Probability with Engineering Applications

ECE 313 Course Notes

Bruce Hajek

Department of Electrical and Computer Engineering
University of Illinois at Urbana-Champaign

August 2014

© 2014 by Bruce Hajek
All rights reserved. Permission is hereby given to freely print and circulate copies of these notes so long as the notes are left intact and not reproduced for commercial purposes. Email to b-hajek@illinois.edu, pointing out errors or hard to understand passages or providing comments, is welcome.

Contents

1 Foundations **3**

1.1 Embracing uncertainty . . . 3
1.2 Axioms of probability . . . 6
1.3 Calculating the size of various sets . . . 10
1.4 Probability experiments with equally likely outcomes . . . 13
1.5 Sample spaces with infinite cardinality . . . 15
1.6 Short Answer Questions . . . 20
1.7 Problems . . . 21

2 Discrete-type random variables **25**

2.1 Random variables and probability mass functions . . . 25
2.2 The mean and variance of a random variable . . . 27
2.3 Conditional probabilities . . . 32
2.4 Independence and the binomial distribution . . . 34
2.4.1 Mutually independent events . . . 34
2.4.2 Independent random variables (of discrete-type) . . . 37
2.4.3 Bernoulli distribution . . . 37
2.4.4 Binomial distribution . . . 37
2.5 Geometric distribution . . . 41
2.6 Bernoulli process and the negative binomial distribution . . . 43
2.7 The Poisson distribution–a limit of binomial distributions . . . 45
2.8 Maximum likelihood parameter estimation . . . 47
2.9 Markov and Chebychev inequalities and confidence intervals . . . 49
2.10 The law of total probability, and Bayes formula . . . 52
2.11 Binary hypothesis testing with discrete-type observations . . . 59
2.11.1 Maximum likelihood (ML) decision rule . . . 60
2.11.2 Maximum a posteriori probability (MAP) decision rule . . . 61
2.12 Reliability . . . 66
2.12.1 Union bound . . . 66
2.12.2 Network outage probability . . . 66
2.12.3 Distribution of the capacity of a flow network . . . 69
2.12.4 Analysis of an array code . . . 71

2.12.5 Reliability of a single backup . . . 73
2.13 Short Answer Questions . . . 73
2.14 Problems . . . 75

3 Continuous-type random variables 93
3.1 Cumulative distribution functions . . . 93
3.2 Continuous-type random variables . . . 98
3.3 Uniform distribution . . . 101
3.4 Exponential distribution . . . 102
3.5 Poisson processes . . . 104
3.5.1 Time-scaled Bernoulli processes . . . 105
3.5.2 Definition and properties of Poisson processes . . . 106
3.5.3 The Erlang distribution . . . 110
3.6 Linear scaling of pdfs and the Gaussian distribution . . . 111
3.6.1 Scaling rule for pdfs . . . 111
3.6.2 The Gaussian (normal) distribution . . . 113
3.6.3 The central limit theorem and the Gaussian approximation . . . 117
3.7 ML parameter estimation for continuous-type variables . . . 121
3.8 Functions of a random variable . . . 122
3.8.1 The distribution of a function of a random variable . . . 122
3.8.2 Generating a random variable with a specified distribution . . . 133
3.8.3 The area rule for expectation based on the CDF . . . 135
3.9 Failure rate functions . . . 136
3.10 Binary hypothesis testing with continuous-type observations . . . 137
3.11 Short Answer Questions . . . 143
3.12 Problems . . . 145

4 Jointly Distributed Random Variables 157
4.1 Joint cumulative distribution functions . . . 157
4.2 Joint probability mass functions . . . 159
4.3 Joint probability density functions . . . 161
4.4 Independence of random variables . . . 170
4.4.1 Definition of independence for two random variables . . . 171
4.4.2 Determining from a pdf whether independence holds . . . 171
4.5 Distribution of sums of random variables . . . 174
4.5.1 Sums of integer-valued random variables . . . 174
4.5.2 Sums of jointly continuous-type random variables . . . 176
4.6 Additional examples using joint distributions . . . 180
4.7 Joint pdfs of functions of random variables . . . 184
4.7.1 Transformation of pdfs under a linear mapping . . . 184
4.7.2 Transformation of pdfs under a one-to-one mapping . . . 186
4.7.3 Transformation of pdfs under a many-to-one mapping . . . 190
4.8 Correlation and covariance . . . 191

4.9 Minimum mean square error estimation . . . 199
4.9.1 Constant estimators . . . 199
4.9.2 Unconstrained estimators . . . 200
4.9.3 Linear estimators . . . 201
4.10 Law of large numbers and central limit theorem . . . 206
4.10.1 Law of large numbers . . . 206
4.10.2 Central limit theorem . . . 208
4.11 Joint Gaussian distribution . . . 211
4.11.1 From the standard 2-d normal to the general . . . 212
4.11.2 Key properties of the bivariate normal distribution . . . 213
4.12 Short Answer Questions . . . 215
4.13 Problems . . . 218

5 Wrap-up **231**

6 Appendix **233**
6.1 Some notation . . . 233
6.2 Some sums . . . 234
6.3 Frequently used distributions . . . 234
6.3.1 Key discrete-type distributions . . . 234
6.3.2 Key continuous-type distributions . . . 235
6.4 Normal tables . . . 237
6.5 Answers to short answer questions . . . 239
6.6 Solutions to even numbered problems . . . 240

Preface

A key objective of these notes is to convey how to deal with uncertainty in both qualitative and quantitative ways. Uncertainty is typically modeled as randomness. We must make decisions with partial information all the time in our daily lives, for instance when we decide what activities to pursue. Engineers deal with uncertainty in their work as well, often with precision and analysis.

A challenge in applying reasoning to real world situations is to capture the main issues in a mathematical model. The notation that we use to frame a problem can be critical to understanding or solving the problem. There are often events, or variables, that need to be given names.

Probability theory is widely used to model systems in engineering and scientific applications. These notes adopt the most widely used framework of probability, namely the one based on Kolmogorov's axioms of probability. The idea is to assume a mathematically solid definition of the model. This structure encourages a modeler to have a consistent, if not completely accurate, model. It also offers a commonly used mathematical language for sharing models and calculations.

Part of the process of learning to use the language of probability theory is learning classifications of problems into broad areas. For example, some problems involve finite numbers of possible alternatives, while others concern real-valued measurements. Many problems involve interaction of physically independent processes. Certain laws of nature or mathematics cause some probability distributions, such as the normal bell-shaped distribution often mentioned in popular literature, to frequently appear. Thus, there is an emphasis in these notes on well-known probability distributions and why each of them arises frequently in applications.

These notes were written for the undergraduate course, *ECE 313: Probability with Engineering Applications,* offered by the Department of Electrical and Computer Engineering at the University of Illinois at Urbana-Champaign. The official prerequisites of the course insure that students have had calculus, including Taylor series expansions, integration over regions in the plane, and the use of polar coordinates.

The author gratefully acknowledges the students and faculty who have participated in this course through the years. He is particularly grateful to Professor D. V. Sarwate, who first introduced the course, and built up much material for it on the website.

B. Hajek
August 2014

Organization

Chapter 1 presents an overview of the many applications of probability theory, and then explains the basic concepts of a probability model and the axioms commonly assumed of probability models. Often probabilities are assigned to possible outcomes based on symmetry. For example, when a six sided die is rolled, it is usually assumed that the probability a particular number i shows is $1/6$, for $1 \leq i \leq 6$. For this reason, we also discuss in Chapter 1 how to determine the sizes of various finite sets of possible outcomes.

Random variables are introduced in Chapter 2 and examined in the context of a finite, or countably infinite, set of possible outcomes. Notions of expectation (also known as mean), variance, hypothesis testing, parameter estimation, multiple random variables, and well known probability distributions–Poisson, geometric, and binomial, are covered. The Bernoulli process is considered–it provides a simple setting to discuss a long, even infinite, sequence of event times, and provides a tie between the binomial and geometric probability distributions.

The focus shifts in Chapter 3 from discrete-type random variables to continuous-type random variables. The chapter takes advantage of many parallels and connections between discrete-type and continuous-type random variables. The most important well known continuous-type distributions are covered: uniform, exponential, and normal (also known as Gaussian). Poisson processes are introduced–they are continuous-time limits of the Bernoulli processes described in Chapter 2. Parameter estimation and binary hypothesis testing are covered for continuous-type random variables in this chapter in close analogy to how they are covered for discrete-type random variables in Chapter 2.

Chapter 4 considers groups of random variables, with an emphasis on two random variables. Topics include describing the joint distribution of two random variables, covariance and correlation coefficient, and prediction or estimation of one random variable given observation of another. Somewhat more advanced notions from calculus come in here, in order to deal with joint probability densities, entailing, for example, integration over regions in two dimensions.

Short answer questions and problems can be found at the end of each chapter with answers to the questions and even numbered problems provided in the appendix. Most of the short answer questions also have video links to solutions which should open if you click on "[video]" in the pdf version of these notes, if you have a browser and internet connection. A brief wrap up is given in Chapter 5. A small number of other videos are provided for examples. These videos are not meant to substitute for reading the notes; it is recommended students attempt solving problems and watch the videos afterwards if needed.

Chapter 1

Foundations

1.1 Embracing uncertainty

We survive and thrive in an uncertain world. What are some uses of probability in everyday life? In engineering? Below is an incomplete list:

- **Call centers and other staffing problems:** Experts with different backgrounds are needed to staff telephone call centers for major financial investment companies, travel reservation services, and consumer product support. Management must decide the number of staff and the mix of expertise so as to meet availability and waiting time targets. A similar problem is faced by large consulting service companies, hiring consultants that can be grouped into multiple overlapping teams for different projects. The basic problem is to weigh future uncertain demand against staffing costs.

- **Electronic circuits:** Scaling down the power and energy of electronic circuits reduces the reliability and predictability of many individual elements, but the circuits must nevertheless be engineered so the overall circuit is reliable.

- **Wireless communication:** Wireless links are subject to fading, interference from other transmitters, doppler spread due to mobility and multipath propagation. The demand, such as the number of simultaneous users of a particular access point or base station, is also time varying and not fully known in advance. These and other effects can vary greatly with time and location, but yet the wireless system must be engineered to meet acceptable call quality and access probabilities.

- **Medical diagnosis and treatment:** Physicians and pharmacologists must estimate the most suitable treatments for patients in the face of uncertainty about the exact condition of the patient or the effectiveness of the various treatment options.

- **Spread of infectious diseases:** Centers for disease control need to decide whether to institute massive vaccination or other preventative measures in the face of globally threatening, possibly mutating diseases in humans and animals.

- **Information system reliability and security:** System designers must weigh the costs and benefits of measures for reliability and security, such as levels of backups and firewalls, in the face of uncertainty about threats from equipment failures or malicious attackers.
- **Evaluation of financial instruments, portfolio management:** Investors and portfolio managers form portfolios and devise and evaluate financial instruments, such as mortgage backed securities and derivatives, to assess risk and payoff in an uncertain financial environment.
- **Financial investment strategies, venture capital:** Individuals raising money for, or investing in, startup activities must assess potential payoffs against costs, in the face of uncertainties about a particular new technology, competitors, and prospective customers.
- **Modeling complex systems:** Models incorporating probability theory have been developed and are continuously being improved for understanding the brain, gene pools within populations, weather and climate forecasts, microelectronic devices, and imaging systems such as computer aided tomography (CAT) scan and radar. In such applications, there are far too many interacting variables to model in detail, so probabilistic models of aggregate behavior are useful.
- **Modeling social science:** Various groups, from politicians to marketing folks, are interested in modeling how information spreads through social networks. Much of the modeling in this area of social science involves models of how people make decisions in the face of uncertainty.
- **Insurance industry:** Actuaries price policies for natural disasters, life insurance, medical insurance, disability insurance, liability insurance, and other policies, pertaining to persons, houses, automobiles, oil tankers, aircraft, major concerts, sports stadiums and so on, in the face of much uncertainty about the future.
- **Reservation systems:** Electronic reservation systems dynamically set prices for hotel rooms, airline travel, and increasingly for shared resources such as smart cars and electrical power generation, in the face of uncertainty about future supply and demand.
- **Reliability of major infrastructures:** The electric power grid, including power generating stations, transmission lines, and consumers is a complex system with many redundancies. Still, breakdowns occur, and guidance for investment comes from modeling the most likely sequences of events that could cause outage. Similar planning and analysis is done for communication networks, transportation networks, water, and other infrastructure.
- **Games, such as baseball, gambling, and lotteries:** Many games involve complex calculations with probabilities. For example, a professional baseball pitcher's choice of pitch has a complex interplay with the anticipation of the batter. For another example, computer rankings of sports teams based on win-loss records is a subject of interesting modeling.
- **Commerce, such as online auctions:** Sellers post items online auction sites, setting initial prices and possibly hidden reserve prices, without complete knowledge of the total demand for the objects sold.

- **Online search and advertising:** Search engines decide which webpages and which advertisements to display in response to queries, without knowing precisely what the viewer is seeking.
- **Personal financial decisions:** Individuals make decisions about major purchases, investments, and insurance, in the presence of uncertainty.
- **Personal lifestyle decisions:** Individuals make decisions about diet, exercise, studying for exams, investing in personal relationships, all in the face of uncertainty about such things as health, finances, and job opportunities.

Hopefully you are convinced that uncertainty is all around us, in our daily lives and in many professions. How can probability theory help us survive, and even thrive, in the face of such uncertainty? Probability theory:

- **provides a language for people to discuss/communicate/aggregate knowledge about uncertainty.** Use of standard deviation is widely used when results of opinion polls are described. The language of probability theory lets people break down complex problems, and to argue about pieces of them with each other, and then aggregate information about subsystems to analyze a whole system.
- **provides guidance for statistical decision making and estimation or inference.** The theory provides concrete recommendations about what rules to use in making decisions or inferences, when uncertainty is involved.
- **provides modeling tools and ways to deal with complexity.** For complex situations, the theory provides approximations and bounds useful for reducing or dealing with complexity when applying the theory.

What does probability mean? If I roll a fair six-sided die, what is the probability a six shows? How do I know? What happens if I roll the same die a million times?

What does it mean for a weather forecaster to say the probability of rain tomorrow is 30%? Here is one system we could use to better understand a forecaster, based on incentives. Suppose the weather forecaster is paid p (in some monetary units, such as hundreds of dollars) if she declares that the probability of rain tomorrow is p. If it does rain, no more payment is made, either to or from the forecaster. If it does not rain, the forecaster keeps the initial payment of p, but she has to pay $-\ln(1-p)$. In view of these payment rules, if the weather forecaster believes, based on all the information she has examined, that the probability of rain tomorrow is q, and if she reports p, her expected total payoff is $p + (1-q)\ln(1-p)$. For q fixed this payoff is maximized at $p = q$. That is, the forecaster would maximize the total payoff she expects to receive by reporting her best estimate. Someone receiving her estimate would then have some understanding about what it meant. (See A.H. Murphy and R.L. Winkler (1984). "Probability Forecasting in Meteorology," *Journal of the American Statistical Association,* 79 (387), 489500.)

1.2 Axioms of probability

There are many philosophies about probability. In these notes we do not present or advocate any one comprehensive philosophy for the meaning and application of probability theory, but we present the widely used axiomatic framework. The framework is based on certain reasonable mathematical axioms. When faced with a real life example, we use a mathematical model satisfying the axioms. Then we can use properties implied by the axioms to do calculations or perform reasoning about the model, and therefore about the original real life example. A similar approach is often taken in the study of geometry. Once a set of axioms is accepted, additional properties can be derived from them.

Before we state the axioms, we discuss a very simple example to introduce some terminology. Suppose we roll a fair die, with each of the numbers one through six represented on a face of the die, and observe which of the numbers shows (i.e. comes up on top when the die comes to rest). There are six possible outcomes to this experiment, and because of the symmetry we declare that each should have equal probability, namely, 1/6. Following tradition, we let Ω (pronounced "omega") denote the *sample space*, which is the set of possible outcomes. For this example, we could take $\Omega = \{1, 2, 3, 4, 5, 6\}$. Performing the experiment of rolling a fair die corresponds to selecting an outcome from Ω. An *event* is a subset of Ω. An event is said to occur or to be true when the experiment is performed if the outcome is in the event. Each event A has an associated probability, $P(A)$. For this experiment, $\{1\}$ is the event that one shows, and we let $P(\{1\}) = 1/6$. For brevity, we write this as $P\{1\} = 1/6$. Similarly, we let $P\{2\} = P\{3\} = P\{4\} = P\{5\} = P\{6\} = 1/6$. But there are other events. For example, we might define B to be the event that the number that shows is two or smaller. Equivalently, $B = \{1, 2\}$. Since B has two outcomes, it's reasonable that $P(B) = 2/6 = 1/3$. And E could be the event that the number that shows is even, so $E = \{2, 4, 6\}$, and $P(E) = 3/6 = 1/2$.

Starting with two events, such as B and E just considered, we can describe more events using "and" and "or," where "and" corresponds to the intersection of events, and "or" corresponds to the union of events. This gives rise to the following two events[1]:

$$\begin{aligned} \text{"the number that shows is two or smaller and even "} &= BE = \{1,2\} \cap \{2,4,6\} = \{2\} \\ \text{"the number that shows is two or smaller or even"} &= B \cup E = \{1,2\} \cup \{2,4,6\} = \{1,2,4,6\}. \end{aligned}$$

The probabilities of these events are $P(BE) = 1/6$ and $P(B \cup E) = 4/6 = 2/3$. Let O be the event that the number that shows is odd, or $O = \{1, 3, 5\}$. Then:

$$\begin{aligned} \text{"the number that shows is even and odd"} &= EO = \{2,4,6\} \cap \{1,3,5\} = \emptyset \\ \text{"the number that shows is even or odd"} &= E \cup O = \{2,4,6\} \cup \{1,3,5\} = \Omega. \end{aligned}$$

Thus, $P(EO) = 0$ and $P(E \cup O) = 1$. This makes sense, because the number that shows can't be both even and odd, but it is always either even or odd.

It is also sometimes useful to refer to an event not being true, which is equivalent to the complement of the event being true, where the complement of an event is the set of outcomes not

[1] Here BE is the intersection of sets B and E. It is the same as $B \cap E$. See Appendix 6.1 for set notation.

in the event. For example, the complement of B, written B^c, is the event $\{3, 4, 5, 6\}$. Then, when the die is rolled, either B is true or B^c is true, but not both. That is, $B \cup B^c = \Omega$ and $BB^c = \emptyset$.

Thus, whatever events we might be interested in initially, we might also want to discuss events that are intersections, unions, or complements of the events given initially. The empty set, $\emptyset$, or the whole space of outcomes, Ω, should be events, because they can naturally arise through taking complements and intersections or unions. For example, if A is any event, then $A \cup A^c = \Omega$ and $AA^c = \emptyset$. The complement of Ω is $\emptyset$, and vice versa.

A bit more terminology is introduced before we describe the axioms precisely. One event is said to *exclude* another event if an outcome being in the first event implies the outcome is not in the second event. For example, the event O excludes the event E. Of course, E excludes O as well. Two or more events $E_1, E_2, \ldots, E_n$, are said to be *mutually exclusive* if at most one of the events can be true. Equivalently, the events $E_1, E_2, \ldots, E_n$ are mutually exclusive if $E_i E_j = \emptyset$ whenever $i \neq j$. That is, the events are disjoint sets. If events $E_1, E_2, \ldots, E_n$ are mutually exclusive, and if $E_1 \cup \cdots \cup E_n = \Omega$, then the events are said to form a *partition* of Ω. For example, if A is an event, then A and A^c form a partition of Ω.

De Morgan's law in the theory of sets is that the complement of the union of two sets is the intersection of the complements. Or vice versa: the complement of the intersection is the union of the complements:

$$(A \cup B)^c = A^c B^c \qquad (AB)^c = A^c \cup B^c. \tag{1.1}$$

De Morgan's law is easy to verify using the *Karnaugh map* for two events, shown in Figure 1.1. The idea of the map is that the events AB, AB^c, A^cB, and A^cB^c form a partition of Ω. The lower half of the figure indicates the events A, B, and $A \cup B$, respectively. The only part of Ω that $A \cup B$

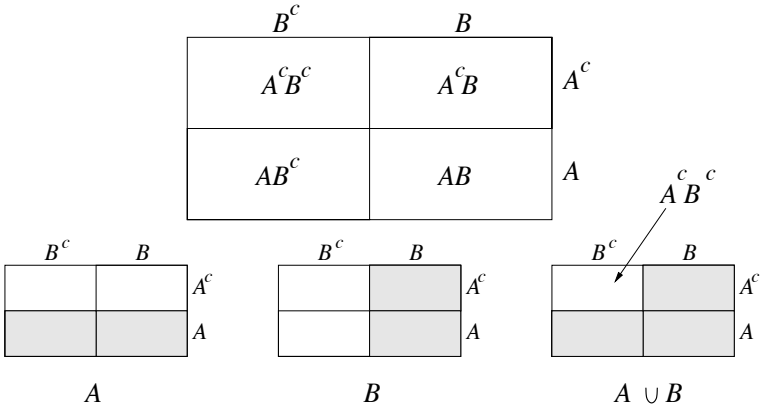


Figure 1.1: Karnaugh map showing how two sets, A and B, partition Ω.

does not cover is A^cB^c, proving the first version of De Morgan's law.

The Axioms of Probability The proof of De Morgan's law outlined above shows that it is a universal truth about sets–it does not need to be assumed. In contrast, the axioms described next

are intuitively reasonable properties that we require to be true of a probability model. The set of axioms together define what we mean by a valid probability model.

An experiment is modeled by a *probability space*, which is a triplet $(\Omega, \mathcal{F}, P)$. We will read this triplet as "Omega, Script F, P." The first component, Ω, is a nonempty set. Each element ω of Ω is called an *outcome* and Ω is called the *sample space*. The second component, $\mathcal{F}$, is a set of subsets of Ω called *events*. The final component, P, of the triplet $(\Omega, \mathcal{F}, P)$, is a *probability measure* on $\mathcal{F}$, which assigns a probability, $P(A)$, to each event A. The axioms of probability are of two types: event axioms, which are about the set of events $\mathcal{F}$, and probability axioms, which are about the probability measure P.

Event axioms The set of events, $\mathcal{F}$, is required to satisfy the following axioms:

Axiom E.1 Ω is an event (i.e. $\Omega \in \mathcal{F}$).

Axiom E.2 If A is an event then A^c is an event (i.e. if $A \in \mathcal{F}$ then $A^c \in \mathcal{F}$).

Axiom E.3 If A and B are events then $A \cup B$ is an event (i.e. if $A, B \in \mathcal{F}$ then $A \cup B \in \mathcal{F}$). More generally, if $A_1, A_2, \ldots$ is a list of events then the union of all of these events (the set of outcomes in at least one of them), $A_1 \cup A_2 \cup \cdots$, is also an event.

One choice of $\mathcal{F}$ that satisfies the above axioms is the set of all subsets of Ω. In fact, in these notes, whenever the sample space Ω is finite or countably infinite (which means the elements of Ω can be arranged in an infinite list, indexed by the positive integers), we let $\mathcal{F}$ be the set of all subsets of Ω. When Ω is uncountably infinite, it is sometimes mathematically impossible to define a suitable probability measure on the set of all subsets of Ω in a way consistent with the probability axioms below. To avoid such problems, we simply don't allow all subsets of such an Ω to be events, but the set of events $\mathcal{F}$ can be taken to be a rich collection of subsets of Ω that includes any subset of Ω we are likely to encounter in applications.

If the Axioms E.1-E.3 are satisfied, the set of events has other intuitively reasonable properties, and we list some of them below. We number these properties starting at 4, because the list is a continuation of the three axioms, but we use the lower case letter "e" to label them, reserving the upper case letter "E" for the axioms.[2]

Property e.4 The empty set, $\emptyset$, is an event (i.e. $\emptyset \in \Omega$). That is because Ω is an event by Axiom E.1, so Ω^c is an event by Axiom E.2. But $\Omega^c = \emptyset$, so $\emptyset$ is an event.

Property e.5 If A and B are events, then AB is an event. To see this, start with De Morgan's law: $AB = (A^c \cup B^c)^c$. By Axiom E.2, A^c and B^c are events. So by Axiom E.3, $A^c \cup B^c$ is an event. So by Axiom E.2 a second time, $(A^c \cup B^c)^c$ is an event, which is just AB.

Property e.6 More generally, if $B_1, B_2, \ldots$ is a list of events then the intersection of all of these events (the set of outcomes in all of them) $B_1 B_2 \cdots$ is also an event. This is true by the same reason given for Property e.5, starting with the fact $B_1 B_2 \cdots = (B_1^c \cup B_2^c \cup \cdots)^c$.

[2] In fact, the choice of which properties to make axioms is not unique. For example, we could have made Property e.4 an axiom instead of Axiom E.1.

Probability axioms The probability measure P is required to satisfy the following axioms:

Axiom P.1 For any event A, $P(A) \geq 0$.

Axiom P.2 If $A, B \in \mathcal{F}$ and if A and B are mutually exclusive, then $P(A \cup B) = P(A) + P(B)$. More generally, if $E_1, E_2, \ldots$ is an infinite list (i.e. countably infinite collection) of mutually exclusive events, $P(E_1 \cup E_2 \cup \cdots) = P(E_1) + P(E_2) + \cdots$.

Axiom P.3 $P(\Omega) = 1$.

If Axioms P.1-P.3 are satisfied (and Axioms E.1-E.3 are also satisfied) then the probability measure P has other intuitively reasonable properties. We list them here:

Property p.4 For any event A, $P(A^c) = 1 - P(A)$. That is because A and A^c are mutually exclusive events and $\Omega = A \cup A^c$. So Axioms P.2 and P.3 yield $P(A) + P(A^c) = P(A \cup A^c) = P(\Omega) = 1$.

Property p.5 For any event A, $P(A) \leq 1$. That is because if A is an event, then $P(A) = 1 - P(A^c) \leq 1$ by Property p.4 and by the fact, from Axiom P.1, that $P(A^c) \geq 0$.

Property p.6 $P(\emptyset) = 0$. That is because $\emptyset$ and Ω are complements of each other, so by Property p.4 and Axiom P.3, $P(\emptyset) = 1 - P(\Omega) = 0$.

Property p.7 If $A \subset B$ then $P(A) \leq P(B)$. That is because $B = A \cup (A^c B)$ and A and $A^c B$ are mutually exclusive, and $P(A^c B) \geq 0$, so $P(A) \leq P(A) + P(A^c B) = P(A \cup (A^c B)) = P(B)$.

Property p.8 $P(A \cup B) = P(A) + P(B) - P(AB)$. That is because, as illustrated in Figure 1.1, $A \cup B$ can be written as the union of three mutually exclusive sets: $A \cup B = (AB^c) \cup (A^c B) \cup (AB)$. So

$$\begin{aligned} P(A \cup B) &= P(AB^c) + P(A^c B) + P(AB) \\ &= (P(AB^c) + P(AB)) + (P(A^c B) + P(AB)) - P(AB) \\ &= P(A) + P(B) - P(AB). \end{aligned}$$

Property p.9 $P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(AB) - P(AC) - P(BC) + P(ABC)$. This is a generalization of Property p.8 and can be proved in a similar way.

Example 1.2.1 (Toss of a fair coin) Suppose the experiment is to flip a coin to see if it shows heads or tails. Using H for heads and T for tails, the experiment is modeled by the following choice of Ω and P :

$$\Omega = \{H, T\} \qquad \mathcal{F} = \{\{H\}, \{T\}, \{H, T\}, \emptyset\}$$
$$P\{H\} = P\{T\} = \frac{1}{2}, \qquad P(\Omega) = P\{H, T\} = 1, \qquad P(\emptyset) = 0.$$

Example 1.2.2 A particular experiment is to observe the color of a traffic signal at the time it is approached by a vehicle. The sample space is $\Omega = \{green, yellow, red\}$ and we let any subset of Ω be an event. What probability measure, P, should we choose? Since there are three colors, we could declare them to be equally likely, and thus have probability 1/3 each. But here is an intuitively more reasonable choice. Suppose when we examine the signal closely, we notice that the color of the signal goes through cycles of duration 75 seconds. In each cycle the signal dwells on green for 30 seconds, then dwells on yellow for 5 seconds, then dwells on red for 40 seconds. Assuming that the arrival time of the vehicle is random, and not at all connected to the signal (in particular, the traffic signal is isolated and not synchronized with other signals that the vehicle passes) then it seems intuitively reasonable to assign probabilities to the colors that are proportional to their dwell times. Hence, we declare that $P\{green\} = \frac{30}{75} = \frac{2}{5}, P\{yellow\} = \frac{5}{75} = \frac{1}{15}$, and $P\{red\} = \frac{40}{75} = \frac{8}{15}$. Note that the three outcomes are not equally likely.

