &= \int_0^1 \theta^k(1-\theta)^{n-k}d\theta
\end{aligned}$$

Now,

$$f(\theta) = \frac{\theta^k\gamma^j}{B}.$$

Take the logarithm.

$$L(\theta) = \ln f(\theta) = k\ln(\theta) + j\ln\gamma - \ln B$$

We are going to expand a power series about $\hat{\theta} = k/n$, so we need $L'(\theta) = dL/d\theta$ and $L''(\theta) = d^2L/d\theta^2$ evaluated at $\hat{\theta}$ (and $\hat{\gamma}$). Note that $d\gamma/d\theta = -1$.

$$\begin{aligned}
L'(\theta) &= \frac{k}{\theta} - \frac{j}{\gamma} \\
L'(\hat{\theta}) &= \frac{k}{k/n} - \frac{j}{j/n} \\
&= n - n \\
&= 0
\end{aligned}$$

This means that the function has its maximum at $\theta = k/n = \hat{\theta}$, which is why we will expand the power series around this point.

$$\begin{aligned}
L''(\theta) &= -\frac{k}{\theta^2} - \frac{j}{\gamma^2} \\
L''(\hat{\theta}) &= -\left(\frac{k}{(k/n)^2} + \frac{j}{(j/n)^2}\right) \\
&= -n^2\left(\frac{1}{k} + \frac{1}{j}\right) \\
&= -n^2\left(\frac{j+k}{kj}\right) \\
&= -n^2\left(\frac{n}{kj}\right) \\
&= -\frac{n^3}{kj} \\
&= -\frac{1}{\sigma^2}
\end{aligned}$$

The power series expansion is given by

$$\begin{aligned}
L(\theta) &= L(\hat{\theta}) + L'(\hat{\theta})(\theta - \hat{\theta}) + L''(\hat{\theta})\frac{(\theta - \hat{\theta})^2}{2} + ... \\
&= L(\hat{\theta}) + 0(\theta - \hat{\theta}) - \frac{(\theta - \hat{\theta})^2}{2\sigma^2} + ... \\
&= L(\hat{\theta}) - \frac{(\theta - \hat{\theta})^2}{2\sigma^2} + ...
\end{aligned}$$

Drop the higher order terms and exponentiate.

$$\begin{aligned}
f(\theta) &\approx f(\hat{\theta}) \exp\left(\frac{-(\theta - \hat{\theta})^2}{2\sigma^2}\right) \\
&\approx \frac{\hat{\theta}^k \hat{\gamma}^j}{B} \exp\left(\frac{-(\theta - \hat{\theta})^2}{2\sigma^2}\right)
\end{aligned}$$

In order for this to be a valid PDF that integrates to 1, the constant $\frac{\hat{\theta}^k \hat{\gamma}^j}{B}$

should be $\frac{1}{\sqrt{2\pi\sigma^2}}$. Let's see if this works out. $1/B = (n+1)\binom{n}{k}$, so

$$\begin{aligned}\frac{\hat{\theta}^k\hat{\gamma}^j}{B} &= (n+1)\binom{n}{k}\hat{\theta}^k\hat{\gamma}^j \\ &= (n+1)\left(\frac{n!}{k!j!}\right)\left(\frac{k}{n}\right)^k\left(\frac{j}{n}\right)^j \\ &= (n+1)\left(\frac{n!}{n^n}\right)\left(\frac{k^k}{k!}\right)\left(\frac{j^j}{j!}\right)\end{aligned}$$

Stirling's formula says $m!/m^m \approx \sqrt{2\pi m}/e^m$.

$$\begin{aligned}\frac{\hat{\theta}^k\hat{\gamma}^j}{B} &\approx (n+1)\left(\frac{\sqrt{2\pi n}}{e^n}\right)\left(\frac{e^k}{\sqrt{2\pi k}}\right)\left(\frac{e^j}{\sqrt{2\pi j}}\right) \\ &\approx (n+1)\frac{\sqrt{n}}{\sqrt{2\pi kj}} \\ &\approx \frac{1}{\sqrt{2\pi(kj/n^3)}} \\ &\approx \frac{1}{\sqrt{2\pi\sigma^2}}\end{aligned}$$

So, here is our approximation:

$$f(\theta) = \frac{\theta^k(1-\theta)^{n-k}}{\int_0^1 \theta^k(1-\theta)^{n-k}d\theta} \approx \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(\frac{-(\theta-\hat{\theta})^2}{2\sigma^2}\right)$$

where

$$\hat{\theta} = k/n$$

$$\sigma^2 = \frac{k(n-k)}{n^3} = \frac{\hat{\theta}(1-\hat{\theta})}{n} = \frac{\hat{\theta}\hat{\gamma}}{n}$$

Side note: An equivalent equation to the above is

$$(n+1)\binom{n}{k}\theta^k(1-\theta)^{n-k} \approx \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(\frac{-(\theta-\hat{\theta})^2}{2\sigma^2}\right)$$

As $n \to \infty$, $n+1 \to n$. Divide both sides by n. Note that $n\hat{\theta} = k$.

