Ma 3/103
Introduction to Probability and Statistics

KC Border
Winter 2021

Lecture 4: Conditional Probability, Trees, and Bayes' Law

Relevant textbook passages:

Pitman [5]: 1.4, 1.5, 1.6, 2.1

Larsen–Marx [4]: 2.4, 3.2

4.1 The effect of new information on the relevant sample space

Pitman [5]: § 1.4
Larsen–Marx [4]: § 2.4

Suppose we acquire new information on the outcome of a random experiment that takes the form, "The sample outcome lies in a set F," or "The event F has occurred." Then we should **update** or **revise** our probabilities to take into account this new information.

4.1.1 Example An urn contains equal numbers of red, white, and blue balls. What is the probability that a randomly drawn ball is red? Obviously it is 1/3. Right?

Now some miscreant comes along before the experiment and removes all the white balls. What is the probability now that a randomly drawn ball is red? Well, now the urn consists of an equal number of red and blue balls, so the probability of red is 1/2. □

4.1.2 Example An urn contains equal numbers of red, white, and blue balls.

Now I draw a ball at random from the urn and truthfully tell you that the ball is not white. Is this any different from Example 4.1.1? I argue that it is not. That is, knowing that the ball is not white means you can eliminate all the white balls from consideration. It is as if the white balls had been removed. So in the new case, the probability of a red ball is again 1/2 . □

What has happened here is the new information, "Someone removed the white balls, or "The ball is not white," led us to renormalize the probability of what is left so that the remainder now has probability one. Remember,

> probability is a measure of our uncertainty, so changes in information can affect probabilities.

4.2 Conditional Probability

More generally, the new information typically takes the form that an event, say F, has occurred.

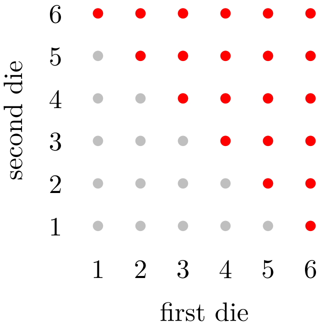
4.2.1 Definition *If $P(F) > 0$, the **conditional probability of E given F**, or **the probability of E conditional on F**, written $P(E \mid F)$, is defined by*

$$P(E \mid F) = \frac{P(EF)}{P(F)}.$$

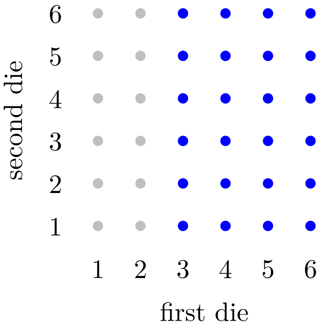
(This only makes sense if $P(F) > 0$.)

Note that $P(F \mid F) = 1$.

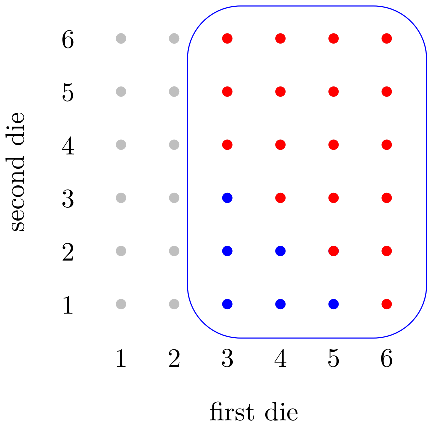
4.2.2 Example You roll two dice. Event E is the event that the sum is $\geqslant 7$, indicated by the red dots below. The Event F is the event that the first die is $\geqslant 3$.



The event $E = (\text{Sum is} \geqslant 7)$. (red)
$P(E) = 21/36 = 7/12$.



The event $F = (\text{First die is} \geqslant 3)$. (blue)
$P(F) = 24/36 = 2/3$.

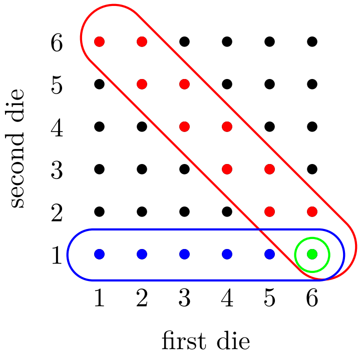


$$E = (\text{Sum is} \geqslant 7.) \qquad P(E) = 21/36.$$
$$F = (\text{First die is} \geqslant 3.) \qquad P(F) = 24/36.$$
$$P(EF) = 18/36.$$
$$P(E \mid F) = \frac{P(EF)}{P(F)} = \frac{18/36}{24/36} = 18/24 = 3/4.$$

□

4.2.1 Another example

The red oval represent the event A that the sum of the two dice is 7 or 8. It contains 11 points, so it has probability 11/36. Let B be the event that the second die is 1. It is outlined in blue, and has 6 points, and so has probability $6/36 = 1/6$. The event BA is circled in green, and consists of the single point $(6, 1)$.



