

Lectures on Probability Theory

Paul Renteln

Department of Physics
California State University
San Bernardino, CA 92407

December 2008

Revised March 2010

Contents

1	The Idea of Probability	1
1.1	Frequentists <i>versus</i> Bayesians	1
1.2	The Sample Space	2
1.3	Properties and Events	3
1.4	Venn Diagrams	3
1.5	The Rules of Probability	4
1.5.1	Laplace's Principle of Insufficient Reason	5
1.5.2	Sum Rules	5
1.5.3	Product Rules	6
1.6	Drawing With and Without Replacement: A Word on Sample Spaces	15
1.7	Exercises	17
2	Counting	18
2.1	The Multiplication Principle	19
2.2	Ordered Sets: Permutations	21
2.3	Unordered Sets: The Binomial Coefficient	23
2.4	Multisets	25
2.5	Multiset Permutations: Multinomial Coefficient	27
2.6	Choosing Collections of Subsets: Multinomial Coefficient Again	28
2.7	Exercises	29
3	Random Variables and Probability Distributions	31
3.1	Discrete Random Variables	31
3.1.1	The Mean or Expectation Value	32
3.1.2	The Variance and the Standard Deviation	34
3.2	Discrete Probability Distributions	36
3.2.1	The Binomial Distribution	36
3.2.2	The Poisson Distribution	38
3.3	Continuous Random Variables	40

3.3.1	The Mean, Average, or Expectation Value	41
3.3.2	The Variance and the Standard Deviation	41
3.4	Continuous Probability Distributions	42
3.4.1	The Uniform Distribution	42
3.4.2	The Gaussian or Normal Distribution	42
3.5	Exercises	45
A	Stirling's Approximation	47
B	The Gaussian Distribution as a Limit of the Binomial Distribution	48

List of Figures

1	A Venn Diagram	4
2	A partition of Ω into four mutually exclusive and exhaustive events	8
3	The binomial coefficients $\binom{8}{k}$	24
4	The first six rows of Pascal's triangle	30
5	The Gaussian distribution	43
6	A Gaussian approximation to a binomial distribution $2^{-8}\binom{8}{4+s}$	50

1 The Idea of Probability

1.1 Frequentists *versus* Bayesians

Imagine a bag of 9 marbles, of which 3 are blue, 3 are green, 2 are red, and 1 is yellow. If you reach in your hand and pull out a marble at random, what are the chances that you get a red marble? Intuitively we would say that the chances of getting a red marble are 2 out of 9, or that the probability of drawing a red marble is $2/9$. But what does this mean? That is, what does the number $2/9$ represent?

There are basically two different answers to this question, depending on whether you are a **frequentist** or a **Bayesian**. According to the frequentist interpretation of probability, to say that the probability of drawing a red marble is $2/9$ is to say that, if you were to repeat this experiment many times and compute the ratio of the number of times you draw a red marble to the total number of drawings, the result would approach $2/9$ in the limit that the number of drawings goes to infinity.

On the other hand, the Bayesian ¹ would say that the probability of an outcome represents our *degree of belief* that the outcome will occur *in a single experiment*. That is, for the Bayesian all probabilities are just numbers that attempt to formalize our intuition about the outcomes of individual experiments, where 0 means that it will never occur, 1 means it will always occur, and anything else is somewhere in-between. Thus, the Bayesian would say that one is 22% sure (because $2/9 \approx 0.22$) that a red marble will be drawn from the bag in any given drawing.

For many years the frequentists and the Bayesians engaged in a war of words, each claiming that his view was superior and that the other's view was nonsensical. So, for example, the frequentist believes that somehow probabilities are *objective*, in the sense that, given enough time and effort, anyone would verify the same answer by performing the same experiment over and over. That is, according to the frequentist, probabilities somehow inhere

¹Thomas Bayes, 1702-1761, British mathematician and Presbyterian minister.

in nature. The frequentist is horrified by the Bayesian approach, because it appears to be entirely subjective. Who is to say whether the next person to come along would not assign a different probability to the same outcome if he has a stronger or weaker degree of belief that the outcome will actually occur?

The Bayesian retorts that it is the frequentist who believes in nonsense, because the frequentist idea is that probabilities somehow refer to an imaginary set of identically performed experiments. One could never perform an experiment an infinite number of times in order to test the frequentist interpretation. Indeed, some experiments can only be performed once! In that case, the Bayesian interpretation seems the most reasonable. Moreover, Bayesians are quick to point out that their view of probability is totally objective, in the sense that any other person, following the correct rules of probability, would come to the same conclusion.

Historically it is the frequentist approach that is taught to most students, but in fact the Bayesian approach is much more flexible and powerful. Still, the war rages on. Fortunately, it does not matter which philosophy one follows, because the rules of probability are the same for both. At the end of the day, both camps compute the same probabilities for the same experiment, even though they may disagree on the interpretation of their final answers. Thus, we shall not take sides, but instead move on to actually computing some probabilities. You may choose whichever point of view you prefer.

1.2 The Sample Space

Suppose we perform an experiment with many possible outcomes. The set of all possible outcomes is called the **sample space** and is usually denoted by the Greek letter Ω . Ω can be finite or infinite, depending on the experiment.² In the marble example discussed above, Ω is a set with 9 elements corresponding to the 9 individual marbles that one could draw from the bag.

²Technically we should distinguish two kinds of infinities, countable and uncountable, but we will gloss over this important distinction here.

On the other hand, if our experiment involved measuring the height of all persons in some country, then Ω would be infinite, corresponding to all possible heights between the smallest height and largest height of a person in that country.³

1.3 Properties and Events

In the marble example above, we were not really interested in the individual marbles, but rather in marbles having the property that their color is, say, red. A **property** is an attribute of an object. In probability theory we are interested, not in properties *per se*, but in outcomes of experiments. So, instead of properties, we speak of events. An **event** is the set of all outcomes of an experiment associated to a certain property. Mathematically, we represent events by subsets of Ω . For example, suppose we label the marbles in our bag with the numbers from 1 to 9, and suppose that marbles 1, 2, and 3 are blue, 4, 5, and 6 are green, 7 and 8 are red, and 9 is yellow. Then $\Omega = \{1, 2, \dots, 9\}$ and the event R corresponding to the outcome that the marble is red is identified with the subset $R = \{7, 8\} \subset \Omega$. Similarly, suppose that the smallest and largest heights of a person in some country are h_1 and h_2 , and suppose we pick people at random. Our sample space would be $\Omega = [h_1, h_2]$, the interval from h_1 to h_2 . The event S corresponding to the outcome that a randomly chosen person has a height somewhere between s_1 and s_2 is then $S = [s_1, s_2] \subset \Omega$.

1.4 Venn Diagrams

All this can be made a bit more clear by means of Venn diagrams. A **Venn diagram** is usually drawn as a rectangle corresponding to the sample space Ω . Events are then subsets of this rectangle, usually drawn as closed curves encompassing all the objects with the corresponding property. In the marble

³The number of persons is clearly finite, but the possible heights lie in an interval of real numbers, and hence are infinite.

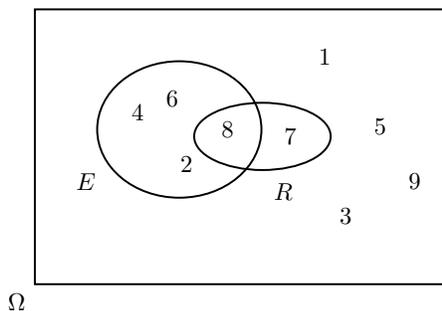


Figure 1: A Venn Diagram

example, a possible Venn diagram is shown in Figure 1. In this figure, marbles 7 and 8 belong to event R corresponding to the property of being red, while marbles 2, 4, 6, and 8 belong to the event E corresponding to the property of being labeled with even numbers. Notice that marble 8 belongs to both events, while marbles 1, 3, 5, and 9 belong to neither event.

We can use this figure to illustrate several other important set-theoretic concepts. Specifically, we write $A \cap B$ (the **intersection** of A and B) for the outcomes that belong to both events A and B , $A \cup B$ (the **union** of A and B) for outcomes that belong to event A or event B or both, and \bar{A} (the **complement** of A) for outcomes that do not belong to event A . In Figure 1, $E \cap R = \{8\}$, $E \cup R = \{2, 4, 6, 7, 8\}$, $\bar{E} = \{1, 3, 5, 7, 9\}$, and $\bar{R} = \{1, 2, 3, 4, 5, 6, 9\}$. These concepts are related by **de Morgan's Laws**:

$$\overline{A \cap B} = \bar{A} \cup \bar{B} \tag{1.1}$$

$$\overline{A \cup B} = \bar{A} \cap \bar{B}. \tag{1.2}$$

For example, $\overline{E \cup R} = \{1, 3, 5, 9\} = \bar{E} \cap \bar{R}$.

1.5 The Rules of Probability

Two questions must be answered before we can begin to compute the probabilities of events. First we must have a rule for assigning probabilities to

outcomes and events, and second we must have some rules for manipulating these probabilities.

1.5.1 Laplace's Principle of Insufficient Reason

The answer to the first question can be somewhat tricky, particularly when the sample space is infinite. When the sample space is finite, however, the answer is provided by **Laplace's Principle of Insufficient Reason: If there is no particular reason to prefer one outcome to another, then all outcomes are equally likely.** The implication of this principle for probability theory is as follows. Let $|X|$ denote the **size** (or **cardinality**) of a finite set X , namely the number of elements it contains.⁴ Then we can formulate

Principle 1.1 (Laplace's Principle). *If every outcome in Ω is equally likely, then for any event A ,*

$$P(A) = \frac{|A|}{|\Omega|} \tag{1.3}$$

That is, the probability that an event A occurs is simply the ratio of the number of outcomes contained in that event to the total number of possible outcomes.

For instance, in the example above, $|\Omega| = 9$, $|R| = 2$, and $P(R) = 2/9$.

1.5.2 Sum Rules

There are several important consequences of Laplace's Principle. For arbitrary events A and B ,

$$0 \leq P(A) \leq 1 \tag{normalization}, \tag{1.4}$$

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) \tag{additivity}, \tag{1.5}$$

$$P(A) + P(\bar{A}) = 1 \tag{complementarity}. \tag{1.6}$$

⁴Note that $|X|$ does not mean the 'absolute value' of X !

The normalization property just says that probabilities are numbers between 0 and 1, which follows immediately from the expression for $P(A)$ as a ratio of something smaller to something larger. If A is the **empty event** \emptyset which contains no events, then $P(\emptyset) = 0$. On the other hand, $P(\Omega) = 1$, which means that some event in Ω always occurs.

To prove additivity, consider two overlapping events A and B . By thinking about Venn diagrams it is clear that

$$|A \cup B| = |A| + |B| - |A \cap B|, \quad (1.7)$$

whereupon Equation (1.5) follows from Laplace's Principle by dividing through by $|\Omega|$. If $A \cap B = \emptyset$ then we say A and B are **disjoint** or **mutually exclusive** events. In that case the additivity property simplifies.

$$P(A \cup B) = P(A) + P(B) \quad (\text{mutually exclusive events}). \quad (1.8)$$

Lastly, the complementarity rule follows immediately from the additivity rule, because for any event A , $A \cap \bar{A} = \emptyset$ and $A \cup \bar{A} = \Omega$.

Example 1 In the marble example given above, find the probability to draw a marble that is both green and odd. Find the probability to draw a marble that is either blue or even or both.

