Probability Theory
Lecture Notes

Patrick White

February 15, 2020

Contents

Preface

These notes were prepared from lectures given in 2019-2020 at Thomas Jefferson High School for Science and Technology. Many thanks to the students of this class and especially the (to be determined) scribes who produced much of this document purely for the love of the science and the benefit of future classes.

1 Introduction

Hello. These are lecture notes in probability theory. This is a collaborative book with student contributors based on Dr. White's lectures.

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

2.1 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]

[1] which we will study in Chapter 2

[2] Never mind the exact calculation; we'll get to it in time.

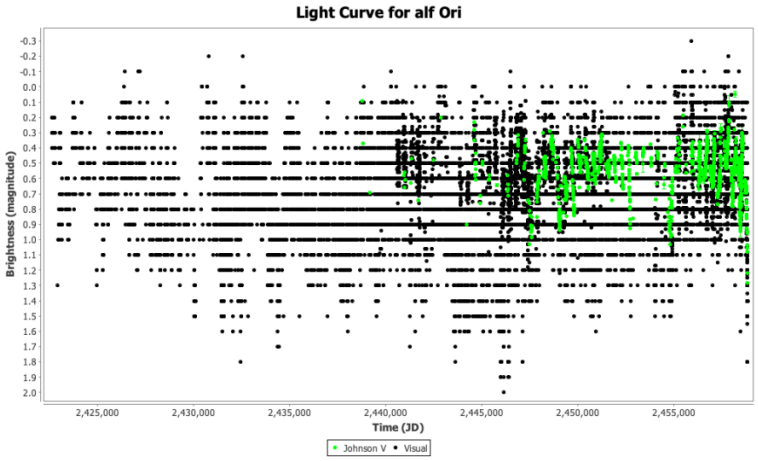

Figure 2.1: Magnitude of α-Ori, 1890-2020

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 long term behaviors can be estimated and predicted.

2.1.1 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.

[3] covered at the end of this course

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 $E \subseteq \Omega$.

2.1.2 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 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 applications: 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

[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.

[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.

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

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. $P(x) = \dfrac{6 - |7 - x|}{36}$

2.1.3 Probability Functions

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

Definition 2.1. *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.

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

Definition 2.2. *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 2.3. *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 $\mathcal{F}$*

2.1.4 Properties of Probability Functions

The proofs of these properties are left as an exercise.

Theorem 2.1. *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. $P(A_1 \cup A_2) = P(A_1) + P(A_2) - P(A_1 \cap A_2)$

2.1.5 Algebra of Sets

Theorem 2.2 (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 2.3 (Principle of Inclusion/Exclusion)**.**

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

[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.

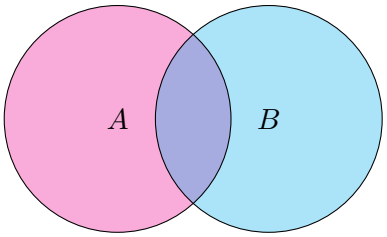

Figure 2.2: Diagram for Theorem 2.3

2.1.6 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 2.3)

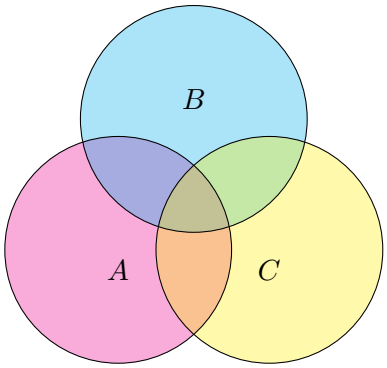

Figure 2.3: Diagram for problem 5

2.2 Examples of Probability Spaces

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

2.2.1 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 $\frac{12}{132} = \frac{6}{66}$. 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!

2.2.2 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 coordinates 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?

Answer $\frac{1}{4}$.

Solution First, consider the probability that $a < b$. By symmetry, this happens with probability $\frac{1}{2}$.

Using the triangle inequality under this assumption gives the three relations $a < \frac{1}{2}$, $b > \frac{1}{2}$, $\frac{1}{2} > b - a$. The probability that a is randomly chosen to be less than $\frac{1}{2}$ is $\frac{1}{2}$, and the probability that b satisfies these constraints is a, so the overall probability is simply $\frac{a}{2}$.

This next step can be made more rigorous with random variables and an expected value argument, but on average, given that $a < \frac{1}{2}$, a is expected to take on the value $\frac{1}{4}$. Hence, the probability the three pieces form a triangle with a, b, $a < b$ is $\frac{1}{8}$. Undoing the condition gives the probability as $\frac{1}{4}$.

2.3 Combinatorics Excursion

2.3.1 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 2.4 (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 $720 = 2^4 \cdot 3^2 \cdot 5$. Any positive divisor d must be of the form $2^{e_2} \cdot 3^{e_3} \cdot 5^{e_5}$ where $e_2 \in \{0, \ldots, 4\}$, $e_3 \in \{0, 1, 2\}$, $e_5 \in 0, 1$. There are 30 such divisors.

2.3.2 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 2.4).

2.3.3 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 $n^{\underline{r}}$.

2.3.4 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 $\dfrac{n^{\underline{r}}}{r!} = \dfrac{n!}{r!(n-r)!} = \binom{n}{r}$, the familiar binomial coefficient.

2.3.5 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.

2.3.6 Exercises

1. If a 3-digit number (000 to 999) is chosen at random, find the probability that exactly 1 digit will be > 5.

2. Find the probability that a five-card poker hand will be: (a) A straight (five cards in sequence regardless of suit; ace may be high but not low). (b) Three of a kind (three cards of the same face value x, plus two cards with face values y and 2, with x, y, z distinct). (c) Two pairs (two cards of face value x, two of face value y, and one of face value z, with x, y, z distinct).

