INFN School of Statistics 2022
Paestum, Italy

Probability Theory
Part 1

Harrison B. Prosper

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

16 May, 2022

Outline

1. Introduction
   - Learning To Count
2. Axioms
   - Boolean Algebra
   - Kolmogorov Axioms
3. Conditional Probability
   - Bayes Theorem
4. Summary

Outline

1. Introduction
   - Learning To Count
2. Axioms
   - Boolean Algebra
   - Kolmogorov Axioms
3. Conditional Probability
   - Bayes Theorem
4. Summary

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, assumed equally probable, then $P(A) = n/m$.
2. relative frequency If n things can happen out of m possible things, then $z = n/m$ is the relative frequency with which the n things have occurred. The probability $P(A)$ is the limit (suitably defined) as $n, m \to \infty$ of the relative frequency. 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 correspondence together with the earlier work of 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 ...*
>
> *...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 with probability q or ≥ 1 sixes with probability p, are exhaustive — i.e., they are the only possible outcomes.
2. There are 6 possible elementary outcomes[2] of which only 1 is a six.
3. The 6 elementary outcomes are equally probable.
4. The elementary outcomes remain equally probable 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 considered equally probable;
- count the total number of possible outcomes, that is, determine the cardinality (size) of the sample space;
- count the number of favorable outcomes
- and take the ratio of the two counts as the probability of the favorable 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 favorable 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 at least 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 is an n-tuple, $(z_1, \ldots, z_n)$, each element of which is an integer $z_i \in \{1, \ldots, 365\}$.

Let M be the cardinality of the set of n-tuples and N the same for n-tuples with $\geq$ two identical entries. In such problems, as in the previous example, it is typically 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 each n-tuple to be 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 in these n-tuples 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 until we reach the nth slot, which can be filled in $365 - (n - 1)$ ways. The size of set A is therefore $K = 365 \times 364 \times \cdots \times (365 - n + 1) = 365!/(365 - n)!$.

Therefore, the probability of at least one duplicate birthday is

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

In general, counting involves permutations and combinations...

[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 each set of k items, there will be $k!$ permutations that consist 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 now

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

This is called the number of combinations.

Example (1.3 Partitions)

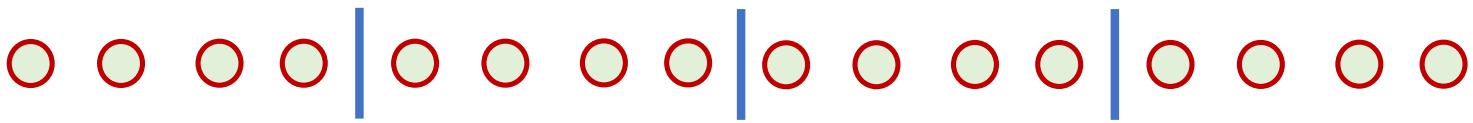
A partition divides objects into groups. Here is an example[a].

16 students, 4 of whom are graduate students, are 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.

Imagine arranging the students in a row



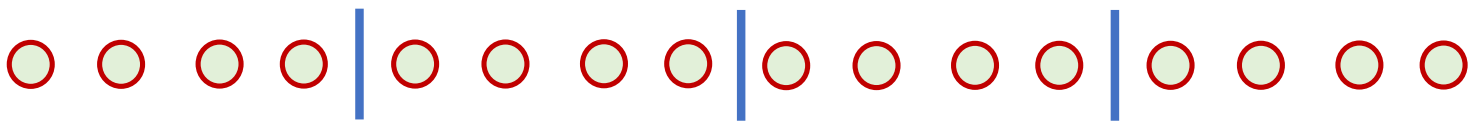
If the order of the students were relevant, there would be 16! ways to arrange the students.

[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 $16!$ overcounts the number of partitions by $4!$ and that this is true for every group.

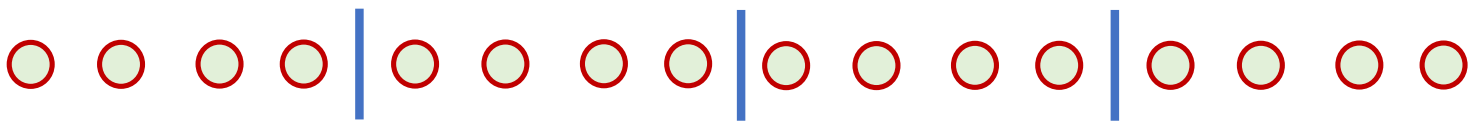
Therefore, the total overcount is $(4!)^4$. Hence, the number of distinct partitions for which the order of students within each group is irrelevant, that is, the cardinality of the sample space, is

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

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.