Let G be the event that the marble is green, B that it is blue, E that it is even, and O that it is odd. Then $P(G \cap O) = 1/9$ because only one marble out of 9 is both green and odd. Also, $P(B \cup E) = P(B) + P(E) - P(B \cap E) = 1/3 + 4/9 - 1/9 = 2/3$, which is correct, because six of the marbles lie in the union of the blue set the even set, namely $\{1, 2, 3, 4, 6, 8\}$.

1.5.3 Product Rules

Suppose we wanted to know the probability that a person's birthday is July 4. In the absence of any additional information, Laplace's Principle would tell us that the probability is $1/365$ (ignoring leap years). But suppose we

happened to know that the person was born sometime in the summer, and say there are 90 days in the summer. Then clearly the probability is only $1/90$. All probabilities are like this. That is, they are all **conditioned** on certain assumptions, or prior knowledge. To account for this, we introduce a new notation.

Let A and B be two events in a sample space Ω . The **conditional probability** $P(A|B)$ is the probability that event A occurs given that event B occurs. Note that conditional probabilities are not symmetric. For instance, in our running example $P(E|R) = 0.5$, because there is a 50% chance that a marble will be even given that it is red. On the other hand, $P(R|E) = 0.25$, because there is only a 25% chance that a marble will be red, given that it is even.

In this notation $P(A)$ just means $P(A|\Omega)$ (*i.e.*, the probability that A occurs is the probability that A occurs given that something occurs), but we will continue to write it simply as $P(A)$. Indeed, $P(A|B)$ can be thought of as the probability of A in the restricted sample space B , and this is often a useful point of view.

We can now state the fundamental product rule for conditional probabilities.

$$P(A \cap B) = P(A|B)P(B) \quad (\text{product rule}). \quad (1.9)$$

In words it says that the probability that A and B both occur equals the probability that A occurs given that B occurs, times the probability that B actually occurs. For instance, in the marble example $P(R \cap E) = P(R|E)P(E) = (1/4)(4/9) = 1/9$, which is correct. Note that, because $A \cap B = B \cap A$ we have $P(A \cap B) = P(B \cap A)$. Exchanging A and B in Equation (1.9) thus yields the very important result known as **Bayes' Rule**:

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)} \quad (\text{Bayes' Rule}). \quad (1.10)$$

If the event A does not depend in any way on the event B and *vice versa*

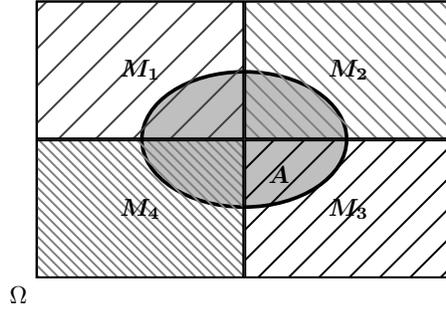


Figure 2: A partition of Ω into four mutually exclusive and exhaustive events

then we say that A and B are **independent**. In that case $P(A|B) = P(A)$, $P(B|A) = P(B)$, and

$$P(A \cap B) = P(A)P(B) \quad (\text{product rule for independent events}). \quad (1.11)$$

It is perhaps worth noting that mutual exclusivity and independence of events are not identical concepts. For example, let A be some event with $P(A) \neq 0, 1$. Then A and \bar{A} are certainly mutually exclusive, yet they are dependent, because $P(A|\bar{A}) = 0$ (the probability that A occurs given that A does not occur is certainly zero). Hence $0 = P(A \cap \bar{A}) \neq P(A)P(\bar{A})$.

Another important consequence of the axioms is the so-called **marginalization rule**. Let A be any event, and suppose that $\{M_k\}_{k=1}^r$ is a set of mutually exclusive and exhaustive events in Ω , meaning that one and only one of the events M_k can occur (see Figure 2).⁵ Clearly,

$$P(A) = \sum_{k=1}^r P(A \cap M_k). \quad (1.12)$$

⁵We also say that the events $\{M_k\}$ **partition** Ω .

Now the product rule gives

$$P(A) = \sum_{k=1}^r P(A|M_k)P(M_k) \quad (\text{marginalization rule}). \quad (1.13)$$

Example 2 [False positives]. Suppose a rare disease occurs by chance in one out of every 100,000 people. A blood test exists for the disease. If you have the disease the test result will be positive 95% of the time, but if you do not have the disease it will be positive 0.5% of the time (false positive). Suppose you take the test and it is positive. What is the probability that you have the disease?

Let D be the event that you have the disease, and T the event that the test is positive. Then we wish to compute $P(D|T)$. This is a classic application of Bayes' Rule. We have

$$P(D|T) = \frac{P(T|D)P(D)}{P(T)} \quad (1.14)$$

By the marginalization rule

$$\begin{aligned} P(T) &= P(T|D)P(D) + P(T|\bar{D})P(\bar{D}) \\ &= (0.95)(0.00001) + (0.005)(0.99999) \approx 0.005 \end{aligned} \quad (1.15)$$

Hence

$$P(D|T) = \frac{(0.95)(0.00001)}{0.005} \approx 0.002 = 0.2\% \quad (1.16)$$

Hence, although the test seems to be pretty accurate at first glance, the result is almost worthless, because there is only a 0.2% chance that you have the disease even though the test says you do. ⁶

Example 3 [The Monty Hall Problem]. This is a classic puzzle, made even more famous when a newspaper columnist, Marilyn vos Savant ⁷ published an answer to the following question posed in her column of 9 September 1990:

⁶Examining the analysis reveals that it is the fact that the disease is so rare that ruins the validity of the test. For very rare diseases one needs a much lower rate of false positives in order to obtain a useful result.

⁷According to the Guinness Book of Records, vos Savant had the highest IQ on the planet between 1986 and 1989!

Suppose you're on a game show, and you're given the choice of three doors. Behind one door is a car, the others, goats. You pick a door, say #1, and the host, who knows what's behind the doors, opens another door, say #3, which has a goat. He says to you: "Do you want to pick door #2 [instead]?" Is it to your advantage to switch your choice of doors?

vos Savant said that it pays to switch doors, because the probability that the car is behind door #2 is $2/3$, whereas the probability that it lies behind door #1 is only $1/3$. Hundreds of readers, including prominent mathematicians and other academics, disagreed strongly and sent in very unflattering letters about her intelligence, not to mention her understanding of probability, saying that the two probabilities are equal to $1/2$. After all, the cars were equally likely to be behind either door #1 or door #2 before the host opened door #3, and nothing the host did could affect this.

Or could it? Well, it turns out that vos Savant was right, and the learned academics wrong, under a certain crucial assumption. The important assumption is that the host does not open doors at random (for, this would make no sense on a game show), but instead only opens a door concealing a goat. In that case, it is better to switch doors, which we can see as follows. Let C_j be the event that the car is behind door j , where $j \in \{1, 2, 3\}$. Let H be the event that the host opens door #3 after the contestant chooses door #1. Then we want to compute $P(C_1|H)$ and $P(C_2|H)$, which we will do using Bayes' rule together with the marginalization rule. If the car is behind door #1, then the host could pick either door #2 or door #3 to reveal a goat. But if the car is behind door #2 then the host must pick door #3, because the contestant has already chosen door #1. Also, the host would not open door #3 if the car were behind it. Hence

$$P(H|C_1) = 1/2, \quad P(H|C_2) = 1, \quad P(H|C_3) = 0. \quad (1.17)$$

Now

$$P(C_1) = P(C_2) = P(C_3) = 1/3, \quad (1.18)$$

because, in the absence of any information, the car is equally likely to be behind

any one of the doors. The marginalization rule (1.13) tells us that

$$P(H) = P(H|C_1)P(C_1) + P(H|C_2)P(C_2) + P(H|C_3)P(C_3) \quad (1.19)$$

$$= (1/2)(1/3) + (1)(1/3) + (0)(1/3) = 1/2, \quad (1.20)$$

so by Bayes' Rule,

$$P(C_1|H) = \frac{P(H|C_1)P(C_1)}{P(H)} = \frac{(1/2)(1/3)}{1/2} = 1/3 \quad (1.21)$$

$$P(C_2|H) = \frac{P(H|C_2)P(C_2)}{P(H)} = \frac{(1)(1/3)}{1/2} = 2/3. \quad (1.22)$$

That is, the probability that the car is behind door #2 is double the probability that it is behind door #1, given the seemingly innocent actions of the host. Clearly, it pays to switch choices in this case.

Example 4 [The Duelists] Two duelists, A and B , take alternate shots at each other, and the duel is over when a shot (fatal or otherwise!) hits its target. Each shot fired by A has a probability α of hitting B , and each shot by B has a probability β of hitting A . Calculate the probability P_1 that A will win the duel assuming that A fires the first shot. Calculate the probability P_2 that A will win the duel assuming that B fires the first shot.

The probability that A hits B on the first shot is α . The probability that A hits B on the third shot is, by the product rule for independent probabilities, $(1 - \alpha)(1 - \beta)\alpha$, because A must miss the first shot, B must miss the second shot, and A must hit the third shot. (Each shot is independent of the last, assuming they are unaffected by the bullets whizzing past their ears!) Each sequence of events (*e.g.*, (miss, miss, hit) and (miss, miss, miss, miss, hit)) is unique, and therefore the collection of sequences of shots constitutes a set of mutually exclusive possibilities. Hence by the additivity rule for mutually exclusive events, the probability that

one of the sequences actually occurs is just the sum of the all the probabilities:

$$P_1 = \alpha + (1 - \alpha)(1 - \beta)\alpha + (1 - \alpha)(1 - \beta)(1 - \alpha)(1 - \beta)\alpha + \dots \quad (1.23)$$

$$= \alpha[1 + (1 - \alpha)(1 - \beta) + (1 - \alpha)^2(1 - \beta)^2 + \dots] \quad (1.24)$$

$$= \frac{\alpha}{1 - [(1 - \alpha)(1 - \beta)]} \quad (1.25)$$

$$= \frac{\alpha}{\alpha + \beta - \alpha\beta}. \quad (1.26)$$

For example, if $\alpha = \beta = 1/2$ then $P_1 = 2/3$. In the penultimate step (1.25) above we used the formula for the sum of a geometric series

$$\sum_{n=0}^{\infty} x^n = \frac{1}{1 - x} \quad \text{if} \quad |x| < 1, \quad (1.27)$$

a result worth remembering.

The probability that A wins given that B fires the first shot is just

$$P_2 = \frac{\alpha(1 - \beta)}{\alpha + \beta - \alpha\beta}, \quad (1.28)$$

which follows from the previous answer in two ways. First, this result is $1 - P_1$ but with α and β interchanged, because the probability that A wins given that B fires first is the same as the probability that B wins given that A fires first, but with the probabilities α and β interchanged. Alternatively, we could simply observe that if B misses with probability $(1 - \beta)$, then the duel reduces to the previous case in which A fires first.

An elegant derivation of the above results uses conditioning instead of an infinite series summation. Let $P(A)$ denote the probability that A wins. (Technically, we should write this as $P(A|F_A)$ to denote the assumption that A fires the first shot, but that is assumed throughout the computation, so to avoid cumbersome notation we drop the F_A .) Then we condition on the event $A_1 = 'A \text{ hits } B \text{ on the first shot}'$ and use the marginalization rule:

$$P(A) = P(A|A_1)P(A_1) + P(A|\overline{A_1})P(\overline{A_1}) \quad (1.29)$$

Now $P(A|A_1) = 1$, as A wins if he hits B on the first shot. Also $P(A_1) = \alpha$

and $P(\overline{A_1}) = 1 - \alpha$. To find $P(A|\overline{A_1})$ we condition on the event B_2 ='B hits A on the second shot of the duel' and using the marginalization rule again to get ⁸

$$P(A|\overline{A_1}) = P(A|\overline{A_1} \cap B_2)P(B_2|\overline{A_1}) + P(A|\overline{A_1} \cap \overline{B_2})P(\overline{B_2}|\overline{A_1}) \quad (1.30)$$

Clearly, $P(A|\overline{A_1} \cap B_2) = 0$, as A loses if he misses the first shot and B hits him on the second shot. Also, $P(A|\overline{A_1} \cap \overline{B_2}) = P(A)$, because if they both miss their shots, the duel begins again as if nothing had happened. Finally, $P(B_2|\overline{A_1}) = P(B_2)$ (and $P(\overline{B_2}|\overline{A_1}) = P(\overline{B_2})$) because the shots are independent of each other. Combining this information with Equations (1.29) and (1.30) yields

$$P(A) = \alpha + P(A)(1 - \alpha)(1 - \beta) \quad (1.31)$$

or

$$P(A) = \frac{\alpha}{\alpha + \beta - \alpha\beta} \quad (1.32)$$

as before.

