INFN School of Statistics 2024
Paestum, Italy

Probability Theory
Part 1

Harrison B. Prosper

Kirby W. Kemper Endowed Professor of Physics
Department of Physics, Florida State University

13 May, 2024

Outline

Outline

Probability is common sense reduced to calculation

Laplace

Consider the statements:

- The odds of drawing four aces in a row from a standard deck of cards (i.e., 52 cards, 4 ranks) is 1 in 270,725.
- The chance that a proton-proton collision at 13 TeV creates a Higgs boson is 10^{-10}.
- It is highly likely that William Shakespeare is the author of the inspired insult:

  You blocks, you stones, you worse than senseless things!

Each statement is about probability. Since probability is the foundational concept in statistics and almost all of contemporary machine learning, it is worth spending a bit of time learning the basics.

So let's get started!

The Stanford Encyclopedia of Philosophy[1] lists six interpretations of probability of which the most common are:

1. classical If n things can happen, called A, out of m possible things then $P(A) = n/m$.
2. relative frequency If n things can happen out of m possible things, then $P(A) = n/m$ in the limit $n, m \to \infty$. This is the basis of the frequentist approach to statistics.
3. degree of belief A measure of the strength of belief of a rational agent in the truth of A. This is the basis of the Bayesian approach to statistics.

We start with a famous problem in probability.

[1] https://plato.stanford.edu/entries/probability-interpret/

Example (1.1 A Gambling Man)

Which of the following is the more probable, getting *at least*

1. one 6 in 4 throws of a single 6-sided die, or
2. a double 6 in 24 throws of two 6-sided dice?

Antoine Gombaud (le chevalier de Méré, 1607-1684) brought this problem to the attention of Blaise Pascal who in 1654 began a correspondence with Pierre de Fermat (1601-1665).

Their work together with earlier work by the Italian mathematician Gerolamo Cardano (1501 - 1576) are the first serious attempt to create a theory of probability.

Blaise Pascal (1623 - 1662)

Example (1.1 A Gambling Man)

What is the probability to get *at least*

1. one 6 in 4 throws of a single 6-sided die?

Let p be the probability to obtain ≥ 1 six in 4 throws of a die and q the probability to get none. We refer to the outcome of each throw as an elementary outcome and the outcomes of the 4 throws of the die as an experimental outcome.

Now to the solution.

> *Alas, there is none!*
>
> *Unless that is one is prepared to make a sufficient number of assumptions to render the problem well-posed. Moreover, because different people may make different assumptions, there may be different answers to the same problem!*

Assumptions

1. The two experimental outcomes, either 0 sixes or ≥ 1 sixes are exhaustive — i.e., they are the only possible outcomes.
2. There are 6 distinct elementary outcomes[2].
3. The elementary outcomes are equally probable.
4. The probability of the elementary outcomes is the same for every throw and every throw is independent of the others.

Assumption 1 $\implies p + q = 1$.
Assumption 2 $\implies$ there are 5 ways not to get a six.
Assumption 3 $\implies$ the probability of each of the elementary outcomes is 1/6, therefore, given assumption 2 the probability not to get a six is 5/6.
Assumption 4 $\implies$ the probability not to get a six in 4 throws of the die is $q = (5/6) \times (5/6) \times (5/6) \times (5/6)$. Therefore, $p = 1 - (5/6)^4$.

[2] From which the experimental outcomes are constructed.

Another way to approach de Méré's problem is by counting outcomes.

Strategy

- Consider outcomes that are assumed to be equally probable;
- count the total number of possible outcomes, that is, determine the cardinality (size) of the sample space;
- count the number of desired outcomes
- and take the ratio of the two counts as the probability of the desired outcome.

The outcome of the experiment can be represented by the 4-tuple (z_1, z_2, z_3, z_4), where $z_i \in \{1, 2, 3, 4, 5, 6\}$. The total number of 4-tuples, that is, experimental outcomes, is $6 \times 6 \times 6 \times 6$. The total number of outcomes without a six is $5 \times 5 \times 5 \times 5$, therefore, the number of outcomes with a six is $6^4 - 5^4$. Consequently, assuming that each experimental outcome is equally probable, the probability of the desired outcome is $p = (6^4 - 5^4)/6^4 = 1 - (5/6)^4$.

Example (1.2 The Birthday Problem)

A crowd of people is randomly assembled. How large must the crowd be so that the chance of finding $\geq$ two people with the same birthday is $\geq 50\%$?

Assumptions

1. There are 365 possible birthdays (ignoring leap years).
2. Every birthday is equally probable.

Consider a crowd of size n. The outcome of an experiment — the random assembling of a crowd — can be modeled as an n-tuple, $(z_1, \ldots, z_n)$, where $z_i \in \{1, \ldots, 365\}$.