3. An urn contains 3 red, 8 yellow, and 13 green balls; another urn contains 5 red, 7 yellow, and 6 green balls. One ball is selected from each urn. Find the probability that both balls will be of the same color.

4. An experiment consists of drawing 10 cards from an ordinary 52-card pack. (a) If the drawing is done with replacement, find the probability that no two cards will have the same face value. (b) If the drawing is done without replacement, find the probability that at least 9 cards will be of the same suit.

5. An urn contains 10 balls numbered from 1 to 10. Five balls are drawn without replacement. Find the probability that the second largest of the five numbers drawn will be 8.

6. m men and w women seat themselves at random in $m+w$ seats arranged in a row. Find the probability that all the women will be adjacent.

7. If a box contains 75 good light bulbs and 25 defective bulbs and 15 bulbs are removed, find the probability that at least one will be defective.

8. Eight cards are drawn without replacement from an ordinary deck. Find the probability of obtaining exactly three aces or exactly three kings (or both).

9. (The game of *recontre*). An urn contains n tickets numbered $1, 2, ..., n$. The tickets are shuffled thoroughly and then drawn one by one without replacement. If the ticket numbered r appears in the rth drawing, this is denoted as a match (French: *rencontre*). Show that the probability of at least one match is

$$1 - \frac{1}{2!} + \frac{1}{3!} - \cdots + \frac{(-1)^{n-1}}{n!} \to 1 - e^{-1} \qquad \text{as } n \to \infty$$

10. A "language" consists of three "words," $W_1 = a, W_2 = ba, W_3 = bb$. Let $N(k)$ be the number of "sentences" using exactly k letters (e.g., $N(1) = 1$ (a), $N(2) = 3$, (aa, ba, bb), $N(3) = 5$, (aaa, aba, abb, baa, bba); no space is allowed between words).
    a) Show that $N(k) = N(k-1) + 2N(k-2), k = 2, 3, ...$ (define $N(O) = 1$).
    b) Show that the general solution to the second-order homogeneous linear difference equation [with $N(O)$ and $N(1)$ specified], is $N(k) = A2^k + B(-1)^k$, where A and B are determined by $N(O)$ and $N(1)$. Evaluate A and B in the present case.

2.3.7 Solutions to Selected Exercises

Solution to 1.3.6, Ex. 9 (The game of *recontre*). An urn contains n tickets numbered $1, 2, ..., n$. The tickets are shuffled thoroughly and then drawn one by one without replacement. If the ticket numbered r appears in the rth drawing, this is denoted as a match (French: *rencontre*). Show that the probability of at least one match is

$$1 - \frac{1}{2!} + \frac{1}{3!} - \cdots + \frac{(-1)^{n-1}}{n!} \to 1 - e^{-1} \qquad \text{as } n \to \infty$$

Remark. This problem is known by many other names – it is sometimes called the *hat problem*, or it is the *derangements* problem.

Solution We proceed with the Principle of Inclusion-Exclusion. First, consider the probability that a ticket matches. Each ticket has a probability $\frac{1}{n}$ of matching its number, and so across all n tickets, we get a probability $\frac{1}{n} \cdot n = 1$.

But wait! We've overcounted the cases where two tickets match. The probability that two given tickets will match is $\frac{1}{n(n-1)}$, which we sum over all $\binom{n}{2}$ pairs. This gives a total of $\frac{1}{2!}$ to subtract off.

But wait once again! We've subtracted off too many times where three tickets match. This time, the probability this triplet of tickets will match is $\frac{1}{n(n-1)(n-2)}$, summed over all $\binom{n}{3}$ pairs, so we add back $\frac{1}{3!}\ldots$

Continuing on, adding back and subtracting off, we indeed get the series

$$1 - \frac{1}{2!} + \frac{1}{3!} - \ldots + \frac{(-1)^{n-1}}{n!},$$

which, if we let $n \to \infty$, becomes the recognizable tail of the Taylor series of e^x, for $x = -1$:

$$\to 1 - e^{-1}.$$

Remark. This gives us an insight into the *Generalized Principle of Inclusion/Exclusion:*

Solution to 1.3.6, Ex. 10

Remark. This is a problem that is better suited for the Concrete Mathematics course – in particular, one learns how to solve this latter recurrence explicitly, which we will briefly show how to do.

Solution We will do this with recursion. Note, of course, that any sentence of length n either begins with a, ba, or bb. If we lop off an a from the front of a word, we must be left with a sentence that has a valid length of $n-1$, and if we remove a ba or bb from the front of a sentence, the remaining words form a sentence of length $n-2$. In particular, *any* sentence of length n can be formed in this way by appending an a to the front of a sentence of length $n-1$, or appending a ba or bb to the front of a sentence of length $n-2$. Hence, we arrive at the recurrence relation

$$N(k) = N(k-1) + 2N(k-2).$$

One rigorous way to arrive at the correct explicit form of $N(k)$ is to use *generating functions*, which will appear with a vengeance later in this course. We will instead appeal to a tactic more suited to a differential equations course – i.e. using an *ansatz*, or a judicious guess.

Suppose $N(k) = r^k$ for some $r \neq 0$. If we plug r into our recurrence, we get

$$r^k = r^{k-1} + 2r^{k-2} \implies r^2 - r - 2 = 0.$$

This is a very nice quadratic! In particular, this gives $r = 2, -1$. Note that any linear combination of 2^k and $(-1)^k$ will give a valid solution to this recurrence, so we arrive at the most general solution

$$N(k) = A2^k + B(-1)^k$$

To solve for the coefficients A, B, we plug in the initial conditions $N(0) = 1, N(1) = 1$, yielding

$$N(k) = \frac{2}{3} \cdot 2^k + \frac{1}{3}(-1)^k.$$

2.4 Conditional Probability and Independence

2.4.1 Conditional Probability

Sometimes we are interested in the outcomes from a subset of the full sample space. For example maybe in a room of 100 persons, 25 of them are wearing a green shirt, while there are 65 males, of whom 15 are wearing a green shirt. Then the probability of a randomly selected person wearing a green shirt depends on if the person is a male or not (or, if we don't know). If we don't know, then the probability is 25/100. But if we know the person to be male, the probability becomes 15/65, which is just a bit lower. This is the essence of **conditional probability**, in which an explicit sample (sub-)space is defined.