1.3 Calculating the size of various sets

An important class of probability spaces are those such that the set of outcomes, Ω, is finite, and all outcomes have equal probability. Therefore, the probability for any event A is $P(A) = \frac{|A|}{|\Omega|}$, where $|A|$ is the number of elements in A and $|\Omega|$ is the number of elements in Ω. This notation "$|A|$" is the same as what we use for absolute value, but the argument is a set, not a value. The number of elements in a set is called the *cardinality* of the set. Thus, it is important to be able to count the number of elements in various sets. Such counting problems are the topic of this section.

Principle of counting Often we are faced with finding the number of vectors or sets satisfying certain conditions. For example, we might want to find the number of pairs of the form (X,N) such that $X \in \{H, T\}$ and $N \in \{1, 2, 3, 4, 5, 6\}$. These pairs correspond to outcomes of an experiment, in which a coin is flipped and a die is rolled, with X representing the side showing on the coin (H for heads or T for tails) and N being the number showing on the die. For example, $\{H, 3\}$, or $H3$ for short, corresponds to the coin showing heads and the die showing three. The set of all possible outcomes can be listed as:

$$\begin{array}{cccccc} H1 & H2 & H3 & H4 & H5 & H6 \\ T1 & T2 & T3 & T4 & T5 & T6. \end{array}$$

Obviously there are twelve possible outcomes. There are two ways to choose X, and for every choice of X, there are six choices of N. So the number of possible outcomes is $2 \times 6 = 12$.

This example is a special case of the *principle of counting*: If there are m ways to select one variable and n ways to select another variable, and if these two selections can be made independently, then there is a total of mn ways to make the pair of selections. The principle extends to more than one variable as illustrated by the next example.

Example 1.3.1 Find the number of possible 8-bit bytes. An example of such a byte is 00010100.

Solution: There are two ways to select the first bit, and for each of those, there are two ways to select the second bit, and so on. So the number of possible bytes is 2^8. (Each byte represents a number in the range 0 to 255 in base two representation.)

The idea behind the principle of counting is more general than the principle of counting itself. Counting the number of ways to assign values to variables can often be done by the same approach, even if the choices are not independent, as long as the choice of one variable does not affect the number of choices possible for the second variable. This is illustrated in the following example.

Example 1.3.2 Find the number of four letter sequences that can be obtained by ordering the letters A, B, C, D, without repetition. For example, $ADCB$ is one possibility.

Solution: There are four ways to select the first letter of the sequence, and for each of those, there are three ways to select the second letter, and for each of those, two ways to select the third letter, and for each of those, one way to select the final letter. So the number of possibilities is $4 \cdot 3 \cdot 2 \cdot 1 = 4!$ (read "four factorial'). Here the possibilities for the second letter are slightly limited by the first letter, because the second letter must be different from the first. Thus, the choices of letter for each position are not independent. If the letter A is used in the first position, for example, it can't be used in the second position. The problem is still quite simple, however, because the choice of letter for the first position does not affect the number of choices for the second position. And the first two choices don't affect the number of choices for the third position.

In general, the number of ways to order n distinct objects is $n! = n \cdot (n-1) \cdots 2 \cdot 1$. An ordering of n distinct objects is called a *permutation,* so the number of permutations of n distinct objects is $n!$. The next example indicates how to deal with cases in which the objects are not distinct.

Example 1.3.3 How many orderings of the letters AAB are there, if we don't distinguish between the two A's?

Solution: There are three orderings: AAB, ABA, BAA. But if we put labels on the two A's, writing them as A_1 and A_2, then the three letters become distinct, and so there are 3!, equal to six, possible orderings:

$$\begin{array}{cc} A_1A_2B & A_2A_1B \\ A_1BA_2 & A_2BA_1 \\ BA_1A_2 & BA_2A_1. \end{array}$$

These six orderings are written in pairs that would be identical to each other if the labels on the A's were erased. For example, A_1A_2B and A_2A_1B would both become AAB if the labels were erased. For each ordering without labels, such as AAB, there are two orderings with labels, because there are two ways to order the labels on the A's.

Example 1.3.4 [video] How many orderings of the letters ILLINI are there, if we don't distinguish the I's from each other and we don't distinguish the L's from each other?[3]

Solution: Since there are six letters, if the I's were labeled as I_1, I_2, I_3 and the L's were labeled as L_1 and L_2, then the six letters would be distinct and there would be 6!=720 orderings, including $I_1I_3L_2NI_2L_1$. Each ordering without labels, such as $IILNIL$ corresponds to $3! \times 2 = 12$ orderings with labels, because there are 3! ways to label the three I's, and for each of those, there are two ways to label the L's. Hence, there are $720/12 = 60$ ways to form a sequence of six letters, using three identical I's, two identical L's, and one N.

Principle of Over Counting The above two examples illustrate the *principle of over counting,* which can be stated as follows: For an integer $K \geq 1$, if each element of a set is counted K times, then the number of elements in the set is the total count divided by K.

Example 1.3.5 Suppose nine basketball players are labeled by the letters A, B, C, D, E, F, G, H, and I. A lineup is a subset consisting of five players. For example, $\{A, B, D, E, H\}$ is a possible lineup. The order of the letters in the lineup is not relevant. That is, $\{A, B, D, E, H\}$ and $\{A, B, E, D, H\}$ are considered to be the same. So how many distinct lineups are there?

Solution: If the order did matter, the problem would be easier. There would be 9 ways to select the first player, and for each of those, 8 ways to select the second player, 7 ways to select the third, 6 ways to select the fourth, and 5 ways to select the fifth. Thus, there are $9 \cdot 8 \cdot 7 \cdot 6 \cdot 5$ ways to select lineups in a given order. Since there are 5! ways to order a given lineup, each lineup would appear 5! times in a list of all possible lineups with all possible orders. So, by the principle of over counting, the number of distinct lineups, with the order of a lineup not mattering, is $9 \cdot 8 \cdot 7 \cdot 6 \cdot 5/5! = 126$.

Example 1.3.6 How many binary sequences of length 9 have (exactly) 5 ones? One such sequence is 110110010.

Solution: If the one's were distinct and the zero's were distinct, there would be 9! choices. But there are 5! ways to order the one's, and for each of those, 4! ways to order the zero's, so there are $5! \cdot 4!$ orderings with labels for each ordering without labels. Thus, the number of binary sequences of length 9 having five ones is $\frac{9!}{5!4!}$. This is the same as the solution to Example 1.3.5. In fact, there is a one-to-one correspondence between lineups and binary sequences of length 9 with 5 ones. The positions of the 1's indicate which players are in the lineup. The sequence 110110010 corresponds to the lineup $\{A, B, D, E, H\}$.

[3] "ILLINI," pronounced "ill LIE nigh," is the nickname for the students and others at the University of Illinois.

In general, the number of subsets of size k of a set of n distinct objects can be determined as follows. There are n ways to select the first object, $n-1$ ways to select the second object, and so on, until there are $n-k+1$ ways to select the kth object. By the principle of counting, that gives a total count of $n(n-1)\cdots(n-k+1)$, but this chooses k distinct objects in a particular order. By definition of the word "set," the order of the elements within a set does not matter. Each set of k objects is counted $k!$ ways by this method, so by the principle of over counting, the number of subsets of size k of a set of n distinct objects (with the order not mattering) is given by $\frac{n(n-1)\cdots(n-k+1)}{k!}$. This is equal to $\frac{n!}{(n-k)!k!}$, and is called "n choose k," and is written as the *binomial coefficient*, $\binom{n}{k} = \frac{n!}{(n-k)!k!}$.

It is useful to keep in mind that $\binom{n}{k} = \binom{n}{n-k}$ and that

$$\binom{n}{k} = \frac{\overbrace{n(n-1)\cdots(n-k+1)}^{k \text{ terms}}}{k!}.$$

For example, $\binom{8}{3}$ and $\binom{8}{5}$ are both equal to $\frac{8\cdot7\cdot6}{3\cdot2\cdot1} = 56$, which is also equal to $\frac{8\cdot7\cdot6\cdot5\cdot4}{5\cdot4\cdot3\cdot2\cdot1}$.

Note that $(a+b)^2 = a^2+2ab+b^2$ and $(a+b)^3 = a^3+3a^2b+3ab^2+b^3$. In general,

$$(a+b)^n = \sum_{k=0}^{n} \binom{n}{k} a^k b^{n-k}. \tag{1.2}$$

Equation (1.2) follows by writing $(a+b)^n = \underbrace{(a+b)(a+b)\cdots(a+b)}_{n \text{ factors}}$, and then noticing that the coefficient of $a^k b^{n-k}$ in $(a+b)^n$ is the number of ways to select k out of the n factors from which to select a.

1.4 Probability experiments with equally likely outcomes

This section presents more examples of counting, and related probability experiments such that all outcomes are equally likely.

Example 1.4.1 Suppose there are nine socks loose in a drawer in a dark room which are identical except six are orange and three are blue. Someone selects two at random, all possibilities being equally likely. What is the probability the two socks are the same color? Also, if instead, three socks were selected, what is the probability that at least two of them are the same color?

Solution For the first question, we could imagine the socks are numbered one through nine, with socks numbered one through six being orange and and socks numbered seven through nine being blue. Let Ω be the set of all subsets of $\{1,2,3,4,5,6,7,8,9\}$ of size two. The number of elements of Ω is given by $|\Omega| = \binom{9}{2} = \frac{9\cdot8}{2} = 36$. The number of ways two orange socks could be chosen is

$\binom{6}{2} = 15$, and the number of ways two blue socks could be chosen is $\binom{3}{2} = 3$. Thus, the probability two socks chosen at random have the same color is $\frac{15+3}{36} = 1/2$.

The second question is trivial. Whatever set of three socks is selected, at least two of them are the same color. So the answer to the second question is one.

Example 1.4.2 Let an experiment consist of rolling two fair dice, and define the following three events about the numbers showing: A =“sum is even,” B =“sum is a multiple of three,” and C =“the numbers are the same.” Display the outcomes in a three-event Karnaugh map, and find $P(ABC)$.

Solution. We write ij for the outcome such that i appears on the first die and j appears on the second die. The outcomes are displayed in a Karnaugh map in Figure 1.2. It is a little tedious to fill in the map, but it is pretty simple to check correctness of the map. For correctness it suffices that

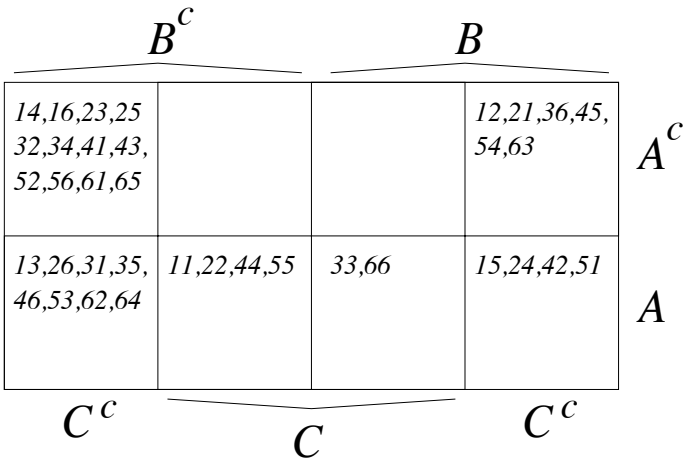


Figure 1.2: Karnaugh map for the roll of two fair dice and events A, B, and C.

for each of the events A, A^c, B, B^c, C, C^c, the correct outcomes appear in the corresponding lines (rows or columns). For example, all 18 outcomes in A are in the bottom row of the map. All 18 outcomes of A^c are in the upper row of the map. Similarly we can check for B, B^c, C, and C^c. There are 36 elements in Ω, and $ABC = \{33, 66\}$, which has two outcomes. So $P(ABC) = 2/36 = 1/18$.

Example 1.4.3 [video] [video] Suppose a deck of playing cards has 52 cards, represented by the set $\mathcal{C}$:

$$\mathcal{C} = \{1C, 2C, \ldots, 13C, 1D, 2D, \ldots, 13D, 1H, 2H, \ldots 13H, 1S, 2S, \ldots, 13S\}.$$

Here C, D, H, or S stands for the *suit* of a card: “clubs,” “diamonds,” “hearts,” or “spades.” Suppose five cards are drawn at random from the deck, with all possibilities being equally likely.

In the terminology of the game of *poker,* a *FULL HOUSE* is the event that three of the cards all have the same number, and the other two cards both have some other number. For example, the outcome $\{1D, 1H, 1S, 2H, 2S\}$ is an element of *FULL HOUSE,* because three of the cards have the number 1 and the other two cards both have the number 2. *STRAIGHT* is the event that the numbers on the five cards can be arranged to form five consecutive integers, or that the numbers can be arranged to get the sequence 10,11,12,13,1 (because, by tradition, the 1 card, or "ace," can act as either the highest or lowest number). For example, the outcome $\{3D, 4H, 5H, 6S, 7D\}$ is an element of *STRAIGHT.* Find $P(FULL\ HOUSE)$ and $P(STRAIGHT)$.

Solution: The sample space is $\Omega = \{A : A \subset \mathcal{C} \text{ and } |A| = 5\}$, and the number of possible outcomes is $|\Omega| = \binom{52}{5}$. To select an outcome in *FULL HOUSE*, there are 13 ways to select a number for the three cards with the same number, and then 12 ways to select a number for the other two cards. Once the two numbers are selected, there are $\binom{4}{3}$ ways to select 3 of the 4 suits for the three cards with the same number, and $\binom{4}{2}$ ways to select 2 of the 4 suits for the two cards with the other number. Thus,

$$\begin{aligned} P(FULL\ HOUSE) &= \frac{13 \cdot 12 \cdot \binom{4}{3}\binom{4}{2}}{\binom{52}{5}} \\ &= \frac{13 \cdot 12 \cdot 4 \cdot 4 \cdot 3 \ \cdot \ 5 \cdot 4 \cdot 3 \cdot 2}{2 \ \cdot \ 52 \cdot 51 \cdot 50 \cdot 49 \cdot 48} \\ &= \frac{6}{17 \cdot 5 \cdot 49} = \frac{6}{4165} \approx 0.0014. \end{aligned}$$

To select an outcome in *STRAIGHT*, there are ten choices for the set of five integers on the five cards that can correspond to *STRAIGHT*, and for each of those, there are 4^5 choices of what suit is assigned to the cards with each of the five consecutive integers. Thus,

$$\begin{aligned} P(STRAIGHT) &= \frac{10 \cdot 4^5}{\binom{52}{5}} \\ &= \frac{10 \cdot 4^5}{\frac{52 \cdot 51 \cdot 50 \cdot 49 \cdot 48}{5 \cdot 4 \cdot 3 \cdot 2}} \approx 0.0039. \end{aligned}$$

1.5 Sample spaces with infinite cardinality

It is appropriate for many experiments to use a sample space Ω with infinite cardinality. For example, an experiment could be to randomly select a number from the interval $[0, 1]$, and there are infinitely many numbers in the interval $[0, 1]$. This section discusses infinite sets in general, and includes examples of sample spaces with infinitely many elements. The section highlights some implications of Axiom P.2.

The previous two sections discuss how to find the cardinality of some finite sets. What about the cardinality of infinite sets? Do all infinite sets have the same number of elements? In a

strong sense, no. The smallest sort of infinite set is called a countably infinite set, or a set with countably infinite cardinality, which means that all the elements of the set can be placed in a list. Some examples of countably infinite sets are the set of nonnegative integers $\mathbb{Z}+ = \{0, 1, 2, \ldots\}$, the set of all integers $\mathbb{Z} = \{0, 1, -1, 2, -2, 3, -3, \ldots\}$, and the set of nonnegative rational numbers: $Q_+ = \{\frac{i}{j} : i \geq 1, j \geq 1, \text{integers}\}$. Figure 1.3 shows a two dimensional array that contains every positive rational number at least once. The zig-zag path in Figure 1.3 shows how all the elements can be placed on an infinite list.

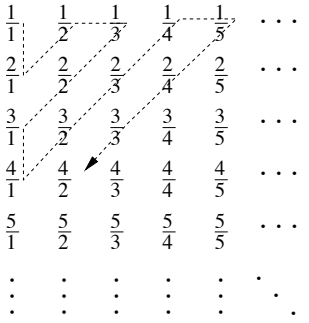


Figure 1.3: An infinite list containing all positive rational numbers

So there are many countably infinite sets, but the set of real numbers is definitely larger, as shown by the following proposition.

Proposition 1.5.1 *The set of real numbers is not countable.*

Proof. We will prove that not even the set of numbers in the interval $[0, 1]$ is countable. It is enough to show that for any list of numbers from the interval $[0, 1]$, there is at least one other number in $[0, 1]$ that is not on the list. So, let $a_1, a_2, \ldots$ be a list of numbers from the interval $[0, 1]$. Consider the decimal expansions of these numbers. For example, it might appear as follows:

$$\begin{aligned} a_1 &= 0.\underline{3}439098\cdots \\ a_2 &= 0.2\underline{4}39465\cdots \\ a_3 &= 0.94\underline{9}3554\cdots \\ a_4 &= 0.334\underline{3}876\cdots \\ a_5 &= 0.9495\underline{2}49\cdots . \end{aligned}$$

For each k, the kth digit (after the decimal point) of a_k is underlined in this list. Let a^* be the number in $[0, 1]$ such that its kth digit is two larger (modulo 10) than the kth digit of a_k. For the example shown, $a^* = 0.56154\cdots$. For any $k \geq 1$, $a^* \neq a_k$ because the kth digit of a^* differs by at least two from the kth digit of a_k. (In particular, a^* and a_k can't be different representations of the same number, like $0.1999\cdots$ and $0.200\cdots$.) So a^* is not in the list. So no list of numbers from the interval $[0, 1]$ can contain all of the numbers from $[0, 1]$. ■

From an engineering perspective, if a set is countably infinite, it means that any element in the set can be referenced by a finite string of bits. The strings for different elements of the set can be

different lengths, but the string for each element is finite. This could be called a variable length encoding of the set. It is enough to represent the index of the element in the list. For example, if π is the twenty sixth element of the list, then the representation of π would be the binary expansion of 26, or 11010. Proposition 1.5.1 means that it is impossible to index the set of real numbers by variable length binary strings of finite length.

Example 1.5.2 Consider an experiment in which a player is asked to choose a positive integer. To model this, let the space of outcomes Ω be the set of positive integers. Thus, Ω is countably infinite. Mathematically, the case that Ω is countably infinite is similar to the case Ω is finite. In particular, we can continue to let $\mathcal{F}$, the set of events, be the set of all subsets of Ω. To name a specific choice of a probability measure P, we use the following rather arbitrary choice. Let

$$(p_1, p_2, p_3, \ldots) = \left(1 - \frac{1}{2}, \frac{1}{2} - \frac{1}{3}, \frac{1}{3} - \frac{1}{4}, \ldots\right) = \left(\frac{1}{2}, \frac{1}{6}, \frac{1}{12}, \frac{1}{20}, \ldots\right),$$

or, equivalently, $p_i = \frac{1}{i(i+1)}$ for $i \geq 1$. Note that $p_1 + \cdots + p_i = 1 - \frac{1}{i+1}$ for $i \geq 1$ and therefore $p_1 + p_2 + \cdots = 1$. Assume that the player chooses integer i with probability p_i. Since the outcomes in any event A can be listed in a sequence, Axiom P.2 requires that for any event A, $P(A) = \sum_{i \in A} p_i$. For example, $P\{1, 4, 5\} = \frac{1}{2} + \frac{1}{20} + \frac{1}{30} \approx 0.8533$. If A is the event that the number chosen is a multiple of five, then $P(A)$ can be calculated numerically:

$$P(A) = \frac{1}{(5)(6)} + \frac{1}{(10)(11)} + \frac{1}{(15)(16)} + \cdots \approx 0.05763$$

This choice of $(\Omega, \mathcal{F}, P)$ satisfies the axioms of probability.

It is often natural and convenient to consider probability experiments involving real-valued outcomes. Even though there are uncountably many possible outcomes, the axioms of probability can still model such situations, as illustrated in the following three examples of probability spaces.

Example 1.5.3 (Standard unit-interval probability space) Take $\Omega = \{\omega : 0 \leq \omega \leq 1\}$. Imagine an experiment in which the outcome ω is drawn from Ω, with no preference for drawing from one interval or another, for two intervals of the same length. This requires the set of events $\mathcal{F}$ to include intervals, and the probability of an interval $[a, b]$ with $0 \leq a \leq b \leq 1$ to be given by

$$P(\ [a, b]\ ) \quad = \quad b - a. \tag{1.3}$$

Taking $a = b$, we see that singleton sets $\{a\}$ are events, and these sets have probability zero. In order for the event axioms to be true, open intervals (a, b) must also be events and $P(\ (a, b)\ ) = b - a$. Any open subset of Ω can be expressed as the union of a finite or countably infinite set of open intervals, and any closed set is the the complement of an open set, so $\mathcal{F}$ should contain all open and all closed subsets of Ω. Thus, $\mathcal{F}$ must contain any set that is the intersection of countably many open sets, and so on.

Example 1.5.3 describes the probability space $(\Omega, \mathcal{F}, P)$ for the experiment of selecting a point from the unit interval, $[0,1]$, such that the probability the outcome is in an interval $[a,b]$ is $b-a$. This is true even if $a=b$, which means that the probability of any particular singleton subset $\{x\}$ of $[0,1]$ is equal to zero. But the entire interval $[0,1]$ is the union of all such sets $\{x\}$, and those sets are mutually exclusive. Why then, doesn't Axiom P.2 imply that $P\{\Omega\} = \sum_{x\in\Omega} P\{x\} = 0$, in contradiction to Axiom P.3? The answer is that Axiom P.2 does not apply to this situation because it only holds for a finite or countably infinite collection of events, whereas the set $[0,1]$ is uncountably infinite.

The next two examples make use of the formula for the sum of a geometric series, so we derive the formula here. A *geometric series* with first term one has the form $1, x, x^2, x^3, \cdots$. Equivalently, the kth term is x^k for $k \geq 0$. Observe that for any value of x :

$$(1-x)(1+x+x^2+\cdots+x^n) = 1-x^{n+1},$$

because when the product on the left hand side is expanded, the terms of the form x^k for $1 \leq k \leq n$ cancel out. Therefore, for $x \neq 1$, the following equivalent expressions hold for the partial sums of a geometric series:

$$1+x+x^2+\cdots+x^n = \frac{1-x^{n+1}}{1-x} \qquad \text{or} \qquad \sum_{k=0}^{n} x^k = \frac{1-x^{n+1}}{1-x}. \tag{1.4}$$

The sum of an infinite series is equal to the limit of the nth partial sum as $n \to \infty$. If $|x| < 1$ then $\lim_{n\to\infty} x^{n+1} = 0$. So letting $n \to \infty$ in (1.4) yields that for $|x| < 1$:

$$1+x+x^2+\cdots = \frac{1}{1-x} \qquad \text{or} \qquad \sum_{k=0}^{\infty} x^k = \frac{1}{1-x}. \tag{1.5}$$

The formula (1.4) for partial sums and the formula (1.5) for infinite sums of a geometric series are used frequently in these notes.

Example 1.5.4 (Repeated binary trials) Suppose we would like to represent an infinite sequence of binary observations, where each observation is a zero or one with equal probability. For example, the experiment could consist of repeatedly flipping a fair coin, and recording a one each time it shows heads and a zero each time it shows tails. Then an outcome ω would be an infinite sequence, $\omega = (\omega_1, \omega_2, \cdots)$, such that for each $i \geq 1$, $\omega_i \in \{0,1\}$. Let Ω be the set of all such ω's. The set of events can be taken to be large enough so that any set that can be defined in terms of only finitely many of the observations is an event. In particular, for any binary sequence $(b_1, \cdots, b_k)$ of some finite length k, the set $\{\omega \in \Omega : \omega_i = b_i \text{ for } 1 \leq i \leq k\}$ should be in $\mathcal{F}$, and the probability of such a set is taken to be 2^{-k}.

There are also events that don't depend on a fixed, finite number of observations. For example, suppose there are two players who take turns performing the coin flips, with the first one to get heads wins. Let F be the event that the player going first wins. Show that F is an event and then find its probability.

Solution: For $k \geq 1$, let E_k be the event that the first one occurs on the kth observation. That is, E_k is the event that the first k observations are given by the binary sequence $(b_1, \ldots, b_k) = (0, 0, \ldots, 0, 1)$, and its probability is given by $P\{E_k\} = 2^{-k}$.

Observe that $F = E_1 \cup E_3 \cup E_5 \cup \ldots$, so F is an event by Axiom E.3. Also, the events $E_1, E_3, \ldots$ are mutually exclusive, so by the full version of Axiom P.3 and (1.5):

$$P(F) = P(E_1) + P(E_3) + \cdots = \frac{1}{2}\left(1 + \left(\frac{1}{4}\right) + \left(\frac{1}{4}\right)^2 + \cdots\right) = \frac{1/2}{1 - (1/4)} = \frac{2}{3}.$$

The player who goes first has twice the chance of winning as the other player.

Example 1.5.5 (Selection of a point in a square) Take Ω to be the square region in the plane,

$$\Omega \quad = \quad \{(x, y) : 0 \leq x < 1, 0 \leq y < 1\}.$$

It can be shown that there is a probability space $(\Omega, \mathcal{F}, P)$ such that any rectangular region that is a subset of Ω of the form $R = \{(u, v) : a \leq u < b, c \leq v < d\}$ is an event, and

$$P(R) = \text{area of } R = (b - a)(d - c).$$

Let T be the triangular region $T = \{(x, y) : x \geq 0, y \geq 0, x + y < 1\}$. Since T is not rectangular, it is not immediately clear whether T is an event. Show that T is an event, and find $P(T)$, using the axioms.

Solution Consider the infinite sequence of square regions shown in Figure 1.4. Square 1 has area

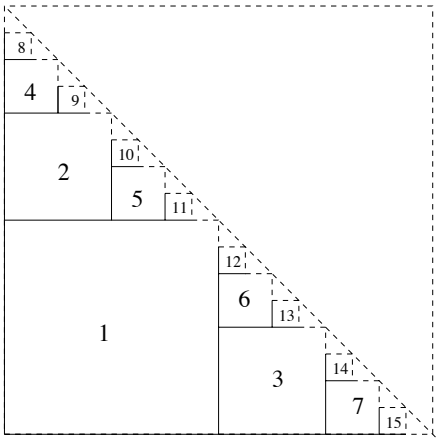


Figure 1.4: Approximation of a triangular region.

1/4, squares 2 and 3 have area 1/16 each, squares 4,5,6, and 7 have area $(1/4)^3$ each, and so on.

The set of squares is countably infinite, and their union is T, so T is an event by Axiom E.3. Since the square regions are mutually exclusive, Axiom P.2 implies that $P(T)$ is equal to the sum of the areas of the squares:

$$\begin{aligned} P(T) &= 1/4 + 2(1/4)^2 + 2^2(1/4)^3 + 2^3(1/4)^4 + \cdots \\ &= (1/4)(1 + 2^{-1} + 2^{-2} + 2^{-3} + \cdots) \\ &= (1/4) \cdot 2 = 1/2. \end{aligned}$$

Of course, it is reasonable that $P(T) = 1/2$, because T takes up half of the area of the square. In fact, any reasonable subset of the square is an event, and the probability of any event A is the area of A.