Let M be the cardinality of the set of n-tuples, the number of possible crowds, and N the cardinality of n-tuples with ≥ 1 duplicate entries. As in the previous example, it is easier to count the number of n-tuples K with *no* duplicates and then compute the desired probability using $p = N/M = (M - K)/M$ assuming n-tuples are equally probable.

M – What is the cardinality of the sample space? Each slot in an n-tuple can be filled in 365 ways. Therefore, there are $M = 365^n$ n-tuples in Ω.[3]

K – What is the cardinality of the set with *no* duplicate birthdays? The 1st slot can be filled in 365 ways. Since duplicates are not allowed, the 2nd slot can be filled in $365 - 1$ ways, the 3rd in $365 - 2$ ways, and so on. So $K = 365 \times 364 \times \cdots \times (365 - (n-1)) = 365!/(365-n)!$.

Therefore, the probability of at least one duplicate birthday is

$$\begin{aligned} p &= (M-K)/M = [365^n - 365!/(365-n)!]/365^n \\ &= 1 - \frac{364}{365} \times \frac{363}{365} \times \cdots \times \frac{(365-n+1)}{365} \geq 0.50. \end{aligned}$$

[3]This is akin to sampling with replacement from a box with 365 distinguishable items.

Permutations

How many ways can n items be arranged in a row with k slots? The first slot can be filled in n ways, the 2nd in $(n-1)$ ways, the third in $(n-2)$ ways and so on until the last slot is reached, which can be filled in $n-k+1$ ways. This yields $n!/(n-k)!$ arrangements. When $k=n$ we get $n!$ permutations.

Combinations

For k items there are $k!$ permutations consisting of rearrangements of the same items in the k slots. If the order of the items is irrelevant (perhaps because the items are indistinguishable) then the number of *distinct* arrangements is smaller by $k!$, that is, the number of arrangements is

$$\frac{n!}{(n-k)!\,k!} \equiv \binom{n}{k}.$$

This is called the number of combinations.

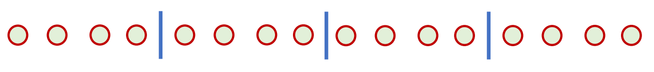
Example (1.3 Partitions)

A partition divides objects into groups. Here is an example[a].

12 undergraduate students and 4 graduate students are randomly divided into 4 groups of 4 students. What is the probability that each group includes a graduate student?

Assumptions

1. Every partition of the students is equally probable.
2. The order of the students within a group is irrelevant.



Note: If the order of the students within groups were relevant, there would be 16! ways to arrange the students.

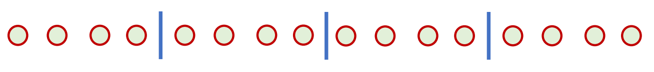
---

[a] Dimitri P. Bertsekas and John N. Tsitsiklis, Introduction to probability, MIT, ISBN 978-1-886529-23-6

Example (1.3 Partitions)

Assumptions

1. Every partition of the students is equally probable.
2. The order of the students within a group is irrelevant.



But Assumption 2 implies that every one of the $4!$ arrangements of students within a group is equivalent. So $16!$ overcounts the number of partitions by $4!^4$.

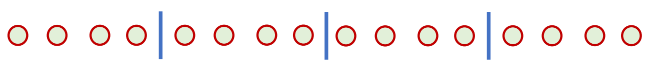
Therefore, the number of *distinct* partitions for which the order of students within each group is irrelevant, that is, the cardinality M of the sample space, is

$$M = \frac{16!}{4!\,4!\,4!\,4!}.$$

Example (1.3 Partitions)

Assumptions

1. Every partition of the students is equally probable.
2. The order of the students within a group is irrelevant.



We now need the number of partitions K with a graduate student in each group. Given this count, the desired probability, $p = K/M$, follows from Assumption 1.

Exercise 1.1

Compute K then compute the probability that each group contains a graduate student.

Outline

In 1933, Andrey Kolmogorov published a highly influential book entitled *Foundations of the Theory of Probability* in which he developed the theory of probability starting from

1. the axioms of Boolean algebra
2. and axioms he introduced.

We first consider the axioms of Boolean algebra, then those of Kolmogorov.

A Boolean algebra is the 4-tuple $(\mathbb{B}, +, \bullet, \bar{\ })$ comprising a collection of sets $\mathbb{B}$ that includes the special sets $\varnothing$ and Ω and the operations OR ($+$), AND ($\bullet$), and NOT ($\bar{\ }$).