If we "**condition on $\boldsymbol{A}$**," we can ask, what is the probability of the event $B =$ (the second die is 1) **given** that we know that A has occurred, denoted $B \mid A$. Thus

$$P(B \mid A) = \frac{P(BA)}{P(A)} = \frac{1/36}{11/36} = \frac{1}{11}.$$

That is, we count the number of points in BA and divide by the number of points in A.

Similarly

$$P(A \mid B) = \frac{P(AB)}{P(B)} = \frac{1/36}{6/36} = \frac{1}{6}.$$

4.3 Independent events

Recall the following definition.

Larsen–Marx [4]: § 2.5
Pitman [5]: § 1.4

4.3.1 Definition *Events E and F are **(stochastically) independent** if*

$$P(EF) = P(E) \cdot P(F).$$

4.3.2 Proposition *If E and F are stochastically independent, then for $P(F) \neq 0$,*

$$P(E \mid F) = P(E),$$

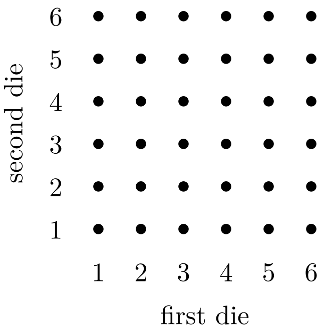
The proof is

$$P(E \mid F) = \frac{P(EF)}{P(F)} = \frac{P(E)P(F)}{P(F)} = P(E).$$

That is, knowing that F has occurred causes no revision of the probability of E. Likewise $P(F \mid E) = P(F)$, provided $P(E) > 0$.

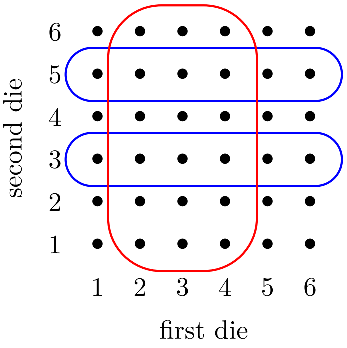
4.3.1 An example

Consider the random experiment of independently rolling two dice. There are 36 equally likely outcomes and the sample space Ω can be represented by the following rectangular array.



The assumption that each outcome is equally likely amounts to assuming that the probability of the product of an event in terms of the first die and one in terms of the second die is the product of the probabilities of each event.

Consider the event E that the second die is 3 or 5, which contains 12 points; and the event F that the first is 2, 3, or 4, which contains 18 points. Thus $P(E) = 12/36 = 1/3$, and $P(F) = 18/36 = 1/2$.

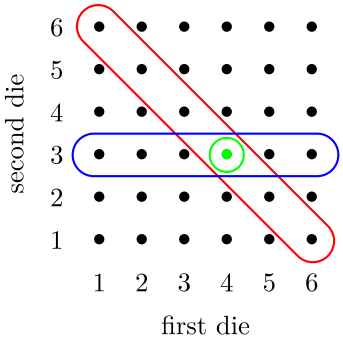


The intersection has 6 points, so $P(EF) = 6/36 = 1/6$. Observe that

$$P(EF) = \frac{1}{6} = \frac{1}{3}\frac{1}{2} = P(E)P(F).$$

4.3.2 Yet another example

To see that not every example of independence involves "rectangular" sets, consider this example involving rolling two dice independently. Event A is the event that the sum is 7, and event B is the event that the second die is 3. The event A is not "rectangular" of the form $E \times F$ (but it is a union of such events), yet they are stochastically independent.



4.4 Conditional probability, as a probability measure

We can think of $P(\cdot \mid A)$ as a "renormalized" probability, with A as the new sample space. That is,

1. $P(A \mid A) = 1$.
2. If $BC = \varnothing$, then
$$P(B \cup C \mid A) = P(B \mid A) + P(C \mid A)$$
3. $P(\varnothing \mid A) = 0$.

Proof of (2):

$$\begin{aligned} P(B \cup C \mid A) &= \frac{P\left((B \cup C)A\right)}{P(A)} \\ &= \frac{P\left((BA) \cup (CA)\right)}{P(A)} \\ &= \frac{P(BA) + P(CA)}{P(A)} \\ &= \frac{P(BA)}{P(A)} + \frac{P(CA)}{P(A)} \\ &= P(B \mid A) + P(C \mid A). \end{aligned}$$

4.5 Bayes' Rule

Now
$$P(B \mid A) = \frac{P(BA)}{P(A)}, \qquad P(A \mid B) = \frac{P(AB)}{P(B)},$$
so we have the following.