1.6 Short Answer Questions

Section 1.2[video]

1. What is $P(AB)$ if $P(A) = 0.5$, $P(B) = 0.46$, and $P(A \cup B) = 2P(AB)$.?
2. What is $P(ABC)$ if $P(A) = P(B) = P(C) = 0.5$, $P(A \cup B) = 0.55$, $P(A \cup C) = 0.7$, $P(BC) = 0.3$ and $P(ABC) = 2P(ABC^c)$?

Section 1.3[video]

1. Find the number of 8 letter sequences made by ordering the letters from BASEBALL
2. Find the number of 6 letter sequences made by deleting two letters from BASEBALL and ordering the remaining six.
3. How many distinct color combinations can be displayed by a set of three marbles drawn from a bag containing six marbles: three orange, two blue, and one white? Suppose the order of the three marbles doesn't matter.
4. How many essentially different ways are there to string eight distinct beads on a necklace? (If one way is the mirror image of another or the same as another up to rotation, they are considered to be the same.)

Section 1.4[video]

1. If the letters of ILLINI are randomly ordered, all orderings being equally likely, what is the probability the three I's are consecutive?
2. If the letters of ILLINI are randomly ordered, all orderings being equally likely, what is the probability no position has the same letter as in the original order?

Section 1.5[video]

1. Consider the ordering of rational numbers indicated in Figure 1.3. What is the 100th rational number on the list?

1.7 Problems

1.1. **[Defining a set of outcomes I]**
Ten balls, numbered one through ten, are initially in a bag. Three balls are drawn out, one at a time, without replacement.

(a) Define a sample space Ω describing the possible outcomes of this experiment. To be definite, suppose the order the three balls are drawn out is important. Explain how the elements of your set correspond to outcomes of the experiment.

(b) What is the cardinality of Ω?

1.2. **[Defining a set of outcomes II]**
Suppose four teams, numbered one through four, play a single-elimination tournament, consisting of three games. Two teams play each game and one of them wins; ties do not occur. The tournament bracket is fixed: teams one and two play each other in the first game and teams three and four play each other in the second game; the winner of the first game plays the winner of the second game in the third game.

(a) Define a set Ω so the elements of Ω correspond to the possible outcomes of the tournament. An element of Ω should specify the entire sequence of outcomes of the games. Explain how the elements of your set correspond to the possible outcomes. (This is an exercise in coming up with good notation.)

(b) How may possible outcomes are there?

(c) Suppose instead that the tournament bracket is not determined ahead of time. Thus, in the first round, team 1 does not need to be paired with team 2. Now how many outcomes are there for the combination of what bracket is used and the game outcomes? (Assume the order the two games are played in the first round does not matter. For example, they could be simultaneous.)

(d) For the setup described in part (c), for what fraction of the outcomes do teams 1 and 2 play each other?

1.3. **[Grouping students into teams]**
Suppose ten students in a class are to be grouped into teams.

(a) If each team has two students, how many ways are there to form teams? (The ordering of students within teams does not matter, and the ordering of the teams does not matter.)

(b) If each team has either two or three students, how many ways are there to form teams?

1.4. **[Possible probability assignments]**
Suppose A and B are events for some probability space such that $P(AB) = 0.3$ and $P(A \cup B) = 0.6$. Find the set of possible values of the pair $(P(A), P(B))$ and sketch this set, as a subset of the plane. Hint: It might help to try filling in the probabilities in a Karnaugh map.

1.5. **[A Karnaugh map for three events]**
Let an experiment consist of rolling two fair dice, and define the following three events about the numbers showing: $A =$"sum is even," $B =$"sum is a multiple of three," and $C =$"the number showing on the first die is (strictly) less than the number showing on the second die."

(a) Display the outcomes in a three-event Karnaugh map, as in Example 1.4.2.

(b) Find $P((A \cup B)C)$.

1.6. **[Displaying outcomes in a two event Karnaugh map]**
Two fair dice are rolled. Let A be the event the sum is even and B be the event at least one of the numbers rolled is three.

(a) Display the outcomes in a Karnaugh map.

(b) Determine $P(AB)$.

1.7. **[A three event Karnaugh puzzle]**
Suppose A, B, and C are events such that: $P(A) = P(B) = P(C) = 0.3$, $P(AB) = 3P(ABC)$, $P(A \cup C) = P(B \cup C) = 0.5$, and $P(A^cB^cC^c) = 0.48$. Sketch a Karnaugh map showing the probabilities of $ABC, ABC^c, \ldots, A^cB^cC^c$. Show your work.

1.8. **[A classification of students in a class]**
[video] Of 30 students in a class,

- 12 are not on Facebook
- 9 are on both Facebook and Twitter
- 2/3 of students not on Twitter don't have iPads
- At least one student is neither on Twitter nor on Facebook
- 3 students on Facebook and having iPads are not on Twitter
- 2 students are on both Facebook and Twitter and have iPads
- 2/3 of the students not on Facebook and without iPads are on Twitter

Find how many students are not on Twitter, not on Facebook and don't have iPads. Show your work. (Hint: Use a Karnaugh map. Fill in numbers or variables, trying to minimize the number of variables and equations needed.)

1.9. **[Selecting socks at random]**
Suppose there are eight socks in a bag, numbered one through eight, which can be grouped into four pairs: $\{1, 2\}$, $\{3, 4\}$, $\{5, 6\}$, or $\{7, 8\}$. The socks of each pair have the same color; different pairs have different colors. Suppose there are four (distinct!) people present, and one at a time, they each draw two socks out of the bag, without replacement. Suppose all socks feel the same, so when two socks are drawn from the bag, all possibilities have equal probability. Let M be the event that each person draws a matching pair of socks.

(a) Define a sample space Ω for this experiment. Suppose that the order that the people draw the socks doesn't matter–all that is recorded is which two socks each person selects.

(b) Determine $|\Omega|$, the cardinality of Ω.

(c) Determine the number of outcomes in M.

(d) Find $P(M)$.

(e) Find a short way to calculate $P(M)$ that does't require finding $|M|$ and $|\Omega|$. (Hint: Write $P(M)$ as one over an integer and factor the integer.)

1.10. **[Two more poker hands I]**
Suppose five cards are drawn from a standard 52 card deck of playing cards, as described in Example 1.4.3, with all possibilities being equally likely.

(a) *TWO PAIR* is the event that two cards both have one number, two other cards both have some other number, and the fifth card has a number different from the other two numbers. Find $P(TWO\ PAIR)$.

(b) *THREE OF A KIND* is the event that three of the cards all have the same number, and the other cards have numbers different from each other and different from the three with the same number. Find $P(THREE\ OF\ A\ KIND)$.

(c) *FOUR OF A KIND* is the event that four of the five cards have the same number. Find $P(FOUR\ OF\ A\ KIND)$.

1.11. **[Two more poker hands II]**
Suppose five cards are drawn from a standard 52 card deck of playing cards, as described in Example 1.4.3, with all possibilities being equally likely.

(a) *FLUSH* is the event that all five cards have the same suit. Find $P(FLUSH)$.

(b) *FOUR OF A KIND* is the event that four of the five cards have the same number. Find $P(FOUR\ OF\ A\ KIND)$.

1.12. **[Some identities satisfied by binomial coefficients]**
[video] Using only the fact that $\binom{n}{k}$ is the number of ways to select a set of k objects from a set of n objects, explain in words why each of the following identities is true. The idea is to identify how to count something two different ways. For example, to explain why $2^n = \sum_{k=0}^{n} \binom{n}{k}$, you could note that 2^n is the total number of subsets of a set of n objects, because there are two choices for each object (i.e. include or not include in the set) and the n choices are independent. The right hand side is also the number of subsets of a set of n objects, with the kth term being the number of such subsets of cardinality k.

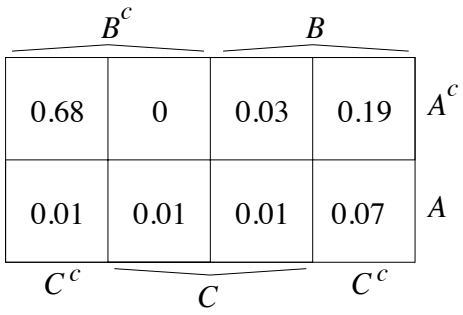
(a) $\binom{n}{k} = \binom{n}{n-k}$ for $0 \leq k \leq n$.

(b) $\binom{n}{k} = \binom{n-1}{k-1} + \binom{n-1}{k}$ for $1 \leq k \leq n-1$.

(c) $\binom{2n}{n} = \sum_{k=0}^{n} \binom{n}{k}^2$. (Hint: $\binom{n}{k}^2 = \binom{n}{k}\binom{n}{n-k}$. Consider a set of $2n$ distinct objects, half orange and half blue.)

(d) $\binom{n}{k} = \sum_{l=k}^{n} \binom{l-1}{k-1}$ for $1 \leq k \leq n$. (Hint: If a set of k objects is selected from among n objects numbered one through n, what are the possible values of the highest numbered object selected? For example, if $n = 5$ and $k = 3$, the equality becomes $\binom{5}{3} = \binom{2}{2} + \binom{3}{2} + \binom{4}{2}$, or $10 = 1 + 3 + 6$. The ten subsets can be divided into three groups: $\{1, 2, 3\}$; $\{1, 2, 4\}, \{1, 3, 4\}, \{2, 3, 4\}$; and $\{1, 2, 5\}, \{1, 3, 5\}, \{1, 4, 5\}, \{2, 3, 5\}, \{2, 4, 5\}, \{3, 4, 5\}$ such that the set in the first group has largest element 3, the sets in the second group have largest element 4, and the sets in the third group have largest element 5.)

1.13. **[Reading a Karnaugh map of probabilities]**
A certain town has three newspapers: A, B, and C. The probability distribution for which newspapers a person in the town reads is shown in the following Karnaugh map:



(a) Find the probability a person reads only newspaper A.

(b) **[5 points]** What is the probability a person reads at least two newspapers?

(c) What is the probability a person doesn't read any newspaper?

(d) If A and C are morning papers and B is an evening paper, what is the probability a person reads at least one morning paper plus an evening paper?

(e) What is the probability a person reads only one morning and one evening paper?

‘

Chapter 2

Discrete-type random variables

2.1 Random variables and probability mass functions

Chapter 1 focuses largely on events and their probabilities. An event is closely related to a binary variable; if a probability experiment is performed, then a particular event either occurs or does not occur. A natural and useful generalization allows for more than two values:

Definition 2.1.1 *A random variable is a real-valued function on* Ω.

Thus, if X is a random variable for a probability space $(\Omega, \mathcal{F}, P)$, if the probability experiment is performed, which means a value ω is selected from Ω, then the value of the random variable is $X(\omega)$. The value $X(\omega)$ is called the *realized value* of X for outcome ω. A random variable can have many possible values, and for each possible value there is some probability that the random variable is equal to that value. A random variable is said to be *discrete-type* if there is a finite set $u_1, \ldots, u_n$ or a countably infinite set $u_1, u_2, \ldots$ such that

$$P\{X \in \{u_1, u_2, \ldots\}\} = 1. \tag{2.1}$$

The probability mass function (pmf) for a discrete-type random variable X, p_X, is defined by $p_X(u) = P\{X = u\}$. Note that (2.1) can be written as:

$$\sum_i p_X(u_i) = 1.$$

The pmf is sufficient to determine the probability of any event determined by X, because for any set A, $P\{X \in A\} = \sum_{i:u_i \in A} p_X(u_i)$.

Example 2.1.2 Let X be the number showing for a single roll of a fair die. Then $p_X(i) = \frac{1}{6}$ for integers i with $1 \leq i \leq 6$.

Example 2.1.3 Let S be the sum of the numbers showing on a pair of fair dice when they are rolled. Find the pmf of S.

Solution: The underlying sample space is $\Omega = \{(i, j) : 1 \leq i \leq 6, 1 \leq j \leq 6\}$, and it has 36 possible outcomes, each having probability $\frac{1}{36}$. The smallest possible value of S is 2, and $\{S = 2\} = \{(1, 1)\}$. That is, there is only one outcome resulting in $S = 2$, so $p_S(2) = \frac{1}{36}$. Similarly, $\{S = 3\} = \{(1, 2), (2, 1)\}$, so $p_S(3) = \frac{2}{36}$. And $\{S = 4\} = \{(1, 3), (2, 2), (3, 1)\}$, so $p_S(4) = \frac{3}{36}$, and so forth. The pmf of S is shown in Fig. 2.1.

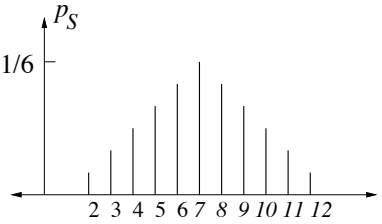


Figure 2.1: The pmf of the sum of numbers showing for rolls of two fair dice.

Example 2.1.4 Suppose two fair dice are rolled and that Y represents the maximum of the two numbers showing. The same set of outcomes, $\Omega = \{(i, j) : 1 \leq i \leq 6, 1 \leq j \leq 6\}$, can be used as in Example 2.1.3. For example, if a 3 shows on the first die and a 5 shows on the second, then $Y = 5$. That is, $Y((3, 5)) = 5$. In general, $Y((i, j)) = \max\{i, j\}$ for $(i, j) \in \Omega$. Determine the pmf of Y.

Solution: The possible values of Y are $1, 2, 3, 4, 5$, and 6. There is only one outcome in Ω such that $Y = 1$. Specifically, $\{Y = 1\} = \{(1, 1)\}$. Similarly, $\{Y = 2\} = \{(2, 2), (1, 2), (2, 1)\}$, $\{Y = 3\} = \{(3, 3), (1, 3), (2, 3), (3, 1), (3, 2)\}$, and so forth. The pmf of Y is shown in Fig. 2.2.

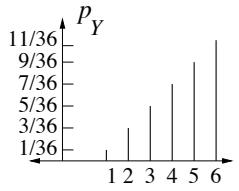


Figure 2.2: The pmf for the maximum of numbers showing for rolls of two fair dice.

2.2 The mean and variance of a random variable

The mean of a random variable is a weighted average of the possible values of the random variable, such that the weights are given by the pmf:

Definition 2.2.1 *The mean (also called expectation) of a random variable X with pmf p_X is denoted by $E[X]$ and is defined by $E[X] = \sum_i u_i p_X(u_i)$, where $u_1, u_2, \ldots$ is the list of possible values of X.*

Example 2.2.2 Let X be the number showing for a roll of a fair die. Find $E[X]$.

Solution Since $p_X(i) = 1/6$ for $1 \leq i \leq 6$, $E[X] = 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + 5 \cdot \frac{1}{6} + 6 \cdot \frac{1}{6} = \frac{1+2+3+4+5+6}{6} = \frac{7}{2}$.

Example 2.2.3 Let Y be the number of distinct numbers showing when three fair dice are rolled. Find the pmf and mean of Y.

Solution The underlying sample space is $\Omega = \{i_1 i_2 i_3 : 1 \leq i_1 \leq 6,\ 1 \leq i_2 \leq 6,\ 1 \leq i_3 \leq 6\}$. There are six ways to choose i_1, six ways to choose i_2, and six ways to choose i_3, so that $|\Omega| = 6 \cdot 6 \cdot 6 = 216$. What are the possible values of Y? Yes, Y takes values $1, 2, 3$. The outcomes in Ω that give rise to each possible value of Y are:

$$\begin{aligned} \{Y = 1\} &= \{111, 222, 333, 444, 555, 666\} \\ \{Y = 2\} &= \{112, 121, 211,\ 113, 131, 311,\ 114, 141, 411, \cdots, 665, 656, 566\} \\ \{Y = 3\} &= \{i_1 i_2 i_3 \in \Omega : i_1, i_2, i_3 \text{ are distinct}\}. \end{aligned}$$

Obviously, $|\{Y = 1\}| = 6$. The outcomes of $\{Y = 2\}$ are listed above in groups of three, such as 112,121,211. There are thirty such groups of three, because there are six ways to choose which number appears twice, and then five ways to choose which number appears once, in the outcomes in a group. Therefore, $|\{Y = 2\}| = 6 \cdot 5 \cdot 3 = 90$. There are six choices for the first number of an outcome in $\{Y = 3\}$, then five choices for the second, and then four choices for the third. So $|\{Y = 3\}| = 6 \cdot 5 \cdot 4 = 120$. To double check our work, we note that 6+90+120=216, as expected. So, $p_Y(1) = \frac{6}{216} = \frac{1}{36}$, $p_Y(2) = \frac{90}{216} = \frac{15}{36}$, and $p_Y(3) = \frac{120}{216} = \frac{20}{36}$. The mean of Y is NOT simply $\frac{1+2+3}{3}$ because the three possible values are not equally likely. The correct value is $E[Y] = 1 \cdot \frac{1}{36} + 2 \cdot \frac{15}{36} + 3 \cdot \frac{20}{36} = \frac{91}{36} \approx 2.527$.

The [video] gives a physical interpretation of the mean of a pmf, and describes how to build and calibrate a nine volt calculator for calculating means.

Example 2.2.4 Suppose X is a random variable taking values in $\{-2, -1, 0, 1, 2, 3, 4, 5\}$, each with probability $\frac{1}{8}$. Let $Y = X^2$. Find $E[Y]$.

Solution. The pmf of X is given by

$$p_X(u) = \begin{cases} \frac{1}{8} & -2 \le u \le 5 \\ 0 & \text{else.} \end{cases}$$

The definition of $E[Y]$ involves the pmf of Y, so let us find the pmf of Y. We first think about the possible values of Y (which are 0, 1, 4, 9, 16, and 25), and then for each possible value u, find $p_Y(u) = P\{Y = u\}$. For example, for $u = 1$, $p_Y(1) = P\{Y = 1\} = P\{X = -1 \text{ or } X = 1\} = 2/8$. The complete list of nonzero values of p_Y is as follows:

u	$p_Y(u)$
0	1/8
1	2/8
4	2/8
9	1/8
16	1/8
25	1/8.

Therefore, $E[Y] = 0 \cdot \frac{1}{8} + 1 \cdot \frac{2}{8} + 4 \cdot \frac{2}{8} + 9 \cdot \frac{1}{8} + 16 \cdot \frac{1}{8} + 25 \cdot \frac{1}{8} = \frac{60}{8} = 7.5$.

You may have noticed, without even thinking about it, that there is another way to compute $E[Y]$ in Example 2.2.4. We can find $E[Y]$ without finding the pmf of Y. Instead of first adding together some values of p_X to find p_Y and then summing over the possible values of Y, we can just directly sum over the possible values of X, yielding

$$E[Y] = (-2)^2 \cdot \frac{1}{8} + (-1)^2 \cdot \frac{1}{8} + 0^2 \cdot \frac{1}{8} + 1^2 \cdot \frac{1}{8} + 2^2 \cdot \frac{1}{8} + 3^2 \cdot \frac{1}{8} + 4^2 \cdot \frac{1}{8} + 5^2 \cdot \frac{1}{8}.$$

The general formula for the mean of a function, $g(X)$, of X, is

$$E[g(X)] = \sum_i g(u_i) p_X(u_i). \tag{2.2}$$

This formula is so natural it is called the *law of the unconscious statistician* (LOTUS).

Example 2.2.5 Suppose two fair dice are rolled. Find the pmf and the mean of the product of the two numbers showing.

Table 2.1: Table of Y vs. X_1 and X_2 for $Y = X_1X_2$.

$X_1 \backslash X_2$	1	2	3	4	5	6
1	1	2	3	4	5	6
2	2	4	6	8	10	12
3	3	6	9	12	15	18
4	4	8	12	16	20	24
5	5	10	15	20	25	30
6	6	12	18	24	30	36.

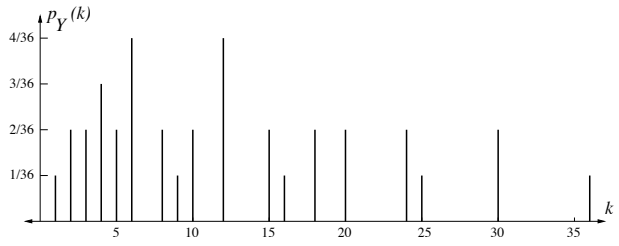


Figure 2.3: The pmf for the product of two fair dice

Solution Let X_1 denote the number showing on the first die and X_2 denote the number showing on the second die. Then $P\{(X_1, X_2) = (i, j)\} = \frac{1}{36}$ for $1 \le i \le 6$ and $1 \le j \le 6$. Let $Y = X_1X_2$. The easiest way to calculate the pmf of Y is to go through all 36 possible outcomes, and add the probabilities for outcomes giving the same value of Y. A table of Y as a function of X_1 and X_2 is shown in Table 2.1. There is only one outcome giving $Y = 1$, so $p_Y(1) = 1/36$. There are two outcomes giving $Y = 2$, so $p_Y(2) = 2/36$, and so on. The pmf of Y is shown in Figure 2.3. One way to compute $E[Y]$ would be to use the definition of expectation and the pmf of Y shown. An alternative that we take is to use LOTUS. We have that $Y = g(X_1, X_2)$, where $g(i, j) = ij$. Each of the 36 possible values of (X_1, X_2) has probability $1/36$, so we have

$$\begin{aligned} E[Y] &= \frac{1}{36}\sum_{i=1}^{6}\sum_{j=1}^{6} ij \\ &= \frac{1}{36}\left(\sum_{i=1}^{6} i\right)\left(\sum_{j=1}^{6} j\right) \\ &= \frac{(21)^2}{36} = \left(\frac{7}{2}\right)^2 = \frac{49}{4} = 12.25. \end{aligned}$$

Example 2.2.6 Let X be a random variable with a pmf p_X and consider the new random variable $Y = X^2 + 3X$. Use LOTUS to express $E[Y]$ in terms of $E[X]$ and $E[X^2]$.

Solution: So $Y = g(X)$ for the function $g(u) = u^2 + 3u$. Using LOTUS yields

$$\begin{aligned} E[X^2 + 3X] &= \sum_i (u_i^2 + 3u_i) p_X(u_i) \\ &= \left(\sum_i u_i^2 p_X(u_i) \right) + \left(\sum_i 3u_i p_X(u_i) \right) \\ &= E[X^2] + 3E[X]. \end{aligned}$$

The LOTUS equation (2.2) illustrates that expectations are weighted averages, with the weighting given by the pmf of the underlying random variable X. An important implication, illustrated in Example 2.2.6, is that expectation is a *linear* operation. A more general statement of this linearity is the following. If $g(X)$ and $h(X)$ are functions of X, and a, b, and c are constants, then $ag(X) + bh(X) + c$ is also a function of X, and the same method as in Example 2.2.6 shows that

$$E[ag(X) + bh(X) + c] = aE[g(X)] + bE[h(X)] + c. \tag{2.3}$$

Variance and standard deviation Suppose you are to be given a payment, with the size of the payment, in some unit of money, given by either X or by Y, described as follows. The random variable X is equal to 100 with probability one, whereas $p_Y(100000) = \frac{1}{1000}$ and $p_Y(0) = \frac{999}{1000}$. Would you be equally happy with either payment? Both X and Y have mean 100. This example illustrates that two random variables with quite different pmfs can have the same mean. The pmf for X is concentrated on the mean value, while the pmf for Y is considerably spread out.

The *variance* of a random variable X is a measure of how spread out the pmf of X is. Letting $\mu_X = E[X]$, the variance is defined by:

$$\mathrm{Var}(X) = E[(X - \mu_X)^2]. \tag{2.4}$$

The difference $X - \mu_X$ is called the deviation of X (from its mean). The deviation is the error if X is predicted by μ_X. By linearly of expectation, the mean of the deviation is zero: $E[X - \mu_X] = E[X] - \mu_X = \mu_X - \mu_X = 0$. Sometimes $\mathrm{Var}(X)$ is called the mean square deviation of X, because it is the mean of the square of the deviation. It might seem a little arbitrary that variance is defined using a power of two, rather than some other power, such as four. It does make sense to talk about the mean fourth power deviation, $E[(X - \mu_X)^4]$, or the mean absolute deviation, $E[|X - \mu_X|]$. However, the mean square deviation has several important mathematical properties which facilitate computation, so it is by far the most commonly used measure of how spread out a distribution is. The variance of X is often denoted by σ_X^2, where $\sigma_X = \sqrt{\mathrm{Var}(X)}$ is called the *standard deviation* of X. If X is in some units, then σ_X is in the same units. For example. if X is in feet, then $\mathrm{Var}(X)$ is in feet^2 and σ_X is again in feet.

As mentioned earlier, the variance of X is a measure of how spread out the pmf of X is. If a constant b is added to X, the pmf of the resulting random variable $X + b$ is obtained by shifting the

pmf of X by b. Adding the constant increases the mean by b, but it does not change the variance, because variance measures spread around the mean. Indeed, by the linearity of expectation, (2.3):

$$\begin{aligned} E[X+b] &= E[X]+b \\ \mathrm{Var}(X+b) &= E[(X+b-E[X+b])^2] = E[(X+b-E[X]-b)^2] = E[(X-E[X])^2] = \mathrm{Var}(X). \end{aligned}$$

If X is multiplied by a constant a, the pmf of the resulting random variable is spread out by a factor a. The mean is multiplied by a, and the variance is multiplied by a^2 :

$$\begin{aligned} E[aX] &= aE[X] \\ \mathrm{Var}(aX) &= E[(aX-E[aX])^2] = E[(aX-aE[X])^2] = a^2E[(X-E[X])^2] = a^2\mathrm{Var}(X). \end{aligned}$$

Combining the two observations just made yields that:

$$E[aX+b] = aE[X]+b \text{ and } \mathrm{Var}(aX+b) = \mathrm{Var}(aX) = a^2\mathrm{Var}(X).$$

The random variable $\frac{X-\mu_X}{\sigma_X}$ is called the *standardized version* of X. The standardized random variable has mean zero and variance one:

$$\begin{aligned} E\left[\frac{X-\mu_X}{\sigma_X}\right] &= \frac{1}{\sigma_X}(E[X]-\mu_X) = 0 \\ \mathrm{Var}\left(\frac{X-\mu_X}{\sigma_X}\right) &= \frac{1}{\sigma_X^2}E[(X-\mu_X)^2] = \frac{\sigma_X^2}{\sigma_X^2} = 1. \end{aligned}$$

Note that even if X is a measurement in some units such as meters, the standardized random variable $\frac{X-\mu_X}{\sigma_X}$ is dimensionless, because the standard deviation σ_X is in the same units as X.

Using the linearity of expectation, we derive another expression for $\mathrm{Var}(X)$:

$$\begin{aligned} \mathrm{Var}(X) &= E[X^2-2\mu_X X+\mu_X^2] \\ &= E[X^2]-2\mu_X E[X]+\mu_X^2 \\ &= E[X^2]-\mu_X^2. \end{aligned} \tag{2.5}$$

For an integer $i \geq 1$, the ith *moment* of X is defined to be $E[X^i]$. Therefore, the variance of a random variable is equal to its second moment minus the square of its first moment. The definition of variance, (2.4), is useful for keeping in mind the meaning and scaling properties of variance, while the equivalent expression (2.5) is often more useful for computing the variance of a random variable from its distribution.

Example 2.2.7 Let X denote the number showing for one roll of a fair die. Find $\mathrm{Var}(X)$ and the standard deviation, σ_X.

Solution: As noted in Example 2.2.2, $\mu_X = 3.5$. Also, $E[X^2] = \frac{1^2+2^2+3^2+4^2+5^2+6^2}{6} = \frac{91}{6}$. So $\mathrm{Var}(X) = \frac{91}{6} - (3.5)^2 \approx 2.9167$ and $\sigma_X = \sqrt{\mathrm{Var}(X)} = 1.7078$.

