1 Probability Spaces

Probability is the best tool we currently have for predicting future events based on past observations. To the extent that we believe physical processes are guided by certain (often hidden) equations, we can turn to probabilistic techniques to help bescribe, extrapolate and infer long-term behaviors.

Probability assumes a predictability to the world while also allowing an element of chance/noise/randomness/chaos. It describes long-term trends and average behavior while also quantifying the extent to which short-term behavior can be expected to deviate. While I may not be able to predict the temperature on March 21 to within 20 degrees, I can fairly certainly predict the average for May to within 1 or 2 degrees.

Even in today's news, it has been observed that the red giant Betelgeuse (α-Orionis) is undergoing a precipitous decrease in apparent brightness. Some have speculated that a supernova is on the horizon. Data exists tracking Betelguese's magnitude for decades so there is some precedent for concluding "this is not normal", but the precise determination of how abnormal this behavior is falls to the realm of probability and statistics. There are physical laws, some of which we know, that govern the brightness of the stars, in addition to other unknown and perhaps unknowable factors.

The end goal of probability and statistics is to formalize a logical

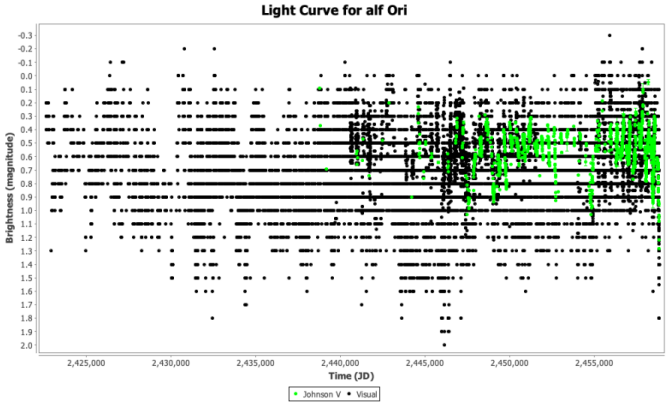

Figure 1.1: Magnitude of α-Ori, 1890-2020

method by which we can reasonably state "given these assumptions and these observations, I draw this conclusion," with a measured amount of confidence.

1.1 Experiments and such

If you flip a coin and it lands on heads, what is the probability it will land on heads if you flip it a second time? One-half you may say? There are at least three reasonable answers, all different.

If you assume the coin to be fair (which was never stated and difficult to prove), then you may believe 1/2 is the right probability and, perhaps, no amount of evidence to the contrary would persuade you to believe otherwise, because the statement of "fairness" is the one that defines your calculation.

Yet there is another method based on Bayesian Inference[1] that says, "I don't know the probability of heads before I flip the coin, other than it may be between 0 and 1, with equal probability. Since I have seen one appearance of heads, I can calculate the probability of a second heads to be 2/3."[2]

And a third method, called maximum likelihood estimation[3], essentially makes the argument "I've seen heads once, and tails never, so I predict the coin will always land on heads" and assigns the value of 1 to the probability.

Which answer is right? Are any of them? All of them? In a sense, it is not the job of probability to decide for you which answer is "right." Each follows from a set of assumptions about the coin and the world in general. Once the underlying assumptions are stated and the goal is defined, then probability can guide us through the calculations.

Of course, to ascertain the true bias of the coin, one would likely try to flip it several times and count heads versus tails. Even then, does 506 heads out of 1000 flips prove fairness? Or bias? Does 10 heads in a row indicate tails is never going to appear? Only approximate conclusions can be drawn from these observations, but the reason we even believe such a process to be informative is one of the assumptions underlying all of probability theory, namely that something about long term behavior...

[1] which we will study in Chapter 2
[2] Never mind the exact calculation; we'll get to it in time.
[3] covered at the end of this course

1.1.1 Definitions

To begin we take as undefined the terms "procedure" and "outcome." Attempts to define them end up circular or mathematically non-rigorous and add nothing to our understanding. Each is understood as you normally understand them!

An **experiment** is defined as a procedure which results in a specific **outcome**. Simple examples include flipping a coin, rolling three dice. More complicated are measuring the stopping distance of a car from a certain speed, or determining the mass of a proton.

The **sample space**, sometimes denoted Ω, is the set of all possible outcomes from a given experiment. We will see specific examples below.

An **event** is a subset of the sample space. It could be as small as one outcome, or as large as all of Ω.