4.5.1 Proposition (Multiplication Rule)

$$P(AB) = P(B \mid A) \cdot P(A) = P(A \mid B) \cdot P(B).$$

This in turn implies

4.5.2 Bayes' Rule *For $P(A) \neq 0$,*

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

We can also discuss odds using Bayes' Law. Recall that the odds against B are $P(B^c)/P(B)$. Now suppose we know that event A has occurred. The **posterior odds** against B are now

$$\frac{P(B^c \mid A)}{P(B \mid A)} = \frac{P(A \mid B^c)\frac{P(B^c)}{P(A)}}{P(A \mid B)\frac{P(B)}{P(A)}}. = \frac{P(A \mid B^c)}{P(A \mid B)}\,\frac{P(B^c)}{P(B)}.$$

The term $P(B^c)/P(B)$ is the **prior odds** ratio. Now let's compare the posterior and prior odds:

$$\frac{P(B^c \mid A)/P(B \mid A)}{P(B^c)/P(B)} = \frac{P(A \mid B^c)}{P(A \mid B)}.$$

The right-hand side term $\frac{P(A \mid B^c)}{P(A \mid B)}$ is called the **likelihood ratio** or the **Bayes factor**.

Aside: According to my favorite authority on such matters, the 13th edition of the *Chicago Manual of Style* [6, ¶ 6.12–6.23], we should write Bayes's Rule, but nobody does.

Pitman [5, § 1.4, p. 41], refers to the next result as the **Law of Average Conditional Probability**.

4.5.3 Proposition (Law of Average Conditional Probability) *Let $B_1, \dots, B_n$ be a **partition** of Ω where $P(B_i) > 0$, for $i = 1, \dots, n$. Then for any $A \in \mathcal{F}$,*

$$P(A) = P(A \mid B_1)P(B_1) + \cdots + P(A \mid B_n)P(B_n).$$

Proof: This follows from the fact than for each i, $P(A \mid B_i)P(B_i) = P(AB_i)$ and the fact that since the B_i's partition Ω, we have

$$A = \bigcup_{i=1}^{n} (AB_i),$$

and for $i \neq j$, $(AB_i)(AB_j) = \varnothing$. Now just use the finite additivity of P. ∎

Pitman [5]: p. 49

We can use this to rephrase Bayes' Rule as follows.

4.5.4 Theorem (Bayes' Law) *Let the events $B_1, \dots, B_n$ be a partition of Ω where $P(B_i) > 0$, for $i = 1, \dots, n$.. The for any event A with $P(A) > 0$ and each B_i,*

$$P(B_i \mid A) = \frac{P(A \mid B_i)P(B_i)}{P(A \mid B_1)P(B_1) + \cdots + P(A \mid B_n)P(B_n)}$$

4.6 Bayes' Law and Medical Diagnoses

"When you hear hoofbeats, think horses, not zebras," is advice commonly given to North American medical students. It means that when you are presented with an array of symptoms, you should think of the most likely, not the most exotic explanation.

Pitman [5]: p. 52

4.6.1 Example (Hoofbeats) Suppose there is a diagnostic test, e.g., CEA (carcinoembryonic antigen) levels, for a particular disease (e.g., colon cancer or rheumatoid arthritis). It is in the nature of human physiology and medical tests that they are imperfect. That is, you may have the disease and the test may not catch it, or you may not have the disease, but the test will suggest that you do.

Suppose we know the following.

- One in a thousand people suffer from the disease D.
- If you have the disease, the test catches it (tests positive) 99% of the time.
- But there is also a 2% chance that it reports falsely that someone has the disease when in fact they do not.
- *What is the probability that a randomly selected individual who is tested and has a positive test result actually has the disease?*
- *What is the probability that someone who tests negative for the disease actually is disease free?*

This is a job for Bayes' Law. Let D denote the event that a randomly selected person has the disease, and $\neg D$ be the complementary event the person does not have the disease. Let $+$ denote the event that the test is positive, and $-$ be the event the test is negative. We are told

$$P(D) = 0.001, \quad P(\neg D) = 0.999.$$

That much is straightforward, but the statements about the test are actually statements about conditional probabilities, namely

$$\operatorname{Prob}\big(+ \bigm| D\big) = 0.99 \text{ and } \operatorname{Prob}\big(+ \bigm| \neg D\big) = 0.02,$$

so

$$\operatorname{Prob}\big(- \bigm| D\big) = 0.01 \text{ and } \operatorname{Prob}\big(- \bigm| \neg D\big) = 0.98.$$

For the first question asks for $P(D \mid +)$. By Bayes' Law (Theorem 4.5.4),

$$\begin{aligned} P(D \mid +) &= \frac{P(+ \mid D)P(D)}{P(+ \mid D)P(D) + P(+ \mid \neg D)P(\neg D)} \\ &= \frac{0.99 \times 0.001}{(0.99 \times 0.001) + (0.02 \times 0.999)} \\ &= 0.047. \end{aligned}$$

In other words, even if the test reports that you have the disease there still is only a 4.7% chance that you do have the disease. That may not seem like a high probability, but recall that the prior probability (before the test) of a random person having the disease was 0.001, so the posterior probability (after a positive test) of the disease is 47 times greater. That sounds scarier.