Example 5 [Gambler's Ruin]. You enter a casino with n dollars, and you repeatedly wager \$1 on a game in which the probability of winning is p . You say to yourself that you will leave the casino if you ever reach a total of N dollars, where $N > n$. What is the probability that you will leave with nothing?

Let $P(k|n)$ be the probability that you amass k dollars given that you begin with n dollars. We want to compute $P(0|n)$. We are given that $P(n+1|n) = p$ and $P(n-1|n) = 1 - p$. You either win or lose the first play, so by the marginalization rule, for any k , ⁹

$$P(k|n) = P(k|n+1)P(n+1|n) + P(k|n-1)P(n-1|n). \quad (1.34)$$

⁸Here we are using the fact that $P(A \cap B|A) = P(B|A)$ for any events A and B , because if we know that A occurs, then the probability that both A and B occur is the same as the probability that B occurs.

⁹We actually need a slight generalization of the marginalization rule, namely that

$$P(A|Q) = \sum_k P(A|M_k)P(M_k|Q), \quad (1.33)$$

which follows easily from the probability rules.

For notational simplicity we set $p_n := P(0|n)$, so (1.34) becomes

$$p_n = p_{n+1}p + p_{n-1}(1 - p) \quad (1.35)$$

If you reach 0 dollars then you definitely leave with nothing, whereas if you reach N dollars then the game ends and you definitely leave with something, so $p_0 = 1$ and $p_N = 0$. These are the boundary conditions for the difference equation (1.35). We can solve this equation by the following trick. Rewrite it as

$$p_n - p_{n-1} = (p_{n+1} - p_{n-1})p = (p_{n+1} - p_n + p_n - p_{n-1})p. \quad (1.36)$$

Define

$$q_n := p_n - p_{n-1}. \quad (1.37)$$

Then (1.36) becomes

$$q_n = (q_{n+1} + q_n)p, \quad (1.38)$$

or, collecting terms,

$$q_{n+1} = \left(\frac{1-p}{p}\right) q_n = \alpha q_n, \quad (1.39)$$

where

$$\alpha := \frac{1-p}{p}. \quad (1.40)$$

Equation (1.39) is a simple recurrence relation, whose solution is

$$q_n = \alpha^{n-1} q_1. \quad (1.41)$$

Next we observe that

$$p_n - p_0 = \sum_{k=1}^n q_k = q_1 \sum_{k=1}^n \alpha^{k-1} = q_1 \sum_{k=0}^{n-1} \alpha^k = (p_1 - p_0) \left(\frac{\alpha^n - 1}{\alpha - 1}\right), \quad (1.42)$$

where the first equality follows because the sum is telescoping and the last equality follows because the sum of a finite geometric series is

$$\sum_{k=0}^m x^k = \frac{x^{m+1} - 1}{x - 1}. \quad (1.43)$$

We find p_1 by setting $n = N$ in (1.42) and using $p_0 = 1$ and $p_N = 0$. This gives

$$0 = 1 + (p_1 - 1) \left(\frac{\alpha^N - 1}{\alpha - 1} \right). \quad (1.44)$$

Plugging this back into (1.42) and using (1.40) yields

$$p_n = \frac{\alpha^N - \alpha^n}{\alpha^N - 1} = \frac{\left(\frac{1-p}{p}\right)^N - \left(\frac{1-p}{p}\right)^n}{\left(\frac{1-p}{p}\right)^N - 1}. \quad (1.45)$$

If the game is fair, meaning that you have an even chance of winning and losing, then $p = 0.5$. In that case, Equation (1.45) says (after applying L'Hospital's rule) that $p_n = 1 - (n/N)$. If decide you want to 'break the bank', then it is almost certain that you will walk out of the casino with nothing, because typically $n \ll N$, so p_n is very close to 1. Your chances are only worsened if $p < 0.5$. The moral of the story is 'never bet against the house'!

1.6 Drawing With and Without Replacement: A Word on Sample Spaces

Consider again our bag of marbles. Suppose that we were to draw *two* marbles out of the bag and ask for the probability that both are red. Before we can answer this question we need to know a little more about how the experiment is performed. We could either draw a marble, replace it in the bag, then draw another ('drawing with replacement'), or else draw a marble, then draw another without replacing the first ('drawing without replacement'). The answer depends on which method we choose.

Firstly let us suppose that we draw with replacement. In this case the two events R_1 (the first marble is red) and R_2 (the second marble is red) are independent. Hence the product rule (1.11) gives

$$P(R_1 \cap R_2) = P(R_1)P(R_2) = (2/9)(2/9) = 4/81 \quad (1.46)$$

Although the result is correct, there is a conceptual difficulty which is worth clearing up: what is our sample space? In our original example our sample space could be represented as a collection of numbers from 1 to 9, corresponding to the possible outcomes of our experiment, which consisted of drawing a single marble from the bag. (We imagine that the marbles all carry identifying numbers.) All our events were subsets of our sample space. But if we were to choose the same sample space for this problem, how would we represent R_1 and R_2 ? The answer is that, strictly speaking, we could not.

The resolution of this conundrum is that our sample space has changed. Recall that the sample space is just the collection of all possible outcomes of an experiment. Our experiment is now drawing *two* marbles from the bag, with replacement. Hence the sample space for drawing two marbles consists of ordered pairs

$$\Omega' = \{(a, b) : 1 \leq a, b \leq 9\}, \quad (1.47)$$

where the first number represents the outcome of the first drawing and the second number represents the outcome of the second drawing. Thus our sample space now has size $9 \times 9 = 81$. There are four ways to get a red marble on both the first and second drawings, namely $\{(7, 7), (7, 8), (8, 7), (8, 8)\}$, so $P(R_1 \cap R_2) = 4/81$, as before.

Now consider the same problem but drawing without replacement. Then the two events R_1 and R_2 are no longer independent. The probability of drawing a red marble on the first draw is still $P(R_1) = 2/9$. The probability of drawing a red marble on the second drawing given that we drew a red marble the first time is $P(R_2|R_1) = 1/8$ because there are only 8 marbles left in the bag after the first drawing, and only one of these is red. Hence

$$P(R_1 \cap R_2) = P(R_2|R_1)P(R_1) = (1/8)(2/9) = 1/36 \quad (1.48)$$

(Clearly, the chances of getting two red marbles in a row are much smaller in this case than in the previous case.) Once again we were able to derive the correct result without too much thought, but once again there is a question

as to what sample space we should use.

The answer is that the sample space now consists of all ordered pairs of numbers from 1 to 9, but with the points of the form (a, a) removed:

$$\Omega'' = \{(a, b) : 1 \leq a, b \leq 9, a \neq b\}, \quad (1.49)$$

because if we do not replace the marbles between drawings we cannot get a particular marble twice in a row. There are $81 - 9 = 72$ such pairs. Two of these pairs have exactly two red marbles, namely the pairs in the set $\{(7, 8), (8, 7)\}$. Hence $P(R_1 \cap R_2) = 2/72 = 1/36$, as before.

The moral of the story is this. To find the probability of a succession of events, each of which takes place in a given sample space Ω , we can proceed in one of two ways. Either we can compute the probabilities of each event separately using Ω and then combine them using the product rule, or else we can carefully define a new sample space Ω' (or Ω'') that treats the succession of events as a single experiment, and use the rules of probability again. Both methods will give you the right answer (provided, of course, that you do the calculations correctly). Which one you choose is a matter of preference, although sometimes one way is easier than the other. (We used the first method in the examples in Section 1.5.) *But if you are ever confused about what you are doing, it is usually a good idea to define your sample space more precisely from the start.*

1.7 Exercises

1. One card is drawn at random from a deck of 52 cards. Let Q be the event that a queen is drawn. Let H be the event that a heart is drawn. Let R be the event that a red card is drawn. (a) Which two of Q , H , and R are independent? (b) What is the probability that the card is a heart? (c) What is the probability that it is a queen? (d) What is the probability that it is the queen of hearts? (e) What is the probability the card is a queen given that it is red? (f) What is the probability that it is a heart given that it is red?

2. What is the probability that a family of two children (a) has two boys given that the first child is a boy, and (b) has two boys given that at least one child is a boy?
3. Andrew, Beatrix, and Charles are playing with a crown. If Andrew has the crown, he throws it to Charles. If Beatrix has the crown, she throws it to Andrew or Charles with equal probabilities. If Charles has the crown, he throws it to Andrew or Beatrix with equal probabilities. At the beginning of the game the crown is given to one of the children with equal probabilities. What are the probabilities that each child has the coin, after the crown is thrown once?
4. A three-man jury has two members, each of whom independently has probability p of making the correct decision, and a third member who flips a coin for each decision. The jury votes according to a majority rule. A one-man jury has probability p of making the correct decision. Which jury has the better probability of making the correct decision? (Of course, this assumes there is a ‘correct’ decision!)
5. An electronics assembly firm buys its microchips from three different suppliers; half of them are bought from firm X , while firms Y and Z supply 30% and 20% respectively. The suppliers use different quality-control procedures and the percentages of defective chips are 2%, 4%, and 4% for X , Y , and Z respectively. The probabilities that a defective chip will fail two or more assembly-line tests are 40%, 60%, and 80% respectively, while all defective chips have a 10% chance of escaping detection. An assembler finds that a chip fails only one test. What is the probability that it came from supplier X ?

2 Counting

There are two parts to every probability problem. First, we must be able to assign probabilities to various events. This is where Laplace’s Principle comes into play. Second, we must be able to manipulate probabilities using the rules discussed in Section 1.5. One might think that the first part would always be trivial, since it just involves counting, but surprisingly this task can sometimes be more difficult than the second.

2.1 The Multiplication Principle

Laplace's Principle is about as simple as possible, for it defines the probability of an event E in terms of a ratio of cardinalities of two sets, E and Ω . The difficulty often lies in computing these cardinalities—that is, in counting. In our marble examples it was easy enough to count the sets involved. For example, in Section 1.6 we counted Ω' by observing that there were 9 possibilities for the first marble and 9 for the second, giving us a total of $9 \times 9 = 81$ possible combinations. It is worth formalizing this as

Principle 2.1 (Principle of Multiplication). *The number of different ordered collections of objects (a, b, \dots, c) such that $a \in A$, $b \in B$, \dots , and $c \in C$ is $|A||B| \cdots |C|$. Alternatively, the number of different ways to choose an ordered collection of objects (a, b, \dots, c) where we are free to choose the first object in $|A|$ ways, the second object in $|B|$ ways, \dots , and the last object in $|C|$ ways is $|A||B| \cdots |C|$.*

Example 6 [The Birthday Problem] What are the chances that two persons have the same birthday in a random group of n people? (Ignore leap years.)