2.3 Conditional probabilities

Let A and B be two events for some probability experiment. The *conditional probability* of B given A is defined by

$$P(B|A) = \begin{cases} \frac{P(AB)}{P(A)} & \text{if } P(A) > 0 \\ \text{undefined} & \text{if } P(A) = 0. \end{cases}$$

It is not defined if $P(A) = 0$, which has the following meaning. If you were to write a computer routine to compute $P(B|A)$ and the inputs are $P(AB) = 0$ and $P(A) = 0$, your routine shouldn't simply return the value zero. Rather, your routine should generate an error message such as "input error–conditioning on event of probability zero." Such an error message would help you or others find errors in larger computer programs which use the routine.

Intuitively, if you know that event A is true for a probability experiment, then you know that the outcome ω of the probability experiment is in A. Conditioned on that, whether B is also true should only depend on the outcomes in B that are also in A, which is the set AB. If B is equal to A, then, given A is true, B should be true with conditional probability one. That is why the definition of $P(B|A)$ has $P(A)$ in the denominator.

One of the very nice things about elementary probability theory is the simplicity of this definition of conditional probability. Sometimes we might get conflicting answers when calculating the probability of some event, using two different intuitive methods. When that happens, inevitably, at least one of the methods has a flaw in it, and falling back on simple definitions such as the definition of conditional probability clears up the conflict, and sharpens our intuition.

The following examples show that the conditional probability of an event can be smaller than, larger than, or equal to, the unconditional probability of the event. (Here, the phrase "unconditional probability of the event" is the same as the probability of the event; the word "unconditional" is used just to increase the contrast with conditional probability.)

Example 2.3.1 Roll two dice and observe the numbers coming up. Define two events by: A="the sum is six," and B="the numbers are not equal." Find and compare $P(B)$ and $P(B|A)$.

Solution: The sample space is $\Omega = \{(i,j) : 1 \le i \le 6, 1 \le j \le 6\}$, which has 36 equally likely events. To find $P(B)$ we count the number of outcomes in B. There are six choices for the number coming up on the first die, and for each of those, five choices for the number coming up on the second die that is different from the first. So B has $6 \times 5 = 30$ outcomes, and $P(B) = 30/36 = 5/6$.

Another way to see that $P(B) = 5/6$ is to notice that whatever the number on the first die is, the probability that the number on the second die will be different from it is $5/6$.

Since $A = \{(1,5),(2,4),(3,3),(4,2),(5,1)\}$ we have $P(A) = 5/36$. Similarly, since $AB = \{(1,5),(2,4),(4,2),(5,1)\}$ we have $P(AB) = 4/36$.

Therefore, $P(B|A) = \frac{P(AB)}{P(A)} = \frac{4/36}{5/36} = \frac{4}{5}$. Note that, for this example, $P(B|A) < P(B)$. That is, if one learns that the sum is six, the chances that different numbers show on the dice decreases from the original probability that different numbers show on the dice.

Example 2.3.2 Continuing with the previous example, find and compare $P(B^c)$ and $P(B^c|A)$.

Solution: Here, B^c is the event that the numbers coming up are the same. Since six outcomes are in B^c, $P(B^c) = \frac{6}{36} = \frac{1}{6}$. Since $AB^c = \{(3,3)\}$, $P(AB^c) = \frac{1}{36}$. As noted above, $P(A) = 5/36$. Therefore, $P(B^c|A) = \frac{P(AB^c)}{P(A)} = \frac{1/36}{5/36} = 1/5$. Another way to compute $P(B^c|A)$ would have been to notice that $P(B|A) + P(B^c|A) = 1$, and use the fact, from Example 2.3.1, that $P(B|A) = 4/5$. Note that, for this example, $P(B^c|A) > P(B^c)$. That is, if one learns that the sum is six, the chances that the numbers coming up are the same number increases from the original probability that the numbers coming up are the same.

Example 2.3.3 Again, consider the rolls of two fair dice. Let E="the number showing on the first die is even," and F="the sum of the numbers showing is seven." Find and compare $P(F)$ and $P(F|E)$.

Solution: Since F has six outcomes, $P(F) = \frac{6}{36} = \frac{1}{6}$. Since $EF = \{(2,5),(4,3),(6,1)\}$, $P(EF) = \frac{3}{36} = \frac{1}{12}$. Since E has 18 outcomes, $P(E) = \frac{1}{2}$. Therefore, $P(F|E) = \frac{P(EF)}{P(E)} = \frac{1/12}{1/2} = \frac{1}{6}$. Note that, for this example, $P(F) = P(F|E)$. That is, if one learns that the number coming up on the first die is even, the conditional probability that the numbers coming up on the dice sum to seven is the same as the original probability that the numbers sum to seven.

We comment briefly on some properties of conditional probabilities. These properties follow easily from the definition of conditional probabilities, and the axioms of probability. Suppose A is an event with $P(A) > 0$, and B and C are also events. Then

1. $P(B|A) \geq 0$.
2. $P(B|A) + P(B^c|A) = 1$. More generally, if $E_1, E_2, \ldots$ are disjoint events, $P(E_1 \cup E_2 \cup \cdots | B) = P(E_1|B) + P(E_2|B) + \cdots$.
3. $P(\Omega|B) = 1$.
4. $P(AB) = P(A)P(B|A)$.

5. $P(ABC) = P(C)P(B|C)P(A|BC)$ (assuming $P(BC) > 0$.)

The first three properties above are equivalent to the statement that, as a function of the argument B for A fixed, the conditional probability $P(B|A)$ has all the properties of an unconditional probability measure P. Compare these properties to the probability axioms in Section 1.2. Intuitively, if one assumes a probability distribution P is given, and then later learns that an event A is true, the conditional probabilities $P(B|A)$ as B varies, is a new probability distribution, giving a new view of the experiment modeled by the probability space.

2.4 Independence and the binomial distribution

2.4.1 Mutually independent events

Let A and B be two events for some probability space. Consider first the case that $P(A) > 0$. As seen in the previous section, it can be that $P(B|A) = P(B)$, which intuitively means that knowledge that A is true does not affect the probability that B is true. It is then natural to consider the events to be independent. If $P(B|A) \neq P(B)$, then knowledge that A is true does affect the probability that B is true, and it is natural to consider the events to be dependent (i.e. not independent). Since, by definition, $P(B|A) = \frac{P(AB)}{P(A)}$, the condition $P(B|A) = P(B)$ is equivalent to $P(AB) = P(A)P(B)$.

Let's consider the other case: $P(A) = 0$. Should we consider A and B to be independent? It doesn't make sense to condition on A, but $P(A^c) = 1$, so we can consider $P(B|A^c)$ instead. It holds that $P(B|A^c) = \frac{P(A^cB)}{P(A^c)} = P(A^cB) = P(B) - P(AB) = P(B)$. Therefore, $P(B|A^c) = P(B)$. That is, if $P(A) = 0$, knowledge that A is *not* true does not affect the probability of B. So it is natural to consider A to be independent of B.

These observations motivate the following definition, which has the advantage of applying whether or not $P(A) = 0$:

Definition 2.4.1 *Event A is* independent *of event B if $P(AB) = P(A)P(B)$.*

Note that the condition in the definition of independence is symmetric in A and B. Therefore, A is independent of B if and only if B is independent of A. Another commonly used terminology for these two equivalent relations is to say that A and B are *mutually independent.* Here, "mutually" means that independence is a property of the two events. It does not make sense to say that a single event A is independent, without reference to some other event.

If the experiment underlying the probability space $(\Omega, \mathcal{F}, P)$ involves multiple physically separated parts, then it is intuitively reasonable that an event involving one part of the experiment should be independent of another event that involves some other part of the experiment that is physically separated from the first. For example, when an experiment involves the rolls of two fair dice, it is implicitly assumed that the rolls of the two dice are physically independent, and an event A concerning the number showing on the first die would be physically independent of any event concerning the number showing on the second die. So, often in formulating a model, it is assumed that if A and B are physically independent, then they should be independent under the probability model $(\Omega, \mathcal{F}, P)$.

The condition for A to be independent of B, namely $P(AB) = P(A)P(B)$ is just a single equation that can be true even if A and B are not physically independent.

Example 2.4.2 Consider a probability experiment related to the experiment discussed in Section 1.3, in which a fair coin is flipped and a die is rolled, with N denoting the side showing on the coin and X denoting the number showing on the die. We should expect the event $\{N = H\}$ to be independent of the event $\{X = 6\}$, because they are physically independent events. This independence holds, assuming all twelve outcomes in Ω are equally likely, because then $P\{N = H\} = 6/12$, $P\{X = 6\} = 2/12$, and $P\{N = H, X = 6\} = 1/12$, so $P\{N = H, X = 6\} = P\{N = H\}P\{X = 6\}$. More generally, any event involving only X is independent of any event involving only N.

Example 2.4.3 Suppose the probability experiment is to roll a single die. Let A be the event that the outcome is even, and let B be the event that the outcome is a multiple of three. Since these events both involve the outcome of a single role of a die, we would not consider them to be physically independent. However, $A = \{2, 4, 6\}$, $B = \{3, 6\}$, and $AB = \{6\}$. So $P(A) = \frac{1}{2}$, $P(B) = \frac{1}{3}$ and $P(AB) = \frac{1}{6}$. Therefore, $P(AB) = P(A)P(B)$, which means that A and B are mutually independent. This is a simple example showing that events can be mutually independent, even if they are not physically independent.

Here is a final note about independence of two events. Suppose A is independent of B. Then

$$P(A^cB) \quad = \quad P(B) - P(AB) \quad = \quad (1 - P(A))P(B) \quad = \quad P(A^c)P(B),$$

so A^c is independent of B. Similarly, A is independent of B^c, and therefore, by the same reasoning, A^c is independent of B^c. In summary, the following four conditions are equivalent: A is independent of B, A^c is independent of B, A is independent of B^c, A^c is independent of B^c.

Let us now consider independence conditions for three events. The following definition simply requires any one of the events to be independent of any one of the other events.

Definition 2.4.4 *Events A, B, and C are* pairwise independent *if $P(AB) = P(A)P(B)$, $P(AC) = P(A)P(C)$ and $P(BC) = P(B)P(C)$.*

Example 2.4.5 Suppose two fair coins are flipped, so $\Omega = \{HH, HT, TH, TT\}$, and the four outcomes in Ω are equally likely. Let

$A = \{HH, HT\}$ =“first coin shows heads,”

$B = \{HH, TH\}$ =“second coin shows heads,”

$C = \{HH, TT\}$ =“both coins show heads or both coins show tails.”

It is easy to check that A, B, and C are pairwise independent. Indeed, $P(A) = P(B) = P(C) = 0.5$ and $P(AB) = P(AC) = P(BC) = 0.25$. We would consider A to be physically independent of B as well, because they involve flips of different coins. Note that $P(A|BC) = 1 \neq P(A)$. That is, knowing that both B and C are true affects the probability that A is true. So A is not independent of BC.

Example 2.4.5 illustrates that pairwise independence of events does not imply that any one of the events is independent of the intersection of the other two events. In order to have such independence, a stronger condition is used to define independence of three events:

Definition 2.4.6 *Events $A, B,$ and C are* independent *if they are pairwise independent and if $P(ABC) = P(A)P(B)P(C)$.*

Suppose A, B, C are independent. Then A (or A^c) is independent of any event that can be made from B and C by set operations. For example, A is independent of BC because $P(A(BC)) = P(ABC) = P(A)P(B)P(C) = P(A)P(BC)$. For a somewhat more complicated example, here's a proof that A is independent of $B \cup C$:

$$\begin{aligned} P(A(B \cup C)) &= P(AB) + P(AC) - P(ABC) \\ &= P(A)[P(B) + P(C) - P(B)P(C)] \\ &= P(A)P(B \cup C). \end{aligned}$$

If the three events A, B, and C have to do with three physically separated parts of a probability experiment, then we would expect them to be independent. But three events could happen to be independent even if they are not physically separated. The definition of independence for three events involves four equalities–one for each pairwise independence, and the final one: $P(ABC) = P(A)P(B)P(C)$.

Finally, we give a definition of independence for any finite collection of events, which generalizes the above definitions for independence of two or three events.

Definition 2.4.7 *Events $A_1, A_2, \ldots, A_n$ are* independent *if*

$$P(A_{i_1} A_{i_2} \cdots A_{i_k}) = P(A_{i_1})P(A_{i_2}) \cdots P(A_{i_k})$$

whenever $2 \leq k \leq n$ and $1 \leq i_1 < i_2 < \cdots < i_k \leq n$.

The definition of independence is strong enough that if new events are made by set operations on nonoverlapping subsets of the original events, then the new events are also independent. That is, suppose $A_1, A_2, \ldots, A_n$ are independent events, suppose $n = n_1 + \cdots + n_k$ with $n_i \geq 1$ for each i, and suppose B_1 is defined by Boolean operations (intersections, complements, and unions) of the first n_1 events $E_1, \ldots, E_{n_1}$, B_2 is defined by Boolean operations on the next n_2 events, $A_{n_1+1}, \ldots, A_{n_1+n_2}$, and so on, then $B_1, \ldots, B_k$ are independent.

2.4.2 Independent random variables (of discrete-type)

Definition 2.4.8 *Random variables X and Y are independent if any event of the form $X \in A$ is independent of any event of the form $Y \in B$.*

If X and Y are independent random variables and i and j are real values, then $P\{X = i, Y = j\} = p_X(i)p_Y(j)$ (this follows by taking $A = \{i\}$ and $B = \{j\}$ in the definition of independence). Conversely, if X and Y are discrete-type random variables such that $P\{X = i, Y = j\} = p_X(i)p_Y(j)$ for all i, j, then for any subsets A and B of the real numbers,

$$\begin{aligned} P\{X \in A, Y \in B\} &= \sum_{(i,j):i\in A, j\in B} P\{X = i, Y = j\} = \sum_{i\in A}\sum_{j\in B} p_X(i)p_Y(j) \\ &= \left(\sum_{i\in A} p_X(i)\right)\left(\sum_{j\in B} p_Y(j)\right) = P\{X \in A\}P\{Y \in B\}, \end{aligned}$$

so that $\{X \in A\}$ and $\{Y \in B\}$ are mutually independent events. Thus, discrete random variables X and Y are independent if and only if $P\{X = i, Y = j\} = p_X(i)p_Y(j)$ for all i, j,

More generally, random variables (not necessarily discrete-type) $X_1, X_2, \ldots, X_n$ are mutually independent if any set of events of the form $\{X_1 \in A_1\}, \{X_2 \in A_2\}, \ldots, \{X_n \in A_n\}$ are mutually independent. Independence of random variables is discussed in more detail in Section 4.4.

2.4.3 Bernoulli distribution

Some distributions arise so frequently that they have names. Two such distributions are discussed in this section: the Bernoulli and binomial distributions. The geometric and Poisson distributions are two other important discrete-type distributions with names, and they are introduced in later sections.

A random variable X is said to have the *Bernoulli distribution* with parameter p, where $0 \le p \le 1$, if $P\{X = 1\} = p$ and $P\{X = 0\} = 1-p$. Note that $E[X] = p$. Since $X = X^2$, $E[X^2] = E[X] = p$. So $\mathrm{Var}(X) = E[X^2] - E[X]^2 = p - p^2 = p(1 - p)$. The variance is plotted as a function of p in Figure 2.4. It is symmetric about $1/2$, and achieves its maximum value, $1/4$, at $p = 1/2$.

2.4.4 Binomial distribution

Suppose n independent Bernoulli trials are conducted, each resulting in a one with probability p and a zero with probability $1 - p$. Let X denote the total number of ones occurring in the n trials. Any particular outcome with k ones and $n - k$ zeros, such as 11010101, if $n = 8$ and $k = 5$, has probability $p^k(1 - p)^{n-k}$. Since there are $\binom{n}{k}$ such outcomes, we find that the pmf of X is

$$p_X(k) = \binom{n}{k} p^k(1 - p)^{n-k} \quad \text{for } 0 \le k \le n.$$

The distribution of X is called the *binomial distribution* with parameters n and p. Figure 2.5 shows the binomial pmf for $n = 24$ and $p = 1/3$. Since a partition for the sample space Ω is the set of

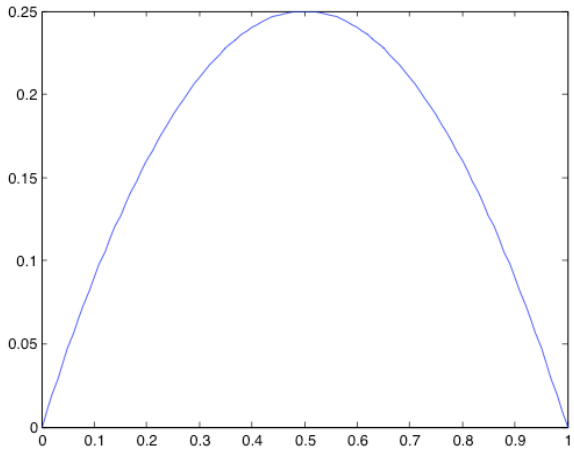


Figure 2.4: The variance of a Bernoulli random variable versus p.

$n+1$ events of the form $\{X = k\}$ for $0 \leq k \leq n$, it follows from the axioms of probability that the pmf p_X just derives sums to one. We will double check that fact using a series expansion. Recall that the Taylor series expansion of a function f about a point x_o is given by

$$f(x) = f(x_o) + f'(x_o)(x - x_o) + f''(x_o)\frac{(x - x_o)^2}{2} + f'''(x_o)\frac{(x - x_o)^3}{3!} + \cdots .$$

The Maclaurin series expansion of a function f is the Taylor series expansion about $x_o = 0$:

$$f(x) = f(0) + f'(0)x + f''(0)\frac{x^2}{2} + f'''(0)\frac{x^3}{3!} + \cdots .$$

The Maclaurin series expansion of $f(x) = (1 + x)^n$ is given by

$$(1 + x)^n = \sum_{k=0}^{n} \binom{n}{k} x^k. \tag{2.6}$$

Substituting $x = p/(1 - p)$ into (2.6) and multiplying through by $(1 - p)^n$ yields that

$$\sum_{k=0}^{n} \binom{n}{k} p^k (1 - p)^{n-k} = 1,$$

so the binomial pmf sums to one, as expected.

Since each trial results in a one with probability p, and there are n trials, the mean number of

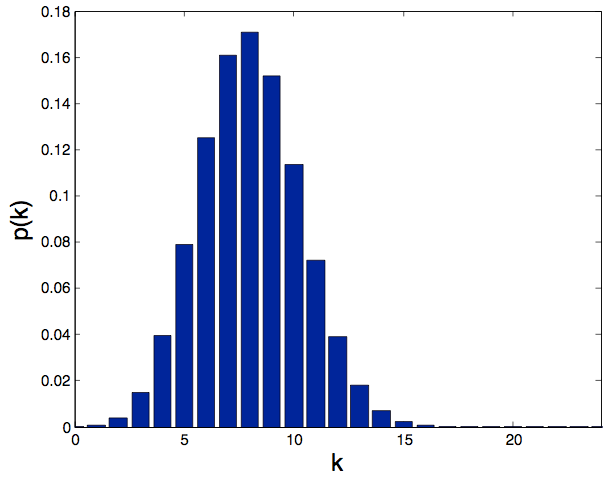


Figure 2.5: The pmf of a binomial random variable with $n = 24$ and $p = 1/3$.

ones is given by $E[X] = np$. This same result can be derived with more work from the pmf:

$$\begin{aligned} E[X] &= \sum_{k=0}^{n} k\binom{n}{k}p^k(1-p)^{n-k} \\ &= \sum_{k=1}^{n} k\binom{n}{k}p^k(1-p)^{n-k} \\ &= np\sum_{k=1}^{n}\frac{(n-1)!}{(n-k)!(k-1)!}p^{k-1}(1-p)^{n-k} \\ &= np\sum_{l=0}^{n-1}\binom{n-1}{l}p^l(1-p)^{n-1-l} \quad (\text{here } l = k-1) \\ &= np. \end{aligned} \tag{2.7}$$

The variance of the binomial distribution is given by $\text{Var}(X) = np(1-p)$. This fact can be shown using the pmf, but a simpler derivation is given in Example 4.8.1.

To explore the shape of the pmf, we examine the ratio of consecutive terms:

$$\begin{aligned} \frac{p(k)}{p(k-1)} &= \frac{\frac{n!}{k!(n-k)!}\ p^k(1-p)^{n-k}}{\frac{n!}{(k-1)!(n-k+1)!}\ p^{k-1}(1-p)^{n-k+1}} \\ &= \frac{(n-k+1)p}{k(1-p)}. \end{aligned}$$

Therefore, $p(k) \geq p(k-1)$ if and only if $(n-k+1)p \geq k(1-p)$, or equivalently, $k \leq (n+1)p$. Therefore, letting $k^* = \lfloor (n+1)p \rfloor$, (that is, k^* is the largest integer less than or equal to $(n+1)p$,

which is approximately equal to np) the following holds:

$$p(0) < \cdots < p(k^*-1) = p(k^*) > p(k^*+1) > \cdots > p(n) \qquad \text{if (n+1)p is an integer}$$
$$p(0) < \cdots < p(k^*-1) < p(k^*) > p(k^*+1) > \cdots > p(n) \qquad \text{if (n+1)p is not an integer.}$$

That is, the pmf p(k) increases monotonically as k increases from 0 to k^* and it decreases monotonically as k increases from k^* to n. Thus, k^* is the value of k that maximizes $p(k)$; which is to say that k^* is the *mode* of the pmf. For example, if $n = 24$ and $p = 1/3$, the pmf is maximized at $k^* = 8$, as seen in Figure 2.5.

A number m is called a *median* of the binomial distribution if $\sum_{k:k\le m} p(k) \ge 0.5$ and $\sum_{k:k\ge m} p(k) \ge 0.5$. It can be shown that any median m of the binomial distribution satisfies $|m - np| \le \max\{p, 1-p\}$.[1], so the median is always close to the mean, np.

Example 2.4.9 [video] Suppose two teams, A and B, play a best-of-seven series of games. Assume that ties are not possible in each game, that each team wins a given game with probability one half, and the games are independent. The series ends once one of the teams has won four games, because after winning four games, even if all seven games were played, that team would have more wins than the other team, and hence it wins the series. Let Y denote the total number of games played. Find the pmf of Y.

Solution. The possible values of Y are 4,5,6, or 7, so let $4 \le n \le 7$. The event $\{Y = n\}$ can be expressed as the union of two events:

$$\{Y = n\} = \{Y = n, A \text{ wins the series}\} \cup \{Y = n, B \text{ wins the series}\}.$$

The events $\{Y = n, A \text{ wins the series}\}$ and $\{Y = n, B \text{ wins the series}\}$ are mutually exclusive, and, by symmetry, they have the same probability. Thus, $p_Y(n) = 2P\{Y = n, A \text{ wins the series}\}$. Next, notice that $\{Y = n, A \text{ wins the series}\}$ happens if and only if A wins three out of the first $n-1$ games, and A also wins the nth game. The number of games that team A wins out of the first $n-1$ games has the binomial distribution with parameters $n-1$ and $p = \frac{1}{2}$. So, for $4 \le n \le 7$,

$$\begin{aligned} p_Y(n) &= 2P\{Y = n, A \text{ wins the series}\} \\ &= 2P\{A \text{ wins 3 of the first } n-1 \text{ games}\}P\{A \text{ wins the } n^{th} \text{ game}\} \\ &= 2\left\{\binom{n-1}{3}2^{-(n-1)}\right\}\frac{1}{2} = \binom{n-1}{3}2^{-(n-1)}. \end{aligned}$$

or $(p_Y(4), p_Y(5), p_Y(6), p_Y(7)) = (\frac{1}{8}, \frac{1}{4}, \frac{5}{16}, \frac{5}{16})$. By the same reasoning, more generally, if A wins each game with some probability p, not necessarily equal to $\frac{1}{2}$:

$$\begin{aligned} p_Y(n) &= P\{Y = n, A \text{ wins the series}\} + P\{Y = n, B \text{ wins the series}\} \\ &= \left\{\binom{n-1}{3}p^3(1-p)^{n-4}\right\}p + \left\{\binom{n-1}{3}p^{n-4}(1-p)^3\right\}(1-p) \end{aligned}$$

for $4 \le n \le 7$.

[1] R. Kaas and J.M. Buhrman, "Mean, median, and mode in binomial distributions," *Statistica Neerlandica*, vol 34, no. 1, pp. 13 - 18, 2008.

2.5 Geometric distribution

Suppose a series of independent trials are conducted, such that the outcome of each trial is modeled by a Bernoulli random variable with parameter p. Here p is assumed to satisfy $0 < p \leq 1$. Thus, the outcome of each trial is one with probability p and zero with probability $1 - p$. Let L denote the number of trials conducted until the outcome of a trial is one. Let us find the pmf of L. Clearly the range of possible values of L is the set of positive integers. Also, $L = 1$ if and only if the outcome of the first trial is one. Thus, $p_L(1) = p$. The event $\{L = 2\}$ happens if the outcome of the first trial is zero and the outcome of the second trial is one. By independence of the trials, it follows that $p_L(2) = (1 - p)p$. Similarly, the event $\{L = 3\}$ happens if the outcomes of the first two trials are zero and the outcome of the third trial is one, so $p_L(3) = (1 - p)(1 - p)p = (1 - p)^2p$. By the same reasoning, for any $k \geq 1$, the event $\{L = k\}$ happens if the outcomes of the first $k - 1$ trials are zeros and the outcome of the kth trial is one. Therefore,

$$p_L(k) = (1 - p)^{k-1}p \text{ for } k \geq 1.$$

The tail probabilities for L have an even simpler form. Specifically, if $k \geq 0$ then the event $\{L > 0\}$ happens if the outcomes of the first k trials are zeros. So

$$P\{L > k\} = (1 - p)^k \quad \text{for } k \geq 0. \tag{2.8}$$

The random variable L is said to have the *geometric distribution* with parameter p. The fact that the sum of the probabilities is one,

$$\sum_{k=1}^{\infty}(1 - p)^{k-1}p = 1,$$

is a consequence of taking $x = 1 - p$ in the formula (1.5) for the sum of a geometric series.

By differentiating each side of (1.5), setting $x = (1 - p)$, and rearranging, we find that $E[L] = \sum_{k=1}^{\infty} kp_L(k) = \frac{1}{p}$. For example, if $p = 0.01$, $E[L] = 100$.

Another, more elegant way to find $E[L]$ is to condition on the outcome of the first trial. If the outcome of the first trial is one, then L is one. If the outcome of the first trial is zero, then L is one plus the number of additional trials until there is a trial with outcome one, which we call $\widetilde{L}$. Therefore, $E[L] = p \cdot 1 + (1-p)E[1+\widetilde{L}]$ However, L and $\widetilde{L}$ have the same distribution, because they are both equal to the number of trials needed until the outcome of a trial is one. So $E[L] = E[\widetilde{L}]$, and the above equation becomes $E[L] = 1 + (1 - p)E[L]$, from which it follows that $E[L] = \frac{1}{p}$.

The variance of L can be found similarly. We could differentiate each side of (1.5) twice, set $x = 1 - p$, and rearrange. Here we take the alternative approach, just described for finding the mean. By the same reasoning as before,

$$E[L^2] = p + (1 - p)E[(1 + \widetilde{L})^2] \text{ or } E[L^2] = p + (1 - p)E[(1 + L)^2].$$

Expanding out, using the linearity of expectation, yields

$$E[L^2] = p + (1-p)(1 + 2E[L] + E[L^2]).$$

Solving for $E[L^2]$, using the fact $E[L] = \frac{1}{p}$, yields $E[L^2] = \frac{2-p}{p^2}$. Therefore, $\mathrm{Var}(L) = E[L^2] - E[L]^2 = \frac{1-p}{p^2}$. The standard deviation of L is $\sigma_L = \sqrt{\mathrm{Var}(L)} = \frac{\sqrt{1-p}}{p}$.

It is worth remembering that if p is very small, the mean $E[L] = \frac{1}{p}$ is very large, and the standard deviation is nearly as large as the mean (just smaller by the factor $\sqrt{1-p}$).

Example 2.5.1 Suppose a fair die is repeatedly rolled until each of the numbers one through six shows at least once. What is the mean number of rolls?