If the false positive rate is reduced twentyfold from 0.02 to 0.001, the posterior probability of disease is 0.4977, basically a coin toss, but still about 500 times higher than a priori.

For the second question, we want to know $P(\neg D \mid -)$, which is

$$\begin{aligned} P(\neg D \mid -) &= \frac{P(- \mid \neg D)P(\neg D)}{P(- \mid \neg D)P(\neg D) + P(- \mid D)P(D)} \\ &= \frac{0.98 \times 0.999}{(0.98 \times 0.999) + (0.01 \times 0.001)} \\ &= 0.999999. \end{aligned}$$

That is, a negative result means it is very unlikely you do have the disease. □

4.7 A family example

Assume that the probability that a child is male or female is 1/2, and that the sex of children in a family are independent trials. So for a family with two children the sample space is

$$\Omega = \{(F,F),(F,M),(M,F),(M,M)\},$$

and each outcome has probability 1/4. Now suppose you are informed that the family has at least one girl. What is the probability that the other child is a boy? Let

$$G = \{(F,F),(F,M),(M,F)\}$$

be the event that there is at least one girl. The event that "the other child is a boy" corresponds to the event

$$B = \{(F,M),(M,F)\}.$$

The probability $P(B \mid G)$ is thus 2/3.

One year a student asked "How does knowing a family has a girl make it more likely to have a boy?" It doesn't. The probability that the family has a boy is not 1/2. It's actually 3/4. So learning that one child is a girl reduces the probability of at least one boy from 3/4 to 2/3.

Now suppose you are told that the elder child is a girl. This is the event

$$E = \{(F,F),(F,M)\}.$$

Now the probability that the other child is a boy is 1/2.

This means that the information that "there is at least one girl" and the information that "the elder is a girl" are really different pieces of information. While it might seem that birth order is irrelevant, a careful examination of the outcome space shows that it is not. The event "the elder is a girl" can happen only two ways, while the event "there is at least one girl" can happen three ways.

4.8 Conditioning and intersections

We already know that $P(AB) = P(A \mid B)P(B)$. This extends by a simple induction argument to the following (Pitman [5, p. 56]):

$$P(A_1A_2\cdots A_n) = P(A_n \mid A_{n-1}\cdots A_1)P(A_{n-1}\cdots A_1)$$

but

$$P(A_{n-1}\cdots A_1) = P(A_{n-1} \mid A_{n-2}\cdots A_1)P(A_{n-2}\cdots A_1),$$

so continuing in this fashion we obtain:

4.8.1 Theorem (Multiplication Rule for Conditional Probabilities)

$$P(A_1A_2\cdots A_n) = P(A_n \mid A_{n-1}\cdots A_1)\,P(A_{n-1} \mid A_{n-2}\cdots A_1)\,\cdots\,P(A_3 \mid A_2A_1)\,P(A_2 \mid A_1)\,P(A_1)$$

It is the basis for computing probabilities in tree diagrams.

4.9 The famous birthday problem

Pitman [5]: § pp. 62–63

Assume that there are only 365 possible birthdays, all equally likely,[1] and assume that in a typical group they are stochastically independent.[2] In a group of size $n \leqslant 365$, what is the probability that at least two people share a birthday? The sample space for this experiment is

$$\Omega = \{1, \dots, 365\}^n,$$

which gets big fast. (The cardinality $\#\,\Omega$ of Ω is about 1.7×10^{51} when $n = 20$.)

This is a problem where it is easier to compute the complementary probability, that is, the probability that all the birthdays are distinct. Number the people from 1 to n. Let A_k be the event that the birthdays of persons 1 through k are distinct. (Note that $A_1 = \Omega$.)

Then

$$A_k \subset A_{k-1}$$

for every k, which means $A_k = A_k A_{k-1} \cdots A_1$ for every k. Thus

$$P(A_{k+1} \mid A_k A_{k-1} \cdots A_1) = P(A_{k+1} \mid A_k).$$

The formula for the probability of an intersection in terms of conditional probabilities implies

$$\begin{aligned} P(A_n) &= P(A_1 \cdots A_n) \\ &= P(A_n \mid A_{n-1} \cdots A_1)\, P(A_{n-1} \mid A_{n-2} \cdots A_1) \cdots P(A_2 \mid A_1)\, P(A_1) \\ &= P(A_n \mid A_{n-1})\, P(A_{n-1} \mid A_{n-2}) \cdots P(A_2 \mid A_1)\, P(A_1) \end{aligned}$$

4.9.1 Claim *For $k < 365$,*

$$P(A_{k+1} \mid A_k) = \frac{365 - k}{365}.$$

While in many ways this claim is obvious, let's plug and chug.