Definition 2.4. *Given a sample space Ω and events $A, B \subseteq \Omega$ then the conditional probability of A given B is defined as*

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \tag{2.1}$$

Example 14. The probability of an applicant to be admitted to a certain college is 0.8. The probability for a student in the college to live on campus is 0.6. What is the probability that an applicant will be admitted to the college and will be assigned a dormitory housing?

Solution The probability of the applicant being admitted and receiving dormitory housing is defined by P(Accepted and Housing) = P(Housing|Accepted)P(Accepted) = (0.6)(0.8) = 0.48

From the definition of conditional probability comes one of the most important identities for intersections, namely

Theorem 2.5.

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

and its generalization:

Theorem 2.6.

$$\begin{aligned} P(A_1 \cap A_2 \cap \cdots \cap A_n) &= P(A_1)P(A_2|A_1)P(A_3|A_2 \cap A_1)\cdots \\ &\qquad P(A_n|A_{n-1} \cap A_{n-2} \cap \cdots \cap A_1) \\ &= P(A_1|A_2 \cap \cdots \cap A_n)P(A_2|A_3 \cap \cdots \cap A_n)\cdots \\ &\qquad P(A_n) \end{aligned}$$

Proof We prove the second statement by induction. The base case is clear for $n = 2$. Suppose the statement is true for any n events, and we wish the same is true for any $n + 1$ events, $A_1, \ldots A_n, A_{n+1}$. Let $B = A_2 \cap \ldots \cap A_n \cap A_{n+1}$. Then,

$$P(A_1 \cap A_2 \cdots \cap A_n \cap A_{n+1}) = P(A_1 \cap B) = P(A_1|B)P(B)$$

Clearly, $P(B)$ can be expanded out in the desired form via the inductive hypothesis, so we are done.

Similarly, an inductive analysis on $A_1 \cap \ldots \cap A_n$ will be sufficient for the first statement.

A very useful form of the definition of conditional probability employs the **Law of Total Probability**

Theorem 2.7 (Law of Total Probability)**.** *If $C_1, \ldots, C_n$ form a partition of the sample space Ω, (that is, the sets are mutually disjoint and their union equals Ω), and A is an event in Ω then*

$$P(A) = \sum_{i=1}^{n} P(A|C_i)P(C_i) \tag{2.2}$$

Now Definition 2.4 becomes

$$P(B|A) = \frac{P(B|A)}{\sum_{i=1}^{n} P(A|C_i)P(C_i)} \tag{2.3}$$

2.4.2 Independence

Given any two random processes, sometimes it is the case that the first process may affect the second; other times there is no relationship whatsoever. For example, if an individual person rolls a die and flips a coin, there is no reason to presuppose any effect of the first process on the second. On the other hand if a random college student is selected and the student is asked to state their SAT score and their family income, studies have shown these two variables to be correlated.

Two random experiments that have no effect on each other are said to be *independent*. Mathematically the definition we impose is that if two events A, B are from the same sample space Ω, then the outcome of one does not affect the probability of the other

Definition 2.5. *If $A, B \subseteq \Omega$ and*

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

then A and B are independent

Corollary 2.8. *If $A, B \subseteq \Omega$ are independent then*

$$P(A \cap B) = P(A)P(B) \tag{2.4}$$

Proof

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

when A and B are independent. ∎

Example 15. Let A and B be two independent events such that $P(B|A \cup B) = \frac{2}{3}$ $P(A|B) = \frac{1}{2}$ What is $P(B)$?

Solution $\dfrac{1}{2}$

When more than two events are considered, independence takes on some added complexity. If each pair of events are independent then we call the set of events pairwise independent.

Definition 2.6. *Given $A_1, A_2, \ldots A_n \subseteq \Omega$, if each pair $i < j$ in $[1, \ldots, n]$ satisfies*

$$P[A_1 \cap A_j] = P[A_i]P[A_j]$$

*then the set of events $\{A_i\}$ are **pairwise-independent**.*

Furthermore a set of events is mutually independent if *every* such intersection can be written as a product, for all subsets of events.

Definition 2.7. *Let $\{A_{i_1}, \ldots, A_{i_k}\}$ be any subset of events in $\{A_1, \ldots A_n\}$. If every such set of size $k \geq 2$ obeys the product principle, then the set of events $\{A_1, \ldots A_n\}$ is said to be **mutually-independent**.*

To keep things interesting, these two types of independence do not always imply each other.

Example 16. Consider the experiment of flipping two fair coins. Consider the three events: A = the first coin shows heads; B = the second coin shows heads, and C = the two coins show the same result. Show that these events are pairwise independent, but not independent.

2.4.3 Exercises

1. An urn contains 22 marbles: 10 red, 5 green, and 7 orange. You pick two at random without replacement. What is the probability that the first is red and the second is orange?

2. You roll two fair dice. Find the (conditional) probability that the sum of the two faces is 6 given that the two dice are showing different faces.

3. A machine produces small cans that are used for baked beans. The probability that the can is in perfect shape is 0.9. The probability of the can having an unnoticeable dent is 0.02. The probability that the can is obviously dented is 0.08. Produced cans get passed through an automatic inspection machine, which is able to detect obviously dented cans and discard them. What is the probability that a can that gets shipped for use will be of perfect shape?