We will compute instead the probability q that no two persons in the group have the same birthday, then use the sum rule (1.6), which tells us that the desired probability is $p = 1 - q$. It is worth doing this problem two different ways to illustrate the two approaches discussed in Section 1.6.

For the first method, we reason as follows. The probability that the first person has a birthday on *some* day is 1, which we write as $\frac{365}{365}$. The probability that the next person has a birthday on some day *other* than that of the first person is $\frac{364}{365}$, because we assume (in the absence of evidence to the contrary) that the second person could have any one of 365 birthdays with equal probability. The probability that the third person has a birthday on a day other than that of the first two is $\frac{363}{365}$, and so on down to the last person, for which the probability is $\frac{365-n+1}{365}$. Hence the probability that all of these events occur is ¹⁰

$$\frac{365}{365} \cdot \frac{364}{365} \cdot \frac{363}{365} \cdots \frac{365-n+1}{365} = \frac{365!}{(365-n)!365^n}. \quad (2.1)$$

¹⁰This follows from the product rule extended to multiple dependent events. For exam-

The second method requires that we define precisely our sample space. To do this, we view the problem as follows. We line up all the persons in order from 1 to n then select one of 365 possible birthdays at random for each person.¹¹ We want to know the probability that all the birthdays in the list are distinct. The sample space Ω consists of all possible sequences of n birthdays, whereas the event D in which we are interested is the subset of Ω corresponding to all those sequences with distinct entries. There are 365 possible birthdays for each person, so by the multiplication principle $|\Omega| = 365^n$. Next we count the sequences in which all the entries are distinct. We can pick the first entry in 365 ways, the next entry in 364 ways (because it must be distinct from the previous entry), the next entry in 363 ways, *etc.*, down to the last entry for which there are $365 - n + 1$ choices. By the multiplication principle $|D| = (365)(364)(363) \cdots (365 - n + 1)$. Hence by Laplace's Principle,

$$P(D) = \frac{(365)(364)(363) \cdots (365 - n + 1)}{365^n} = \frac{365!}{(365 - n)!365^n} \quad (2.2)$$

as before.

It is an interesting exercise to plug in a few numbers. To do this we will employ **Stirling's approximation**,¹² which is that, for large N ,

$$N! \approx \sqrt{2\pi N} N^N e^{-N} = \sqrt{2\pi N} \left(\frac{N}{e}\right)^N \quad (2.3)$$

ple, if we have three events A , B , and C , then applying Equation (1.9) twice gives

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

Applied to the birthday example with three persons, A is the event that the first person has *some* birthday, B is the event that the second person has some birthday distinct from that of person one, and C is the event that the third person has some birthday distinct from that of persons one and two.

¹¹This may seem a little strange, given that we already picked the n persons, and they already come equipped with a birthday, so to speak. But if you think about it, our reformulation of the problem is equivalent, because before we know what a person's birthday is, each birthday is equally likely.

¹²A proof of Stirling's approximation is given in Appendix A.

Then, assuming n is not too large compared to 365, we get

$$q \approx \left(\frac{365}{365 - n} \right)^{365 - n + 1/2} e^{-n} \quad (2.4)$$

Recall, though, that we really want $p = 1 - q$. Plugging in a few numbers we find that, for $n = 23$, $p = 0.507$ while for $n = 50$, $p = 0.970$. In other words, in a room of 23 people, the chances are better than 50% that at least two persons will share the same birthday, while in a room of 50 people, the chances are around 97% that at least two persons will share the same birthday. Would you have expected this? How many times have you been out to dinner in a large restaurant when two groups of persons spontaneously break into the ‘happy birthday’ song?

2.2 Ordered Sets: Permutations

Recall that a **permutation** of n objects is simply an ordered rearrangement of the objects. The set of all permutations of n objects is denoted S_n . For example, if $n = 3$ there are six permutations given by 123, 132, 213, 231, 312, and 321, so $|S_n| = 6$. Suppose we wanted $|S_{10}|$. It would take too long to actually list all the different possibilities, so we need a better way. To write down a permutation of n objects we have n choices for the first entry, $n - 1$ choices for the second entry (because we already used up one number for the first entry), $n - 2$ for the third, and so on until the very end, for which we have only one choice remaining. By the multiplication principle then, there are $n(n - 1)(n - 2) \cdots 1 = n!$ different permutations on n objects. For $n = 10$ the answer is $10! = 3,628,800$ different permutations. Incidentally, because the empty set can be permuted in only one way, we agree to set $0! = 1$.

Example 7 [Montmort’s Problem, or The Hatcheck Problem]. n men attending a banquet check their hats with an absent-minded hatcheck person who fails to attach identifying labels to the hats. The men all get drunk at the banquet, so when they retrieve their hats, each one accepts some hat given to him at random. What is the probability that no one ends up with his own hat?

We translate the problem into one involving permutations. Label the hats from 1 to n and the men from 1 to n . An assignment of hats to men is just a permutation $\sigma = (\sigma_1, \sigma_2, \dots, \sigma_n)$, where $\sigma_j = k$ if the j^{th} man receives the k^{th} hat. For example, suppose we had 4 men and 4 hats. Then one possible assignment of hats to men is 3214, which means that man 1 received hat 3, man 2 received hat 2, man 3 received hat 1, and man 4 received hat 4. If $\sigma_j = j$ we say that j is a **fixed point** of σ . For instance, the permutation 3214 has two fixed points, namely 2 and 4. The problem asks for the probability that a random element of S_n has no fixed points.

Let A_j be the event that a permutation $\sigma \in S_n$ fixes j , so that $\overline{A_j}$ is the event that σ does not fix j . Let T be the event that σ has no fixed points. Then using de Morgan's laws and Equation (1.6) we have

$$\begin{aligned} P(T) &= P(\overline{A_1} \cap \overline{A_2} \cap \dots \cap \overline{A_n}) \\ &= P(\overline{A_1 \cup A_2 \cup \dots \cup A_n}) \\ &= 1 - P(A_1 \cup A_2 \cup \dots \cup A_n), \end{aligned} \tag{2.5}$$

because T is the event that a random permutation possesses none of the properties A_j for any j .

The next step is to compute $P(A_1 \cup A_2 \cup \dots \cup A_n)$. To do this, we employ another basic counting technique known as the **principle of inclusion-exclusion**, which is a natural extension of Equation (1.5) to more than two sets:

$$\begin{aligned} P(A_1 \cup A_2 \cup \dots \cup A_n) &= \sum_i P(A_i) - \sum_{i < j} P(A_i \cap A_j) \\ &\quad + \sum_{i < j < k} P(A_i \cap A_j \cap A_k) \\ &\quad - \dots + (-1)^{n-1} P(A_1 \cap A_2 \cap \dots \cap A_n). \end{aligned} \tag{2.6}$$

The terms in this sum are computed as follows. How many permutations fix 1? Well, if we fix 1 then we can choose the rest of the numbers in our permutation in $(n-1)!$ ways, so $P(A_1) = (n-1)!/n! = 1/n$. Since there was nothing special about 1, it follows that $P(A_j) = 1/n$ for all j . There are n summands in $\sum_i P(A_i)$, so the first term is just 1. Next, consider $P(A_1 \cap A_2)$. This is the probability that a given

permutation fixes both 1 and 2. There is one way to choose 1 in the first position, one way to choose 2 in the next position, and $(n - 2)!$ ways to choose the rest of the numbers, so $P(A_1 \cap A_2) = (n - 2)!/n! = 1/n(n - 1)$. Again, there was nothing special about the pair $(1, 2)$, so $P(A_i \cap A_j) = 1/n(n - 1)$ for all $1 \leq i < j \leq n$. There are $n(n - 1)/2$ summands in $\sum_{i < j} P(A_i \cap A_j)$, because there are this many ordered pairs (i, j) with $1 \leq i < j \leq n$, so the second term contributes $-1/2!$ to the total. Continuing in this way we find

$$P(A_1 \cup A_2 \cup \dots \cup A_n) = 1 - \frac{1}{2!} + \frac{1}{3!} - \dots + (-1)^{n-1} \frac{1}{n!}. \quad (2.7)$$

Combining Equations (2.5) and (2.7) yields

$$P(T) = 1 - 1 + \frac{1}{2!} - \frac{1}{3!} + \dots + (-1)^n \frac{1}{n!} = \sum_{i=0}^n \frac{(-1)^i}{i!}. \quad (2.8)$$

This gives the probability that a randomly selected permutation has no fixed points.¹³ We recognize the expression on the right as the first n terms in the Taylor series expansion of $1/e \approx 0.37$. Hence, in the limit that $n \rightarrow \infty$, the probability that no person receives his own hat is roughly 37%.

2.3 Unordered Sets: The Binomial Coefficient

A slightly more difficult question is, how many different *unordered* collections of objects are there? More precisely, how many different ways are there to choose a k element set from an n element set? (The word ‘set’ means that the collection is unordered.) This number, called a **binomial coefficient** for reasons that will become clear later, is denoted $\binom{n}{k}$. For small n and k this number can be computed just by listing all the different possibilities. For example, if $n = 3$ and $k = 2$ the choices are $\{1, 2\}$, $\{1, 3\}$, and $\{2, 3\}$, so $\binom{3}{2} = 3$. We seek a formula valid for all n and k .

We proceed using a popular technique called **double counting**. We will

¹³A permutation with no fixed points is also called a **derangement**, so the number of derangements of S_n is $d_n = n! \sum_{i=0}^n (-1)^i / i!$.

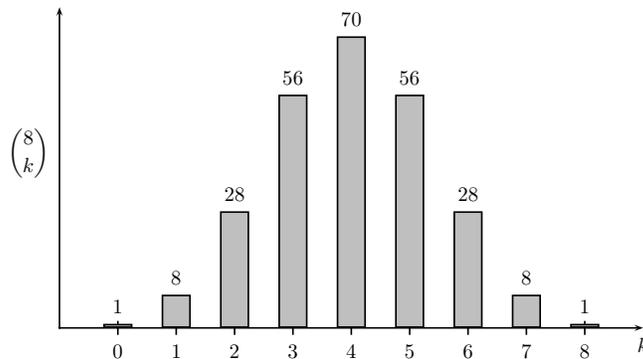


Figure 3: The binomial coefficients $\binom{8}{k}$

introduce a new set of objects and count it in two different ways. Equating the two answers will give us our result. So, let $N(n, k)$ denote the number of *ordered* collections of k distinct numbers, each chosen from the set $\{1, 2, \dots, n\}$. There are n ways to choose the first number, $n - 1$ ways to choose the second number, and so on, ending up with $n - k + 1$ ways to choose the k^{th} number, so by the multiplication principle $N(n, k) = n(n - 1) \cdots (n - k + 1) = n! / (n - k)!$. On the other hand, by definition there are $\binom{n}{k}$ ways to choose an unordered collection of k objects from n objects, and $k!$ ways to order the k objects in each collection, so $N(n, k) = \binom{n}{k} k!$. Equating these two expressions yields the important result

$$\binom{n}{k} = \frac{n!}{k!(n - k)!} \quad (\text{binomial coefficient}). \quad (2.9)$$

For example, there are $\binom{5}{3} = 10$ ways to choose 3 objects from a collection of 5 objects without regard to order. They are 123, 124, 125, 134, 135, 145, 234, 235, 245, and 345. (We should really have written $\{1, 2, 3\}$, $\{1, 2, 4\}$, *etc.*, but we have dropped the brackets and commas to save writing.) The binomial coefficients $\binom{8}{k}$ are illustrated in Figure 3.