Proof: By definition,

$$P(A_{k+1} \mid A_k) = \frac{P(A_{k+1}A_k)}{P(A_k)} = \frac{P(A_{k+1})}{P(A_k)}.$$

Now there are 365^k equally likely possible lists of birthdays for the k people, since it is quite possible to repeat a birthday. How many give distinct birthdays? There are 365 possibilities for the first person, but after that only 364 choices remain for the second, etc. Thus there are $365!/(365-k)!$ lists of distinct birthdays for k people. So for each $k \leqslant 365$,

$$P(A_k) = \frac{365!}{(365-k)!\,365^k},$$

which in turn implies

$$P(A_{k+1} \mid A_k) = \frac{P(A_{k+1})}{P(A_k)} = \frac{\frac{365!}{(365-k-1)!\,365^{k+1}}}{\frac{365!}{(365-k)!\,365^k}} = \frac{365-k}{365},$$

as claimed. ∎

[1] It is unlikely that all birthdays are equally likely. For instance, it was reported that many parents scheduled C-sections so their children would be born on 12/12/2012. There are also special dates, such as New Year's Eve, on which children are more likely to be conceived. It also matters how the group is selected. Malcolm Gladwell [3, pp. 22–23] reports that a Canadian psychologist named Roger Barnsley discovered the "iron law of Canadian hockey: in *any* elite group of hockey players—the very best of the best—40% of the players will be born between January and March, 30% between April and June, 20% between July and September, and 10% between October and December." Can you think of an explanation for this?

[2] If this were a convention of identical twins, the stochastic independence assumption would have to be jettisoned.

Thus

$$P(A_n) = \prod_{k=1}^{n-1} \frac{365-k}{365}.$$

The probability that at least two share a birthday is 1 minus this product. Here is some Mathematica code to make a table

```
TableForm[
  Table[{n, N[1 - Product[(365 - k)/365, {k, 1, n - 1}]]}, {n, 20, 30}]
  ]  // TeXForm
```

to produce this table:[3]

n	Prob. of sharing
20	0.411438
21	0.443688
22	0.475695
23	0.507297
24	0.538344
25	0.5687
26	0.598241
27	0.626859
28	0.654461
29	0.680969
30	0.706316

Pitman [5, p. 63] gets 0.506 for $n = 23$, but Mathematica gets 0.507. Hmmm.

4.10 Multi-stage experiments

Bayes' Law is useful for dealing with multi-stage experiments. The first example is inferring from which urn a ball was drawn. While this may seem like an artificial example, it is a simple example of Bayesian statistical inference.

Pitman [5]: § 1.5

4.10.1 Example (Guessing urns) There are n urns filled with black and white balls. Let f_i be the fraction of white balls in urn i. (N.B. This is the fraction, not the number of white balls!) In stage 1 an urn is chosen at random (each urn has probability $1/n$). In stage 2 a ball is drawn at random from the urn. Thus the sample space is $\Omega = \{1, \dots, n\} \times \{B, W\}$. Let $\mathcal{F}$ be the set of all subsets of Ω.

Suppose a white ball is drawn from the chosen urn. What can we say about the event that Urn i was chosen. (This is the subset $E = \{(i, W), (i, B)\}$.) According to Bayes' Rules this is:

$$P(i \mid W) = \frac{P(W \mid i)P(i)}{P(W \mid 1)P(1) + \cdots + P(W \mid n)P(n)} = \frac{f_i}{f_1 + \cdots + f_n}.$$

(Note that $P(W \mid i) = f_i$. Also, by assumption each $P(i) = 1.n$. This is crucial.))

It is traditional to call the distribution with which the urn is selected the **prior probability distribution**, or simply the **prior**, on the urns. (In this case each urn was equally likely, but that need be the case in general.) After a ball is drawn, we have more information on the likelihood of which urn was used. This probability distribution, which is found using Bayes' Law, is known as the **posterior probability distribution**, or simply the **posterior**. □

[3] I did edit the TeX code to make the table look better.

Sometimes, in a multi-stage experiment, such as in the urn problem, it is easier to specify conditional probabilities than the probabilities of every point in the sample space. A **tree diagram** is then useful for describing the probability space. **Read section 1.6 in Pitman [5].**

In a tree diagram, the probabilities labeling a branch are actually the conditional probabilities of choosing that branch conditional on reaching the node. (This is really what Section 4.8 is about.) Probabilities of final nodes are computed by multiplying the probabilities along the path.