4. A box of television tubes contains 20 tubes, of which five are defective. If three of the tubes are selected at random and removed from the box in succession without replacement, what is the probability that all three tubes are defective?

5. Bowl I contains eight red balls and six blue balls. Bowl II is empty. Four balls are selected at random, without replacement, and transferred from bowl I to bowl II. One ball is then selected at random from bowl II. Calculate the conditional probability that two red balls and two blue balls were transferred from bowl I to bowl II, given that the ball selected from bowl II is blue.

6. A machine has two parts labeled A and B: The probability that part A works for one year is 0.8 and the probability that part B works for one year is 0.6. The probability that at least one part works for one year is 0.9. Calculate the probability that part B works for one year, given that part A works for one year.

7. A public health researcher examines the medical records of a group of 937 men who died in 1999 and discovers that 210 of the men died from causes related to heart disease. Moreover, 312 of the 937 men had at least one parent who suffered from heart disease, and, of these 312 men, 102 died from causes related to heart disease. Determine the probability that a man randomly selected from this group died of causes related to heart disease, given that neither of his parents suffered from heart disease.

8. An insurance company examines its pool of auto insurance customers and gathers the following information:
   a) All customers insure at least one car.
   b) 70% of the customers insure more than one car.
   c) 20% of the customers insure a sports car.
   d) Of those customers who insure more than one car, 15% insure a sports car.

   Calculate the probability that a randomly selected customer insures exactly one car and that car is not a sports car.

9. An actuary is studying the prevalence of three health risk factors, denoted by A, B, and C within a population of women. For each of the three factors, the probability is 0.1 that a woman in the population has only this risk factor (and no others). For any two of the three factors, the probability is 0.12 that she has exactly these two risk factors (but not the other). The probability that a woman has all three risk factors, given that she has A and B, is $\frac{1}{3}$. What is the probability that a woman has none of the three risk factors, given that she does not have risk factor A?

10. Prove that if A and B are independent, then so are A and B^C.

11. Prove that if A and B are independent, then so are A^C and B^C.

12. One urn contains 4 red balls and 6 blue balls. A second urn contains 16 red balls and x blue balls. A single ball is drawn from each urn. The probability that both balls are the same color is 0.44. Calculate x.

13. Assume A and B are independent events with $P(A) = 0.2$ and $P(B) = 0.3$ Let C be the event that neither A nor B occurs, let D be the event that exactly one of A or B occurs. Find $P(C)$ and $P(D)$

14. Throw a dice twice. Let A be the event the first throw came up 1, 2, or 3. Let B be the event that the first throw came up 3,4, or 5. Let C be the event that the sum of the two throws is 9. Show that $P(A \cap B \cap C) = P(A)P(B)P(C)$ but A,B, and C are not pairwise independent.

15. In a certain game of chance, a square board with area 1 is colored with sectors of either red or blue. A player, who cannot see the board, must specify a point on the board by giving an x-coordinate and a y-coordinate. The player wins the game if the specified point is in a blue sector. The game can be arranged with any number of red sectors, and the red sectors are designed so that

$$R_i = \left(\frac{9}{20}\right)^i$$

where R_i is the area of the ith red sector. Calculate the minimum number of red sectors that makes the chance of a player winning less than 20%.

2.5 Bayes' Theorem

Bayes' Theorem can be thought of a direct corollary of the definitions and theorems we already have. Recall our definition for conditional probability on two events A and B, written both ways:

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

We can thus find equivalent expressions for $P(A \cap B)$ –

$$P(B|A)P(A) = P(A|B)P(B),$$

and thus we have

Theorem 2.9 (Bayes' Theorem)**.**

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

This is often used to "flip" the direction of the conditionality. If we know something about $P(B|A)$, we are able to infer something about $P(A|B)$ given that we know information about $P(A)$ and $P(B)$.

In deep learning, Bayes' Theorem can also be interpreted in another way. If we know $P(A)$, and we can find out $P(B)$, we can recompute and thus "learn" $P(A|B)$. For example, if we believe something about an underlying probability distribution, such as believing a die is fair, and we learn new data about that distribution, we can update what our beliefs are about that probability distribution based on our empirical results.

If $P(B)$ is not directly computable, we can use the Law of Total Probability:

Corollary 2.10. *If $A_1, \dots A_n$ is a partition of the sample space Ω, then*

$$P(A_k|B) = \frac{P(B|A_k)P(A_k)}{\sum_{i=1}^{n} P(B|A_i)P(A_i)}$$

We can use this to do an exercise from the previous section:

Alt. solution to Sec. 1.4.3, Ex. 6 Let event A_i be the probability i blue marbles are drawn and put into the second bucket, and B the probability a blue marble is then drawn from the second bucket. By Bayes' Theorem,

$$P(A_2|B) = \frac{P(B|A_2)P(A_2)}{\sum_{i=1}^{4} (B|A_i)P(A_i)}$$

How is this easier to compute? Note that $P(A_i)$ can be easily computed:

$$P(A_i) = \frac{\binom{8}{4-i}\binom{6}{i}}{\binom{14}{4}}$$

In general, $P(B|A_i)$ is also easily computed – if i blue marbles are put into the second bucket, then clearly the probability is $\frac{i}{4}$.