Solution. The total number of rolls, R, can be expressed as $R = R_1 + \ldots + R_6$, where for $1 \le i \le 6$, R_i is the number of rolls made after $i-1$ distinct numbers have shown, up to and including the roll such that the ith distinct number shows. For example, if the sequence of numbers showing is $2, 4, 2, 3, 4, 4, 3, 5, 3, 5, 4, 4, 6, 2, 3, 3, 4, 1$, insert a vertical bar just after each roll that shows a new distinct number to get $2|4|2, 3|4, 4, 3, 5|3, 5, 4, 4, 6|2, 3, 3, 4, 1|$. Then R_i is the number of numbers in the ith part of the sequence, or, for this example, $R_1 = R_2 = 1$, $R_3 = 2$, $R_4 = 4$, $R_5 = R_6 = 5$. After $i-1$ numbers have shown, the probability each subsequent roll is distinct from those $i-1$ numbers is $\frac{6-i+1}{6}$. Therefore, R_i has the geometric distribution with parameter $\frac{6-i+1}{6}$. Therefore, $E[R_i] = \frac{6}{6-i+1}$. Hence,

$$E[R] = E[R_1] + \cdots + E[R_6] = \frac{6}{6} + \frac{6}{5} + \frac{6}{4} + \frac{6}{3} + \frac{6}{2} + \frac{6}{1} = 6\left(\frac{1}{1} + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \frac{1}{5} + \frac{1}{6}\right)$$

This is a special case of the *coupon collector problem,* with $n = 6$ coupon types. In general, if there are n coupon types, the expected number of coupons needed until at least one coupon of each type is obtained, is $n\left(\frac{1}{1} + \frac{1}{2} + \cdots + \frac{1}{n}\right) \approx n\ln(n)$.

Memoryless property of geometric distribution Suppose L has the geometric distribution with parameter p. By the definition of conditional probability and (2.8), for $n \ge 0$ and $k \ge 0$:

$$\begin{aligned} P(L > k+n | L > n) &= \frac{P\{L > k+n, L > n\}}{P\{L > n\}} \\ &= \frac{P\{L > k+n\}}{P\{L > n\}} \\ &= \frac{(1-p)^{k+n}}{(1-p)^n} \\ &= (1-p)^k = P\{L > k\}. \end{aligned}$$

That is, $P(L > k+n | L > n) = P\{L > k\}$, which is called the memoryless property in discrete time. Think of this from the perspective of an observer waiting to see something, such that the

total waiting time for the observer is L time units, where L has a geometric distribution. The memoryless property means that given the observer has not finished waiting after n time units, the conditional probability that the observer will still be waiting after k additional time units, is equal to the unconditional probability that the observer will still be waiting after k time units from the beginning.

2.6 Bernoulli process and the negative binomial distribution

Recall that a random variable has the Bernoulli distribution with parameter p if it is equal to one with probability p and to zero otherwise. A Bernoulli process is an infinite sequence, $X_1, X_2, \ldots$, of Bernoulli random variables, all with the same parameter p, and independent of each other. Therefore, for example, $P\{X_5 = 1, X_6 = 1, X_7 = 0, X_{12} = 1\} = p^3(1-p)$. The kth random variable X_k indicates the (random) outcome of the kth trial in an infinite sequence of trials. For any ω in the underlying probability space, the Bernoulli process has a corresponding realized value $X_k(\omega)$ for each time k, and that function of time is called the *sample path* of the Bernoulli process for outcome ω. A sample path of a Bernoulli process is illustrated in Figure 2.6. The figure indicates

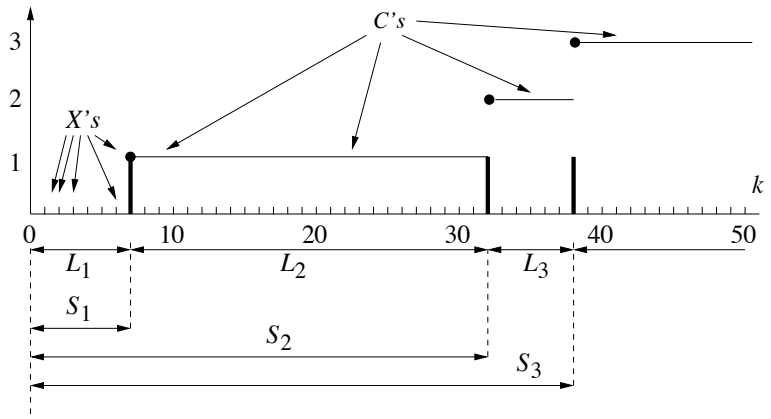


Figure 2.6: Depiction of a sample path of a Bernoulli process.

some additional random variables associated with a Bernoulli process, described next.

Let L_1 be the number of trials needed until the outcome of a trial is one. As seen in Section 2.5, L_1 has the geometric distribution with parameter p. Let L_2 denote the number of trials, after the first L_1 trials, until again the outcome of a trial is one. Then L_2 also has the geometric distribution with parameter p. In general, let L_j denote the number of trials needed after the first $L_1 + \cdots + L_{j-1}$ trials, until again the outcome of a trial is one. The random variables $L_1, L_2, \ldots$ are independent random variables, each geometrically distributed with parameter p. Note that the L's are determined by the X's. The converse is also true; the values of all the L's determine the values of all the X's.

We shall give two more ways to describe the process. Let S_j denote the total number of

trials, counting from the very first one, until a total of j trials have outcome one. Equivalently, $S_j = L_1 + L_2 + \cdots L_j$, for $j \geq 1$. The L's determine the S's, and the converse is also true: $L_j = S_j - S_{j-1}$ for $j \geq 1$, with the understanding $S_0 = 0$.

Let C_k denote the cumulative number of ones in the first k trials. That is, $C_k = X_1 + X_2 + \cdots + X_k$. By convention, $C_0 = 0$. The sequence $(C_k : k \geq 0)$ is sometimes called the counting sequence of the Bernoulli process, because it counts the number of ones vs. the number of trials. Clearly, the counting sequence is determined by the X's. Conversely, the X's are determined by the counting sequence: $X_k = C_k - C_{k-1}$ for $k \geq 0$. If $0 \leq k \leq l$, the difference $C_l - C_k$ is called the *increment* of C over the interval $(k, l] = \{k+1, k+2, \cdots, l\}$. It is the number of trials in the interval with outcome equal to one.

To summarize, there are four ways to describe the same random sequence:

- The underlying Bernoulli sequence $(X_1, X_2, \ldots)$. The random variables $X_1, X_2, \cdots$ are independent Bernoulli random variables with parameter p.
- The numbers of additional trials required for each successive one to be observed: $L_1, L_2, \cdots$. The random variables $L_1, L_2, \cdots$ are independent, geometrically distributed random variables with parameter p.
- The cumulative number of ones in k trials, for $k \geq 0$, $(C_0, C_1, C_2, \ldots)$. For k fixed, C_k is the number of ones in k independent Bernoulli trials, so it has the binomial distribution with parameters k and p. More generally, for $0 \leq k < l$, the increment $C_l - C_k$ is the number of ones in $k - l$ Bernoulli trials, so it has the binomial distribution with parameters $l - k$ and p. Also, the increments of C over nonoverlapping intervals are independent.
- The cumulative numbers of trials for j ones, for $j \geq 0$: $(S_0, S_1, S_2, \ldots)$. As discussed below, for integers $r \geq 1$, S_r has the negative binomial distribution with parameters r and p.

Negative binomial distribution In the remainder of this section we discuss the distribution of S_r, which is the number of trials required for r ones, for $r \geq 1$. The possible values of S_r are $r, r+1, r+2, \ldots$. So let $n \geq r$, and let $k = n - r$. The event $\{S_r = n\}$ is determined by the outcomes of the first n trials. The event is true if and only if there are $r - 1$ ones and k zeros in the first $k + r - 1$ trials, and trial n is a one. There are $\binom{n-1}{r-1}$ such sequences of length n, and each has probability $p^{r-1}(1-p)^{n-r}p$. Therefore, the pmf of S_r is given by

$$p(n) = \binom{n-1}{r-1} p^r (1-p)^{n-r} \quad \text{for } n \geq r.$$

This is called the *negative binomial distribution* with parameters r and p. To check that the pmf sums to one, begin with the Maclaurin series expansion (i.e. Taylor series expansion about zero) of $(1-x)^{-r}$:

$$(1-x)^{-r} = \sum_{k=0}^{\infty} \binom{k+r-1}{r-1} x^k.$$

Set $x = 1 - p$ and use the change of variables $k = n - r$ to get:

$$\sum_{n=r}^{\infty} \binom{n-1}{r-1} p^r (1-p)^{n-r} = p^r \sum_{k=0}^{\infty} \binom{k+r-1}{r-1} (1-p)^k = 1.$$

Use of the expansion of $(1-x)^{-r}$ here, in analogy to the expansion of $(1+x)^n$ used for the binomial distribution, explains the name "negative binomial distribution." Since $S_r = L_1 + \ldots L_r$, where each L_j has mean $\frac{1}{p}$, $E[S_r] = \frac{r}{p}$. It is shown in Example 4.8.1 that $\mathrm{Var}(S_r) = r\mathrm{Var}(L_1) = \frac{r(1-p)}{p^2}$.

2.7 The Poisson distribution–a limit of binomial distributions

By definition, the Poisson probability distribution with parameter $\lambda > 0$ is the one with pmf $p(k) = \frac{e^{-\lambda}\lambda^k}{k!}$ for $k \geq 0$. In particular, 0! is defined to equal one, so $p(0) = e^{-\lambda}$. The next three terms of the pmf are $p(1) = \lambda e^{-\lambda}$, $p(2) = \frac{\lambda^2}{2} e^{-\lambda}$, and $p(3) = \frac{\lambda^3}{6} e^{-\lambda}$. The Poisson distribution arises frequently in practice, because it is a good approximation for a binomial distribution with parameters n and p, when n is very large, p is very small, and $\lambda = np$. Some examples in which such binomial distributions occur are:

- **Radio active emissions in a fixed time interval:** n is the number of uranium atoms in a rock sample, and p is the probability that any particular one of those atoms emits a particle in a one minute period.
- **Incoming phone calls in a fixed time interval:** n is the number of people with cell phones within the access region of one base station, and p is the probability that a given such person will make a call within the next minute.
- **Misspelled words in a document:** n is the number of words in a document and p is the probability that a given word is misspelled.

Binomial distributions have two parameters, namely n and p, and they involve binomial coefficients, which can be cumbersome. Poisson distributions are simpler–having only one parameter, λ, and no binomial coefficients. So it is worthwhile using the Poisson distribution rather than the binomial distribution for large n and small p. We now derive a limit result to give evidence that this is a good approximation. Let $\lambda > 0$, let n and k be integers with $n \geq \lambda$ and $0 \leq k \leq n$, and let $p_b(k)$ denote the probability mass at k of the binomial distribution with parameters n and $p = \lambda/n$.

We first consider the limit of the mass of the binomial distribution at $k = 0$. Note that

$$\ln p_b(0) = \ln(1-p)^n = n \ln(1-p) = n \ln\left(1 - \frac{\lambda}{n}\right).$$

By Taylor's theorem, $\ln(1+u) = u + o(u)$ where $o(u)/u \to 0$ as $u \to 0$. So, using $u = -\frac{\lambda}{n}$,

$$\ln p_b(0) = n\left(-\frac{\lambda}{n} + o\left(-\frac{\lambda}{n}\right)\right) \to -\lambda \quad \text{as } n \to \infty.$$

Therefore,

$$p_b(0) = \left(1 - \frac{\lambda}{n}\right)^n \to e^{-\lambda} \quad \text{as } n \to \infty, \tag{2.9}$$

so the probability mass at zero for the binomial distribution converges to the probability mass at zero for the Poisson distribution. Similarly, for any integer $k \geq 0$ fixed,

$$\begin{aligned} p_b(k) &= \binom{n}{k} p^k (1-p)^{n-k} \\ &= \frac{n \cdot (n-1) \cdots (n-k+1)}{k!} \left(\frac{\lambda}{n}\right)^k \left(1 - \frac{\lambda}{n}\right)^{n-k} \\ &= \frac{\lambda^k p_b(0)}{k!} \left[\frac{n \cdot (n-1) \cdots (n-k+1)}{n^k}\right] \left[\left(1 - \frac{\lambda}{n}\right)^{-k}\right] \qquad (2.10) \\ &\to \frac{\lambda^k e^{-\lambda}}{k!} \quad \text{as } n \to \infty, \end{aligned}$$

because (2.9) holds, and the terms in square brackets in (2.10) converge to one.

Checking that the Poisson distribution sums to one, and deriving the mean and variance of the Poisson distribution, can be done using the pmf as can be done for the binomial distribution. This is not surprising, given that the Poisson distribution is a limiting form of the binomial distribution. The Maclaurin series for e^x plays a role:

$$e^x = \sum_{k=0}^{\infty} \frac{x^k}{k!}. \tag{2.11}$$

Letting $x = \lambda$, and dividing both sides of (2.11) by e^λ, yields

$$1 = \sum_{k=0}^{\infty} \frac{\lambda^k e^{-\lambda}}{k!},$$

so the pmf does sum to one. Following (2.7) line for line yields that if Y has the Poisson distribution with parameter λ,

$$\begin{aligned} E[Y] &= \sum_{k=0}^{\infty} k \frac{\lambda^k e^{-\lambda}}{k!} \\ &= \sum_{k=1}^{\infty} k \frac{\lambda^k e^{-\lambda}}{k!} \\ &= \lambda \sum_{k=1}^{\infty} \frac{\lambda^{k-1} e^{-\lambda}}{(k-1)!} \\ &= \lambda \sum_{l=0}^{\infty} \frac{\lambda^l e^{-\lambda}}{l!} \quad (\text{here } l = k-1) \\ &= \lambda. \end{aligned}$$

Similarly, it can be shown that $\mathrm{Var}(Y) = \lambda$. The mean and variance can be obtained by taking the limit of the mean and limit of the variance of the binomial distribution with parameters n and $p = \lambda/n$, as $n \to \infty$, as follows. The mean of the Poisson distribution is $\lim_{n\to\infty} n\frac{\lambda}{n} = \lambda$, and the variance of the Poisson distribution is $\lim_{n\to\infty} n\frac{\lambda}{n}\left(1 - \frac{\lambda}{n}\right) = \lambda$.

2.8 Maximum likelihood parameter estimation

Sometimes when we devise a probability model for some situation we have a reason to use a particular type of probability distribution, but there may be a parameter that has to be selected. A common approach is to collect some data and then estimate the parameter using the observed data. For example, suppose we decide that an experiment is accurately modeled by a probability model with a random variable X, and that the pmf of X is p_θ, where θ is a parameter, but the value of the parameter is not known before the experiment is performed. When the experiment is performed, suppose we observe a particular value k for X. According to the probability model, the probability of k being the observed value for X, before the experiment was performed, would have been $p_\theta(k)$. It is said that the *likelihood* that $X = k$ is $p_\theta(k)$. The *maximum likelihood estimate* of θ for observation k, denoted by $\widehat{\theta}_{ML}(k)$, is the value of θ that maximizes the likelihood, $p_\theta(k)$, with respect to θ. Intuitively, the maximum likelihood estimate is the value of θ that best explains the observed value k, or makes it the least surprising.

A way to think about it is that $p_\theta(k)$ depends on two variables: the parameter θ and the value of the observed value k for X. It is the likelihood that $X = k$, which depends on θ. For parameter estimation, the goal is to come up with an estimate of the parameter θ for a given value k of the observed value. It wouldn't make sense to maximize the likelihood with respect to the observed value k, because k is assumed to be given. Rather, the maximization is performed with respect to the unknown parameter value; the maximum likelihood estimate is the value of the parameter that maximizes the likelihood of the observed value.

In the following context it makes sense to maximize $p_\theta(k)$ with respect to k for θ fixed. Suppose you know θ and then you enter into a guessing game, in which you guess what the value of X will be, before the experiment is performed. If your guess is k, then your probability of winning is $p_\theta(k)$, so you would maximize your probability of winning by guessing the value of k that maximizes $p_\theta(k)$. For parameter estimation, the value k is observed and the likelihood is maximized with respect to θ; for the guessing game, the parameter θ is known and the likelihood is maximized with respect to k.

Example 2.8.1 Suppose a bent coin is given to a student. The coin is badly bent, but the student can still flip the coin and see whether it shows heads or tails. The coin shows heads with probability p each time it is flipped. The student flips the coin n times for some large value of n (for example, $n = 1000$ is reasonable). Heads shows on k of the flips. Find the ML estimate of p.

Solution: The parameter to be estimated here is p, so p plays the role of θ in the definition of ML estimation. We use the letter "p" in this example instead of "θ" because "p" is commonly used in connection with Bernoulli trials. Let X be the number of times heads shows for n flips of

the coin. It has the binomial distribution with parameters n and p. Therefore, the likelihood that $X = k$ is $p_X(k) = \binom{n}{k}p^k(1-p)^{n-k}$. Here n is known (the number of times the coin is flipped) and k is known (the number of times heads shows). Therefore, the maximum likelihood estimate, $\widehat{p}_{ML}$, is the value of p that maximizes $\binom{n}{k}p^k(1-p)^{n-k}$. Equivalently, $\widehat{p}_{ML}$, is the value of p that maximizes $p^k(1-p)^{n-k}$. First, assume that $1 \le k \le n-1$. Then

$$\frac{d(p^k(1-p)^{n-k})}{dp} = \left(\frac{k}{p} - \frac{n-k}{1-p}\right)p^k(1-p)^{n-k} = (k-np)p^{k-1}(1-p)^{n-k-1}.$$

This derivative is positive if $p \le \frac{k}{n}$ and negative if $p \ge \frac{k}{n}$. Therefore, the likelihood is maximized at $p = \frac{k}{n}$. That is, $\widehat{p}_{ML}(k) = \frac{k}{n}$. A slightly different approach to the computation here would be to note that $\widehat{p}_{ML}$ is also the maximum of the log-likelihood: $\ln p_X(k)$. We assumed that $1 \le k \le n-1$. But if $k = 0$ then the likelihood is $(1-p)^n$, which is maximized at $p = 0$, and if $k = n$ then the likelihood is p^n, which is maximized at $p = 1$, so the formula $\widehat{p}_{ML}(k) = \frac{k}{n}$ is true for $0 \le k \le n$.

Example 2.8.2 Suppose it is assumed that X is drawn at random from the numbers 1 through n, with each possibility being equally likely (i.e. having probability $\frac{1}{n}$). Suppose n is unknown but that it is observed that $X = k$, for some known value k. Find the ML estimator of n given $X = k$ is observed.

Solution: The pmf can be written as $p_X(k) = \frac{1}{n}I_{\{1\le k\le n\}}$. Recall that $I_{\{1\le k\le n\}}$ is the indicator function of $\{1 \le k \le n\}$, equal to one on that set and equal to zero elsewhere. The whole idea now is to think of $p_X(k)$ not as a function of k (because k is the given observation), but rather, as a function of n. An example is shown in Figure 2.7. It is zero if $n \le k-1$; it jumps up to $\frac{1}{k}$

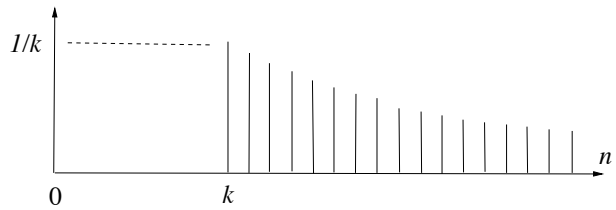


Figure 2.7: The likelihood of $\{X = k\}$ for k fixed, as a function of n.

at $n = k$; as n increases beyond k the function decreases. It is thus maximized at $n = k$. That is, $\widehat{n}_{ML}(k) = k$.

Example 2.8.3 Suppose X has the geometric distribution with some parameter p which is unknown. Suppose a particular value k for X is observed. Find the maximum likelihood estimate, $\widehat{p}_{ML}$.

Solution: The pmf is given by $p_X(k) = (1-p)^{k-1}p$ for $k \geq 1$. If $k = 1$ we have to maximize $p_X(1) = p$ with respect to p, and $p = 1$ is clearly the maximizer. So $\widehat{p}_{ML} = 1$ if $k = 1$. If $k \geq 2$ then $p_X(k) = 0$ for $p = 0$ or $p = 1$. Since $((1-p)^{k-1}p)' = (-(k-1)p-(1-p))(1-p)^{k-2} = (1-kp)(1-p)^{k-2}$, we conclude that $p_X(k)$ is increasing in p for $0 \leq p \leq \frac{1}{k}$ and decreasing in p for $\frac{1}{k} \leq p \leq 1$. Therefore, $\widehat{p}_{ML} = \frac{1}{k}$ if $k \geq 2$. This expression is correct for $k = 1$ as well, so for any $k \geq 1$, $\widehat{p}_{ML} = \frac{1}{k}$.

Example 2.8.4 It is assumed that X has a Poisson distribution with some parameter λ with $\lambda \geq 0$, but the value of λ is unknown. Suppose it is observed that $X = k$ for a particular integer k. Find the maximum likelihood estimate of X.

Solution: The likelihood of observing $X = k$ is $\frac{\lambda^k e^{-\lambda}}{k!}$; the value of λ maximizing such likelihood is to be found for k fixed. Equivalently, the value of λ maximizing $\lambda^k e^{-\lambda}$ is to be found. If $k \geq 1$,

$$\frac{d(\lambda^k e^{-\lambda})}{d\lambda} = (k-\lambda)\lambda^{k-1}e^{-\lambda},$$

so the likelihood is increasing in λ for $\lambda < k$ and decreasing in λ for $\lambda > k$; the likelihood is maximized at $\lambda = k$. Thus, $\widehat{\lambda}_{ML}(k) = k$. If $k = 0$, the likelihood is $e^{-\lambda}$, which is maximized by $\lambda = 0$, so $\widehat{\lambda}_{ML}(k) = k$ for all $k \geq 0$.

2.9 Markov and Chebychev inequalities and confidence intervals

The mean and variance of a random variable are two key numbers that roughly summarize the distribution of the random variable. In some applications, the mean, and possibly the variance, of a random variable are known, but the distribution of the random variable is not. The following two inequalities can still be used to provide bounds on the probability a random variable takes on a value far from its mean. The second of these is used to provide a confidence interval on an estimator of the parameter p of a binomial distribution.

First, the *Markov inequality* states that if Y is a nonnegative random variable, then for $c > 0$,

$$P\{Y \geq c\} \leq \frac{E[Y]}{c}.$$

To prove Markov's inequality for a discrete-type nonnegative random variable Y with possible values $u_1, u_2, \ldots$, note that for each i, u_i is bounded below by zero if $u_i < c$, and u_i is bounded below by c if $u_i \geq c$. Thus,

$$\begin{aligned} E[Y] &= \sum_i u_i p_Y(u_i) \\ &\geq \sum_{i:u_i<c} 0 \cdot p_Y(u_i) + \sum_{i:u_i \geq c} c p_Y(u_i) \\ &= c \sum_{i:u_i \geq c} p_Y(u_i) = cP\{Y \geq c\}, \end{aligned}$$

which implies the Markov inequality. Equality holds in the Markov inequality if and only if $p_Y(0) + p_Y(c) = 1$.

Example 2.9.1 Suppose 200 balls are distributed among 100 buckets, in some particular but unknown way. For example, all 200 balls could be in the first bucket, or there could be two balls in each bucket, or four balls in fifty buckets, etc. What is the maximum number of buckets that could each have at least five balls?

Solution: This question can be answered without probability theory, but it illustrates the essence of the Markov inequality. If asked to place 200 balls within 100 buckets to maximize the number of buckets with exactly five balls, you would naturally put five balls in the first bucket, five in the second bucket, and so forth, until you ran out of balls. In that way, you'd have 40 buckets with five or more balls; that is the maximum possible; if there were 41 buckets with five or more balls then there would have to be at least 205 balls. This solution can also be approached using the Markov inequality as follows. Suppose the balls are distributed among the buckets in some particular way. The probability experiment is to select one of the buckets at random, with all buckets having equal probability. Let Y denote the number of balls in the randomly selected bucket. Then Y is a nonnegative random variable with

$$E[Y] = \sum_{i=1}^{100} (0.01)(\text{number of balls in } i^{th} \text{ bucket}) = (0.01)(200) = 2,$$

so by Markov's inequality, $P\{Y \geq 5\} \leq \frac{2}{5} = 0.4$. That is, the fraction of buckets with five or more balls is less than or equal to 0.4. Equality is achieved if and only if the only possible values of Y are zero and five, that is, if and only if each bucket is either empty or has exactly five balls.

Second, the *Chebychev inequality* states that if X is a random variable with finite mean μ and variance σ^2, then for any $d > 0$,

$$P\{|X - \mu| \geq d\} \leq \frac{\sigma^2}{d^2}. \tag{2.12}$$

The Chebychev inequality follows by applying the Markov inequality with $Y = |X-\mu|^2$ and $c = d^2$. A slightly different way to write the Chebychev inequality is to let $d = a\sigma$, for any constant $a > 0$, to get

$$P\{\big|X - \mu| \geq a\sigma\} \leq \frac{1}{a^2}. \tag{2.13}$$

In words, this form of the Chebychev inequality states that the probability that a random variable differs from its mean by a or more standard deviations is less than or equal to $\frac{1}{a^2}$.

Confidence Intervals The Chebychev inequality can be used to provide confidence intervals for estimators. Confidence intervals are often given when some percentages are estimated based on samples from a large population. For example, an opinion poll report might state that, based on a survey of some voters, 64% favor a certain proposal, with polling accuracy ±5%. In this case, we

would call $[59\%, 69\%]$ the confidence interval. Also, although it is not always made explicit, there is usually a level of confidence associated with the confidence interval. For example, a 95% confidence in a confidence interval means that, from the perspective of someone before the data is observed, the probability the (random) confidence interval contains the fixed, true percentage is at least 95%. (It does *not* mean, from the perspective of someone who knows the value of the confidence interval computed from observed data, that the probability the true parameter is in the interval is at least 95%.)

In practice, the true fraction of voters that favor a certain proposition would be a given number, say p. In order to estimate p we could select n voters at random, where n is much smaller than the total population of voters. For example, a given poll might survey $n = 200$ voters to estimate the fraction of voters, within a population of several thousand voters, that favor a certain proposition. The resulting estimate of p would be $\widehat{p} = \frac{X}{n}$, where X denotes the number of the n sampled voters who favor the proposition. We shall discuss how a confidence interval with a given level of confidence could be determined for this situation. Assuming the population is much larger than n, it is reasonable to model X as a binomial random variable with parameters n and p. The mean of X is np so Chebychev's inequality yields that for any constant $a > 0$:

$$P\{|X - np| \geq a\sigma\} \leq \frac{1}{a^2}.$$

Another way to put it, is:

$$P\left\{\left|\frac{X}{n} - p\right| \geq \frac{a\sigma}{n}\right\} \leq \frac{1}{a^2} \quad \text{or, equivalently,} \quad P\left\{\left|\frac{X}{n} - p\right| < \frac{a\sigma}{n}\right\} \geq 1 - \frac{1}{a^2}.$$

Still another way to put this, using $\widehat{p} = \frac{X}{n}$ and $\sigma = \sqrt{np(1-p)}$, is:

$$P\left\{p \in \left(\widehat{p} - a\sqrt{\frac{p(1-p)}{n}}, \widehat{p} + a\sqrt{\frac{p(1-p)}{n}}\right)\right\} \geq 1 - \frac{1}{a^2}. \tag{2.14}$$

In (2.14), p is the fixed proportion to be estimated. It is treated as a constant, even though we don't know its value. The variable $\widehat{p}$ is random, and therefore the interval $\left(\widehat{p} - a\sqrt{\frac{p(1-p)}{n}}, \widehat{p} + a\sqrt{\frac{p(1-p)}{n}}\right)$ is also random. In fact, such an interval is sometimes called an *interval estimator* to emphasize that it is the interval itself that is random (from the perspective of someone before the data for the probability experiment is observed), not the parameter p. Likewise, $\widehat{p}$ is sometimes called a point estimator. For example, if $a = 5$, then, before we start taking the poll, we would have 96% confidence that p will be in this random interval. This interval is not quite suitable for use as a confidence interval, because it depends on the unknown parameter p, so we wouldn't know the interval. However, $p(1-p) \leq 0.25$ for any $p \in [0, 1]$, so if we replace $p(1-p)$ by 0.25 in defining the confidence interval, it makes the confidence interval larger, and therefore it can only increase the probability that the true parameter is in the interval. In summary, the final form of the confidence interval with *confidence level* $1 - \frac{1}{a^2}$ is $\left(\widehat{p} - \frac{a}{2\sqrt{n}}, \widehat{p} + \frac{a}{2\sqrt{n}}\right)$, and it has the property that

$$P\left\{p \in \left(\widehat{p} - \frac{a}{2\sqrt{n}}, \widehat{p} + \frac{a}{2\sqrt{n}}\right)\right\} \geq 1 - \frac{1}{a^2}. \tag{2.15}$$

Again, in (2.15), it is the interval, not p, that is random. The larger the constant a is, the greater the probability that p is in the random interval. Table 2.2 shows some possible choices of a and the associated upper bound in (2.13), which the confidence level associated with the interval. The

Table 2.2: Some choices of a and resulting values of $1 - \frac{1}{a^2}$.

a	$1-(1/a^2)$
2	75%
5	96%
10	99%

confidence interval given by (2.15) is on the conservative side; the probability on the left-hand side of (2.15) may be much closer to one than the right-hand side. Often in practice other confidence intervals are used which are based on different assumptions (see Example 3.6.10).