The reason for the terminology ‘binomial coefficient’ is because the num-

bers $\binom{n}{k}$ are precisely the numbers appearing in the binomial expansion:

$$(a + b)^n = \sum_{k=0}^n \binom{n}{k} a^k b^{n-k}. \quad (2.10)$$

The way to see this is to expand the left side:

$$(a + b)^n = \underbrace{(a + b)(a + b) \cdots (a + b)}_{n \text{ times}}. \quad (2.11)$$

How many monomials of the form $a^k b^{n-k}$ are there on the right hand side? This is just asking for the number of ways we can select a set of k a 's (without regard to order) from a collection of n a 's, which is precisely the binomial coefficient $\binom{n}{k}$.

Example 8 [Poker Hands] A poker hand of 5 playing cards is dealt from a well-shuffled pack of 52 cards. What is the probability that the hand contains a pair of aces?

There are $\binom{52}{5}$ possible poker hands, and each one is equally likely. There are $\binom{4}{2}$ pairs of aces, and $\binom{52-4}{3}$ ways to complete a hand containing such a pair (because we must choose the remaining three cards from all the cards that are not aces), so there are $\binom{48}{3} \binom{4}{2}$ hands containing a pair of aces. The probability is therefore

$$\frac{\binom{48}{3} \binom{4}{2}}{\binom{52}{5}} = \frac{48!4!47!5!}{45!3!2!2!52!} = \frac{48 \cdot 47 \cdot 46 \cdot 4 \cdot 3 \cdot 5 \cdot 4 \cdot 3 \cdot 2}{3 \cdot 2 \cdot 2 \cdot 52 \cdot 51 \cdot 50 \cdot 49 \cdot 48} = 0.040 \quad (2.12)$$

2.4 Multisets

Although we did not define the notion of a set, we intuitively understand that it is a collection of distinct objects. A **multiset** is a generalization of the idea of a set in which we allow repetitions of objects. So, for instance, $\{1, 3, 5\}$ is a set, while $\{1, 1, 2, 2, 2, 4\}$ is a multiset. (Note that, as with sets,

the order of the elements is irrelevant.) The number of multisets of k objects chosen from a collection of n objects is denoted $\binom{n}{k}$. Again we can consider a small example with $n = 3$ and $k = 2$. Then (using the same abbreviated notation as for sets) we have the multisets 11, 12, 13, 22, 23, and 33, so $\binom{3}{2} = 6$. We wish to find an expression for this number as a function of n and k .

To enumerate multisets we first transform the problem into an equivalent one. We can represent multisets using something called multiplicity notation, which is best illustrated by an example. In multiplicity notation the multiset $\{1, 1, 2, 2, 2, 4\}$ is represented by $1^2 2^3 3^0 4^1$, where the exponent encodes the number of times the corresponding number appears in the multiset. Observe that the sum of the exponents is $2 + 3 + 0 + 1 = 6$, which is the total number of elements in the multiset. In general, a multiset of size k on an n element set is represented by the expression $1^{s_1} 2^{s_2} \cdots n^{s_n}$ where for each $1 \leq i \leq n$, $s_i \geq 0$ and $\sum_i s_i = k$. The problem of enumerating these multisets thus becomes one of finding all solutions to the equation

$$s_1 + s_2 + \cdots + s_n = k, \tag{2.13}$$

where each s_i is a nonnegative integer.

To count the number of solutions to Equation (2.13) we employ a technique affectionately known as **stars and bars**. Again, an example will help to illustrate the basic idea. Suppose $n = 4$ and $k = 6$. Then we write the following string of $k = 6$ stars and $n - 1 = 3$ bars:

$$** \mid *** \mid \mid * \tag{2.14}$$

which represents the sum

$$2 + 3 + 0 + 1. \tag{2.15}$$

In general, there is a one-to-one correspondence between sums of the form (2.13) and strings of k stars and $n - 1$ bars. How many such strings are there? We can imagine writing down a string of $k + n - 1$ stars, and selecting

$n - 1$ of them to change into bars. (In the above example we would start with 9 stars, then select 3 of the stars to change into bars.) But thanks to our work above, we know that the number of ways to do this is precisely $\binom{k+n-1}{n-1}$ because we are asking for the number of unordered sets of $n - 1$ objects chosen from a set containing $k + n - 1$ objects. Hence

$$\left(\binom{n}{k} \right) = \binom{k+n-1}{n-1} \quad (\# \text{ of multisets of size } k \text{ on } n \text{ objects}). \quad (2.16)$$

To illustrate the formula, we have

$$\left(\binom{3}{2} \right) = \binom{4}{2} = 6, \quad (2.17)$$

which agrees with our enumeration of this case given above.

2.5 Multiset Permutations: Multinomial Coefficient

A multiset is unordered, but we can ask for the number of different ways to order a multiset. Such an object is called a **multiset permutation** and is written with parentheses instead of curly brackets to indicate that we are now specifying an order. So, for example $(1, 2, 1, 2, 4, 2)$ is a multiset permutation corresponding to the multiset $\{1, 1, 2, 2, 2, 4\}$.

Multiset permutations are a bit more subtle than ordinary permutations, because if we interchange two of the same symbols (*e.g.*, two ones) the result is indistinguishable from the multiset permutation with which we started. But there is a simple trick to get around this. We imagine putting subscripts on each of the numbers in order to distinguish them temporarily. So, for the example above, we would consider the set $\{1_1, 1_2, 2_1, 2_2, 2_3, 4_1\}$. Now we consider all possible permutations of these six distinguishable objects. There are $6!$ such permutations. But, because the subscripts are artificial, we have actually overcounted the number of distinguishable permutations by the number of ways we can order each of the numbers. For example, if there were no subscripts, $(1_1, 2_2, 1_2, 2_1, 4_1, 2_3)$ and $(1_1, 2_3, 1_2, 2_1, 4_1, 2_2)$ would

be the same. A little thought shows that we must divide $6!$ by the product $2!3!0!1!$ to avoid overcounting. In general, the number of permutations of a multiset represented in multiplicity notation by $1^{s_1}2^{s_2}\cdots n^{s_n}$ is given by the **multinomial coefficient**

$$\binom{k}{s_1, s_2, \dots, s_n} = \frac{k!}{s_1!s_2!\cdots s_n!} \quad (\text{multinomial coefficient}), \quad (2.18)$$

where it is always understood that $k = \sum_i s_i$. For example, the number of multiset permutations of $\{1, 1, 2, 2, 2, 4\}$ is $\binom{6}{2,3,0,1} = 60$.

2.6 Choosing Collections of Subsets: Multinomial Coefficient Again

The multinomial coefficient reduces to the binomial coefficient in the case $n = 2$. But the binomial coefficient had a nice interpretation in terms of choosing subsets of a set. We are therefore led to ask whether there is another interpretation of the multinomial coefficient that is similar to that of the binomial coefficient. The answer is yes. Consider the binomial coefficient again. To form a set of k objects from a set of n objects we choose k of the n elements, leaving $n - k$ elements. That is, there are $\binom{n}{k}$ ways to partition n into two sets, one of size k and one of size $n - k$.

It turns out that $\binom{k}{s_1, s_2, \dots, s_n}$ counts the number of ways to partition a set of k objects into n sets of sizes (s_1, s_2, \dots, s_n) . The proof goes as follows. We start with a set of k objects, from which we choose s_1 objects. There are $\binom{k}{s_1}$ ways to do this. Then, from the remaining $k - s_1$ objects we choose s_2 objects in $\binom{k-s_1}{s_2}$ ways. We continue in this way until we have chosen all the objects from our original set. Using the formula for the binomial coefficient and the multiplication principle this gives

$$\frac{k!}{s_1!(k-s_1)!} \cdot \frac{(k-s_1)!}{s_2!(k-s_1-s_2)!} \cdots \frac{(k-s_1-s_2-\cdots-s_{n-1})!}{s_n!0!} \quad (2.19)$$

different ways to accomplish our task. After canceling terms from the nu-

erator and denominator we are left with the multinomial coefficient, as claimed.

As was the case for the binomial coefficient, the multinomial coefficient also has an interpretation in terms of the expansion of a product. Specifically, we have

$$(x_1 + x_2 + \cdots + x_n)^k = \sum_{\substack{s_i \geq 0 \\ s_1 + s_2 + \cdots + s_n = k}} \binom{k}{s_1, s_2, \dots, s_n} x_1^{s_1} x_2^{s_2} \cdots x_n^{s_n}. \quad (2.20)$$

The proof is left to the reader.

Example 9 Eight marbles are selected one at a time from a bag of differently colored marbles. How many distinct ways are there of getting three red, two green, and three blue marbles?

$$\binom{8}{3, 2, 3} = \frac{8!}{3!2!3!} = 560. \quad (2.21)$$

Example 10 [Quantum States] Suppose we have a collection of N identical quantum systems (for instance, N identical particles). Each system can be in one of m possible states $\{s_1, s_2, \dots, s_m\}$. If there are n_i systems in state s_i for $1 \leq i \leq m$ we say that the collection of N systems is in a **configuration** of type (n_1, n_2, \dots, n_m) . (Note that we must have $\sum_i n_i = N$.) It follows immediately from the above discussion that there are

$$\binom{N}{n_1, n_2, \dots, n_m} \quad (2.22)$$

different configurations of type (n_1, n_2, \dots, n_m) .

2.7 Exercises

1. Prove that the product of any n consecutive integers is divisible by $n!$.

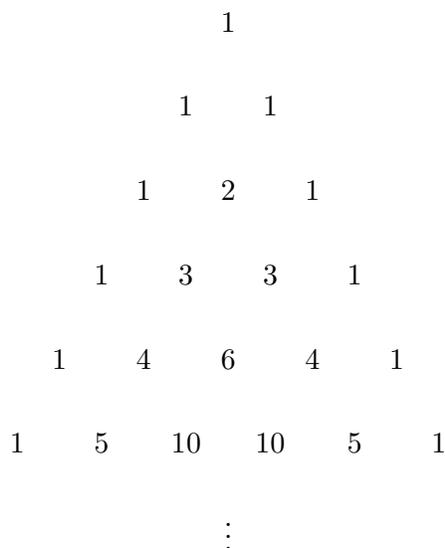


Figure 4: The first six rows of Pascal's triangle

2. Define a row vector $\mathbf{a}(n) = \left(\binom{n}{0}, \binom{n}{1}, \dots, \binom{n}{n}\right)$. If we stack these vectors on top of one another, starting with $\mathbf{a}(0)$, then $\mathbf{a}(1)$, and so on, we get a triangular array of numbers known as **Pascal's Triangle** (see Figure 4.) Prove that an entry in of any given row can be obtained by adding the two entries that lie nearest and in the row directly above. That is, show that

$$\binom{n}{k} = \binom{n-1}{k-1} + \binom{n-1}{k}. \quad (2.23)$$

3. Prove that the sum of the entries in the n^{th} row Pascal's triangle is 2^n .
4. A **lattice path** on a two dimensional grid is a sequence consisting of upward steps (from (a, b) to $(a, b + 1)$) and rightward steps (from (a, b) to $(a + 1, b)$). Show that the number of lattice paths from $(0, 0)$ to (n, n) is $\binom{2n}{n}$.
5. A club with twenty members decides to form three committees, with four, five, and six members, respectively. How many ways can this be done? [Leave your answer in terms of factorials.]

3 Random Variables and Probability Distributions

Having explored some basic counting techniques we now develop some other useful tools that will allow us to investigate various probabilistic phenomena in more detail. A **random variable** is just a map $X : \Omega \rightarrow \mathbb{R}$. That is, it is an assignment of a real number to every element of the sample space. Sometimes the sample space already comes equipped with a natural random variable. For example, in the game of craps two dice are thrown and one is interested in the total number showing on both dice. The sample space $\Omega = \{(a, b) : 1 \leq a, b \leq 6\}$ is the set of all pairs of numbers from 1 to 6, and the random variable of interest is $X(a, b) = a + b$, the sum of the numbers showing on the two dice. Similarly, if we want to know the distribution of heights in a population, then the appropriate random variable X is the height of each person. In other problems we may decide to assign a random variable.