It's actually more intuitive than it sounds.

4.10.2 Example (A numerical example)
For concreteness say there are two urns and urn 1 has 10 white and 5 black balls ($f_1 = 10/15$), and urn 2 has 3 white and 12 black balls ($f_2 = 3/15$). (It's easier to leave these over a common denominator.) Each Urn is equally likely to be selected. Figure 4.1 gives a tree diagram for this example.

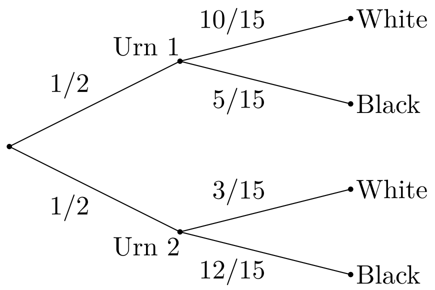


Figure 4.1. Tree diagram for urn selection

Then

$$P(\text{Urn } 1 \mid W) = \frac{\frac{10}{15} \cdot \frac{1}{2}}{\frac{10}{15} \cdot \frac{1}{2} + \frac{3}{15} \cdot \frac{1}{2}} = \frac{10}{13}.$$

and

$$P(\text{Urn } 1 \mid B) = \frac{\frac{5}{15} \cdot \frac{1}{2}}{\frac{5}{15} \cdot \frac{1}{2} + \frac{12}{15} \cdot \frac{1}{2}} = \frac{5}{17}.$$

Sometimes it easier to think in terms of posterior odds. Recall (Section 4.5) that the **posterior odds** against B given A are

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

Letting B be the event that Urn 1 was chosen and A be the event that a White ball was drawn we have

$$\frac{P(\text{Urn } 2 \mid \text{White})}{P(\text{Urn } 1 \mid \text{White})} = \frac{P(\text{White} \mid \text{Urn } 2)}{P(\text{White} \mid \text{Urn } 1)} \frac{P(\text{Urn } 2)}{P(\text{Urn } 1)} = \frac{\frac{3}{15}}{\frac{10}{15}} \frac{\frac{1}{2}}{\frac{1}{2}}.$$

The odds against Urn 1 are $3 : 10$, so the probability is $10/(3 + 10) = 10/13$.

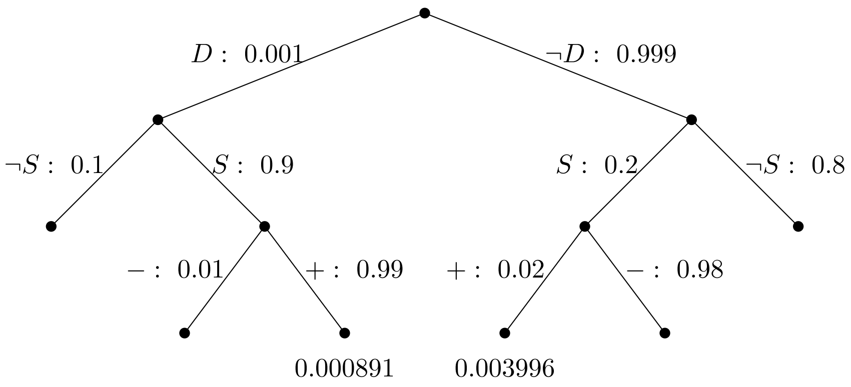
□

4.10.1 Hoofbeats revisited

There is something wrong with the medical testing example. The conclusions would be correct if the person being tested were randomly selected from the population at large. But this is rarely the case. Usually someone is tested only if they exhibit symptoms or there are other reasons to believe the person may have the disease or have been exposed to it.

So let's modify this example by introducing a symptom S. Suppose that if you have the disease there is a 90% chance you have the symptom, but if you do not have the disease, there is only a 20% chance you have the symptom. Suppose further that only those exhibiting the symptom are tested, but that the symptom has no influence on the outcome of the test.

Let's draw tree diagram for this case.



Now the probability of having the disease given a positive test (and the symptom) is

$$\frac{0.000891}{0.000891 + 0.003996} = 0.18232.$$

While low, this is still much higher than without the symptoms.[4]

4.11 The Monty Hall Problem

This problem is often quite successful at getting smart people into fights with each other. Here is the back story.