We can now easily compute the numerator:

$$P(B|A_2)P(A_2) = \frac{2}{4} \cdot \frac{\binom{8}{2}\binom{6}{2}}{\binom{14}{4}}$$

We now condition the numerator over all the terms in the denominator. Note that every factor in the denominator has this $\binom{14}{4}$ term, so we neglect it. This gives us the answer we had before for this problem:

$$P(A_2|B) = \frac{\frac{2}{4}\binom{8}{2}\binom{6}{2}}{\frac{0}{4}\binom{8}{4}\binom{6}{0} + \frac{1}{4}\binom{8}{3}\binom{6}{1} + \frac{2}{4}\binom{8}{2}\binom{6}{2} + \frac{3}{4}\binom{8}{1}\binom{6}{3} + \frac{4}{4}\binom{8}{0}\binom{6}{4}} = \frac{70}{173} = 0.4895$$

Let's look at another example:

Example 17. Car Insurance, 10.2
An auto insurance company insures drivers of all ages. An actuary compiled the following statistics on the company's insured drivers:

Age of Driver	Probability of Accident	Portion of Company's Insured Drivers
16-20	0.06	0.08
21-30	0.03	0.15
31-65	0.02	0.49
66-99	0.04	0.28

A randomly selected driver that the company insures has an accident. Calculate the probability that the driver was age 16-20.

Answer. $\frac{16}{101} = 0.158$

Example 18. Hospital, 10.4 Upon arrival at a hospital's emergency room, patients are categorized ac- cording to their condition as critical, serious, or stable. In the past year:

(a) 10% of the emergency room patients were critical;
(b) 30% of the emergency room patients were serious;
(c) the rest of the emergency room patients were stable;
(d) 40% of the critical patients died;
(e) 10% of the serious patients died; and
(f) 1% of the stable patients died.

Given that a patient survived, what is the probability that the patient was categorized as serious upon arrival?

Answer. $\frac{45}{154} = 0.292$

Bayes' theorem also gives us a more cautious analysis on medical tests. Generally, if we believe a test has a 95% chance of correctly identifying whether a given person has the disease, we'd think that's a pretty decent test! Not necessarily, says Bayes' Theorem...

Let's define a few terms to describe how "good" a test is. For a test, its **specificity** is the ratio of its true positives to the total number of positives it gives – i.e. it measures how good a test is at finding positives. A test's **sensitivity** is the the ratio of the test's true negatives to the total number of negatives – i.e. it measures how good a test is at finding negatives.

With this in mind, consider the following:

Example 19. Disease, 10.7 A blood test indicates the presence of a particular disease 95% of the time when the disease is actually present. The same test indicates the presence of the disease 0.5% of the time when the disease is not present. One percent of the population actually has the disease. Calculate the probability that a person has the disease given that the test indicates the presence of the disease.

Answer. $\frac{190}{289} = 0.657$

Example 20. A variant: for the above test, is it possible that we can improve probability one has the disease given that they get a positive test result to be greater than 90%?

Answer No! Even if the test is able to perfectly identify true positives (i.e. its specificity is 100%), we still won't achieve the desired 90% cutoff line.

Here is a final example on posterior/prior probabilities, and updating our beliefs on a probability distribution based on new data:

Example 21. Consider a six-sided die, but the side labeled "1" might have been changed to any number from 1-6. Assume the die is still fair, and assume that the probability of changing the 1 to any of $1-6$ is also even. Find the probability distribution of the unknown side given that the die comes up as 6 on its first roll.

More mathematically, find $P(1 \text{ changed to k}|\text{roll a } 6)$ for all $k = 1, 2, \ldots 6$.

Answer. $\frac{1}{7}$ for $k = 1, 2, 3, 4, 5$, $\frac{2}{7}$ for $k = 6$.

Solution

Notice that having been provided with this additional information, we have to update our initial assumption based this result.

3 Probability Distributions

3.1 Random Variables

Definition 3.1. *A **random variable** is a function $X : \Omega \to R$. It assigns a real number to each outcome.*

(In other words, a random variable is like a vending machine- it dispenses numbers according to a possibly unknown function.)

A random variable X might output some concrete number x, with some probability. If this probability is $\frac{1}{2}$, then we write this

$$\Pr[X = x] = \frac{1}{2}$$

For the sake of continuing the "vending machine" analogy, we will temporarily say

$$\Pr[X \hookrightarrow x] = \frac{1}{2},$$

where $\hookrightarrow$ is pronounced "dispenses," i.e. X dispenses x with probability $\frac{1}{2}$.

3.1.1 Probability Mass Function

These random variables have probabilities associated with them, just as outcomes and events in the sample space. The probability function associated with a discrete random variable is called a **probability mass function**

Definition 3.2. *A **probability mass function** (pmf) is a function $f : \mathbb{R} \to [0, 1]$ from the reals to the unit interval such that $f(x) = \Pr[X \hookrightarrow x]$, that is the probability that a random variable dispenses a given value. It must satisfy the following criteria*

1. $\sum_{x_i} f(x_i) = 1$, *where the sum is over every x_i in the range of X.*
2. $f(x_i) = 0$ *for every x_i not in the range of X.*

We will elucidate these ideas with a number of examples

Example 1. Let a 6-sided die be constructed such that the probability of rolling a 4 is twice that of rolling any other value. Describe this in terms of a random variable X and a pmf $f(x)$. Next, let Y be the number of prime factors of X. Give the pmf of Y.

Solution Let X be a random variable that dispenses the value the die shows (in $1, 2, \ldots 6$). $f(x)$ is a function such that $f(1) = f(2) = f(3) = f(5) = f(6) = \frac{1}{7}$, $f(4) = \frac{2}{7}$, and $f(x)$ is 0 otherwise.

Suppose Y is the number of prime divisors of X. We can calculate what Y outputs for given values that X outputs:

X	Y
1	0
2	1
3	1
4	1
5	1
6	2

This means that for the pmf of Y, $g(y)$, $g(0) = \Pr[Y = 0] = \frac{1}{7}$, $g(1) = \Pr[Y = 1] = \frac{5}{7}$, $g(2) = \Pr[Y = 2] = \frac{1}{7}$, and $g(y) = 0$ otherwise.