Example 2.9.2 Suppose the fraction p of telephone numbers that are busy in a large city at a given time is to be estimated by $\widehat{p} = \frac{X}{n}$, where n is the number of phone numbers that will be tested and X will be the number of the tested numbers found to be busy. If p is to be estimated to within 0.1 with 96% confidence, how many telephone numbers should be sampled, based on (2.15)?

Solution: For 96% confidence we take $a = 5$, so the half-width of the confidence interval is $\frac{a}{2\sqrt{n}} = \frac{2.5}{\sqrt{n}}$, which should be less than or equal to 0.1. This requires $n \geq (\frac{2.5}{0.1})^2 = 625$.

2.10 The law of total probability, and Bayes formula

Events $E_1, \ldots, E_k$ are said to form a *partition* of Ω if the events are mutually exclusive and $\Omega = E_1 \cup \cdots \cup E_k$. Of course for a partition, $P(E_1) + \cdots + P(E_k) = 1$. More generally, for any event A, the *law of total probability* holds because A is the union of the mutually exclusive sets $AE_1, AE_2, \ldots, AE_k$:

$$P(A) \quad = \quad P(AE_1) + \cdots + P(AE_k).$$

If $P(E_i) \neq 0$ for each i, this can be written as

$$P(A) \quad = \quad P(A|E_1)P(E_1) + \cdots + P(A|E_k)P(E_k).$$

Figure 2.8 illustrates the conditions of the law of total probability.

The definition of conditional probability and the law of total probability lead to *Bayes' formula* for $P(E_i|A)$ (if $P(A) \neq 0$) in simple form:

$$P(E_i|A) = \frac{P(AE_i)}{P(A)} = \frac{P(A|E_i)P(E_i)}{P(A)}, \tag{2.16}$$

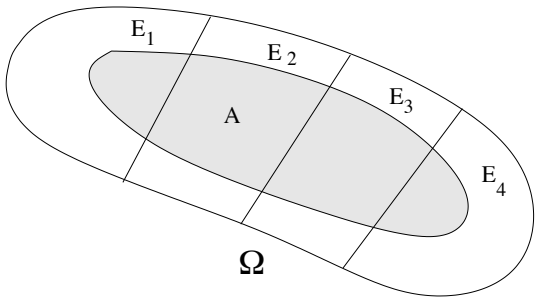


Figure 2.8: Partitioning a set A using a partition of Ω.

or in expanded form:

$$P(E_i|A) = \frac{P(A|E_i)P(E_i)}{P(A|E_1)P(E_1) + \cdots + P(A|E_k)P(E_k)}. \tag{2.17}$$

An important point about (2.16) is that it is a formula for $P(E_i|A)$, whereas the law of total probability uses terms of the form $P(A|E_i)$. In many instances, as illustrated in the following examples, the values $P(A|E_i)$ are specified by the probability model, and (2.16) or (2.17) are used for calculating $P(E_i|A)$.

Example 2.10.1 There are three dice in a bag. One has one red face, another has two red faces, and the third has three red faces. One of the dice is drawn at random from the bag, each die having an equal chance of being drawn. The selected die is repeatedly rolled.
(a) What is the probability that red shows on the first roll?
(b) Given that red shows on the first roll, what is the conditional probability that red shows on the second roll?
(c) Given that red shows on the first three rolls, what is the conditional probability that the selected die has red on three faces?

Solution: Let E_i be the event that the die with i red faces is drawn from the bag, for i =1, 2, or 3. Let R_j denote the event that red shows on the jth role of the die.
(a) By the law of total probability,

$$\begin{aligned} P(R_1) &= P(R_1|E_1)P(E_1) + P(R_1|E_2)P(E_2) + P(R_1|E_3)P(E_3) \\ &= \frac{1}{6}\frac{1}{3} + \frac{2}{6}\frac{1}{3} + \frac{3}{6}\frac{1}{3} = \frac{1}{3}. \end{aligned}$$

(b) We need to find $P(R_2|R_1)$, and we'll begin by finding $P(R_1R_2)$. By the law of total probability,

$$\begin{aligned} P(R_1R_2) &= P(R_1R_2|E_1)P(E_1) + P(R_1R_2|E_2)P(E_2) + P(R_1R_2|E_3)P(E_3) \\ &= \left(\frac{1}{6}\right)^2\frac{1}{3} + \left(\frac{2}{6}\right)^2\frac{1}{3} + \left(\frac{3}{6}\right)^2\frac{1}{3} = \frac{7}{54}. \end{aligned}$$

Therefore, $P(R_2|R_1) = \frac{P(R_1R_2)}{P(R_1)} = \frac{7}{18}$. (Note: we have essentially derived/used Bayes formula.)
(c) We need to find $P(E_3|R_1R_2R_3)$, and will do so by finding the numerator and denominator in the definition of $P(E_3|R_1R_2R_3)$. The numerator is given by $P(E_3R_1R_2R_3) = P(E_3)P(R_1R_2R_3|E_3) = \frac{1}{3}\left(\frac{3}{6}\right)^3 = \frac{1}{24}$. Using the law of total probability for the denominator yields

$$\begin{aligned} P(R_1R_2R_3) &= P(R_1R_2R_3|E_1)P(E_1) + P(R_1R_2R_3|E_2)P(E_2) + P(R_1R_2R_3|E_3)P(E_3) \\ &= \left(\frac{1}{6}\right)^3\frac{1}{3} + \left(\frac{2}{6}\right)^3\frac{1}{3} + \left(\frac{3}{6}\right)^3\frac{1}{3} = \frac{1}{18}. \end{aligned}$$

Therefore, $P(E_3|R_1R_2R_3) = \frac{P(E_3R_1R_2R_3)}{P(R_1R_2R_3)} = \frac{18}{24} = \frac{3}{4}$. (Note: we have essentially derived/used Bayes formula.)

Example 2.10.2 Consider a two stage experiment. First roll a die, and let X denote the number showing. Then flip a fair coin X times, and let Y denote the total number of times heads shows. Find $P\{Y = 3\}$ and $P(X = 3|Y = 3)$.

Solution: By the law of total probability,

$$\begin{aligned} P\{Y=3\} &= \sum_{j=1}^{6} P(Y=3|X=j)p_X(j) \\ &= \frac{1}{6}\left[0+0+\binom{3}{3}2^{-3}+\binom{4}{3}2^{-4}+\binom{5}{3}2^{-5}+\binom{6}{3}2^{-6}\right] \\ &= \frac{1}{6}\left[0+0+\frac{1}{8}+\frac{1}{4}+\frac{10}{32}+\frac{20}{64}\right] \\ &= \frac{1}{6}. \end{aligned}$$

Now $P\{X = 3, Y = 3\}$ is the term in the above sum for $j = 3$: $P\{X = 3, Y = 3\} = \frac{2^{-3}}{6} = \frac{1}{48}$. Therefore,

$$P(X=3|Y=3) = \frac{P\{X=3,Y=3\}}{P\{Y=3\}} = \frac{1}{8}.$$

Example 2.10.3 Consider two boxes as shown in Figure 2.9. Box 1 has three black and two white balls, while Box 2 has two black and two white balls. Consider the following two step experiment.

Step 1: Select a ball from Box 1, all five having equal probability, and transfer it to Box 2.

Step 2: Remove a ball from Box 2, with each of the five balls having the same chance of being removed.

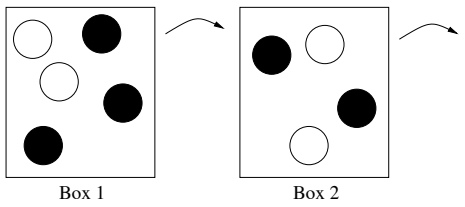


Figure 2.9: Initial state of boxes.

Define the following two events:

W="the ball that is transferred is white"

B="the ball drawn from Box 2 is black"

Find $P(W), P(B)$, and $P(W|B)$?

Solution: This problem is difficult to work out in one's head. But following the definition of conditional probability, it is pretty simple. First, $P(W) = \frac{2}{5}$ because the five possibilities in Step 1 have equal probability, and W is consists of two of the five possibilities. Second, by the law of total probability,

$$\begin{aligned} P(B) &= P(B|W)P(W) + P(B|W^c)P(W^c) \\ &= \frac{2}{5}\cdot\frac{2}{5} + \frac{3}{5}\cdot\frac{3}{5} = \frac{13}{25}. \end{aligned}$$

Third,

$$\begin{aligned} P(W|B) &= \frac{P(WB)}{P(B)} \\ &= \frac{P(W)P(B|W)}{P(B)} \\ &= \frac{\frac{2}{5}\cdot\frac{2}{5}}{\frac{13}{25}} = \frac{4}{13}. \end{aligned}$$

(We just used Bayes formula, perhaps without even realizing it.) For this example it is interesting to compare $P(W|B)$ to $P(W)$. We might reason about the ordering as follows. Transferring a white ball (i.e. event W) makes removing a black ball in Step 2 (i.e. event B) less likely. So given that B is true, we would expect the conditional probability of W to be smaller than the unconditional probability. That is indeed the case ($\frac{4}{13} < \frac{2}{5}$).

As we've seen, the law of total probability can be applied to calculate the probability of an event, if there is a partition of the sample space. The law of total probability can also be used to compute the mean of a random variable. The conditional mean of a discrete-type random variable

X given an event A is defined the same way as the original unconditional mean, but using the conditional pmf:

$$E[X|A] = \sum_i u_i P(X = u_i|A),$$

and also the conditional version of LOTUS holds:

$$E[g(X)|A] = \sum_i g(u_i) P(X = u_i|A).$$

These equations were used implicitly in Section to derive the mean and variance of a geometrically distributed random variable. The law of total probability extends from probabilities to conditional expectations as follows. If $E_1, \ldots, E_J$ is a partition of the sample space, and X is a random variable,

$$E[X] = \sum_{j=1}^{J} E[X|E_j] P(E_j).$$

Example 2.10.4 Let $0 < p < 0.5$. Suppose there are two biased coins. The first coin shows heads with probability p and the second coin shows heads with probability q, where $q = 1 - p$. Consider the following two stage experiment. First, select one of the two coins at random, with each coin being selected with probability one half, and then flip the selected coin n times. Let X be the number of times heads shows. Compute the pmf, mean, and standard deviation of X.

Solution: Let A be the event that the first coin is selected. By the law of total probability,

$$\begin{aligned} p_X(k) &= P\{X = k\} = P(\{X = k\}A) + P(\{X = k\}A^c) \\ &= P(X = k|A)P(A) + P(X = k|A^c)P(A^c) \\ &= \frac{1}{2}\binom{n}{k} p^k q^{n-k} + \frac{1}{2}\binom{n}{k} q^k p^{n-k}. \end{aligned}$$

The two conditional pmfs and the unconditional pmf of X are shown in Figure 2.10 for $n = 24$ and $p = 1/3$.

The two plots in the top of the figure are the binomial pmf for $n = 24$ and parameter $p = 1/3$ and the binomial pmf for $n = 24$ and the parameter $p = 2/3$. The bottom plot in the figure shows the resulting pmf of X. The mean of X can be calculated using the law of total probability:

$$\begin{aligned} E[X] &= E[X|A]P(A) + E[X|A^c]P(A^c) \\ &= \frac{np}{2} + \frac{nq}{2} = \frac{n(p+q)}{2} = \frac{n}{2}. \end{aligned}$$

To calculate $\mathrm{Var}(X)$ we will use the fact $\mathrm{Var}(X) = E[X^2] - (E[X])^2$, and apply the law of total probability:

$$E[X^2] = E[X^2|A]P(A) + E[X^2|A^c]P(A^c).$$

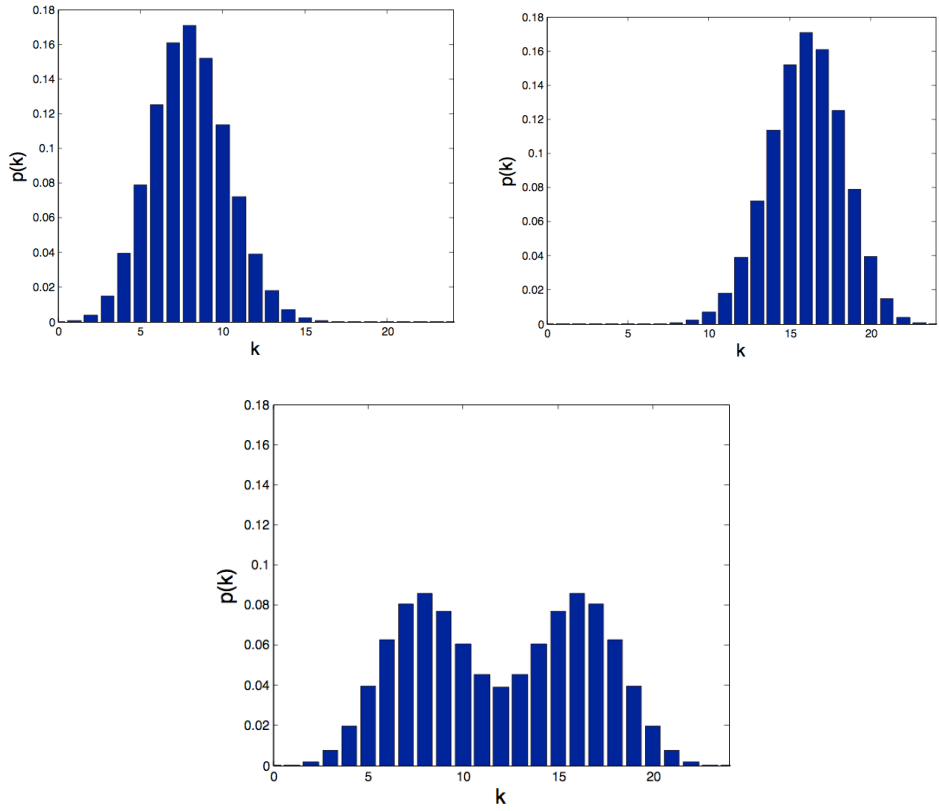


Figure 2.10: The pmfs of X for $n = 24$ and $p = 1/3$. Top left: conditional pmf of X given the first coin is selected. Top right: conditional pmf of X given the second coin is selected. Bottom: The unconditional pmf of X.

Here $E[X^2|A]$ is the second moment of the binomial distribution with parameters n and p, which is equal to the mean squared plus the variance: $E[X^2|A] = (np)^2 + npq$. Similarly, $E[X^2|A^c] = (nq)^2 + nqp$. Therefore,

$$E[X^2] = \frac{(np)^2 + npq}{2} + \frac{(nq)^2 + nqp}{2} = \frac{n^2(p^2+q^2)}{2} + npq,$$

so

$$\begin{aligned} \mathrm{Var}(X) &= E[X^2] - E[X]^2 \\ &= n^2\left(\frac{p^2+q^2}{2} - \frac{1}{4}\right) + npq \\ &= \left(\frac{n(1-2p)}{2}\right)^2 + np(1-p). \end{aligned}$$

Note that

$$\sigma_X = \sqrt{\mathrm{Var}(X)} \geq \frac{(1-2p)n}{2},$$

which for fixed p grows linearly with n. In comparison, the standard deviation for a binomial random variable with parameters n and p is $\sqrt{np(1-p)}$, which is proportional to $\sqrt{n}$.

2.11 Binary hypothesis testing with discrete-type observations

The basic framework for binary hypothesis testing is illustrated in Figure 2.11. It is assumed that

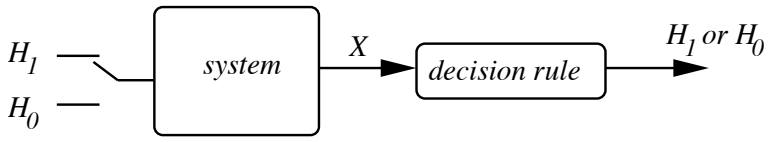


Figure 2.11: The framework for binary hypothesis testing.

either hypothesis H_1 is true or hypothesis H_0 is true, as indicated by the position of the switch at the left end of the figure. Based on which hypothesis is true, a system generates an observation X. The observation is fed into a decision rule, which then declares either H_1 or H_0. The system is assumed to be random, so the decision rule can sometimes declare the true hypothesis, or it can make an error and declare the other hypothesis. For example, the data could be from a computer aided tomography (CAT) scan system, and the hypotheses could be H_1 : a tumor is present; H_0 : no tumor is present. Here we model the observed data by a discrete-type random variable X. Suppose if hypothesis H_1 is true, then X has pmf p_1 and if hypothesis H_0 is true then X has pmf p_0. The *likelihood matrix* is an array with one row for each of the two hypotheses, and one column for each possible value of X. The entries in the row for hypothesis H_i are values of the corresponding pmf, p_i. For example, the likelihood matrix might be the following:

	$X=0$	$X=1$	$X=2$	$X=3$
H_1	0.0	0.1	0.3	0.6
H_0	0.4	0.3	0.2	0.1

In practice, the numbers in the table might be based on data accumulated from past experiments when either one or the other hypothesis is known to be true. As mentioned above, a decision rule specifies, for each possible observation, which hypothesis is declared. A decision rule can be conveniently displayed on the likelihood matrix by underlining one entry in each column, to specifying which hypothesis is to be declared for each possible value of X. An example of a decision rule is shown below, where H_1 is declared whenever $X \geq 1$. For example, if $X = 2$ is observed, then H_1 is declared, because the entry underlined under $X = 2$ is in the H_1 row of the likelihood matrix.

	$X=0$	$X=1$	$X=2$	$X=3$		
H_1	0.0	$\underline{0.1}$	$\underline{0.3}$	$\underline{0.6}$	$\leftarrow$	underlines indicate the decision rule
H_0	$\underline{0.4}$	0.3	0.2	0.1		used for this example.

Since there are two possibilities for which hypothesis is true, and two possibilities for which hypothesis is declared, there are four possible outcomes:

Hypothesis H_0 is true and H_0 is declared.

Hypothesis H_1 is true and H_1 is declared.

Hypothesis H_0 is true and H_1 is declared. This is called a *false alarm.*[2]

Hypothesis H_1 is true and H_0 is declared. This is called a *miss.*

By convention, $p_{\text{false alarm}}$ is defined to be the conditional probability:

$$p_{\text{false alarm}} = P(\text{declare } H_1 \text{ true } | H_0).$$

Note that $p_{\text{false alarm}}$ is the sum of the entries in the H_0 row of the likelihood matrix that are not underlined. For the decision rule given above $p_{\text{false alarm}} = P(\text{declare } H_1 \text{ true } | H_0) = 0.3+0.2+0.1 = 0.6$. Note that $p_{\text{false alarm}}$ is rather large for this rule.

A similar analysis can be carried out for the case when H_1 is the true hypothesis and we declare that H_0 is the true hypothesis. By convention, p_{miss} is defined to be the conditional probability:

$$p_{\text{miss}} = P(\text{declare } H_0 \text{ true } | H_1).$$

Note that p_{miss} is the sum of the entries of the H_1 row of the likelihood matrix that are not underlined. The decision rule above declares H_1 unless $X = 0$, and $P(X = 0|H_1) = 0.0$. Therefore $p_{\text{miss}} = 0.0$, which is unusually good. Of course this small value p_{miss} is earned at the expense of the large value of $p_{\text{false alarm}}$ noted above.

It is important to keep in mind the convention that both $p_{\text{false alarm}}$ and p_{miss} are defined as *conditional* probabilities.

Any decision rule can be indicated by underlining one element in each column of the likelihood matrix. For a given rule, p_{miss} is the sum of entries not underlined in the H_1 row of the likelihood matrix, and $p_{\text{false alarm}}$ is the sum of the entries not underlined in the H_0 row of the likelihood matrix. This representation allows us to illustrate trade-offs between the two types of error probabilities. If the underlining in some column is moved from one row to the other, then one error probability increases (because there is one more entry not underlined to include in the sum) and correspondingly, the other error probability decreases (because there is one fewer entry not underlined to include in the other sum.) For example, if we were to modify the sample decision rule above by declaring H_0 when $X = 1$, then $p_{\text{false alarm}}$ would be reduced from 0.6 to just 0.3 while p_{miss} would increase from 0.0 to 0.1. Whether this pair of error probabilities is better than the original pair depends on considerations not modeled here.

So far, we have discussed decision rules in general. But which decision rule should be used? The two most widely used methods for coming up with decision rules are described next. Each decision rule has a corresponding pair $p_{\text{false alarm}}$ and p_{miss}. The philosophy we recommend for a given real-world application is to try to evaluate the pair of conditional error probabilities $(p_{\text{false alarm}}, p_{\text{miss}})$ for multiple decision rules and then make a final selection of decision rule.

2.11.1 Maximum likelihood (ML) decision rule

The ML decision rule declares the hypothesis which maximizes the probability (or likelihood) of the observation. Operationally, the ML decision rule can be stated as follows: Underline the larger

[2]Use of "false alarm" is common in the signal processing literature. This is called a "type I error" in the statistics literature. The word "miss" is used in the signal processing literature for the other type of error, which is called a "type II error" in the statistics literature.

entry in each column of the likelihood matrix. If the entries in a column of the likelihood matrix are identical, then either can be underlined. The choice may depend on other considerations such as whether we wish to minimize $p_{\text{false alarm}}$ or p_{miss}. The ML rule for the example likelihood matrix above is the following:

	$X=0$	$X=1$	$X=2$	$X=3$
H_1	0.0	0.1	$\underline{0.3}$	$\underline{0.6}$
H_0	$\underline{0.4}$	$\underline{0.3}$	0.2	0.1

$\leftarrow$ underlines indicate the ML decision rule .

It is easy to check that for the ML decision rule, $p_{\text{false alarm}} = 0.2{+}0.1 = 0.3$ and $p_{\text{miss}} = 0.0{+}0.1 = 0.1$.

There is another way to express the ML decision rule. Note that for two positive numbers a and b, the statement $a > b$ is equivalent to the statement that $\frac{a}{b} > 1$. Thus, the ML rule can be rewritten in a form called a *likelihood ratio test* (LRT) as follows. Define the likelihood ratio $\Lambda(k)$ for each possible observation k as the ratio of the two conditional probabilities:

$$\Lambda(k) = \frac{p_1(k)}{p_0(k)}.$$

The ML rule is thus equivalent to deciding that H_1 is true if $\Lambda(X) > 1$ and deciding H_0 is true if $\Lambda(X) < 1$. The ML rule can be compactly written as

$$\Lambda(X) \begin{cases} > 1 & \text{declare } H_1 \text{ is true} \\ < 1 & \text{declare } H_0 \text{ is true.} \end{cases}$$

We shall see that the other decision rule, described in the next section, can also be expressed as an LRT, but with the threshold 1 changed to different values. An LRT with threshold τ can be written as

$$\Lambda(X) \begin{cases} > \tau & \text{declare } H_1 \text{ is true} \\ < \tau & \text{declare } H_0 \text{ is true.} \end{cases}$$

Note that if the threshold τ is increased, then there are fewer observations that lead to deciding H_1 is true. Thus, as τ increases, $p_{\text{false alarm}}$ decreases and p_{miss} increases. For most binary hypothesis testing problems there is no rule that simultaneously makes both $p_{\text{false alarm}}$ and p_{miss} small. In a sense, the LRT's are the best possible family of rules, and the parameter τ can be used to select a given operating point on the tradeoff between the two error probabilities. As noted above, the ML rule is an LRT with threshold $\tau = 1$.

2.11.2 Maximum a posteriori probability (MAP) decision rule

The other decision rule we discuss requires the computation of joint probabilities such as $P(\{X = 1\} \cap H_1)$. For brevity we write this probability as $P(H_1, X = 1)$. Such probabilities cannot be deduced from the likelihood matrix alone. Rather, it is necessary for the system designer to assume some values for $P(H_0)$ and $P(H_1)$. Let the assumed value of $P(H_i)$ be denoted by π_i, so $\pi_0 = P(H_0)$ and $\pi_1 = P(H_1)$. The probabilities π_0 and π_1 are called *prior* probabilities, because they are the probabilities assumed prior to when the observation is made.

Together the conditional probabilities listed in the likelihood matrix and the prior probabilities determine the joint probabilities $P(H_i, X = k)$, because $P(H_i, X = k) = \pi_i p_i(k)$. The *joint probability matrix* is the matrix of joint probabilities $P(H_i, X = k)$. For our first example, suppose $\pi_0 = 0.8$ and $\pi_1 = 0.2$. Then the joint probability matrix is given by

	$X=0$	$X=1$	$X=2$	$X=3$
H_1	0.00	0.02	0.06	0.12
H_0	0.32	0.24	0.16	0.08.

Note that the row for H_i of the joint probability matrix is π_i times the corresponding row of the likelihood matrix. Since the row sums for the likelihood matrix are one, the sum for row H_i of the joint probability matrix is π_i. Therefore, the sum of all entries in the joint probability matrix is one. The joint probability matrix can be viewed as a Venn diagram.

Conditional probabilities such as $P(H_1|X = 2)$ and $P(H_0|X = 2)$ are called *a posteriori probabilities*, because they are probabilities that an observer would assign to the two hypotheses after making the observation (in this case observing that $X = 2$). Given an observation, such as $X = 2$, the maximum a posteriori (MAP) decision rule chooses the hypothesis with the larger conditional probability. By Bayes' formula, $P(H_1|X = 2) = \frac{P(H_1,X=2)}{P(X=2)} = \frac{P(H_1,X=2)}{P(H_1,X=2)+P(H_0,X=2)} = \frac{0.06}{0.06+0.16}$ That is, $P(H_1|X = 2)$ is the top number in the column for $X = 2$ in the joint probability matrix divided by the sum of the numbers in the column for $X = 2$. Similarly the conditional probability $P(H_0|X = 2)$ is the bottom number in the column for $X = 2$ divided by the sum of the numbers in the column for $X = 2$. Since the denominators are the same (both denominators are equal to $P\{X = 2\}$) it follows that whether $P(H_1|X = 2) > P(H_0|X = 2)$ is equivalent to whether the top entry in the column for $X = 2$ is greater than the bottom entry in the column for $X = 2$.

Thus, the MAP decision rule can be specified by underlining the larger entry in each column of the joint probability matrix. For our original example, the MAP rule is given by

	$X=0$	$X=1$	$X=2$	$X=3$
H_1	0.00	0.02	0.06	$\underline{0.12}$
H_0	$\underline{0.32}$	$\underline{0.24}$	$\underline{0.16}$	0.08

$\leftarrow$ underlines indicate the MAP decision rule .