Axioms of Boolean Algebra (Huntington)

For all ($\forall$) $A, B, C \in \mathbb{B}$:

$$A + B = B + A \quad (1)$$
$$A + (BC) = (A + B)(A + C) \quad (2)$$
$$A + \varnothing = A \quad (3)$$
$$A + \bar{A} = \Omega \quad (4)$$
$$AB = BA \quad (5)$$
$$A(B + C) = AB + AC \quad (6)$$
$$A\Omega = A \quad (7)$$
$$A\bar{A} = \varnothing \quad (8)$$

With a slight abuse of notation, we'll use 0 as a synonym for $\varnothing$ and 1 for Ω.

We also assume the "meta" axiom $\overline{(\overline{A})} = A$ and that we can make assignments, e.g., $A = B$.

Here are a few useful lemmas and theorems:

$$A + A = A$$
$$A + 1 = 1$$
$$A0 = 0$$
$$AA = A$$
$$(A + B) + C = A + (B + C)$$

$$\bar{0} = 1$$
$$\bar{1} = 0$$
$$A + AB = A$$
$$A(A + B) = A$$
$$(AB)C = A(BC)$$

De Morgan's Laws

$$\overline{A + B} = \bar{A}\,\bar{B}$$
$$\overline{AB} = \bar{A} + \bar{B}$$

Lemma (1)

$A + A = A$

Proof.

$$
\begin{aligned}
A + (BC) &= (A+B)(A+C) && \text{(axiom 2)} \\
A + (CB) &= (A+B)(A+C) && \text{(axiom 5)} \\
A + (A\bar{A}) &= (A+\bar{A})(A+A) && \text{let } C = A, B = \bar{A} \\
A + 0 &= 1(A+A) && \text{(axioms 8, 4)}, (0) \rightarrow 0, (1) \rightarrow 1 \\
A &= 1(A+A) && \text{(axiom 3)} \\
A &= A + A && \text{(axioms 6, 5, 7)}
\end{aligned}
$$

□

Lemma (2)

$A + 1 = 1$

Proof.

$$
\begin{aligned}
A + (BC) &= (A + B)(A + C) && \text{(axiom 2)} \\
A + (B1) &= (A + B)(A + 1) && \text{let } C = 1 \\
A + B &= (A + B)(A + 1) && \text{(axiom 7)}, (B) \to B \\
A + \bar{A} &= (A + \bar{A})(A + 1) && B = \bar{A} \\
1 &= 1(A + 1) && \text{(axiom 4)}, (1) \to 1 \\
1 &= A + 1 && \text{(axioms 6, 5, 7)}
\end{aligned}
$$

□

Exercise 1.2

Prove

$$\begin{aligned} A0 &= 0, \\ A + AB &= A, \\ A(A+B) &= A. \end{aligned}$$

Justify every step with one or more axioms including the meta axiom $(A) \to A$.

Kolmogorov Axioms

Let Ω be a set of elementary events E and S a collection of subsets of Ω, called events, including the empty set $\varnothing$ and the set Ω. Probability P is a real number assigned to all events $A, B \in S$ such that

$$P(A) \geq 0 \tag{9}$$

$$P(\Omega) = 1 \tag{10}$$

$$P(A+B) = P(A) + P(B) \quad \forall AB = \varnothing \tag{11}$$

If $AB = \varnothing$, A and B are said to be mutually exclusive.

Here are a few basic theorems that can be derived from the two sets of axioms, Eqs. (1 – 11):

$$P(\varnothing) = 0$$

$$P(A+B) = P(A) + P(B) - P(AB)$$

$$P(A_1 + \cdots + A_n) = P(A_1) + \cdots P(A_n) \quad \forall A_i B_j = \varnothing, i \neq j.$$

Lemma (3)

$P(\varnothing) = 0$

Proof.

The lemma $A\varnothing = \varnothing$ implies $\Omega\varnothing = \varnothing$, that is, the startling conclusion that events Ω and $\varnothing$ are mutually exclusive! Therefore,

$$
\begin{aligned}
P(\Omega + \varnothing) &= P(\Omega) + P(\varnothing) && \text{(axiom 11)} \\
P(\Omega) &= P(\Omega) + P(\varnothing) && \text{(axiom 3)} \\
1 &= 1 + P(\varnothing) && \text{(axiom 10)} \\
\therefore\ P(\varnothing) &= 0.
\end{aligned}
$$

□

Exercise 1.3

Prove

$$P(A) + P(\bar{A}) = 1,$$
$$P(A + B) = P(A) + P(B) - P(AB).$$

Justify every step with one or more axioms.