Now, we need the number of partitions with a graduate student in each group. Given this count, the desired probability follows from Assumption 1.

Exercise 1.1

Compute the cardinality of this set and show that the probability that each group includes a graduate student is

$$\frac{12 \times 8 \times 4}{15 \times 14 \times 13}$$

Example (1.4 An Act of Desperation)

On October 19, 1900, at a meeting of the German Physical Society, Max Planck presented a successful formula for the black body radiation spectrum. The challenge now was to derive his formula from first principles.

About a month later, he had a derivation based on the following assumptions:

Assumptions

1. The radiation and oscillators within a cavity are in thermal equilibrium at temperature T.
2. Radiation is absorbed or emitted by an oscillator in quanta of size $\epsilon = h\nu$. He considered the latter to be a *purely formal assumption* that he hoped he could dispense with later.
3. Given this assumption, the system comprises m indistinguishable quanta with energy ϵ in thermal equilibrium with n indistinguishable oscillators whose energies can be zero up to $m\epsilon$ quanta.

Example (1.4 An Act of Desperation)

The derivation required computing the average energy per oscillator, which reduced to a counting problem: how many ways Ω can m indistinguishable quanta be distributed among the n indistinguishable oscillators?

Given this number, the entropy S of the system can be calculated using Ludwig Boltzmann's formula

$$S = k \log \Omega,$$

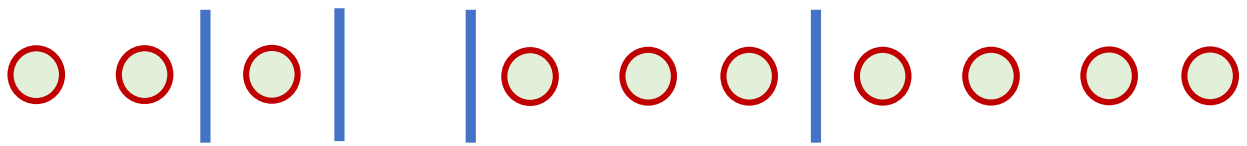
which, through the relation,

$$\frac{\partial S}{\partial E} = \frac{1}{T},$$

provides the link to thermodynamics.

Example (1.4 An Act of Desperation)

Like the student problem, imagine arranging the m quanta in a row with $n-1$ walls between them. In the figure below, we have $m = 10$ and $n - 1 = 4$.



The walls divide the quanta into n oscillators and yield a total of $m + n - 1$ items that can be arranged in $(m + n - 1)!$ ways. But, we need to count only the *distinguishable* permutations. Since the quanta and the walls are indistinguishable, we must divide by $m! \times (n-1)!$. Therefore, we conclude that

$$\Omega = \frac{(m+n-1)!}{m!\,(n-1)!}.$$

For large n and m this yields an expression for the entropy per oscillator which Planck compared with the entropy he derived from his radiation formula...and the rest, as they say, is history!

Outline

1. Introduction
   - Learning To Count
2. Axioms
   - Boolean Algebra
   - Kolmogorov Axioms
3. Conditional Probability
   - Bayes Theorem
4. Summary

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 $\mathbb{B}$, including the special sets 0 and 1, equipped with an equivalence relation $(=)$ and the operations OR $(+)$, AND $(\bullet)$, and NOT $(\bar{})$.

An equivalence relation $\sim$ obeys the rules:

1. $a \sim a$
2. $a \sim b$ and $b \sim a$
3. $a \sim b$, $b \sim c$, and $a \sim c$.

Example (Crazy Boris)

"as crazy as" is an equivalence relation $\sim$[a]

1. Boris $\sim$ Boris
2. Boris $\sim$ Aardvark and Aardvark $\sim$ Boris
3. Boris $\sim$ Aardvark, Aardvark $\sim$ Zorg, and Boris $\sim$ Zorg.

[a] With apologies to all persons named Boris, Aardvark, or Zorg!

Axioms of Boolean Algebra (Huntington)

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

$$A + B = B + A \quad (1)$$
$$A + (BC) = (A + B)(A + C) \quad (2)$$
$$A + 0 = A \quad (3)$$
$$A + \bar{A} = 1 \quad (4)$$
$$AB = BA \quad (5)$$
$$A(B + C) = AB + AC \quad (6)$$
$$A1 = A \quad (7)$$
$$A\bar{A} = 0 \quad (8)$$

We also assume the "meta" axiom $(A) = A$ and we can assign values, e.g., $A = B$.

Here are some 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.

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

□

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

$$A0 = 0,$$
$$A + AB = A,$$
$$A(A + B) = A.$$

Justify every step with one or more axioms including assumptions such as $(A) \to A$.