Finally, a **probability function** assigns a number $P(E)$ to each event $E \subseteq \Omega$.

1.1.2 Examples

Example 1. One coin toss Toss a fair coin. The experiment is to record which side lands up: heads or tails. The sample space is $\{H, T\}$. The probability function assigned to a *fair* coin would be $P(H) = P(T) = \frac{1}{2}$.

Example 2. Three coin tosses Toss a coin three times and record which side lands up. The sample space is

$$\Omega = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}.$$

Examples of *events* are 'two heads', 'an odd number of tails', 'more heads than tails'. One probability function would assign $1/8$ to each outcome, making the coin fair. Another function could assign $\frac{1}{2}$ to the outcome HHH and $\frac{1}{16}$ to each of the other outcomes[4].

Example 3. Proton mass Measure the mass of a proton, in grams. The sample space is, perhaps surprisingly, all non-negative reals $\Omega = [0, \infty)$. While we all know a proton will never weigh 1 gram, it is mathematically more pleasing to leave the upper-bound unspecified

[4] "This doesn't make sense," you may protest, because if $P(HHH) = \frac{1}{2}$ then $P(H)$ must be $\frac{1}{\sqrt[3]{2}}$. You'd be right if the flips were known to be independent, which in practice they usually are. But it's not strictly required for our probability function to assume independence of the individual coin tosses. More on independence in a later section.

and allow the probability to fade away to negligible amounts, rather than abruptly stop the domain at some pre-determined amount. The same occurs in many applcations: actuaries preparing life-expectancy tables will allow for a person to live for 1000 years, with ridiculously low probability, just as a traffic analyst will consider equally negligible the case that one million cars cross an intersection during a 5-minute interval.

An event in this sample space could be an interval such as "the mass is between $1.6726219 \times 10^{-24}$ and $1.6726220 \times 10^{-24}$ grams." Another event is "the mass is an even number," although this event has zero probability.

Example 4. Roll two dice In this experiment you roll two identical, six-sided, fair dice simultaneously and record the numbers that appear on the top of each[5]. In this example, the sample space depends on what you do with those two numbers. You have at least three choices

1. Label one die A and one B and record the numbers appearing on each
2. Record the two numbers appearing, without distinguishing the two dice
3. Record the sum of the two numbers

In the first case the sample space is all ordered pairs

$$\Omega = \{(x, y) \quad | \quad 1 \leq x, y \leq 6\}$$

and thus has order (*size*) 36. In the second case the sample space is *unordered* pairs

$$\Omega = \{(x, y) \quad | \quad 1 \leq x \leq y \leq 6\}$$

where we adopt the convention of recording the smaller of the two results first. This space has order 21.[6] Finally in the third case the sample space is simply the 11-element set

$$\Omega = \{2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12\}$$

If the die are fair then the appropriate probability functions to assign to each of the sample spaces are

[5] Actually you're recording the number of "pips" present on the top of each die, a term I found out most of my students didn't know when I included it on a unit test in the fall term and fielded a dozen questions about it.

[6] Add the $\binom{6}{2}$ pairs where $x \neq y$ to the 6 pairs where $x = y$

1. $P(\omega) = \frac{1}{36}$ for all $\omega \in \Omega$
2. $P(x, y) = \begin{cases} \frac{1}{18} & x < y \\ \frac{1}{36} & x = y \end{cases}$
3. $P(x) = \dfrac{6 - |7 - x|}{36}$

1.1.3 Probability Functions

We have seen examples of probability functions in the preceding section. Now we'll develop a rigorous definition.

Definition Given a sample space Ω, the **class of events** $\mathcal{F}$ is a class of subsets of Ω that form a *sigma algebra*. That is, they are closed under complementation and countable union.

This is a bit of a mathematical formality that we need to give a good definition of a probability function. In just about every case we encounter (if not *really every* case), the class of events $\mathcal{F}$ is just the powerset of Ω, that is, the set of all subsets of Ω. The important thing about sigma-algebras is the closure properties.