Outline

Conditional Probability

Consider events A and B. The conditional probability of A given B, written as $P(A|B)$ and assuming $P(B) > 0$, is defined by

$$P(A|B) = \frac{P(AB)}{P(B)}. \tag{12}$$

This definition implies

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

provided that $P(A) > 0$. Note that $BA = AB$ implies $P(BA) = P(AB)$. Therefore, we arrive at

Bayes' Theorem

$$P(B|A) = \frac{P(A|B)\,P(B)}{P(A)}. \tag{13}$$

Example (1.4 Two Dice)

Suppose that an elementary event consists of rolling two dice. The outcome can be modeled as a 2-tuple (z_1, z_2) with $z_i \in \{1, 2, 3, 4, 5, 6\}$. Let Ω be the set of all possible outcomes whose cardinality is $|\Omega| = 36$. The probability associated with Ω is $P(\Omega) = 1$ (axiom 9). If every elementary outcome is equally likely, then the probability of event E is $P(E) = |E|/|\Omega|$ where $|E|$ is the cardinality of E and where the event E is an element of the power set S[a] of Ω.

Example: The power set of $\Omega = \{A, B, C\}$ contains the sets

$$\begin{array}{ccc} \{\} & \{A,B,C\} & \\ \{A\} & \{B\} & \{C\} \\ \{A,B\} & \{A,C\} & \{B,C\} \end{array}$$

[a] The set of all subsets including $\varnothing$ and Ω with cardinality 2^n.

Example (1.4 Two Dice)

Consider the two events

$A = \{(2,2),(2,4),(2,6),(4,2),(4,4),(4,6),(6,2),(6,4),(6,6)\}$ and
$B = \{(2,6),(3,5),(4,4),(5,3),(6,2)\}$

$\in S$. What is the probability, $P(A|B)$, of A given B?

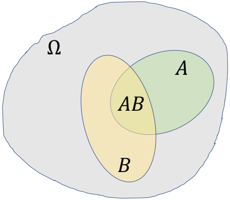
- A is the event in which each die yields an even number.
- B is the event in which the two numbers sum to 8.

What is the operational meaning of $P(A|B) = P(AB)/P(B)$?

It tells us to restrict the set of elementary events to those in event B. In effect, B assumes the rôle of Ω, but with fewer possibilities. Then, determine the fraction of the events in B that are also in A, that is, in the event AB.

The probabilities of the events A, B, and AB given our assumptionsl are:

$$P(A) = 9/36$$
$$P(B) = 5/36$$
$$P(AB) = 3/36$$



Therefore,

$$P(A|B) = P(AB)/P(B), \\ = (3/36)/(5/36), \\ = 3/5.$$

$$A = \{(2,2),(2,4),(\mathbf{2},\mathbf{6}), (4,2),(\mathbf{4},\mathbf{4}),(4,6), (\mathbf{6},\mathbf{2}),(6,4),(6,6)\}$$

$$B = \{(\mathbf{2},\mathbf{6}),(3,5),(\mathbf{4},\mathbf{4}),(5,3),(\mathbf{6},\mathbf{2})\}$$

$$AB = \{(2,6),(4,4),(6,2)\}$$

Example (1.5 Is he doomed?)

In 2016 a man returned to Italy from a trip to the United States. Just to be sure, he asked his doctor to test him for Ebola!

The test result was positive $(+)$. Should he have worried?

Here are a few pertinent facts:

1. During the 2014 - 2016 Ebola outbreak in West Africa, 4 cases of Ebola infection were reported in the United States.
2. At the time, according to World Health Organization (WHO) there was a test that correctly identified 92% of people with Ebola and correctly identified 85% who are Ebola free.

Example (1.5 Is he doomed?)

Consider the following mutually exclusive events and associated probabilities:

$$\text{event } D = \text{You are Diseased}$$
$$\text{event } H = \text{You are Healthy}$$
$$P(+|D) = 0.92 \qquad P(D) = 4/320,000,000$$
$$P(+|H) = 0.15 \qquad P(H) = 1 - P(D)$$

Example (1.5 Is he doomed?)

Bayes' theorem $P(B|A) = P(A|B)\,P(B)/P(A)$ can be generalized to

$$P(B_i|A) = \frac{P(A|B_i)\,P(B_i)}{\sum_i P(A|B_i)\,P(B_i)}, \quad \sum_i B_i = 1,$$

for mutually exclusive and exhaustive events B_i. For this problem, we can write

$$P(D|+) = \frac{P(+|D)\,P(D)}{P(+|D)\,P(D) + P(+|H)\,P(H)} \approx \frac{P(+|D)}{P(+|H)}\,P(D)$$
$$\approx 1/13,000,000.$$

Our conclusion? No he was not doomed! Just exceedingly cautious! *A priori* our cautious friend would have had a 10 times higher chance of drowning in his hotel bathtub in the US than catching Ebola in 2016!

Outline

- The modern theory of probability defines the latter as a measure that satisfies the Kolmogorov axioms.
- However, if a problem can be broken down into outcomes that are judged to be equally probable, then the classical approach to probability can be used. This typically involves subtle combinatorial reasoning.
- But to be useful probability must be interpreted.
- The two most common interpretations are: relative frequency and degree of belief.