$$\begin{aligned}\binom{n}{k}\theta^k(1-\theta)^{n-k} &\approx \frac{1}{\sqrt{2\pi n^2\sigma^2}}\exp\left(\frac{-(n\theta-n\hat{\theta})^2}{2n^2\sigma^2}\right) \\ &\approx \frac{1}{\sqrt{2\pi n^2\sigma^2}}\exp\left(\frac{-(n\theta-k)^2}{2n^2\sigma^2}\right)\end{aligned}$$

Substitute $n\hat{\theta}\hat{\gamma}$ for $n^2\sigma^2$.

$$\binom{n}{k}\theta^k(1-\theta)^{n-k} \approx \frac{1}{\sqrt{2\pi n\hat{\theta}\hat{\gamma}}}\exp\left(\frac{-(n\theta-k)^2}{2n\hat{\theta}\hat{\gamma}}\right)$$

If I assume that $n\hat{\theta}\hat{\gamma}$ approaches $n\theta\gamma$ as $n \to \infty$, then

$$\binom{n}{k}\theta^k(1-\theta)^{n-k} \approx \frac{1}{\sqrt{2\pi n\theta\gamma}}\exp\left(\frac{-(k-n\theta)^2}{2n\theta\gamma}\right)$$

Doesn't this provide the normal approximation to the binomial?

#9 Mean of the Beta PDF

The mean of $f(\theta)$ is calculated as follows:

$$E(\theta) = \int_0^1 \theta f(\theta) d\theta$$

$$= \int_0^1 \theta \frac{\theta^k(1-\theta)^{n-k}}{B(k+1, n-k+1)} d\theta$$

$$= \frac{\int_0^1 \theta^{k+1}(1-\theta)^{n-k}}{B(k+1, n-k+1)} d\theta$$

Remember that $B(k+1, n-k+1) = \frac{1}{\binom{n}{k}(n+1)}$, so

$$E(\theta) = \binom{n}{k}(n+1)\int_0^1 \theta^{k+1}(1-\theta)^{n-k} d\theta$$

Using what we have learned about integrals like this,

$$\int_0^1 \theta^{k+1}(1-\theta)^{n-k} d\theta = \frac{1}{\binom{n+1}{k+1}(n+2)},$$

so

$$E(\theta) = \frac{\binom{n}{k}(n+1)}{\binom{n+1}{k+1}(n+2)}$$

$$= \frac{k+1}{n+2}$$

Just as $\hat{\theta} = k/n$ is not the mean of the Beta PDF, $\sigma = \sqrt{k(n-k)/n^3}$ is not the standard deviation Beta PDF. The standard deviation of the Beta PDF is given by

$$\sqrt{\frac{(k+1)(n-k+1)}{(n+2)^2(n+3)}} \quad .$$

10 References

Bayes, T. 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. 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."

Bernoulli, Daniel. 1738 "Exposition of a New Theory on the Measurement of Risk." Papers of the Imperial Academy of Sciences in St. Petersburg. Translated from the Latin by Louise Sommer in *Econometrica*, Vol 22, 1954, pp. 23-36.
Daniel's uncle, Jacob Bernoulli, wrote Ars Conjectandi, published posthumously in 1713. Daniel distinguishes between the monetary value of an event and its utility. Bayes isn't concerned with this distinction, but he clearly means to define the probability of an event as the ratio of its expected utility to the utility realized if it occurs.

Blitzstein JK, Hwang J. 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 this essay. Except for Jaynes (see below), they all use the notation of set theory.

The following books all identify $P(A \cap B) = P(A)P(B|A)$ as the "multiplication rule".
Berry, Donald A. Statistics: A Bayesian Perspective. Duxbury; 1996. p. 133
Freund JE, Walpole RE. Mathematical Statistics. Third edition. Prentice-Hall; 1980. p. 53.
Hogg RV, Craig AT. Introduction to Mathematical Statistics. Fourth edition. MacMillan; 1978. p. 63.
Ross, Sheldon. A First Course in Probability. 10th Edition. Pearson; 2020. p. 73.

Dale AI. Bayes or Laplace? An Examination of the Origin and Early Applications of Bayes Theorem. Archive for History of Exact Sciences 1982;27:2347.
As we will see, Bayes couldn't find a good approximate solution to his problem when the number of successes and failures are both large (>15). Laplace found one 20 years later. Dale shows that Price, building on Bayes's analysis, got close.

Jaynes ET, Bretthorst GL. Probability theory: the logic of science. Cambridge, UK; New York, NY: Cambridge University Press; 2003. 727 p.
Presents probability as an extension of logic. 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, but since that is the approach taken by Blitzstein and the other textbooks on my shelf, I will use set notation in my comments. 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., A Treatise on Probability. Macmillan & Co., London, 1921.
Before his "General Theory of Employment, Interest and Money", Keynes published this book on probability theory, which according to Dale (see next reference) presents the odds form of Bayes's Rule.

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.

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

Stigler SM. 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.