Definition Given a sample space Ω and an event class $\mathcal{F}$, a **probability function** on $\mathcal{F}$ is a function that assigns a real number to each element $E \in \mathcal{F}$ such that

1. $P(E) \geq 0$ for every $E \in \mathcal{F}$
2. $P(\Omega) = 1$
3. If $E_1, E_2, \ldots$ are disjoint sets in $\mathcal{F}$ then
$$P(E_1 \cup E_2 \cup \cdots) = P(E_1) + P(E_2) + \cdots$$

Notice that nowhere in the definition do we claim that P says anything about the "probability" of anything occurring. That would get us into a big mess trying to define probability in terms of probability and also giving it a tangible interpretation. For now, it is simply a type of function.[7]

Finally we can give a formal definition of a Probability Space.

Definition A **probability space** is a triple $(\Omega, \mathcal{F}, P)$ where Ω is a set, $\mathcal{F}$ is a sigma-algebra of subsets of Ω and P is a probability function on $\mathcal{F}$

[7] But we will see that according to provable theorems like the Law of Large Numbers, this definition of probability implies that it behaves like we want a "probability" function to behave.

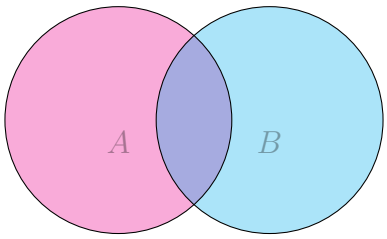

Figure 1.2: Diagram for Theorem 1.3

1.1.4 Properties of Probability Functions

The proofs of these properties are left as an exercise.

Theorem 1.1. *The following properties can be proven from the above definition of a probability function.*

1. $P(A^C) + P(A) = 1$ *where* A^C *is the complement of* A*, that is, everything in the set* $\Omega - A$
2. $P(\emptyset) = 0$
3. *If* $A_1 \subseteq A_2$ *then* $P(A_1) \leq P(A_2)$
4. $P(A_1 \cup A_2) = P(A_1) + P(A_2) - P(A_1 \cap A_2)$

1.1.5 Algebra of Sets

Theorem 1.2 (DeMorgan's Laws)**.** *For sets* A, B *the following are true*

- $(A \cup B)^C = A^C \cap B^C$
- $(A \cap B)^C = A^C \cup B^C$

Theorem 1.3 (Principle of Inclusion/Exclusion)**.**

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

1.1.6 Exercises

1. Verify the following relations
   a) $(A \cup B)' = A'B'$
   b) $AA = A = A \cup A$

c) $(A \cup B) - AB = AB' \cup A'B$

d) $(A \cup B)C = AC \cup BC$

e) $(A \cup B) - B = A - AB = AB'$

f) $(A - AB) \cup B = A \cup B$

g) $A' \cup B' = (AB)'$

2. Find simple expressions for

a) $(A \cup B)(A \cup B')$

b) $(A \cup B)(A' \cup B)(A \cup B')$

c) $(A \cup B)(B \cup C)$

3. State which of the following are correct and which are incorrect

a) $(A \cup B) - C = A \cup (B - C)$

b) $ABC = AB(C \cup B)$

c) $A \cup B \cup C = A \cup (B - AB) \cup (C - AC)$

d) $A \cup B = (A - AB) \cup B$

e) $AB \cup BC \cup CA \supset ABC$

f) $(AB \cup BC \cup CA) \subset (A \cup B \cup C)$

g) $(A \cup B) - A = B$

h) $AB'C \subset A \cup B$

i) $(A \cup B \cup C)' = A'B'C'$

j) $(A \cup B)'C = A'C \cup B'C$

k) $(A \cup B)'C = A'B'C$

l) $(A \cup B)'C = C - C(A \cup B)$

4. Prove Theorem 1.1

5. Give an expression for $P(A \cup B \cup C)$ analogous to the one given for two sets in the text. (See fig 1.3)

1.2 Examples of Probability Spaces

Before developing more theory, let's get our hands dirty with some simple examples of discrete and continuous probability spaces.

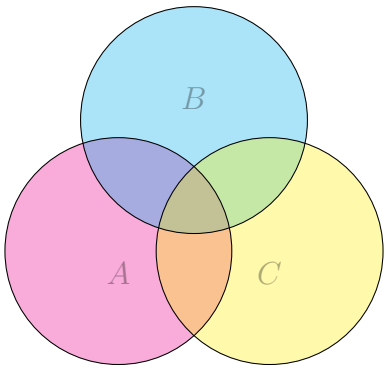

Figure 1.3: Diagram for problem 5

1.2.1 Discrete equiprobable spaces

These are the bread-and-butter spaces of basic probability, and, at the same time, the field admits problems that can get quite complicated. Classic problems about rolling dice, flipping coins, pulling marbles out of bags, etc. all are examples of discrete equiprobable spaces because the sample space of possible outcomes is discrete and each outcome is ideally assumed to be equally likely. That is, if Ω contains n points, then $P(\omega) = \frac{1}{n}$ for all $\omega \in \Omega$. Similarly if an event E contains r events, then $P(E) = \frac{r}{n}$.