Given a random variable X and a real number x we define the event χ_x in Ω by

$$\chi_x = \{\omega \in \Omega : X(\omega) = x\}. \quad (3.1)$$

That is, χ_x is the collection of all points in Ω where X achieves the value x . The **probability distribution** $f_X(x)$ gives us the probabilities of all the events χ_x :

$$f_X(x) := P(\chi_x) = P(X = x), \quad (3.2)$$

where the last expression is just a convenient shorthand notation. Probability distributions come in two types, discrete and continuous, and we need to understand both. We begin with discrete distributions.

3.1 Discrete Random Variables

A random variable that assumes only the discrete values x_1, x_2, \dots , with probabilities p_1, p_2, \dots , is called a **discrete random variable**. In this case

the probability distribution f_X is also discrete, because it is defined by the discrete list of numbers $f_X(x_i) = p_i$. In the game of craps, for example, the random variable X representing the sum of the two dice only assumes integer values from 2 to 12, with probabilities $p_2 = 1/36$, $p_3 = 1/18$, $p_4 = 1/12$, $p_5 = 1/9$, $p_6 = 5/36$, $p_7 = 1/6$, $p_8 = 5/36$, $p_9 = 1/9$, $p_{10} = 1/12$, $p_{11} = 1/18$, and $p_{12} = 1/36$.¹⁴

3.1.1 The Mean or Expectation Value

Probability distributions contain a lot of information, often too much, so that it is necessary to be able to characterize the general features of a distribution. Arguably the most important property of a probability distribution of a random variable is its **mean** (also known as the **expectation value**)

$$\langle X \rangle = \sum_i x_i p_i. \quad (3.3)$$

For example, for the distribution given above,

$$\begin{aligned} \langle X \rangle &= 2 \cdot \frac{1}{36} + 3 \cdot \frac{1}{18} + 4 \cdot \frac{1}{12} + 5 \cdot \frac{1}{9} + 6 \cdot \frac{5}{36} + 7 \cdot \frac{1}{6} \\ &\quad + 8 \cdot \frac{5}{36} + 9 \cdot \frac{1}{9} + 10 \cdot \frac{1}{12} + 11 \cdot \frac{1}{18} + 12 \cdot \frac{1}{36} \\ &= 7 \end{aligned} \quad (3.4)$$

If we throw the dice many times, the average value of all the throws (*i.e.*, the sum of all the values divided by the number of throws) converges to

¹⁴There are a few subtleties here that may be worth discussing. The index i here does not refer to the points of the sample space Ω , which consists of the 36 possible pairs of numbers that could appear on the dice. Instead it indexes the possible values of X . The sample space is split into eleven distinguished events $\chi_i := \chi_{x_i}$ corresponding to the pairs of dice that have the value x_i . For instance, $\chi_4 = \{(1, 3), (2, 2), (3, 1)\}$. The indexing is arbitrary, but is chosen in a way that is natural for the problem. So, instead of setting $x_1 = 2, x_2 = 3, \dots, x_{11} = 12$, we choose to set $x_2 = 2, x_3 = 3, \dots, x_{12} = 12$. This makes it easier to see what is going on with our random variable X , but is a potential source of confusion as well, because for many problems it is not possible to perform the indexing in this way. Probabilists often make these kinds of choices without comment, as shall we.

the mean of the distribution.¹⁵ This is the sense in which it is the ‘expected value’. Note that the mean is *not* necessarily the **most probable value** (also known as the **mode** of the distribution). For example, if the dice were weighted so that snake eyes showed up more often than any other number, then the most probable value would be 2, but the mean would be higher than 2 (simply because some numbers higher than 2 show up on occasion). Yet another statistic that is often discussed is the **median** of the distribution, which is the number m such that half the values of X lie above m and half lie below m , but this one is not as informative as the other two.

Example 11 An American roulette wheel has pockets numbered from 1 to 36 and two extra pockets labeled 0 and 00 for a total of 38 places where the roulette ball can land. A wide variety of bets are possible. One of the most popular bets is even or odd, which pays 1 to 1. If the player bets that the ball will land in, say, an even numbered pocket and it does, he keeps his wager and wins an amount equal to his original wager. If it lands in an odd pocket or in the special numbers 0 and 00, he loses his wager. What is the expected gain per wager?

The probability that the player wins on one spin is $18/38=9/19$, so the probability of a loss on one spin is $20/38=10/19$. We want the expected gain per wager, so we assume a wager of 1 unit. The amount gained for a win is $w = 1$ while the amount gained for a loss is $w = -1$. In this case w is our random variable, and the expected gain is

$$(9/19)(1) + (10/19)(-1) = -1/19 = -0.053 \quad (3.5)$$

Hence, in the long run, the player will lose about five cents for every dollar wagered.

Example 12 Another player, more of a gambler than the first, decides to play the 0, which pays at 35 to 1. This means that, if the ball lands in the 0 he keeps his wager and wins 35 times his original wager. If it lands anywhere else, he loses his wager. What is the expected loss per wager?

¹⁵This is a theorem in probability called the **Law of Large Numbers**.

Now the probability of winning is $1/38$, while the probability of losing is $37/38$. The amount gained for a wager of 1 is $w = 35$ while the amount gained for a loss is $w = -1$. The expected gain is now

$$(1/38)(35) + (37/38)(-1) = -2/38 = -1/19 = -0.053 \quad (3.6)$$

which is exactly the same as before! In the long run, each player expects to win (or, more to the point, lose) the same amount per wager. Of course, the difference between the two players is that the former player's holdings tend to stay near zero, while the latter player's holdings tend to fluctuate much more wildly. (See the next example.)

3.1.2 The Variance and the Standard Deviation

The **variance** of a distribution is

$$\sigma_X^2 = \langle (X - \langle X \rangle)^2 \rangle = \langle X^2 \rangle - \langle X \rangle^2, \quad (3.7)$$

where

$$\langle X^2 \rangle = \sum_i x_i^2 p_i. \quad (3.8)$$

The **standard deviation** σ_X is the square root of the variance. In our dice example we have

$$\begin{aligned} \langle X^2 \rangle &= 4 \cdot \frac{1}{36} + 9 \cdot \frac{1}{18} + 16 \cdot \frac{1}{12} + 25 \cdot \frac{1}{9} + 36 \cdot \frac{5}{36} + 49 \cdot \frac{1}{6} \\ &\quad + 64 \cdot \frac{5}{36} + 81 \cdot \frac{1}{9} + 100 \cdot \frac{1}{12} + 121 \cdot \frac{1}{18} + 144 \cdot \frac{1}{36} \\ &= 54.83 \end{aligned} \quad (3.9)$$

and $\langle X \rangle = 7$, so the standard deviation is

$$\sigma_X = \sqrt{54.83 - 49} = 2.4. \quad (3.10)$$

The standard deviation is a measure of how far the values of X are distributed on either side of the mean. If the standard deviation is small we say that the distribution is ‘narrowly peaked’ about the mean; if not, we say the distribution is ‘wide’. The number 2.4 is somewhere in between narrow and wide for our example. (See Section 3.4.2 below for a more precise statement.)

Example 13 Let us return to the case of our two roulette players. Even though both have the same expected loss per wager, our intuitions tell us that something is different between the two situations. We would expect that the conservative gambler who bets on even or odd would see his holdings rise or fall (but on average, fall) by only a little bit each time, whereas we expect to see larger fluctuations in the holdings of the more reckless gambler. This can be seen in the standard deviation of the corresponding random variables. In the case of the conservative gambler, the standard deviation is obtained as follows:

$$\langle w^2 \rangle = (9/19)(1)^2 + (10/19)(-1)^2 = 1 \quad (3.11)$$

$$\langle w \rangle^2 = (-1/19)^2 = 0.0028 \quad (3.12)$$

$$\sigma_w = \sqrt{1 - 0.0028} \approx 1. \quad (3.13)$$

The value of 1 just confirms our intuition that the gain for this gambler tends to stay within plus or minus 1 of the mean, which is to say between -1.053 and 0.947. For the reckless gambler the standard deviation is

$$\langle w^2 \rangle = (1/38)(35)^2 + (37/38)(-1)^2 = 33.2 \quad (3.14)$$

$$\langle w \rangle^2 = (-1/19)^2 = 0.0028 \quad (3.15)$$

$$\sigma_w = \sqrt{33.2 - 0.0028} \approx 5.8, \quad (3.16)$$

which shows that this gambler’s gain fluctuates more wildly.

3.2 Discrete Probability Distributions

3.2.1 The Binomial Distribution

The simplest type of experiment one can perform is one in which there are only two possible outcomes, say A and B , with different probabilities of occurrence. By convention we usually set $P(A) = p$ and $P(B) = q = 1 - p$. For example, we might flip a coin that is weighted, so that it shows heads with probability $p = 0.6$ and tails with probability $q = 0.4$. If we repeat this experiment n times, then A will occur X times, where $0 \leq X \leq n$. The **binomial distribution** $f(n, k) = P(X = k)$ gives the distribution of the discrete random variable X .¹⁶

Computing these probabilities is another exercise in counting. We will call the outcome of any particular set of n experiments a *configuration of length n* . For example, one possible configuration of length 8 is $ABBAAABA$. For this configuration $X = 5$. The first question is, how many different configurations of length n are there? Well, this is easy. There are two choices for the first outcome, two for the next, and so on, so by the multiplication principle there are 2^n possible configurations of length n . Of these, how many have $X = k$? That is, how many have k A 's? The answer is $\binom{n}{k}$. One way to see this is to number the elements of a configuration from 1 to n . We choose k of these to be A 's by choosing a subset of size k from these n numbers, and there are $\binom{n}{k}$ ways to do this.¹⁷

In the very special case that each configuration is equally likely, Laplace's Principle tells us that the probability distribution is simply

$$f(n, k) = \frac{1}{2^n} \binom{n}{k} \quad (\text{binomial distribution, equal probabilities}) \quad (3.17)$$

¹⁶For notational convenience we now drop the subscript X from the distribution function $f_X(x)$, as the random variable is usually clear from context. In the case of the binomial distribution the new notation also serves to remind us that the distribution depends on the values of both n and k .

¹⁷The fact that all our choices 'look alike'—that is, they are all the same letter A —means that we want to choose our sets without regard to order.

But in general this is not true. Indeed, by the product rule for probabilities, we know that the probability that any particular configuration occurs is just the product of the probabilities for each of the independent events. For example, the probability to get the configuration $ABBAAABA$ is $pqqppppp = p^5q^3$. This means that different configurations have different probabilities. Fortunately, the situation is simplified by the fact that the probability depends only on the *number* of A 's and B 's that appear, and not *where* they appear in the configuration. Each configuration with k A 's and $n - k$ B 's has the same probability of occurrence, namely p^kq^{n-k} . We already agreed that there were $\binom{n}{k}$ such configurations,¹⁸ so the probability that such a configuration appears is, by the additivity property of probabilities for mutually exclusive events,¹⁹

$$f(n, k) = \binom{n}{k} p^k q^{n-k} \quad (\text{binomial distribution, general case}) \quad (3.18)$$

Observe that Equation (3.18) reduces to Equation (3.17) in the case $p = q = 1/2$. The binomial distribution is normalized so that the sums of all the probabilities is unity:

$$\sum_{k=0}^n f(n, k) = 1. \quad (3.19)$$

Example 14 We flip a fair coin 5 times. What is the probability that heads shows up 3 times?