Example 2. Let X be the sum of the pips showing on 2 rolled, fair, 10-sided die. Find the pmf for X.

Solution The **support set** of X, or the set of possible values of X, is $\{2, 3, \ldots 20\}$, so we can write out a few values of f:

$$f(2) = \frac{1}{100}, \quad f(3) = \frac{2}{100}, \quad \ldots \quad f(20) = \frac{1}{100}$$

We can write out a nice closed form for f:

$$f(x) = \begin{cases} \frac{10-|11-x|}{100} & x \in \{2, 3, \ldots 20\} \\ 0 & \text{otherwise} \end{cases}$$

Example 3. 5 Juniors and 5 seniors take a test and are ranked 1-10 according to their test score (1 = highest score). Assume all scores are distinct and that all 10! student rankings are equally likely. Let X be the highest rank (smallest integer value) of a junior in the class. Find the pmf $f(x)$ for X.

Solution There are $\binom{10}{5}$ ways to arrange the seniors and juniors into distinct ranking orders. To calculate the individual values of f, we proceed with casework.

If X = 1, a junior must have taken the highest rank, and then the remaining juniors and seniors can fill in the ranks in any order. This can be accomplished in $\binom{9}{4}$ ways, so $f(1) = \frac{\binom{9}{4}}{\binom{10}{5}} = \frac{1}{2}$.

If X = 2, a senior takes rank 1, a junior takes rank 2, and the remaining juniors and seniors can fill in the rest of the ranks. This can be accomplished in $\binom{8}{4}$ ways, so $f(2) = \frac{\binom{8}{4}}{\binom{10}{5}} = \frac{5}{18}$.

A similar analysis can be done for the cases where $X = 3, 4, 5, 6$. The highest rank of a junior can't be lower than 6, as that would require more than 5 seniors to fill in rankings. In general, a closed form could be

$$f(x) = \begin{cases} \frac{\binom{10-x}{4}}{\binom{10}{5}} & x \in \{1, 2, \dots 6\} \\ 0 & \text{otherwise.} \end{cases}$$

Remark Note that if we try ensure that the sum of all of the outputs of $f(x)$ is 1, we get a version of the famed *hockey stick identity*:

$$\binom{4}{4} + \binom{5}{4} + \binom{6}{4} + \binom{7}{4} + \binom{8}{4} + \binom{9}{4} = \binom{10}{5}$$

This can be generalized to arbitrary $r = 4, n = 9$:

$$\sum_{k=r}^{n} \binom{k}{r} = \binom{n+1}{r+1}$$

Example 4. Let $f(0) = f(1)$ and $f(k+1) = \frac{1}{k} f(k)$. If you know that f is a pmf over the non-negative integers, then find $f(0)$.

Solution We write out a few of the first few values of $f(k)$ in terms of $f(0)$:

$$\begin{aligned} f(2) &= f(1) = f(0) \\ f(3) &= \frac{1}{2} f(2) = \frac{1}{2} f(0) \\ f(4) &= \frac{1}{3} f(3) = \frac{1}{6} f(0) \dots \end{aligned}$$

In order for the pmf to satisfy $\sum_{x_i} f(x_i) = 1$, we get that

$$\begin{aligned} f(0) + f(1) + f(2) + f(3) + f(4) + \dots &= f(0) + f(0) + f(0) + \frac{1}{2} f(0) + \frac{1}{6} f(0) + \dots \\ &= f(0) \left(1 + \sum_{n=0}^{\infty} \frac{x^n}{n!} \right) \\ &= f(0)(1 + e) = 1 \end{aligned}$$

Therefore, $f(0) = \frac{1}{1+e}$.

Example 5. Find k if $f(x) = \dfrac{k}{x^2}$ is a pmf over positive integers.

Solution In order for the pmf to be valid, we require the pmf to be *normalized*, i.e. $\sum_{x_i} f(x_i) = 1$, so

$$k \sum_{n=1}^{\infty} \frac{1}{n^2} = 1 = \frac{\pi^2 k}{6} \implies k = \frac{6}{\pi^2}.$$

Example 6. In Example 5, let X be a random variable over positive integers with pmf f. Let Y be a random variable that equals 1 if X is even and 2 if X is odd. Find the pmf of Y.

Solution We can evaluate $g(1)$ and $g(2)$ independently. $g(1)$ is the sum of the probabilities that X dispenses an odd number:

$$g(1) = \sum_{n=0}^{\infty} \Pr[X = 2n+1] = \frac{6}{\pi^2} \sum_{n=0}^{\infty} \frac{1}{(2n+1)^2}$$

and $g(2)$ is the sum of the probabilities that X dispenses an even number:

$$g(2) = \sum_{n=1}^{\infty} \Pr[X = 2n] = \frac{6}{\pi^2} \sum_{n=1}^{\infty} \frac{1}{4n^2}$$

This latter sum is easier to evaluate – it becomes $\frac{1}{4} \cdot \frac{\pi^2}{6} = \frac{\pi^2}{24}$, so $g(2) = \frac{6}{\pi^2} \cdot \frac{\pi^2}{24} = \frac{1}{4}$. The sum of squares of odd reciprocals is $\frac{\pi^2}{8}$, so $g(1) = \frac{3}{4}$, which is perfectly consistent.

Example 7. Find k if $f(x) = \dfrac{k}{x}$ is a pmf over positive integers.

Solution This is not a valid pmf – in trying to normalize the pmf, we require

$$\sum_{n=1}^{\infty} \frac{k}{n} = 1,$$

but the left hand side of the equation diverges. Such a pmf does not exist.