When I was young there was a very popular TV game show called *Let's Make A Deal*, hosted by one Monty Hall, hereinafter referred to as MH. At the end of every show, a contestant was offered the choice of a prize behind one of three numbered doors. Behind one of the doors was a very nice prize, often a new car, and behind each of the other two doors was a booby prize, often a goat. Once the contestant had made his or her selection, MH would often open one of the *other* doors to reveal a goat. (He always knew where the car was.) He would then try to buy back the door selected by the contestant. Reportedly, on one occasion, the contestant asked to trade for the unopened door. *What is the probability that the car is behind the unopened door?*

A popular, but incorrect answer to this question runs like this. Since there are two doors left, and since we know that there is always a goat to show, the opening of the door with the goat conveys no information as to the whereabouts of the car, so by Laplace's dictum the probability of the car being behind either of the two remaining doors must each be one-half. This is wrong, even though intelligent people have argued at great length that it is correct. (See the Wikipedia article, which claims that even Paul Erdös (pioneer of random graph theory, among other things) believed it was fifty-fifty until he was shown simulations.)

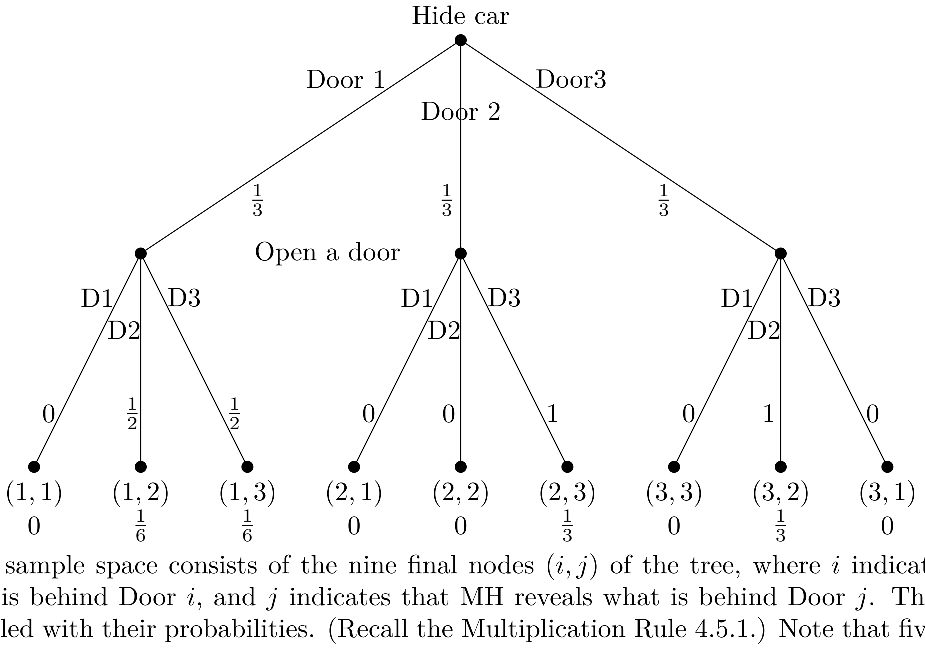
[4] I cavalierly said that 18% was a low probability, but you might not think it was small if your doctor told you that you had an 18% probability of dying before the end of the year.

To answer this question, we must first carefully describe the random experiment, which is a little ambiguous. This has two parts, one is to describe the sample space, and the other is to describe MH's decision rule, which governs the probability distribution. I claim that by rearranging the numbers we may safely assume that the contestant has chosen door number 1. [5] We now assume that the car has been placed at random, so it is equally likely to be behind each door. We now make the following assumption on MH's behavior (which seems to be borne out by the history of the show): We assume that MH will always reveal a goat (and never the car), and that if he has a choice of doors to reveal a goat, he chooses randomly between them with equal probability.

Let C_i denote the event that the car is behind Door i, and let R_j denote the event that MH reveals what is behind Door j. Our probability model is that $P(C_i) = \frac{1}{3}$, $i = 1, 2, 3$, and that MH's behavior provides the following *conditional* probabilities:

$$\begin{aligned}
P(R_1 \mid C_i) &= 0, \qquad i = 1,2,3 \\
P(R_2 \mid C_1) &= \frac{1}{2} \\
P(R_2 \mid C_2) &= 0 \\
P(R_2 \mid C_3) &= 1 \\
P(R_3 \mid C_1) &= \frac{1}{2} \\
P(R_3 \mid C_2) &= 1 \\
P(R_3 \mid C_3) &= 0.
\end{aligned}$$

Here is a tree diagram for the experiment:



The sample space consists of the nine final nodes (i, j) of the tree, where i indicates that the car is behind Door i, and j indicates that MH reveals what is behind Door j. The nodes are labeled with their probabilities. (Recall the Multiplication Rule 4.5.1.) Note that five of the points in the sample space have probability zero.

The contestant is interested in conditional probabilities of the form

$$P\big(\text{Car is behind Door 2} \mid \text{MH reveals Door 3}\big).$$

[5] The numbers on the doors are irrelevant. The doors do not even have to have numbers, they are only used so we can refer to them in a convenient fashion. That, and the fact that they did have numbers and MH referred to them as "Door number one," "Door number two," and "Door number three."