Thus, if the observation is $X = k$, the MAP rule declares hypothesis H_1 is true if $\pi_1 p_1(k) > \pi_0 p_0(k)$, or equivalently if $\Lambda(k) > \frac{\pi_0}{\pi_1}$, where Λ is the likelihood ratio defined above. Therefore, the MAP rule is equivalent to the LRT with threshold $\tau = \frac{\pi_0}{\pi_1}$.

If $\pi_1 = \pi_0$, the prior is said to be *uniform*, because it means the hypotheses are equally likely. For the uniform prior the threshold for the MAP rule is one, and the MAP rule is the same as the ML rule. Does it make sense that if $\pi_0 > \pi_1$, then the threshold for the MAP rule (in LRT form) is greater than one? Indeed it does, because a larger threshold value in the LRT means there are fewer observations leading to deciding H_1 is true, which is appropriate behavior if $\pi_0 > \pi_1$.

The MAP rule has a remarkable optimality property, as we now explain. The average error probability, which we call p_e, for any decision rule can be written as $p_e = \pi_0 p_{\text{false alarm}} + \pi_1 p_{\text{miss}}$. A decision rule is specified by underlining one number from each column of the joint probability matrix. The corresponding p_e is the sum of all numbers in the joint probability matrix that are not underlined. From this observation it easily follows that, among all decision rules, the MAP

decision rule is the one that minimizes p_e. That is why some books call the MAP rule the minimum probability of error rule.

Examples are given in the remainder of this section, illustrating the use of ML and MAP decision rules for hypothesis testing.

Example 2.11.1 Suppose you have a coin and you know that either : H_1: the coin is biased, showing heads on each flip with probability 2/3; or H_0: the coin is fair. Suppose you flip the coin five times. Let X be the number of times heads shows. Describe the ML and MAP decision rules, and find $p_{\text{false_alarm}}, p_{\text{miss}}$, and p_e for both of them, using the prior $(\pi_0, \pi_1) = (0.2, 0.8)$ for the MAP rule and for defining p_e for both rules.

Solution: The rows of the likelihood matrix consist of the pmf of the binomial distribution with $n = 5$ and $p = 2/3$ for H_1 and $p = 1/2$ for H_0 :

	$X=0$	$X=1$	$X=2$	$X=3$	$X=4$	$X=5$
H_1	$\left(\frac{1}{3}\right)^5$	$5\left(\frac{2}{3}\right)\left(\frac{1}{3}\right)^4$	$10\left(\frac{2}{3}\right)^2\left(\frac{1}{3}\right)^3$	$10\left(\frac{2}{3}\right)^3\left(\frac{1}{3}\right)^2$	$5\left(\frac{2}{3}\right)^4\left(\frac{1}{3}\right)$	$\left(\frac{2}{3}\right)^5$
H_0	$\left(\frac{1}{2}\right)^5$	$5\left(\frac{1}{2}\right)^5$	$10\left(\frac{1}{2}\right)^5$	$10\left(\frac{1}{2}\right)^5$	$5\left(\frac{1}{2}\right)^5$	$\left(\frac{1}{2}\right)^5$.

In computing the likelihood ratio, the binomial coefficients cancel, so

$$\Lambda(k) = \frac{\left(\frac{2}{3}\right)^k \left(\frac{1}{3}\right)^{5-k}}{\left(\frac{1}{2}\right)^5} = 2^k \left(\frac{2}{3}\right)^5 \approx \frac{2^k}{7.6}.$$

Therefore, the ML decision rule is to declare H_1 whenever $\Lambda(X) \geq 1$, or equivalently, $X \geq 3$. For the ML rule,

$$\begin{aligned} p_{\text{false_alarm}} &= 10\left(\frac{1}{2}\right)^5 + 5\left(\frac{1}{2}\right)^5 + \left(\frac{1}{2}\right)^5 = 0.5 \\ p_{\text{miss}} &= \left(\frac{1}{3}\right)^5 + 5\left(\frac{2}{3}\right)\left(\frac{1}{3}\right)^4 + 10\left(\frac{2}{3}\right)^2\left(\frac{1}{3}\right)^3 = \frac{51}{243} \approx 0.201 \\ p_e &= (0.2)p_{\text{false_alarm}} + (0.8)p_{\text{miss}} \approx 0.26. \end{aligned}$$

The MAP decision rule is to declare H_1 whenever $\Lambda(X) \geq 0.25$, or equivalently, $X \geq 1$. That is, the MAP rule declares H_0 only if $X = 0$. For the MAP rule,

$$\begin{aligned} p_{\text{false_alarm}} &= 1 - \left(\frac{1}{2}\right)^5 \approx 0.97 \\ p_{\text{miss}} &= \left(\frac{1}{3}\right)^5 = \frac{1}{243} \approx 0.041 \\ p_e &= (0.2)p_{\text{false_alarm}} + (0.8)p_{\text{miss}} \approx 0.227. \end{aligned}$$

As expected, p_e for the MAP rule (designed with the correct prior probabilities in mind) is smaller than p_e for the ML rule.

Example 2.11.2 (Detection problem with Poisson distributed observations) A certain deep space transmitter uses on-off modulation of a laser to send a bit, with value either zero or one. If the bit is zero, the number of photons, X, arriving at the receiver has the Poisson distribution with mean $\lambda_0 = 2$; and if the bits is one, X has the Poisson distribution with mean $\lambda_1 = 6$. A decision rule is needed to decide, based on observation of X, whether the bit was a zero or a one. Describe (a) the ML decision rule, and (b) the MAP decision rule under the assumption that sending a zero is a priori five times more likely than sending a one (i.e. $\pi_0/\pi_1 = 5$). Express both rules as directly in terms of X as possible.

Solution: The ML rule is to decide a one is sent if $\Lambda(X) \geq 1$, where Λ is the likelihood ratio function, defined by

$$\Lambda(k) = \frac{P(X = k|\text{one is sent})}{P(X = k|\text{zero is sent})} = \frac{e^{-\lambda_1}\lambda_1^k/k!}{e^{-\lambda_0}\lambda_0^k/k!} = \left(\frac{\lambda_1}{\lambda_0}\right)^k e^{-(\lambda_1-\lambda_0)} = 3^k e^{-4} \approx \frac{3^k}{54.6}.$$

Therefore, the ML decision rule is to decide a one is sent if $X \geq 4$.

The MAP rule is to decide a one is sent if $\Lambda(X) \geq \frac{\pi_0}{\pi_1}$, where Λ is the likelihood ratio already found. So the MAP rule with $\pi_0 = 5\pi_1$ decides a one is sent if $\frac{3^X}{54.6} \geq 5$, or equivalently, if $X \geq 6$. Note that when $X = 4$ or $X = 5$, the ML rule decides that a one was transmitted, not a zero, but because zeroes are so much more likely to be transmitted than ones, the MAP rule decides in favor of a zero in this case.

Example 2.11.3 (Sensor fusion) Two motion detectors are used to detect the presence of a person in a room, as part of an energy saving temperature control system. The first sensor outputs a value X and the second sensor outputs a value Y. Both outputs have possible values $\{0, 1, 2\}$, with larger numbers tending to indicate that a person is present. Let H_0 be the hypothesis a person is absent and H_1 be the hypothesis a person is present. The likelihood matrices for X and for Y are shown:

	$X=0$	$X=1$	$X=2$
H_1	0.1	0.3	0.6
H_0	0.8	0.1	0.1

	$Y=0$	$Y=1$	$Y=2$
H_1	0.1	0.1	0.8
H_0	0.7	0.2	0.1

For example, $P(Y = 2|H_1) = 0.8$. Suppose, given one of the hypotheses is true, the sensors provide conditionally independent readings, so that
$P(X = i, Y = j|H_k) = P(X = i|H_k)P(Y = j|H_k)$ for $i, j \in \{0, 1, 2\}$ and $k \in \{0, 1\}$.
(a) Find the likelihood matrix for the observation (X, Y) and indicate the ML decision rule. To be definite, break ties in favor of H_1.
(b) Find $p_{\text{false_alarm}}$ and p_{miss} for the ML rule found in part (a).
(c) Suppose, based on past experience, prior probabilities $\pi_1 = P(H_1) = 0.2$ and $\pi_0 = P(H_0) = 0.8$ are assigned. Compute the joint probability matrix and indicate the MAP decision rule.

(d) For the MAP decision rule, compute $p_{\text{false_alarm}}$, p_{miss}, and the unconditional probability of error $p_e = \pi_0 p_{\text{false_alarm}} + \pi_1 p_{\text{miss}}$.
(e) Using the same priors as in part (c), compute the unconditional error probability, p_e, for the ML rule from part (a). Is it smaller or larger than p_e found for the MAP rule in (d)?

Solution: (a) The likelihood matrix for observation (X, Y) is the following.

$(X,Y) \to$	$(0,0)$	$(0,1)$	$(0,2)$	$(1,0)$	$(1,1)$	$(1,2)$	$(2,0)$	$(2,1)$	$(2,2)$
H_1	0.01	0.01	<u>0.08</u>	0.03	<u>0.03</u>	<u>0.24</u>	0.06	<u>0.06</u>	<u>0.48</u>
H_0	<u>0.56</u>	<u>0.16</u>	0.08	<u>0.07</u>	0.02	0.01	<u>0.07</u>	0.02	0.01.

The ML decisions are indicated by the underlined elements. The larger number in each column is underlined, with the tie in case $(0, 2)$ broken in favor of H_1, as specified in the problem statement. Note that the row sums are both one.

(b) For the ML rule, $p_{\text{false_alarm}}$ is the sum of the entries in the row for H_0 in the likelihood matrix that are not underlined. So $p_{\text{false_alarm}} = 0.08 + 0.02 + 0.01 + 0.02 + 0.01 = 0.14$.
For the ML rule, p_{miss} is the sum of the entries in the row for H_1 in the likelihood matrix that are not underlined. So $p_{\text{miss}} = 0.01 + 0.01 + 0.03 + 0.06 = 0.11$.

(c) The joint probability matrix is given by

$(X,Y) \to$	$(0,0)$	$(0,1)$	$(0,2)$	$(1,0)$	$(1,1)$	$(1,2)$	$(2,0)$	$(2,1)$	$(2,2)$
H_1	0.002	0.002	0.016	0.006	0.006	<u>0.048</u>	0.012	0.012	<u>0.096</u>
H_0	<u>0.448</u>	<u>0.128</u>	<u>0.064</u>	<u>0.056</u>	<u>0.016</u>	0.008	<u>0.056</u>	<u>0.016</u>	0.008.

(The matrix specifies $P(X = i, Y = j, H_k)$ for each hypothesis H_k and for each possible observation value (i, j). The 18 numbers in the matrix sum to one. The MAP decisions are indicated by the underlined elements in the joint probability matrix. The larger number in each column is underlined.)

(d) For the MAP rule,

$$p_{\text{false_alarm}} = P\left[(X,Y) \in \{(1,2),(2,2)\} | H_0\right] = 0.01 + 0.01 = 0.02,$$

and

$$p_{\text{miss}} = P((X,Y) \notin \{(1,2),(2,2)\} | H_1) = 1 - P((X,Y) \in \{(1,2),(2,2)\} | H_1) = 1 - 0.24 - 0.48 = 0.28.$$

Thus, for the MAP rule, $p_e = (0.8)(0.02) + (0.2)(0.28) = 0.072$. (This p_e is also the sum of the probabilities in the joint probability matrix that are not underlined.)

(e) Using the conditional probabilities found in (a) and the given values of π_0 and π_1 yields that for the ML rule: $p_e = (0.8)(0.14) + (0.2)(0.11) = 0.134$, which is larger than the value 0.072 for the MAP rule, as expected because of the optimality of the MAP rule for the given priors.

2.12 Reliability

Reliability of complex systems is of central importance to many engineering design problems. Extensive terminology, models, and graphical representations have been developed within many different fields of engineering, from construction of major structures to logistics of complex operations. A common theme is to try to evaluate the reliability of a large system by recursively evaluating the reliability of its subsystems. Often no more underlying probability theory is required beyond that covered earlier in this chapter. However, intuition can be sharpened by considering the case that many of the events have very small probabilities.

2.12.1 Union bound

A general tool for bounding failure probabilities is the following. Given two events A and B, the *union bound* is

$$P(A \cup B) \leq P(A) + P(B).$$

A proof of the bound is that $P(A) + P(B) - P(A \cup B) = P(AB) \geq 0$. If the bound, $P(A) + P(B)$, is used as an approximation to $P(A \cup B)$, the error or gap is $P(AB)$. If A and B have large probabilities, the gap can be significant; $P(A) + P(B)$ might even be larger than one. However, in general, $P(AB) \leq \min\{P(A), P(B)\}$, so the bound is never larger than two times $P(A \cup B)$. Better yet, if A and B are independent and have small probabilities, then $P(AB) = P(A)P(B) \ll P(A \cup B)$ (here "$\ll$" means "much smaller than"). For example, if $P(A) = P(B) = 0.001$ and A and B are independent, then $P(A \cup B) = 0.002 - 0.000001$ which is very close to the union bound, 0.002.

The union bound can be extended from a bound on the union of two events to a bound on the union of m events for any $m \geq 2$. The *union bound* on the probability of the union of a set of m events is

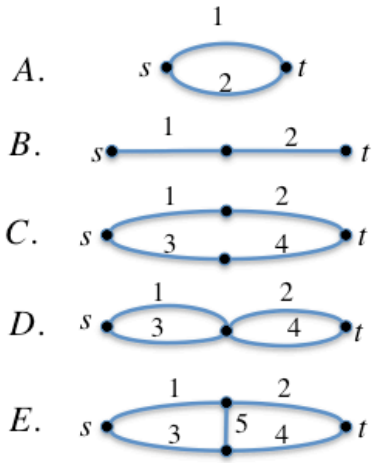
$$P(A_1 \cup A_2 \cup \cdots \cup A_m) \leq P(A_1) + P(A_2) + \cdots + P(A_m). \tag{2.18}$$

The union bound for m events follows from the union bound for two events and argument by induction on m. The union bound is applied in the following sections.

2.12.2 Network outage probability

The network outage problem addressed here is one of many different ways to describe and analyze system reliability involving serial and parallel subsystems. An $s - t$ *network* consists of a source node s, a terminal node t, possibly some additional nodes, and some links that connect pairs of nodes. Figure 2.12 pictures five such networks, network A through network E. Network A has two links, link 1 and link 2, in parallel. Network B has two links in series. Network C has two parallel paths, with each path consisting of two links in series. Network D has two stages of two parallel links. Network E has five links.

Suppose each link i fails with probability p_i. Let F_i be the event that link i fails. So $p_i = P(F_i)$. Assume that for a given network, the links fail independently. That is, the F_i's are independent events. *Network outage* is said to occur if at least one link fails along every $s - t$ path, where an $s - t$ path is a set of links that connect s to t. Let F denote the event of network outage. We will



	$P(F)$
A.	p_1p_2
B.	$p_1 + p_2 - p_1p_2$
C.	$(p_1 + p_2 - p_1p_2)(p_3 + p_4 - p_3p_4)$
D.	$p_1p_3 + p_2p_4 - p_1p_2p_3p_4$
E.	$p_1p_3 + p_2p_4 - p_1p_2p_3p_4 + p_1q_2q_3p_4p_5 + q_1p_2p_3q_4p_5$

Figure 2.12: Five $s-t$ networks with associated outage probabilities.

find $P(F)$ for each of the five networks, and then apply the union bound to get a simple upper bound to $P(F)$ for networks B through E.

Exact calculation of network outage Network A has two $s-t$ paths: the path consisting of link 1 alone, and the path consisting of link 2 alone. Therefore, network outage happens if and only if both links fail. Thus, for network A, $F = F_1F_2$. By the assumption that F_1 and F_2 are independent, $P(F) = p_1p_2$. For example, if $p_1 = p_2 = 10^{-3}$, then $P(F) = 10^{-6}$. That is, if each link fails with probability one in a thousand, the network outage probability is one in a million.

Network B has only one $s-t$ path, consisting of both links 1 and 2. Therefore, outage occurs in network B if link 1 fails or if link 2 fails. So for network B, $F = F_1 \cup F_2$, and $P(F) = p_1 + p_2 - p_1p_2$. Notice that if p_1 and p_2 are both very small, then p_1p_2 is small compared to $p_1 + p_2$. For example, if $p_1 = p_2 = 10^{-3}$, then $P(F) = 2 \times 10^{-3} - 10^{-6} = 0.001999 \approx 0.002$.

Network C is similar to network A. Like network A, network C has two $s-t$ paths in parallel, but for network C, each path consists of two links in series. The upper path has links 1 and 2, and the event the upper path fails is $F_1 \cup F_2$, which has probability $p_1 + p_2 - p_1p_2$. Similarly, the lower path has links 3 and 4, and the event the lower path fails is $F_3 \cup F_4$, which has probability $p_3 + p_4 - p_3p_4$. Since the two $s-t$ paths involve separate sets of links, the event the upper path fails is independent of the event the lower path fails. So the network outage probability is the product of the failure probabilities for the upper and lower paths. That is, for network C, $P(F) = (p_1 + p_2 - p_1p_2)(p_3 + p_4 - p_3p_4)$. If $p_i = 0.001$ for all i, $P(F) = (0.001999)^2 \approx .000004$.

Network D is similar to network B. Like network B, network D has two stages in series, but each stage of network D consists of two links in parallel instead of only a single link per stage. Outage occurs in network D if and only if either the first stage fails or the second stage fails. The first stage has links 1 and 3, and the event the first stage fails can be written as F_1F_3, which has probability p_1p_3. Similarly, the second stage has links 2 and 4, and the event that the

second stage fails can be written as F_2F_4, which has probability p_2p_4. Since the two stages involve separate sets of links, the first stage fails independently of the second stage. So the network outage probability is $P\{\text{first stage fails}\}+P\{\text{second stage fails}\}-P\{\text{both stages fail}\}$. That is, for network C, $P(F) = p_1p_3+p_2p_4-p_1p_2p_3p_4$. If $p_i = 0.001$ for all i, $P(F) \approx 0.000002$, or about half the outage probability for network C.

Finally, consider network E. One way to approach the computation of $P(F)$ is to use the law of total probability, and consider two cases: either link 5 fails or it doesn't fail. So $P(F) = P(F|F_5)P(F_5) + F(F|F_5^c)P(F_5^c)$. Now $P(F|F_5)$ is the network outage probability if link 5 fails. If link 5 fails we can erase it from the picture, and the remaining network is identical to network C. So $P(F|F_5)$ for Network E is equal to $P(F)$ for network C. If link 5 does not fail, the network becomes equivalent to network D. So $P(F|F_5^c)$ for Network E is equal to $P(F)$ for network D. Combining these yields:

$$P(\text{outage in network } E) = p_5P(\text{outage in network } C) + (1-p_5)P(\text{outage in network } D).$$

Another way to calculate $P(F)$ for network E is by closer comparison to network D. If the same four links used in network D are used along with link 5 in network E, then network E fails whenever network D fails. Moreover, some thought shows there are exactly two ways network E can fail such that network D does not fail, namely, links 1,4, and 5 are the only ones that fail, or links 2, 3, and 5 are the only ones that fail. So $P(F)$ for network D is equal to $P(F)$ for network D plus $p_1q_2q_3p_4p_5 + q_1p_2p_3q_4p_5$, where $q_i = 1-p_i$ for all i. This yields the outage probability given in Figure 2.12. For example, if $p_i = 0.001$ for $1 \le i \le 5$, then $P(F) = 0.000002001995002 \approx 0.000002$. The network outage probability is determined mainly by the probability that both links 1 and 3 fail or both links 2 and 4 fail. Thus, the outage probability of network E is about the same as for network D.

Applying the union bound The union bound doesn't apply and isn't needed for network A.

For network B, $F = F_1 \cup F_2$. The union bound for two events implies that $P(F) \le p_1 + p_2$. The numerical value is 0.002 if $p_1 = p_2 = 0.001$.

For network C, $F = F_1F_3 \cup F_1F_4 \cup F_2F_3 \cup F_2F_4$ so that the union bound for $m = 4$ yields $P(F) \le p_1p_3 + p_1p_4 + p_2p_3 + p_2p_4 = (p_1+p_2)(p_3+p_4)$. This upper bound could also be obtained by first upper bounding the probability of failure of the upper branch by p_1+p_2 and the lower branch by $p_3 + p_4$ and multiplying these bounds because the branches are independent. The numerical value is 0.000004 in case $p_i = 0.001$ for $1 \le i \le 4$.

For network D, $F = F_1F_3 \cup F_2F_4$. The union bound for two events yields $P(F) \le p_1p_3 + p_2p_4$. The numerical value is 0.000002 in case $p_i = 0.001$ for $1 \le i \le 4$.

Fo network E, $F = F_1F_3 \cup F_2F_4 \cup F_1F_5F_4 \cup F_3F_5F_2$. The union bound for four events yields $P(F) \le p_1p_3 + p_2p_4 + p_1p_5p_4 + p_3p_5p_2$. The numerical value is 0.000002002 in case $p_i = 0.001$ for $1 \le i \le 5$, which is approximately 0.000002.

Note that for all four networks, B through E, the numerical value of the union bound, in case $p_i = 0.001$ for all i, is close to the exact value of network outage computed above.

Calculation by exhaustive listing of network states Another way to calculate the outage probability of an $s-t$ network is to enumerate all the possible network states, determine which ones correspond to network outage, and then add together the probabilities of those states. The method becomes intractable for large networks,[3] but it is often useful for small networks and it sharpens intuition about most likely failure modes.

For example, consider network C. Since there are four links, the network state can be represented by a length four binary string such as 0110, such that the ith bit in the string is one if link i fails, This computation is illustrated in Table 2.3, where we let $q_i = 1 - p_i$. For example, if the network

Table 2.3: A table for calculating $P(F)$ for network C.

State	Network fails?	probability
0000	no	
0001	no	
0010	no	
0011	no	
0100	no	
0101	yes	$q_1p_2q_3p_4$
0110	yes	$q_1p_2p_3q_4$
0111	yes	$q_1p_2p_3p_4$
1000	no	
1001	yes	$p_1q_2q_3p_4$
1010	yes	$p_1q_2p_3q_4$
1011	yes	$p_1q_2p_3p_4$
1100	no	
1101	yes	$p_1p_2q_3p_4$
1110	yes	$p_1p_2p_3q_4$
1111	yes	$p_1p_2p_3p_4$

state is 0110, meaning that links 2 and 3 fail and links 1 and 4 don't, then the network fails. The probability of this state is $q_1p_2p_3q_4$. For given numerical values of p_1 through p_4, numerical values can be computed for the last column of the table, and the sum of those values is $P(F)$.

2.12.3 Distribution of the capacity of a flow network

An $s-t$ flow network is an $s-t$ network such that each link has a capacity. Two $s-t$ flow networks are shown in Figure 2.13. We assume as before that each link i fails with probability p_i. If a link fails, it cannot carry any flow. If a link does not fail, it can carry flow at a rate up to

[3]From a computational complexity point of view, the problem of computing the outage probability, also called the reliability problem, has been shown to be P-space complete, meaning exact solution for large networks is probably not feasible.

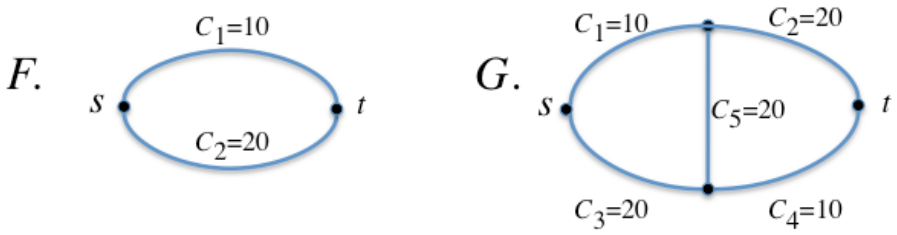


Figure 2.13: Two $s-t$ flow networks.

the link capacity. Link capacities and flows are given in units of some quantity per unit time. For example, the units could be gallons per minute (for an irrigation system), packets per second (for a communication network), or vehicles per hour (for a highway transportation system).

For simplicity we have chosen to work with undirected links, and we assume that the flow can be in either direction on a link, up to the total capacity of the link. We assume that flow can be split among multiple paths. The *capacity* of an $s-t$ flow network is the maximum flow rate from s to t. The capacity is random because of the possible link failures. We will compute the pmf of the capacity of each of the two networks.

We begin with Network F. If both of the links are working, then the capacity is 30, because the top link can carry flow at rate 20 and the bottom link can carry flow at rate 10. If link 2 fails and link 1 does not, then the total maximum flow rate is 10. By considering all four network states, we arrive at the following expression for the pmf of X, the capacity of Network F:

$$p_X(0) = p_1p_2 \quad p_X(10) = q_1p_2 \quad p_X(20) = p_1q_2 \quad p_X(30) = q_1q_2.$$

Let Y denote the capacity of $s-t$ flow network G. We will find the pmf of Y. If none of the links fail, the network capacity is 30, because flow can be carried from s to t at rate 30 by sending flow over links 1, 2, 3, and 4 from left to right at the link capacities, and sending flow up the middle link at rate 10. If one or more of links from $\{1,2,3,4\}$ fails, then $Y < 30$, because either the total flow out of s would be less than 30 or the total flow into t would be less than 30. If link 5 fails, then the total flow must be less than or equal to 20, because at most 10 units of flow could be relayed over links 1 and 2 in series and at most 10 units of flow could be relayed over links 3 and 4 in services. So if any of the links fail then Y is less than 30. Thus, $p_Y(30) = q_1q_2q_3q_4q_5$. At the other extreme, $Y = 0$ if and only if every $s-t$ path has at least one failed link. Thus, $p_Y(0)$ is the same as the outage probability of $s-t$ network E considered previously. The other two possible values of Y are 10 or 20. In order to have $Y = 20$, to begin with, both links 2 and 3 must work, or else there wouldn't be enough capacity for flow out of s and enough capacity for flow into t. If links 2 and 3 work, then $Y = 20$ if link 5 fails and links 1 and 4 work, or if link 5 works and at least one of links 1 or 4 fails. Therefore, $p_Y(20) = q_2q_3(p_5q_1q_4 + q_5(p_1 + p_4 - p_1p_4))$. Finally, $p_Y(10) = 1 - p_Y(0) - p_Y(20) - p_Y(30)$.

A more systematic way to calculate the pmf of Y is to enumerate all the network states, calculate the capacity of each state, and then for each capacity value, add together all probabilities of all

states with that capacity. The first few lines of such a table are shown in Table 2.4. For example, $p_Y(10)$ is the sum of all probabilities in rows such that the capacity shown in the row is 10.

Table 2.4: A table for calculating the distribution of capacity of network G.

State	capacity	probability
00000	30	$q_1q_2q_3q_4q_5$
00001	20	$q_1q_2q_3q_4p_5$
00010	20	$q_1q_2q_3p_4q_5$
00011	10	$q_1q_2q_3p_4p_5$
00100	10	$q_1q_2p_3q_4q_5$
00101	10	$q_1q_2p_3q_4p_5$
00110	10	$q_1q_2p_3p_4q_5$
00111	10	$q_1q_2p_3p_4p_5$
⋮	⋮	⋮
11111	0	$p_1p_2p_3p_4p_5$

2.12.4 Analysis of an array code

The array shown in Figure 2.12.4 below illustrates a two-dimensional error detecting code for use in digital systems such as computer memory or digital communication systems. There are 49 data

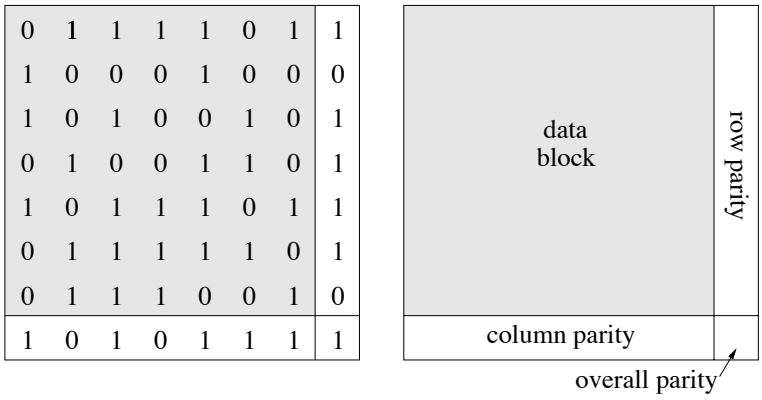


Figure 2.14: Illustration of an eight by eight array code.

bits, located in the shaded region. The array is filled out to a total of 64 bits by the addition of a row and column of parity bits, selected so that all rows and columns have even parity. The idea is that all 64 bits should be stored or transmitted. During storage or transmission, one or more bits might be changed due to a physical deviation. A bit that has been changed is said to be in error,

and the binary array with ones at the locations of the bits in error is called the error pattern. There are 2^{64} error patterns, including the pattern with no errors. When the data is needed, the reader or receiver first checks to see if all rows and columns have even parity. If yes, the data is deemed to be correct. If no, the data is deemed to be corrupted, and we say the errors were detected. Since there may be backup copies in storage, or the data can possibly be retransmitted, the effect of errors is not severe if they are detected. But the effect of errors can be severe if they are not detected, which can happen for some nonzero error patterns.