The coin is fair, so $p = q = 1/2$. Hence

$$f(5, 3) = P(X = 3) = 2^{-5} \binom{5}{3} = \frac{10}{32} = \frac{5}{16} \quad (3.20)$$

¹⁸This depended only on a counting argument and not on any probabilities.

¹⁹Each configuration is unique, so the configurations are mutually exclusive events.

Example 15 We throw a six-sided die six times in a row. What is the probability we get two sixes?

Now $p = 1/6$, because this is the probability of obtaining a six on a single throw. Hence

$$f(6, 2) = P(X = 2) = \binom{6}{2} \left(\frac{1}{6}\right)^2 \left(\frac{5}{6}\right)^4 = 15 \cdot \frac{625}{46656} = 0.20 \quad (3.21)$$

3.2.2 The Poisson Distribution

Many times we have a situation in which a certain event occurs at random but with a definite average rate, and we want to know that probability that X events happen in a particular time interval τ .²⁰ This kind of situation applies to cars passing over a bridge, telephone calls coming into a switching center, people lining up at the post office, radioactive decays of particles, and so on. All these processes are governed by the **Poisson distribution** $f_\mu(k) = P_\mu(X = k)$, where μ is the average rate of the process.

The Poisson distribution can be viewed as a kind of limiting distribution of the binomial distribution in which the number n of trials goes to infinity and the probability p of ‘success’ goes to zero in such a way that the average $np = \mu$ remains finite. We can see this as follows. Suppose we have a sample of a certain radioactive material which is known to have an average decay rate of μ decays per second.²¹ We divide up the time interval τ (in this case, one second) into a large number n of very small time intervals in which the probability of a single atomic decay is p (which is therefore small). Assuming all the atoms act independently, we are just asking for the probability that we get k decays in n separate decay events. This is the precisely what the binomial distribution gives us, provided we set the mean number μ of decays

²⁰Note that we are using the word ‘event’ in the colloquial sense.

²¹For a radioactive decay process, the average number of particles remaining after a time t is given by $N = N_0 e^{-\lambda t}$, where λ is the radioactive decay constant and N_0 is the initial number of particles. The average decay rate μ , called the **activity**, is $|dN/dt| = \lambda N$.

in a time τ to the value np , which is the mean of the binomial distribution. This gives

$$f_\mu(k) = \lim_{\substack{n \rightarrow \infty \\ p \rightarrow 0 \\ np = \mu}} \binom{n}{k} p^k q^{n-k} \quad (3.22)$$

$$= \lim_{n \rightarrow \infty} \frac{n!}{k!(n-k)!} \cdot \left(\frac{\mu}{n}\right)^k \left(1 - \frac{\mu}{n}\right)^{n-k}. \quad (3.23)$$

Now

$$\lim_{n \rightarrow \infty} \frac{n!}{(n-k)!} = \lim_{n \rightarrow \infty} n(n-1)(n-2) \cdots (n-k+1) \approx n^k \quad (3.24)$$

and

$$\lim_{n \rightarrow \infty} \left(1 - \frac{\mu}{n}\right)^{n-k} = \lim_{n \rightarrow \infty} \left(1 - \frac{\mu}{n}\right)^n = e^{-\mu}, \quad (3.25)$$

so the Poisson probability distribution function is

$$f_\mu(k) = e^{-\mu} \frac{\mu^k}{k!} \quad (\text{Poisson distribution}). \quad (3.26)$$

The mean and variance of the Poisson process are both equal to μ . The latter fact implies that, for a Poisson process, the random variable k is distributed about the mean with a standard deviation of $\sqrt{\mu}$. This is the famous ‘square root rule’ of Poisson statistics. It implies that, if we perform a real experiment where the probability p that a particular event occurs is small, and we count ν occurrences of our event in a time interval τ , then our best estimate for μ is ν , and our uncertainty is $\pm\sqrt{\nu}$. The Poisson distribution function is normalized to unity:

$$\sum_{k=0}^{\infty} f_\mu(k) = 1. \quad (3.27)$$

Example 16 A certain radioactive material is measured to have an average decay rate of 0.5 particles per second. We count decays for 4 seconds. What is the

expected number of decays? What is the probability of actually measuring this number of decays?

The expected number of decays is just $(0.5)(4) = 2$. The probability of actually measuring this many decays is

$$f_2(2) = e^{-2} \frac{2^2}{2!} = 0.27. \quad (3.28)$$

This means that there is only a 27% chance of observing exactly 2 decays in the 4 second time interval.

3.3 Continuous Random Variables

Suppose we measure the height X of all persons in a given population. Then X is a continuous random variable, because it can assume any real value. We can still ask for the probability that $X = x$, but now the answer is always zero! This is because there is no chance to get exactly the height of, say, 6 feet, because no person's height is exactly 6 feet. A person may have a height of 6 ft and 0.00002 inches, but that is not precisely 6 feet.²² Instead, the only meaningful idea is that the height of person lies somewhere in a certain range, say, between 5'9" and 6'1". For this reason, we must be a little more careful in how we define our probability distribution.

If X is a continuous random variable, we define the probability distribution function $f(x)$ by

$$f(x) dx = \text{probability that } X \text{ lies in the interval } [x, x + dx]. \quad (3.29)$$

This means that the distribution function $f(x)$ is not a probability *per se*, but rather a **probability density**, more specifically, a probability per unit interval. In order to specify such a probability distribution, we need a continuous

²²Indeed, 'height' is actually an ill defined concept, because if we continue to make finer and finer measurements of the height of an object we eventually get down to the microscopic level where the surface disappears in a quantum haze. But for now we ignore this complication.

function of x .²³

3.3.1 The Mean, Average, or Expectation Value

In the discrete case the mean value of the random variable X was a weighted sum of the possible values of X times the probability of their occurrence. As you might imagine, in the continuous case we must replace a sum by an integral. Thus, we define the mean value of the continuous probability distribution $f(x)$ to be

$$\langle X \rangle = \int_a^b x f(x) dx, \quad (3.30)$$

where the probability distribution function is defined on the interval $[a, b]$.

Example 17 What is the mean value of the **exponential probability distribution** $f(x) = \lambda e^{-\lambda x}$ on the interval $[0, \infty)$?

We compute

$$\begin{aligned} \langle X \rangle &= \lambda \int_0^{\infty} x e^{-\lambda x} dx = -\lambda \frac{d}{d\lambda} \int_0^{\infty} e^{-\lambda x} dx \\ &= -\lambda \frac{d}{d\lambda} \left(-\frac{1}{\lambda} e^{-\lambda x} \Big|_0^{\infty} \right) = -\lambda \frac{d}{d\lambda} \left(\frac{1}{\lambda} \right) = \frac{1}{\lambda}. \end{aligned} \quad (3.31)$$

3.3.2 The Variance and the Standard Deviation

The variance of a continuous random variable is obtained using the exact same formula as in the discrete case, namely

$$\sigma_X^2 = \langle X^2 \rangle - \langle X \rangle^2. \quad (3.32)$$

and the standard deviation is σ_X .

²³Even though the Poisson probability distribution is a continuous function of k , it is still a *discrete* distribution, because k is supposed to be an integer.

Example 18 What is the variance of the exponential probability distribution $f(x) = \lambda e^{-\lambda x}$ on the interval $[0, \infty]$?

We compute

$$\begin{aligned} \langle X^2 \rangle &= \lambda \int_0^\infty x^2 e^{-\lambda x} dx = \lambda \frac{d^2}{d\lambda^2} \int_0^\infty e^{-\lambda x} dx \\ &= \lambda \frac{d^2}{d\lambda^2} \left(-\frac{1}{\lambda} e^{-\lambda x} \Big|_0^\infty \right) = \lambda \frac{d^2}{d\lambda^2} \left(\frac{1}{\lambda} \right) = \frac{2}{\lambda^2}, \end{aligned} \quad (3.33)$$

so

$$\sigma_X^2 = \langle X^2 \rangle - \langle X \rangle^2 = \frac{2}{\lambda^2} - \frac{1}{\lambda^2} = \frac{1}{\lambda^2}. \quad (3.34)$$

3.4 Continuous Probability Distributions

3.4.1 The Uniform Distribution

Undoubtedly, the most important of all probability distributions is also the most boring of all distributions, namely the **uniform distribution**, which assigns equal probabilities to every interval of the same length. If the random variable X takes values between a and b , then the uniform distribution is given by

$$f(x) = \begin{cases} (b-a)^{-1} & \text{for } a \leq x \leq b, \\ 0 & \text{otherwise.} \end{cases} \quad (3.35)$$

This distribution has mean $(a+b)/2$ and variance $(b-a)^2/12$.

3.4.2 The Gaussian or Normal Distribution

The second most important probability distribution is the **Gaussian distribution** (also known as the **normal distribution** or **bell shaped curve**), defined as

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp \left\{ -\frac{1}{2} \left(\frac{x-\mu}{\sigma} \right)^2 \right\} \quad -\infty \leq x \leq \infty. \quad (3.36)$$

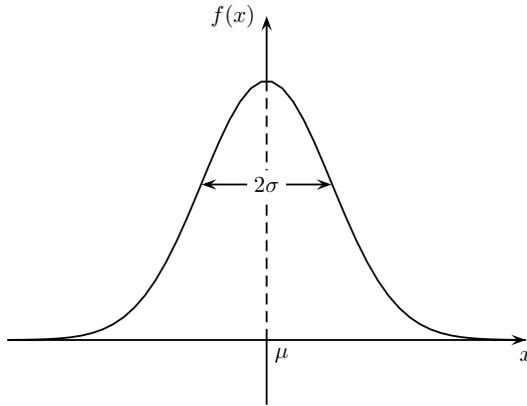


Figure 5: The Gaussian distribution

The mean of the distribution is μ and the standard deviation is σ . The factor in front is called a **normalization factor** and is present so that the sum of all the probabilities is unity:

$$\int_{-\infty}^{\infty} f(x) dx = 1. \quad (3.37)$$

A Gaussian distribution is sketched in Figure 5. From the figure we see that it is symmetric about $x = \mu$ and that σ governs the width of the distribution. A more precise statement is given below.

The Gaussian distribution is ubiquitous in all areas of mathematics and physics. The fundamental reason for its prominence is because of the Central Limit Theorem, which very roughly says that enough random systems put together always follow a Gaussian distribution, even if the individual systems do not. This immediately implies that the Gaussian distribution arises as the limit of the binomial distribution in which the probability p of success in a single trial stays finite but the number n of trials goes to infinity (so that

np also goes to infinity).²⁴ Specifically, in the limit of large n ,

$$f(n, k) \rightarrow \frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{npq}} \exp \left\{ -\frac{1}{2} \frac{(k - np)^2}{npq} \right\}. \quad (3.38)$$

A proof when $p = q = 1/2$ is supplied in Appendix B.