Example 5. Select a card at random from a deck of 52 cards. Let A be the event 'the card is a spade' and B be the event 'the card is a face card (J,Q,K).' Compute $P(A), P(B), P(A \cup B), P(A \cap B)$

Solution A has size 13 and B has size 16. So $P(A) = \frac{13}{52}, P(B) = \frac{12}{52}$. $A \cap B$ has 3 cards in it $\{J\spadesuit, Q\spadesuit, K\spadesuit\}$ so $P(A \cap B) = \frac{3}{52}$. Finally

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) = \frac{22}{52}$$

Example 6. Select two items at random from a lot containing 12 items, of which four are defective. Compute the probability that

- Neither is defective
- Both are defective
- At least one is defective

Solution The sample space is every possible way of selecting two light bulbs from a lot of 12, namely

$$\binom{12}{2} = 66.$$

- Two non-defective light bulbs can be selected in $\binom{8}{2} = 28$ ways, giving a probability of $\frac{28}{66}$
- Two defective light bulbs can be selected in $\binom{4}{2} = 6$ ways, giving a probability of $\frac{6}{66}$
- This event is the complement of "neither is defective", so the probability is $1 - \frac{28}{66} = \frac{38}{66}$

It is worth pointing out at this time that this solution method considers an unordered sample space or, in other words, the light bulbs are un-labeled so the only way to distinguish two events is by the *number* of non-defective and defective light bulbs and not the order in which they were chosen. It is possible to solve this same problem with an *ordered* sample space, which would have size $12 \cdot 11 = 132$. Then the event 'two defective' corresponds to $4 \cdot 3 = 12$ elements, giving a probability of $\dfrac{12}{132} = \dfrac{6}{66}$. You see that ordered sample spaces provide the same correct answer, as long as you are consistent.

Example 7. Birthday Problem In a classroom of 27 students, what is the probability that at least two people have the same birthday?

Solution We will ignore leap-years and determine the size of the sample space is the number of ways to make 27 selections from 365 days, which is 365^{27}. (This is called selecting with repetition).

Now the event 'at least two are the same' is again complementary to the event 'all birthdays are distinct.' To have all birthdays distinct, the first person may have any of 365 birthdays, the second only can be from 364 remaining days, the third from 363, etc. So there are

$$365 \cdot 364 \cdot 363 \cdots (365 - 27 + 1) = \frac{365!}{338!}$$

events corresponding to all birthdays distinct.

So the probability of at least two birthdays the same is

$$P = 1 - \frac{365!}{338! \cdot 365^{27}} = 0.627,$$

which is pretty good odds!

1.2.2 Continuous Sample Spaces

Example 8. Two points a and b are selected at random such that $b \in [-2, 0]$ and $a \in [0, 3]$. Find the probability that $|a - b| > 3$

Solution In the ab plane, the sample space consists of the 3×2 rectangle between $(0, 0)$ and $(3, -2)$. The event desired is the subset of points for which $a - b > 3$, which defines a triangle below $a - b = 3$ and inside the rectangle. The ratio of the areas gives the probability:

$$p = \frac{2}{6} = \frac{1}{3}$$

Example 9. A point is selected at random inside a circle. Find the probability that the point is closer to the center than the circumference.

Answer $\frac{1}{4}$

Example 10. Let X denote the lattice of points in the cartesian plane where both co-ordinates are integers. A coin of diameter $\frac{1}{2}$ is tossed onto the (infinite) plane. What is the probability that the coin covers a point in X?

Answer $\frac{\pi}{16}$

Example 11. Three points a, b and c are selected at random from the circumference of a circle. Find the probability that the points lie on a semicircle.

Answer $\frac{3}{4}$

Example 12. A stick of unit length is broken randomly into three pieces. (Specifically, two points a, b are chosen at random on the stick and it is cut at these two points.) What is the probability that the three stick pieces can be formed into a triangle?