3.1.2 Cumulative Mass Functions

A probability mass function over a random variable X gives the probability that x equals a certain value. In many instances it will prove quite helpful to work with instead the probability that X is less than or equal to a certain value. This is called the *cumulative mass function*.

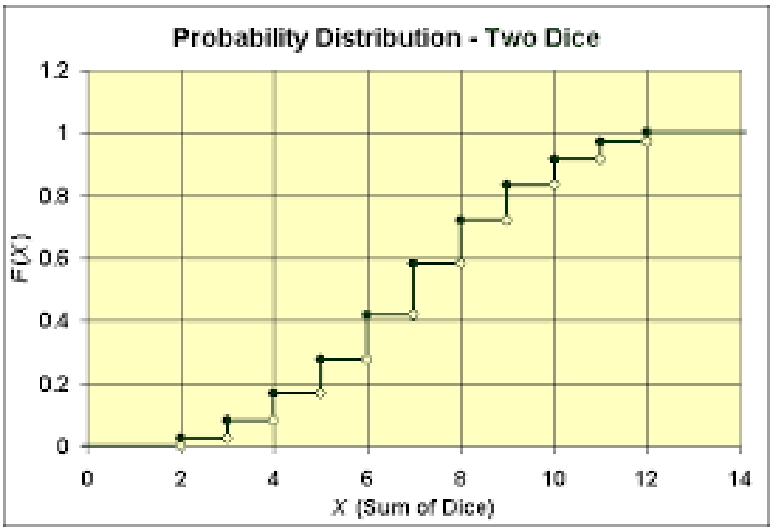

Figure 3.1: Cumulative distribution function for the sum obtained by rolling two dice

Definition 3.3. *The **cumulative mass function (cmf)** of a random variable X with pmf f is defined as a function $F : \mathbb{R} \to [0, 1]$ such that*

$$F(x) = \Pr[X \leq x] = \sum_{-\infty}^{x} f(t)$$

Example 8. Let X be the sum of the pips on the roll of 2 fair six-sided die. Find the cmf of X.

Solution It is impossible to have a sum of 1 or less on two dice rolls, so $F(1) = 0$. We calculate $F(2)$ by noting the only non-zero contribution is if two 1's show up on both of the dice, so $F(2) = \Pr[X \leq 2] = f(2) = \frac{1}{36}$. Similarly, $F(3) = \Pr[X \leq 3] = f(2) + f(3) = \frac{1}{36} + \frac{2}{36} = \frac{3}{36}$, and $F(4) = \Pr[X \leq 4] = f(2) + f(3) + f(4) = \frac{1}{36} + \frac{2}{36} + \frac{3}{36} = \frac{6}{36}$. We can continue computing in this way until we eventually reach $F(11) = \sum_{k=1}^{11} f(k) = 1 - f(12) = \frac{35}{36}$. See Figure 3.1 for a graph of the cmf – note how it monotonically increases until it reaches a final maximum value at 1.

Example 9. Let $f(x) = c\left(\frac{1}{4}\right)^x$ be the pmf of the random variable X, where the support set is $\mathbb{Z}_{\geq 0}$

1. Find the appropriate constant c
2. Determine the cmf $F(x)$

3. Use the cmf to compute $\Pr[2 < X \leq 8]$
4. Write formulas involving F and f for the following
   a) $\Pr[X > a]$
   b) $\Pr[X \geq a]$
   c) $\Pr[a < X < b]$
   d) $\Pr[a \leq X \leq b]$

Solution Most of the following analysis follows from definition:

1. Using the sum of an infinite geometric series formula, we arrive at $c = \frac{3}{4}$.
2. Using the sum of a finite geometric series formula, we arrive at $F(x) = \frac{3}{4}\left(\frac{1-(1/4)^{x+1}}{1-1/4}\right) = 1 - (1/4)^{x+1}$
3. $\Pr[2 < X \leq 8] = F(8) - F(2) = \frac{4095}{262144}$, from the cmf in b
4. This is a more general question that applies to other cmfs other than this one:
   a) $\Pr[X > a] = 1 - F(a)$
   b) $\Pr[X \geq a] = 1 - F(a) + f(a) = 1 - F(a) + \lim_{\Delta t \to 0}(F(a) - F(a - \Delta t))$
   c) $\Pr[a < X < b] = F(b) - F(a) - f(b) = F(b) - F(a) - \lim_{\Delta t \to 0}(F(b) - F(b - \Delta t))$
   d) $\Pr[a \leq X \leq b] = F(b) - F(a) + f(a) = F(b) - F(a) + \lim_{\Delta t \to 0}(F(a) - F(a - \Delta t))$

Remark Please note that what we are calling the cmf is very often called a **distribution function** in other texts and even by us, later on. The reasons will become more apparent when we extend pmf and cmf to continuous functions and partly-continuous functions.

3.2 Discrete Joint Probability Functions

When two or more random experiments occur simultaneously, the outcomes can be analyzed with the use of a *joint probability function*. A common example is the height, X, and weight, Y, of randomly selected subjects. If the subjects are humans, the sample space of X (in feet) could be $\Omega_X = [0, 10]$ and weight $\Omega_Y = [0, 1000]$ in pounds.[1]

[1] These sample spaces are discrete technically because of the finite limitation of measurement, but for all practical purposes would best be treated as continuous. The type doesn't concern us here; it's still a nice example.

If enough sample data were collected, one could approximate $\Pr[X \hookrightarrow x \cap Y \hookrightarrow y]$ for (x, y) in the joint sample space $\Omega_X \times \Omega_Y$. This probability defines the joint probability function, or joint pmf:

$$f(x, y) = \Pr[X \hookrightarrow x, Y \hookrightarrow Y]$$

where the "comma" in the probability implies intersection and is usually read as "and." Be careful to **not** equate this with $\Pr[X \hookrightarrow x] \cdot \Pr[Y \hookrightarrow y]$, which would equal $\Pr[X \hookrightarrow x, Y \hookrightarrow Y]$ only when X and Y are independent. In fact, we should all be able to agree that height and weight of humans (or any type of object) are almost always correlated and, therefore, *not* independent.