If we have a continuous random variable X obeying a Gaussian distribution with mean μ and variance σ^2 , the probability that X lies between x and $x + dx$ is

$$P(x < X < x + dx) = f(x) dx = \frac{1}{\sigma\sqrt{2\pi}} \exp \left\{ -\frac{1}{2} \left(\frac{x - \mu}{\sigma} \right)^2 \right\} dx, \quad (3.39)$$

so the probability that X can be found in some finite interval $[a, b]$ is

$$P(a < X < b) = \int_a^b f(x) dx = \frac{1}{\sigma\sqrt{2\pi}} \int_a^b \exp \left\{ -\frac{1}{2} \left(\frac{x - \mu}{\sigma} \right)^2 \right\} dx. \quad (3.40)$$

This integral cannot be evaluated analytically so must be computed numerically. Instead of computing these integrals for every different Gaussian distribution they are all compared to the **standard Gaussian distribution** with $\mu = 0$ and $\sigma = 1$. The values of the **cumulative probability function**

$$F(z) := \frac{1}{\sqrt{2\pi}} \int_{-\infty}^z e^{-t^2/2} dt \quad (3.41)$$

are tabulated in standard references and can be used to evaluate integrals of the type given in (3.40) by observing that, if we define a new random variable by $Z := (X - \mu)/\sigma$ then

$$P(a < X < b) = P \left(\frac{a - \mu}{\sigma} < Z < \frac{b - \mu}{\sigma} \right) = F \left(\frac{b - \mu}{\sigma} \right) - F \left(\frac{a - \mu}{\sigma} \right). \quad (3.42)$$

²⁴Contrast this with the Poisson distribution, which arises from the binomial distribution when n goes to infinity but p goes to zero in such a way that np remains constant.

Example 19 If X is normally distributed with mean μ and standard deviation σ , what is the probability that X lies within one, two and three standard deviations of the mean?

From Equation (3.42) we have

$$\begin{aligned} P(\mu - m\sigma < X < \mu + m\sigma) &= F(m) - F(-m) \\ &= F(m) - (1 - F(m)) \\ &= 2F(m) - 1. \end{aligned} \tag{3.43}$$

By consulting a standard table we find

$$P(\mu - \sigma < X < \mu + \sigma) = 2F(1) - 1 = 0.6826 \tag{3.44}$$

$$P(\mu - 2\sigma < X < \mu + 2\sigma) = 2F(2) - 1 = 0.9544 \tag{3.45}$$

$$P(\mu - 3\sigma < X < \mu + 3\sigma) = 2F(3) - 1 = 0.9974 \tag{3.46}$$

Thus, one says that the probability of finding X within one standard deviation of the mean is 68% (the ‘one-sigma limit’), while the probability of finding X within two standard deviations of the mean is 95% (the ‘two-sigma limit’), and the probability of finding X within three standard deviations of the mean is 99.8% (the ‘three-sigma limit’). These results play a prominent role in hypothesis testing in experiments, because if you perform some experiment that you know (or perhaps only suspect ²⁵) obeys a normal distribution, and if your result lies outside, say, the two-sigma limit, then it is highly unlikely that it happened by chance alone, which means that something else must be a work.

3.5 Exercises

1. On the average, how many times must a die be thrown until one gets a 6? [Hint: Consider the derivative of the geometric series $\sum_k x^k$.]
2. Chuck-a-Luck is a gambling game often played at carnivals and gambling

²⁵Other statistical tests can be performed to check this.

houses. A player may bet on any one of the numbers 1, 2, 3, 4, 5, 6. Three six-sided dice are rolled. If a player's number appears on one, two, or three of the dice, he receives respectively one, two, or three times his original stake plus his own money back; otherwise he loses his stake. What is the player's expected loss per stake? (Actually, the player may distribute stakes on several numbers, but each such stake can be regarded as a separate bet.)

3. Show that the two expressions for the variance given in Equation (3.7) are indeed equal.
4. A bag contains five red, three blue, three green, and four yellow marbles. Six children in turn each pick one marble then return it to the bag. What is the probability that the six children choose two red, two blue, and two green marbles? [Hint: One approach to this problem is to generalize the discussion in Section 3.2.1 to an experiment with three possible outcomes, A , B , and C . This will require the use of multinomial coefficients.]
5. Show that the mean of the binomial distribution is np . [Hint: Consider the binomial expansion of $(1 + x)^n$. Apply the differential operator $x(d/dx)$ to both sides and relate the resulting expressions to $\langle X \rangle$.]
6. Show that the mean of the Poisson distribution is μ .
7. During each one second interval a switchboard operator receives one call with probability $p = 0.01$ and no calls with probability $1 - p = 0.99$. Use the Poisson distribution to estimate the probability that the operator misses at most one call if she takes a 5 minute coffee break.
8. An expert witness in a paternity suit testifies that the length (in days) of a pregnancy (from conception to delivery), is approximately normally distributed with a mean of $\mu = 270$ and a standard deviation of $\sigma = 10$. The defendant in the suit is able to prove that he was out of the country during the period from 290 to 250 days before the birth of the child. What is the probability that the defendant was in the country when the child was conceived?

A Stirling's Approximation

Recall the definition of the **gamma function**

$$\Gamma(x) = \int_0^{\infty} t^{x-1} e^{-t} dt \quad (\text{A.1})$$

The fundamental recurrence property of the gamma function, which is proved by integration by parts, is

$$\Gamma(x+1) = x\Gamma(x) \quad (\text{A.2})$$

When $x = n$ is an integer, this gives

$$n! = \Gamma(n+1) = \int_0^{\infty} t^n e^{-t} dt = \int_0^{\infty} e^{n \ln t - t} dt \quad (\text{A.3})$$

If we set $t = n + s$ then we can write

$$\ln t = \ln n + \ln(1 + s/n) = \ln n + \frac{s}{n} - \frac{1}{2} \left(\frac{s}{n}\right)^2 + \frac{1}{3} \left(\frac{s}{n}\right)^3 - \dots \quad (\text{A.4})$$

Hence

$$n! = \int_{-n}^{\infty} \exp \left[n \left(\ln n + \frac{s}{n} - \frac{s^2}{2n^2} + \dots \right) - n - s \right] ds \quad (\text{A.5})$$

For large n we may approximate this by

$$n! \approx e^{n \ln n - n} \int_{-\infty}^{\infty} e^{-s^2/2n} ds = \sqrt{2\pi n} n^n e^{-n} \quad (\text{A.6})$$

as claimed.

B The Gaussian Distribution as a Limit of the Binomial Distribution

In this appendix we prove that the equal probability binomial distribution Equation (3.17) approaches a symmetric Gaussian distribution in the limit that the number of trials approaches infinity.

Start with the binomial coefficient

$$\binom{n}{k} = \frac{n!}{k!(n-k)!} \tag{B.1}$$

and define

$$s := k - \frac{n}{2}, \tag{B.2}$$

so that the binomial coefficient becomes a function of s :

$$g(n, s) := \frac{n!}{\left(\frac{n}{2} + s\right)! \left(\frac{n}{2} - s\right)!}. \tag{B.3}$$

Now take the logarithm of both sides:

$$\log g(n, s) = \log n! - \log \left(\frac{n}{2} + s\right)! - \log \left(\frac{n}{2} - s\right)!. \tag{B.4}$$

Next use Stirling's approximation (A.6)

$$\log n! \approx \frac{1}{2} \log 2\pi + \left(n + \frac{1}{2}\right) \log n - n. \tag{B.5}$$

Plugging this into Equation (B.4) and canceling some terms gives

$$\begin{aligned} \log g(n, s) = & -\frac{1}{2} \log 2\pi + \left(n + \frac{1}{2}\right) \log n - \left(\frac{n}{2} + s + \frac{1}{2}\right) \log \left(\frac{n}{2} + s\right) \\ & - \left(\frac{n}{2} - s + \frac{1}{2}\right) \log \left(\frac{n}{2} - s\right). \end{aligned} \tag{B.6}$$

Next we add zero to the right side of Equation (B.6) in the form

$$\left(\frac{n}{2} + s + \frac{1}{2}\right) \log n + \left(\frac{n}{2} - s + \frac{1}{2}\right) \log n - n \log n - \log n, \quad (\text{B.7})$$

and use the properties of logarithms to write

$$\begin{aligned} \log g(n, s) = & \frac{1}{2} \log \left(\frac{1}{2\pi n} \right) - \left(\frac{n}{2} + s + \frac{1}{2} \right) \log \left(\frac{1}{2} \left(1 + \frac{2s}{n} \right) \right) \\ & - \left(\frac{n}{2} - s + \frac{1}{2} \right) \log \left(\frac{1}{2} \left(1 - \frac{2s}{n} \right) \right). \end{aligned} \quad (\text{B.8})$$

From the properties of binomial coefficients we know that $g(n, s)$ is always peaked around $s = 0$. This peak only becomes more pronounced as n gets larger. Expanding about $s = 0$ using Taylor's theorem gives

$$\log \left(1 + \frac{2s}{n} \right) = \frac{2s}{n} - \frac{1}{2} \left(\frac{2s}{n} \right)^2 + \frac{1}{3} \left(\frac{2s}{n} \right)^3 + \dots, \quad (\text{B.9})$$

so

$$\log \left(\frac{1}{2} \left(1 + \frac{2s}{n} \right) \right) = -\log 2 + \frac{2s}{n} - \frac{1}{2} \left(\frac{2s}{n} \right)^2 + \dots. \quad (\text{B.10})$$

Plugging this equation (and an identical one with s replaced by $-s$) into Equation (B.8) we get

$$\begin{aligned} \log g(n, s) & \approx \frac{1}{2} \log \left(\frac{1}{2\pi n} \right) - \left(\frac{n}{2} + s + \frac{1}{2} \right) \left[-\log 2 + \frac{2s}{n} - \frac{1}{2} \left(\frac{2s}{n} \right)^2 \right] \\ & \quad - \left(\frac{n}{2} - s + \frac{1}{2} \right) \left[-\log 2 - \frac{2s}{n} - \frac{1}{2} \left(\frac{2s}{n} \right)^2 \right]. \\ & = \frac{1}{2} \log \left(\frac{2}{\pi n} \right) + n \log 2 - \frac{2s^2}{n}. \end{aligned} \quad (\text{B.11})$$

Thus

$$g(n, s) \approx g(n, 0) e^{-2s^2/n}, \quad (\text{B.12})$$

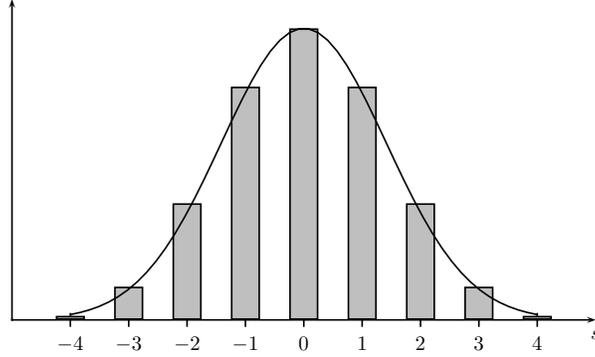


Figure 6: A Gaussian approximation to a binomial distribution $2^{-8} \binom{8}{4+s}$

where

$$g(n, 0) = \left(\frac{2}{\pi n} \right)^{1/2} 2^n. \quad (\text{B.13})$$

Using Equation (B.2) we can write the equal probability binomial distribution as

$$f(n, s) = \frac{n!}{\left(\frac{n}{2} + s\right)! \left(\frac{n}{2} - s\right)!} 2^{-n}. \quad (\text{B.14})$$

Hence we conclude that, in the limit $n \rightarrow \infty$,

$$f(n, s) \rightarrow \sqrt{\frac{2}{\pi n}} e^{-2s^2/n}. \quad (\text{B.15})$$

which is a Gaussian distribution centered on 0 with variance $\sigma^2 = n/4$. The limit of the general binomial distribution (3.18) is more complicated to prove, and we shall simply state the result, known as the **de Moivre-Laplace theorem**:

$$f(n, s) \rightarrow \frac{1}{\sqrt{2\pi n}} \frac{1}{\sqrt{pq}} \exp \left\{ -\frac{1}{2} \frac{(s + n((1/2) - p))^2}{npq} \right\}. \quad (\text{B.16})$$

The approximation is illustrated graphically in Figure 6.