Solution TBD

1.3 Combinatorics Excursion

1.3.1 The Fundamental Principle of Counting

Combinatorics is the study of counting arrangements or structures. We'll review just a small bit of combinatorics in the section in case the reader needs a refresher, or perhaps a first introduction.

It begins with the following theorem about selecting items from sets

Theorem 1.4 (Fundamental Principle of Counting)**.** *Let $A_1, A_2, \ldots, A_n$ be a collection of non-empty sets. The number of ways, n, of selecting one item from each set is equal to*

$$n = |A_1| \cdot |A_2| \cdots |A_n|$$

Proof Consider a tree with a root R, at level 0. At level 1, place each of the elements of A_1, and make each a descendant of R. The tree now has A_1 leaves corresponding to the ways to select one item from A_1. Underneath each $a \in A_1$, now add a leaf at level 2 for each element in A_2. The tree now has $|A_1| \cdot |A_2|$ leaves and each leaf corresponds to a selection of two elements: one each from A_1 and A_2. Continue this process through A_n and the n-level tree's leaf-count completes the proof

Example 13. How many positive divisors does 720 have?

Solution 720 can be factored into $720 = 2^4 3^2 5$. Any positive divisor d must be of the form $2^{e_2} 3^{e_3} 5^{e_5}$ where $e_2 \in \{0, \ldots, 4\}$, $e_3 \in \{0, 1, 2\}$, $e_5 \in 0, 1$. There are 30 such divisors.

1.3.2 Ordered samples with replacement

Given a set of n distinct elements (like numbered marbles), an ordered selection of size r with replacement corresponds to selecting one of the n elements uniformly at random, recording its value in a list, replacing it and selecting another elements and recording its value as the second element in the list, and so on, until r elements are listed. The list constitutes an ordered sample. There are n^r such lists, according to the fundamental principle of counting (Theorem 1.4).

1.3.3 Ordered samples without replacement

Given a set of n distinct elements (like numbered marbles), an ordered selection of size r without replacement corresponds to selecting one of the n elements uniformly at random, putting the element in a list and not returning it to the set, selecting a second, and so on until r elements are in the list. In this case, the fundamental principle tell us there are n selections for the first element, $(n-1)$ for the second and so on until $(n-r+1)$ for the r-th element. The number of these lists is given by $n(n-1)(n-2)\cdots(n-r+1) = \dfrac{n!}{(n-r)!}$. A common notation for this is P_r^n and also $n^{\underline{r}}$.

1.3.4 Unordered samples without replacement

Unordered sampling corresponds to putting the selected elements into a bag, or set, instead of a list. With ordered sampling, $[6,4,1]$ is distinct from $[4,6,1]$ but now the two are the same as they both form the set $\{1,4,6\}$. Since any set of r distinct elements can be arranged into $r!$ different lists, the number of ordered samples with replacement of size r must be a factor of $r!$ larger than the number of unordered samples with replacement. Therefore the number we seek is $n^{\underline{r}}/r! = \dfrac{n!}{r!(n-r)!} = \binom{n}{r}$, the familiar binomial coefficient.

1.3.5 Unordered samples with replacement

Let's develop this idea with a specific example. Given the set a,b,c,d,e we want to select an unordered sample of size 12, with replacement. Since the set is unordered we can assume it to be sorted, *e.g*

$$(a,a,a,b,c,d,d,d,d,e,e,e).$$

This selection is equivalent to the list $(3,1,1,4,3)$, where each number counts the occurrence of the corresponds letters in the sorted original set. Let's now describe this list with symbols "XXX.X.X.XXXX.XXX", that is 12 X's, and 5-1=4 dots. The number of X's give the frequency of each letter. Now we claim that any permutation of 12 X's and 4 dots corresponds to a list of numbers (n_a,n_b,n_c,n_d,n_e) which encodes exactly one unordered sample of size 12. Furthermore this encoding is reversible – each sample corresponds to a string that is a permutation of "XXXXXXXXXXXX....". A permutation of this string is equivalent to selecting, from among 16 locations, where the 4 dots will go, and there are $\binom{16}{4}$ ways to do that.

By analogy this argument can be extended easily to show the number of unordered samples with replacement from n elements is given by

$$\binom{n-1+r}{n-1} = \binom{n-1+r}{r}$$

where the equality follows from the symmetry of the binomial coefficient.

1.4 Conditional Probability and Independence

Sometimes we are interested in the outcomes from a subset of the sample space.

1.5 Bayes Theorem