Kolmogorov Axioms

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

$$P(A) \geq 0 \tag{9}$$
$$P(\Omega) = 1 \tag{10}$$
$$P(A + B) = P(A) + P(B) \quad \forall AB = \emptyset. \tag{11}$$

If $AB = \emptyset$, 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:

$$P(\emptyset) = 0$$
$$P(A + B) = P(A) + P(B) - P(AB)$$
$$P(A_1 + \cdots + A_n) = P(A_1) + \cdots P(A_n) \quad \forall A_i B_j = \emptyset, i \neq j.$$

Lemma (3)

$P(\emptyset) = 0$

Proof.

The lemma $A0 = 0$ implies $\Omega\emptyset = \emptyset$, the startling conclusion that events Ω and $\emptyset$ are mutually exclusive! Therefore,

$$\begin{aligned} P(\Omega + \emptyset) &= P(\Omega) + P(\emptyset) && \text{(axiom 11)} \\ P(\Omega) &= P(\Omega) + P(\emptyset) && \text{(axiom 3)} \\ \therefore P(\emptyset) &= 0 && \text{(since } P \text{ is a finite real number)} \end{aligned}$$

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

1. Introduction
   - Learning To Count
2. Axioms
   - Boolean Algebra
   - Kolmogorov Axioms
3. Conditional Probability
   - Bayes Theorem
4. Summary

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 $P(A) > 0$. Since $BA = AB$, $P(BA) = P(AB)$. Therefore, we arrive at

Bayes' Theorem

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

Example (1.5 Two Dice)

The elementary events consist of rolling two 6-sided dice with outcomes modeled as 2-tuples (z_1, z_2), $z_i \in \{1, 2, 3, 4, 5, 6\}$ that form the set Ω, that is, the sampling space of cardinality $|\Omega| = 36$. The only probability associated with Ω is $P(\Omega) = 1$ (axiom 9). However, if we assume each outcome of the two dice to be equally likely, then to each event E in the power set S^a of Ω we can assign the probability $P(E) = |E|/|\Omega|$, where $|E|$ denotes the cardinality of a set within the power set of Ω.

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

Example (The power set of $\{A, B, C\}$ contains)

$$\{\} \quad \{A, B, C\}$$
$$\{A\} \quad \{B\} \quad \{C\}$$
$$\{A, B\} \quad \{A, C\} \quad \{B, C\}$$

Given events

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

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

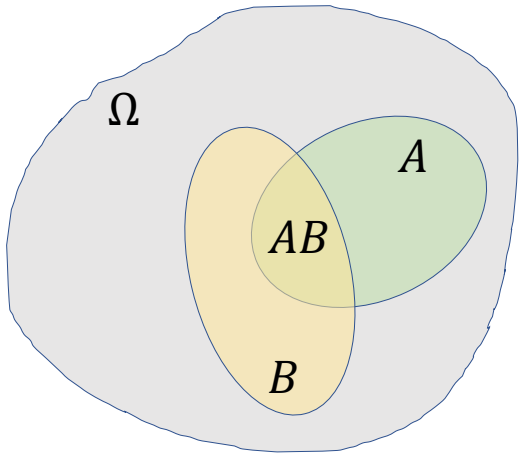
- A is the event in which each die yields an even number.
- B is the event in which the two numbers sum to 8.

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 what fraction of the events in B are also in A, that is, in the event AB.

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

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



Therefore,

$$\begin{aligned} P(A|B) &= P(AB)/P(B), \\ &= (3/36)/(5/36), \\ &= 3/5. \end{aligned}$$

$$\begin{aligned} A = \{&(2,2),(2,4),\mathbf{(2,6)}, \\ &(4,2),\mathbf{(4,4)},(4,6), \\ &\mathbf{(6,2)},(6,4),(6,6)\} \end{aligned}$$

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

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

Example (1.6 Ebola)

You've just returned to Italy from a visit to the US. Just to be sure, you aske your doctor to test you for Ebola! The test result is positive ($+$).
Should you worry? Well, it depends...
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. According to WHO there is a test that correctly identifies 92% of people with Ebola and correctly identifies 85% who are Ebola free.

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 (1.6 Are You Doomed?)

Bayes' theorem can be generalized to

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

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

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

So we conclude that **No, you are not doomed! You are merely paranoid!** Had you been in the US in 2016, *a priori*, your chance of drowning in a bathtub would have been 10 times higher!

[a] $\sum_i B_i = 1$

Outline

1. Introduction
   - Learning To Count
2. Axioms
   - Boolean Algebra
   - Kolmogorov Axioms
3. Conditional Probability
   - Bayes Theorem
4. Summary

- 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.