This is where Bayes' Law comes in. By Bayes Law 4.5.4,

$$
\begin{aligned}
P(C_2 \mid R_3) &= \frac{P(R_3 \mid C_2)P(C_2)}{P(R_3 \mid C_1)P(C_1) + P(R_3 \mid C_2)P(C_2) + P(R_3 \mid C_3)P(C_3)} \\
&= \frac{1 \cdot \frac{1}{2}}{\frac{1}{2} \cdot \frac{1}{2} + 1 \cdot \frac{1}{2} + 0 \cdot \frac{1}{2}} = \frac{2}{3}.
\end{aligned}
$$

A similar argument shows that $P(\text{car behind } 3 \mid \text{MH opens } 2) = 2/3$.

While it is true that MH is certain to reveal a goat, opening Door 3 to reveal a goat does shed light on where the car is—we now know that it is not behind Door 3. What opening the door does not shed light on is whether it is behind Door 1. Since nothing has happened to cause the contestant to change his assessment of that probability, it is still 1/3, so the probability that the car is behind Door 2 is now 2/3.

To understand this more fully, consider a different behavioral rule for MH: open the highest numbered door that has a goat. The difference between this and the previous rule is that if the cars is behind 1, then MH will always open door 3. The only time he opens door 2 is when the car is behind door 3. The new probability space is:

$$
\Omega = \{ \underbrace{(1,2)}_{0}, \underbrace{(1,3)}_{\frac{1}{3}}, \underbrace{(2,3)}_{\frac{1}{3}}, \underbrace{(3,2)}_{\frac{1}{3}} \}.
$$

In this case, $P(\text{car behind } 3 \mid \text{MH opens } 2) = 1$, and $P(\text{car behind } 2 \mid \text{MH opens } 3) = 1/2$.

Bertrand's Boxes

Monty Hall dates back to my youth, but similar problems have been around longer. Joseph Bertrand, a member of the Académie Francaise, in fact, the Secrétaire Perpétual of the Académie des Sciences, and the originator of the Bertrand model of oligopoly [1], struggled with explaining a similar situation.

In his treatise on probability [2],[6] he gave the following rather unsatisfactory discussion (pages 2–3, loosely translated with a little help from Google):

> 2. Three boxes/cabinets are identical in appearance. Each has two drawers, and each drawer contains a coin/medal. The coins of the first box are gold; those of the second box are silver; the third box contains one gold coin and one silver coin.
>
> One chooses a box; what is the probability of finding one gold coin and one silver coin?
>
> There are three possibilities and they are equally likely since the boxes look identical:
>
> One possibility is favorable. The probability is 1/3.
>
> Now choose a box and open a drawer. Whatever coin one finds, only two possibilities remain. The unopened drawer contains a coin of the same metal as the first or it doesn't. Of the two possibilities, only one is favorable for the box being the one with the different coins. The probability of this is therefore 1/2.
>
> How can we believe that simply opening a drawer raises the probability from 1/3 to 1/2?
>
> The reasoning cannot be right. In fact, it is not.
>
> After opening the first drawer there are two possibilities. Of these two possibilities, only one is favorable, that much is true, but the two possibilities are not equally likely.
>
> If the first coin is gold, the other may be silver, but a better bet is that it is gold.
>
> Suppose that instead of three boxes, we have three hundred, one hundred with two gold medals, etc. Out of each box, open a drawer and examine the three hundred medals. One hundred will be gold and one hundred will be silver, for sure. The other hundred are in doubt, chance governs their numbers.

[6] Once again, Lindsay Cleary found this for me in the basement of the Sherm (SFL).

> We should expect on opening three hundred drawers to find fewer than two hundred gold pieces. Therefore the probability of a gold coin belonging to one of the hundred boxes with two gold coins is greater than 1/2.

That is true, as far as it goes, but you can now do better. If a randomly selected coin is gold, the probability is 2/3 that it came from a box with two gold coins.

Bibliography

[1] J. L. F. Bertrand. 1883. Théorie mathématique de la richesse sociale. *Journal des Savants* 48:499–508.

[2] ——— . 1907. *Calcul des probabilités*, second ed. Paris: Gauthier–Villars.

[3] M. Gladwell. 2008. *Outliers: The story of success.* New York, Boston, London: Little, Brown.

[4] R. J. Larsen and M. L. Marx. 2012. *An introduction to mathematical statistics and its applications*, fifth ed. Boston: Prentice Hall.

[5] J. Pitman. 1993. *Probability*. Springer Texts in Statistics. New York, Berlin, and Heidelberg: Springer.

[6] University of Chicago Press Editorial Staff, ed. 1982. *The Chicago manual of style*, 13th ed. Chicago: University of Chicago Press.