The joint pmf of a set of discrete random variables $\{X_1, \ldots, X_n\}$ satisfies the following properties:

Theorem 3.1. *If f is a function then f can be the pmf of a set of random variables if and only if*

$$f(x_1, \ldots, x_n) \geq 0 \tag{3.1}$$

$$\sum_{x_1} \cdots \sum_{x_n} f(x_1, \ldots, x_n) = 1 \tag{3.2}$$

Where in both items the sum is taken over all values in the domain of f.

Example 10. Let $f(x, y) = kxy$ be a pmf for $x = 1, 2, 3$ and $y = 1, 2, 3$. Determine the value of k.

Solution In order to normalize this joint pmf, we force $\sum_x \sum_y kxy = 1$. We can "pull apart" the sum:

$$k \left(\sum_x x \right) \left(\sum_y y \right) = 1 \implies k \cdot 6 \cdot 6 = 1 \implies k = \frac{1}{36}.$$

Example 11. A jar contains 3 red, 2 green and 4 blue marbles. Two marbles are drawn simultaneously at random. Let R be the number of red and G the number of green marbles drawn. Determine the joint pmf $f(r, g)$

Solution In this case, R can be $0, 1, 2$, and G can be $0, 1, 2$ as well, but these can't be satisfied simultaneously – in particular, R and G cannot both dispense 2 at the same time.

To calculate this, we could consider this case-by-case, starting with $\Pr[R = 0, G = 0]$, say. The probability in this case is

$$f(0, 0) = \frac{\binom{3}{0}\binom{2}{0}\binom{4}{2}}{\binom{9}{2}}$$

and in general, $f(r,g)$ has a closed form:

$$f(r,g) = \frac{\binom{3}{r}\binom{2}{g}\binom{4}{2-r-g}}{\binom{9}{2}}$$

Example 12. Given the pmf $f(x,y,z)$ of random variables X, Y, Z

$$f(x,y,z) = \frac{(x+y)z}{63} \qquad x = 1,2; y = 1,2,3; z = 1,2$$

calculate $\Pr[X \hookrightarrow 2, Y + Z \leq 3]$.

Solution We can compute this probability by casework. The only possible triples (x,y,z) that work are $(2,1,2)$, $(2,2,1)$, and (2,1,1). If we plug in directly, we get the total probability as

$$\frac{(2+1)2}{63} + \frac{(2+2)1}{63} + \frac{(2+1)1}{63} = \frac{13}{63}$$

The last example hints at a definition for the cumulative distribution function for a pmf f. In the bivariate case the definition is

Definition 3.4. *Let $f(x,y)$ be the pmf of two random variables X and Y. Then the distribution function $F(x,y)$ is defined by*

$$F(x,y) = \Pr[X \leq x, Y \leq y] = \sum_{u=-\infty}^{x} \sum_{v=-\infty}^{y} f(u,v)$$

Example 13. Determine the distribution function F for the pmf defined in Example 11.

Solution

Example 14. Write an expression for $\Pr[a < X \leq b, c < Y \leq d]$ in terms of F. careful to consider all the cases!

Solution It's best to think of this geometrically:

3.2.1 Marginal and Conditional Distributions

Marginal distributions and conditional distributions reduce the number of variables in joint distribution functions and d

Definition 3.5. *Given X, Y, and joint pmf $f(x,y)$, then $f_X(x)$ is the X-marginal distribution of f, where $f_X(x) = \Pr[X \hookrightarrow x]$.*

We can write out what this means explicitly, in terms of a sum:

Theorem 3.2.

$$f_X(x) = \sum_y f(x, y)$$

and

$$f_Y(y) = \sum_x f(x, y)$$

It follows clearly that

Example 15. Consider the following joint distribution function Compute the marginal distributions for

Definition 3.6. *Given random variables X, Y, and joint pmf $f(x, y)$, the conditional distribution function $f_{X|y}(x|y)$ is equal to*

$$f_{X|y}(x|y) = \frac{\Pr[X = x \cap Y = y]}{\Pr[Y = y]} = \frac{f(x, y)}{f_Y(y)}$$

This is rather reminiscent of the Law of Total Probability

Example 16. For the joint distribution function above, compute $f_{X|y}(1|Y = 1)$.

Answer. $\frac{0.4}{0.4+0.2} = \frac{0.4}{0.6} = \frac{2}{3}$.

Example 17. Suppose we flip a fair coin 4 times. Let X be the number of heads in the first three tosses, and Y be the number of heads in the last three tosses. Find $f(x, y)$, f_X, f_Y, $f_{X|y}$, and $f_{Y|x}$.

3.2.2 Independent Random Variables

3.3 Continuous Random Variables

The concepts we have studied so far with discrete random variables and mass functions transfer almost immediately to continuous random variables, where sums are "replaced" with integrals. The fundamental underlying difference in interpretation is that a probability mass function becomes, in the continuous case, a probability *density* function.

Definition 3.7. *Let X be a random variable. A function $f(x)$ from the domain of X into $\mathbb{R}$ is a probability density function (pdf) if it satisfies the following*

1. *$f(x) \geq 0$ for all x in the domain of f*
2. *$\Pr[x \in A] = \Pr[A] = \int_A f(x)\ dx$ gives the probability that the random variable X dispenses an element of the set A*
3. *$\int_{\mathcal{A}} f(x)\ dx = 1$ where $\mathcal{A}$ is the entire domain of the random variable X*