It is explained next that any error pattern with either one, two, or three bit errors is detected, but certain error patterns with four bit errors are not detected. An error pattern is undetected if and only if the parity of each row and column is even, which requires that the parity of the entire error pattern be even. Thus, any error pattern with an odd number of bit errors is detected. In particular, error patterns with one or three bit errors are detected. If there are two bit errors, they are either in different rows or in different columns (or both). If they are in different rows then those rows have odd parity, and if they are in different columns then those columns have odd parity. So, either way, the error pattern is detected. A pattern with four bit errors is undetected if and only if there are two rows and two columns such that the bit errors are at the four intersections of the two rows and two columns.

We now discuss how to bound the probability of undetected error, assuming that each bit is in error with probability 0.001, independently of all other bits. Let Y denote the number of bits in error. As already discussed, there can be undetected bit errors only if $Y \geq 4$. So the probability there exist undetected bit errors is less than or equal to $P\{Y \geq 4\}$. We have $Y = X_1 + \cdots + X_{64}$, where X_i is one if the ith bit in the array is in error, and $X_i = 0$ otherwise. By assumption, the random variables $X_1, \ldots, X_{64}$ are independent Bernoulli random variables with parameter $p = 0.001$, so Y is a binomial random variable with parameters $n = 64$ and $p = 0.001$. The event $\{Y \geq 4\}$ can be expressed as the union of $\binom{64}{4}$ events of the form $\{X_i = 1 \text{ for all } i \in A\}$, indexed by the subsets A of $\{1, \ldots, n\}$ with $|A| = 4$. For such an A, $P\{X_i = 1 \text{ for all } i \in A\} = \prod_{i \in A} P\{X_i = 1\} = p^4$. So the union bound with $m = \binom{64}{4}$ yields $P\{Y \geq 4\} \leq \binom{64}{4} p^4$. (This is in contrast to the fact $P\{Y = 4\} = \binom{64}{4} p^4 (1-p)^{60}$.) Thus,

$$P(\text{undetected errors}) \leq P\{Y \geq 4\} \leq \binom{64}{4} p^4 = (635376) p^4 = 0.635376 \times 10^{-6}.$$

A considerably tighter bound can be obtained with a little more work. Four errors go undetected if and only if they are at the intersection points of two columns and two rows. There are $\binom{8}{2} = 28$ ways to choose two columns and $\binom{8}{2} = 28$ ways to choose two rows, so there are $28^2 = 784$ error patterns with four errors that go undetected. The probability of any given one of those patterns occurring is less than p^4, so by the union bound, the probability that the actual error pattern is an undetected pattern of four bit errors is less than or equal to $(784)p^4$. Any pattern of five bit errors is detected because any pattern with an odd number of bit errors is detected. So the only other way undetected errors can occur is if $Y \geq 6$, and $P\{Y \geq 6\}$ can again be bounded by the union

bound. Thus,

$$\begin{aligned} P(\text{undetected errors}) &\leq \binom{8}{2}^2 10^{-12} + \binom{64}{6} 10^{-18} \\ &= (784 + 74.974368)10^{-12} \leq (0.859)10^{-9}. \end{aligned}$$

2.12.5 Reliability of a single backup

Suppose there are two machines (such as computers, generators, vehicles, etc.) each used to back up the other. Suppose each day, given a machine was up the day before or was finished being repaired the day before or that it is the first day, the machine is up on the given day with probability 0.999 and it goes down with probability 0.001. If a machine goes down on a given day, then the machine stays down for repair for a total of five days. The two machines go up and down independently. We would like to estimate the mean time until there is a day such that both machines are down.

This is a challenging problem. Can you show that the mean time until outage is between 200 and 400 years? Here is a reasonably accurate solution. Given both machines were working the day before, or were just repaired, or it is the first day, the probability that at least one of them fails on a given day is about 0.002. That is, the waiting time until at least one machine goes down is approximately geometrically distributed with parameter $p = 0.002$. The mean of such a distribution is $1/p$, so the mean time until at least one of the machines goes down is about 500 days. Given that one machine goes down, the probability the other machine goes down within the five day repair time is about 5/1000=1/200. That is, the number of repair cycles until a double outage occurs has a geometric distribution with parameter about 1/200, so on average, we'll have to wait for 200 cycles, each averaging about 500 days. Therefore, the mean total waiting time until double outage is about $500 \times 200 = 100,000$ days, or approximately 274 years.

2.13 Short Answer Questions

Section 2.2[video]

1. Ten balls, numbered one through ten, are in a bag. Three are drawn out at random without replacement, all possibilities being equally likely. Find $E[S]$, where S is the sum of numbers on the three balls.

2. Find $\text{Var}(X)$ if the pmf of X is $p_X(i) = \frac{i}{10}$ for $1 \leq i \leq 4$.

3. Find $\text{Var}(3X + 20)$ if the pmf of X is $p_X(i) = \frac{i}{10}$ for $1 \leq i \leq 4$.

Section 2.3[video]

1. Two fair dice are rolled. Find the conditional probability that doubles are rolled given the sum is even.

2. Two fair dice are rolled. Find the conditional probability the sum is six given the product is even.

3. What is the maximum number of mutually exclusive events that can exist for a given probability experiment, if each of them has probability 0.3?

Section 2.4[video]

1. Find the probability a one is rolled exactly four times in six rolls of a fair die.

2. Find $\text{Var}(3X + 5)$ assuming X has the binomial distribution with parameters $n = 12$ and $p = 0.9$.

3. Each step a certain robot takes is forward with probability 2/3 and backward with probability 1/3, independently of all other steps. What is the probability the robot is two steps in front of its starting position after taking 10 steps.

Section 2.5[video]

1. Find the median of the geometric distribution with parameter $p = .02$.

2. Suppose each trial in a sequence of independent trials is a success with probability 0.5. What is the mean number of trials until two consecutive trials are successful?

Section 2.6[video]

1. Suppose a Bernoulli process starts with 0,0,0,0,0,0,1,0,1,0,0,0,0,1,0,0,0,1 Identify the corresponding values of L_1 through L_4.

2. Suppose a Bernoulli process starts with 0,0,0,0,0,0,1,0,1,0,0,0,0,1,0,0,0,1 Identify the corresponding values of C_{10} through C_{14}.

3. Suppose a Bernoulli process starts with 0,0,0,0,0,0,1,0,1,0,0,0,0,1,0,0,0,1 Identify the corresponding values of S_1 through S_4.

Section 2.7[video]

1. Suppose X is a Poisson random variable such that $P\{X = 0\} = 0.1$. Find $E[X]$.

2. Suppose X is a Poisson random variable with $E[X^2] = 12$. Find $P\{X = 2\}$.

Section 2.8[video]

1. Suppose X has a Poisson distribution with mean θ^2. Find $\widehat{\theta}_{ML}(10)$, the maximum likelihood estimate of θ if it is observed that $X = 10$.

2. Suppose $X = 2Y + 4$, where Y has the geometric distribution with parameter p. Find $\widehat{p}_{ML}(10)$, the maximum likelihood estimate of p if it is observed that $X = 10$.

Section 2.9[video]

1. Let X have mean 100 and standard deviation $\sigma_X = 5$. Find the upper bound on $P\{|X - 100| \geq 20\}$ provided by the Chebychev inequality.

2. Suppose X denotes the number of heads showing in one million flips of a fair coin. Use the Chebychev inequality to identify d large enough that $P\{|X - 500,000| \geq d\} \leq 0.1$.

Section 2.10[video]

1. Two coins, one fair and one with a 60% vs. 40% bias towards heads, are in a pocket, one is drawn out randomly, each having equal probability, and flipped three times. What is the probability that heads shows on all three flips?

2. The number of passengers for a limousine pickup is thought to be either 1, 2, 3, or 4, each with equal probability, and the number of pieces of luggage of each passenger is thought to be 1 or 2, with equal probability, independently for different passengers. What is the probability that there will be six or more pieces of luggage?

3. Two fair dice are rolled. What is the conditional probability the sum is a multiple of three, given it is a multiple of two?

Section 2.11[video]

1. Suppose observation X has pmf $p_1(i) = \frac{i}{10}$ for $1 \leq i \leq 4$ if H_1 is true, and pmf $p_0(i) = 0.25$ for $1 \leq i \leq 4$ if H_0 is true. Find p_{miss} and $p_{false\ alarm}$, respectively, for the ML decision rule.

2. Suppose observation X has pmf $p_1(i) = \frac{i}{10}$ for $1 \leq i \leq 4$ if H_1 is true, and pmf $p_0(i) = 0.25$ for $1 \leq i \leq 4$ if H_0 is true. Suppose H_0 is twice as likely as H_1, a priori (i.e. $\pi_0/\pi_1 = 2$. Find the minimum possible average probability of error.

3. Suppose X has the Poisson distribution with parameter 10 if H_1 is true and the Poisson distribution with parameter 3 if H_0 is true. The ML decision rule declares that H_1 is true if $X \geq \tau$. Find the threshold τ.

Section 2.12[video]

1. Consider a storage system with five servers which does not fail as long as three or more servers is functioning. Suppose the system fails if three or more of the servers crash, and suppose each server crashes independently in a given day with probability 0.001. Find a numerical upper bound, provided by the union bound, on the probability of system failure in a given day.

2. Suppose each corner of a three dimensional cube burns out with probability 0.001, independently of the other corners. Find an upper bound, using the union bound, on the probability that there exist two neighboring corners that both burn out.

2.14 Problems

Discrete random variables (Sections 2.1 and 2.2)

2.1. **[The difference of two dice]**
Suppose two students each roll a standard, six-sided die. Let X_1 be the number showing for the first student and X_2 be the number showing for the second student. Let Y be the absolute value of the difference: $Y = |X_1 - X_2|$.

(a) Identify the set of possible values of Y.

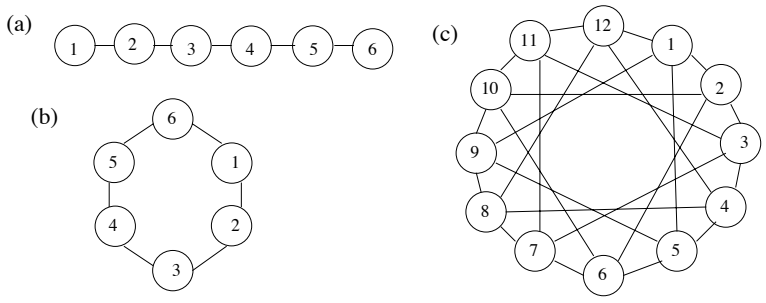
(b) Find and carefully sketch the pmf of Y.

(c) Find $E[Y]$.

(d) Find $\text{Var}(Y)$. (You may have two terms in your answer–you don't need to simplify to the end.)

(e) Suppose $Z = 100Y$, where Y is defined above. Express the mean and variance of Z in terms of the mean and variance of Y. (Note: This does NOT require you to solve the previous parts of this problem.)

2.2. **[Distance between two randomly selected vertices]**
Solve the following problem for each of the three undirected graphs below. For a given graph, two vertices, i and j, are selected at random, with all possible values of (i, j) having equal probability, including the cases with $i = j$. Let D denote the distance between i and j, which is the minimum number of edges that must be crossed to walk in the graph from i to j. If $i = j$ then $D = 0$. Find and sketch the pmf of D, and find its mean and variance. (Hint: For (b) and (c), by symmetry, it can be assumed that $i = 1$ and only j is selected at random.)



2.3. **[Distribution of number of matches]**
Suppose four people write their names on slips of paper; the slips of paper are randomly shuffled and then each person gets back one slip of paper; all possibilities of who gets what slip are equally likely. Let X denote the number of people who get back the slip with their own name (i.e. the number of matches).

(a) Describe a suitable sample space Ω to describe the experiment. How many elements does it have?

(b) Find the pmf of X.

(c) Find $E[X]$.

(d) Find the probability that a given person gets her/his own name. Explain how this question is related to part (c).

(e) Find Var(X).

2.4. **[A problem on sampling without replacement]**
A bag contains n pairs of shoes in distinct styles and sizes. You pick two shoes at random from the bag. Note that this is sampling *without* replacement.

(a) What is the probability that you get a pair of shoes?

(b) What is the probability of getting one left shoe and one right shoe?

Suppose now that $n \geq 2$ and that you choose 3 shoes at random from the bag.

(c) What is the probability that you have a pair of shoes among the three that you have picked?

(d) What is the probability that you picked at least one left shoe and at least one right shoe?

2.5. **[Mean and standard deviation of a complicated random variable]**
(Use a spreadsheet program, Matlab, programmable calculator, or your favorite computer language for this problem.) Suppose two fair dice are rolled independently, so the sample space is $\Omega = \{(i,j) : 1 \leq i \leq 6, 1 \leq j \leq 6\}$ and all outcomes are equally likely. Let X be the random variable defined by $X(i,j) = (i-j)^2 - |2i-j|$.

(a) Calculate $E[X]$. (Hint: This is just the average of the 36 values of X.)

(b) Calculate $E[X^2]$. (Hint: This is just the average of the 36 values of X^2.)

(c) Using your answers to (a) and (b), calculate the standard deviation of X. (Note: If you were to apply the STDEV function of the Excel spreadsheet to the list of 36 values of X, you would not get the standard deviation of X, because the STDEV function is for *estimating* the standard deviation from n samples, and it uses what is called the $n-1$ rule. The STDEVP function returns the correct value, and can be used for you to check your answer.)

(d) Find the pmf of X, $p_X(k)$, for $k = 0, 1, 2$. (Note: In principle, you could compute the complete pmf of X and use it to do (a) and (b) above, but the hints for (a) and (b), based on the law of the unconscious statistician (LOTUS), give a much simpler way to do (a) and (b).)

2.6. **[Mean and standard deviation of two simple random variables]**
Suppose two fair dice are rolled independently, so the sample space is
$\Omega = \{(i,j) : 1 \leq i \leq 6, 1 \leq j \leq 6\}$ and all outcomes are equally likely. Let X be the number showing on the first die, $X(i,j) = i$, and let Y be the minimum of the two numbers showing, $Y(i,j) = \min\{i,j\}$.

(a) Derive the pmf of X and sketch it.

(b) Find the mean, $E[X]$, and standard deviation, σ_X, of X. Correct numerical answers are fine, but show your work.

(c) Derive the pmf of Y and sketch it.

(d) Find the mean $E[Y]$ and standard deviation, σ_Y, of Y. Correct numerical answers are fine, but show your work. (Hint: It may be helpful to use a spreadsheet or computer program.)

(e) Which is larger, σ_X or σ_Y? Is that consistent with your sketches of the pmfs?

(f) The random variable X takes values in the set $\{1, 2, 3, 4, 5, 6\}$. Specify another pmf on the same set which has a larger standard deviation than X.

2.7. **[Up to five rounds of double or nothing]**
A gambler initially having one chip participates in up to five rounds of gambling. In each round, the gambler bets all of her chips, and with probability one half, she wins, thereby doubling the number of chips she has, and with probability one half, she loses all her chips. Let X denote the number of chips the gambler has after five rounds, and let Y denote the maximum number of chips the gambler ever has (with stopping after five rounds). For example, if the gambler wins in the first three rounds and loses in the fourth, $Y = 8$. If the gambler loses in the first round, $Y = 1$.

(a) Find the pmf, mean, and variance of X.

(b) Find the pmf, mean, and variance of Y.

2.8. **[Selecting supply for a random demand]**
A reseller reserves and prepays for L rooms at a luxury hotel for a special event, at a price of a dollars per room, and the reseller sells the rooms for b dollars per room, for some known, fixed values a and b with $0 < a \leq b$. Letting U denote the number of potential buyers, the actual number of buyers is $\min\{U, L\}$ and the profit of the reseller is $b \min\{U, L\} - aL$. The reseller must declare L before observing U, but when L is selected the reseller assumes U takes on the possible values $1, 2, \cdots, M$, each with probability $\frac{1}{M}$ for some known $M \geq 1$.

(a) Express the expected profit of the reseller in terms of M, L, a, and b. Simplify your answer as much as possible. For simplicity, without loss of generality, assume $0 \leq L \leq M$. (Hint: $1 + \cdots + L = \frac{L(L+1)}{2}$. Your answer should be valid whenever $0 \leq L \leq M$; for $L = 3$ and $M = 5$ it should reduce to $E[\text{profit}] = \frac{12b}{5} - 3a$.)

(b) Determine the value of L as a function of a, b and M that maximizes the expected profit.

2.9. **[The Zipf distribution]**
The Zipf distribution has been found to model well the distribution of popularity of items such as books or videos in a library. The Zipf distribution with parameters M and α, is the distribution supported on $\{1, \ldots, M\}$ with pmf $p(k) = \frac{k^{-\alpha}}{Z}$ for $1 \leq k \leq M$, where Z is the constant chosen to make the pmf sum to one, i.e. $Z = \sum_{k=1}^{M} k^{-\alpha}$. The interpretation is that

if the videos in a library are numbered 1 through M, in order of decreasing popularity, then a random request to view a video would be for video k with probability $p(k)$. (Light use of a computer or programable calculator is recommended for this problem.)

(a) If X has the Zipf distribution with parameters $M = 2000$ and $\alpha = 0.8$ (suitable for a typical video library), what is $P\{X \leq 500\}$?

(b) If Y has the Zipf distribution with parameters $M = 2000$ and $\alpha > 0$, for what numerical value of α is it true that $P\{Y \leq 100\} = 30\%$?

2.10. **[First and second moments of a ternary random variable]**
This problem focuses on the possible mean and variance of a random variable X with support set $\{-1, 0, 1\}$. Let $p_X(-1) = a$ and $p_X(1) = b$ and $p_X(0) = 1 - a - b$. This is a valid pmf if and only if $a \geq 0$, $b \geq 0$, and $a + b \leq 1$. Let $\mu = E[X]$, $\sigma^2 = \text{Var}(X)$, and $m_2 = E[X^2]$.

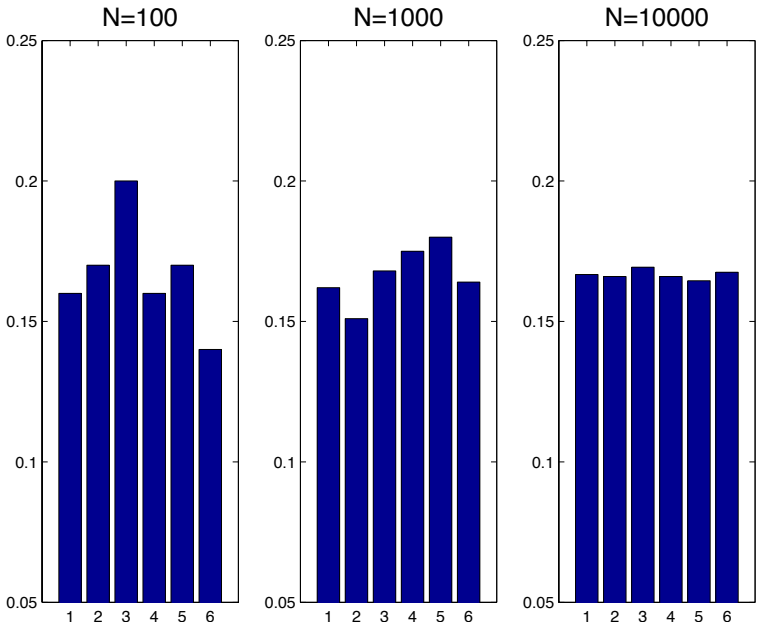
(a) Find (a, b) so that $\mu = \frac{1}{2}$ and $\sigma^2 = \frac{3}{4}$.

(b) Express a and b in terms of μ and m_2. Determine and sketch the region of (μ, m_2) pairs for which there is a valid choice of (a, b).

(c) Determine and sketch the set of (μ, σ^2) pairs for which there is a valid choice of (a, b). (Hint: For μ fixed, the set of possible values of σ^2 is the set of possible values of $m_2 - \mu^2$.)

2.11. **[Producing some simple empirical distributions by computer simulation]**
This problem is intended to be solved using a computer.

(a) Write a computer program that can simulate N independent rolls of a fair die for any integer value of N that you put into the program. Display the empirical distribution, where the *empirical distribution* for such an experiment is the fraction of rolls that are one, the fraction of rolls that are two, and so on, up to the fraction of rolls that are six, for $N = 100, N = 1000$ and $N = 10,000$. Your program output should look something like the histograms shown.



(b) Run your program at least twenty times; the displayed triplet of empirical distributions should change each time. See how much the distributions change from one simulation to the next, for each of the three values of N. The empirical distribution for the case $N = 100$ can often be far from uniform. Here is what to turn in for this part: Find, print out, and turn in the printout, for an example triplet of plots produced by the program such that the first empirical distribution (i.e. the one for $N = 100$) is far from the uniform one.

(c) Produce a modification of the program so that for each N, two dice are rolled N times, and for each time the sum of the two dice is recorded. The empirical distributions will thus have support over the range from 2 to 12. Print out and turn in a figure showing empirical distributions for $N = 100$, $N = 1000$, and $N = 10000$, and also turn in a copy of the computer code you used. (Hint: Even though the sum of two dice is always at least 2, it might be easier to make the plot if one of the bins for the histogram is centered at one; the count for such bin will always be zero.)

Conditional probability, independence, and the binomial distribution Sections 2.3–2.4

2.12. **[Conditional probability]**
Two fair dice are rolled.

(a) Find the conditional probability doubles are rolled, given the sum is eight.

(b) Find the conditional probability the sum is eight, given doubles are rolled.

(c) Find the probability at least one die lands on six.

(d) Find the conditional probability at least one die lands on six, given that doubles are *not* rolled.

2.13. **[Coincidence of birth months]**
There are three persons in a room. Find the probability that *at least* two persons have the same *birth-month*. (For simplicity, assume that each of the twelve possibilities for the birth-month of a given person are equally likely.)

2.14. **[Independence]**
Suppose two dice, one orange and one blue, are rolled. Define the following events:
A: The product of the two numbers that show is 12
B: The number on the orange die is strictly larger than the number on the blue die.
C: The sum of the numbers is divisible by four.
D: The number on the orange die is either 1 or 3.

(a) List all pairs of events from the set A, B, C, and D that are independent.

(b) List all triplets of events, if any, from the set of A, B, C, and D, that are mutually independent.

2.15. **[A Karnaugh map for independent events]**
Suppose A, B, and C are events for a probability experiment such that A and B are mutually independent, $P(A) = P(B) = P(C) = 0.5$, $P(AC) = P(BC) = 0.3$, and $P(ABC) = 0.1$. Fill in the probabilities of all events in a Karnaugh map. Show your work.

2.16. **[A team selection problem]**
Suppose Alice and Bob are among seven people on a debate team. Suppose four of the team members are selected to travel to a debate, with all sets of four having equal probability. Let A be the event that Alice is among the four selected, and B be the event Bob is among the four selected. Find the probabilities indicated. *Briefly explain your reasoning.*

(a) $P(A)$.

(b) $P(A|B)$.

(c) $P(A \cup B)$.

2.17. **[Binomial distribution I]**
Five cars start out on a cross-country race. The probability that a car breaks down and drops out of the race is 0.2. Cars break down independently of each other.

(a) What is the probability that exactly two cars finish the race?

(b) What is the probability that at most two cars finish the race?

(c) What is the probability that at least three cars finish the race?

2.18. **[Binomial distribution II]**
A New Yorker runs an investment management service that has the stated goal of doubling the value of his clients' investments in a week via day trading. His brochure boasts that, "On average, my clients triple their money in five weeks!" After poring over back issues of the *Wall Street Journal* you learn the truth: at the end of any week, the investments of his clients will have doubled with probability 0.5, and will have decreased by 50% with probability 0.5. Thus, at the end of the first week, an initial investment of \$32 will be worth either \$64 or \$16, each with probability 0.5. Performance in any week is independent of performance during the other weeks. Anxious to apply your new skills in probability theory, you decide to invest \$32, and to let that investment ride for five weeks (in fact, you decide not to even look at the stock prices until the five weeks are over). Let the random variable X denote the value in dollars of your investment at the end of a five week period.

(a) What are the possible values of X?

(b) What is the pmf of the random variable X?

(c) What is the expected value of X? Is the TV commercial accurate?

(d) What is the probability that you will lose money on your investment?

2.19. **[The power of sampling before assigning]**
A job arriving in a certain cloud computing system must be routed to one of eight servers. Due to the loads, the servers all have different rates. Routing the job to the server with the highest rate would require sampling the rates of all eight servers. Instead, the rates of three randomly selected distinct servers are sampled (all choices being equally likely) and the job is routed to the *sampled* server with the highest service rate.

(a) Let A be the event the job is assigned to the server with the highest service rate. Find $P(A)$.

(b) Let B be the event the job is assigned to one of the four slowest servers. Find $P(B)$.

(c) Let C be the event the job is assigned to one of the two servers with the highest rates. Find $P(C)$.

2.20. **[Binomial Random Variable]**
An aircraft has four major subsystems that determine its safety: mechanical, electrical, hydraulics, and communications. The aircraft will crash if two or more systems fail at the same time. Assume that each system fails independently with a component probability of failure during a flight of $p_c = 10^{-3}$.

(a) What is the probability that the aircraft will crash during a flight?

(b) Assuming that each flight is independent of the others, and each flight has a probability of 10^{-9} of crashing independent of other flights, how many flights are needed for the probability of the aircraft crashing at least once to reach 0.01%?

(c) What should be the component probability of failure p_c in order for the aircraft failure probability to equal 10^{-9} as in part (b).

Geometric and Poisson distributions, Bernoulli processes, ML parameter estimation and confidence intervals Sections 2.5–2.9

2.21. [**Probability of a tough committee**]
A committee of three judges is randomly selected from among ten judges. Four of the ten judges are tough; the committee is tough if at least two of the judges on the committee are tough. A committee decides whether to approve petitions it receives. A tough committee approves 50% of petitions and a committee that is not tough approves 80% of petitions.

(a) Find the probability a committee is tough.

(b) Find the probability a petition is approved.

(c) Suppose a petition can be submitted many times until it is approved. If a petition is approved with probability 3/4 each time, what is the mean number of times it has to be submitted until it is approved?

(d) If instead, a petition can be submitted a maximum of three times, and the probability of approval each time is 1/4, find the probability a petition is eventually approved.

2.22. [**Repeated rolls of four dice**]

(a) Suppose four dice are simultaneously rolled. What is the probability that two even and two odd numbers show?

(b) Suppose four dice are repeatedly simultaneously rolled. What is the probability that strictly more than three rolls are needed until two even and two odd numbers show on the same roll? (A simultaneous roll of all four dice is counted as one roll.)

2.23. [**Time to first repetition**]
A fair six-sided die is rolled repeatedly. Each time the die is rolled, the number showing is written down. Let X be the number of rolls until the first time a number shows that was already written down. The possible values of X are 2,3,4,5,6, or 7. For each of the following parts, explain your reasoning and express your answer as a fraction in reduced form.

(a) Find $P\{X > 3\}$.

(b) (7 points) Find $P\{X = 5\}$.

(c) Find $P\{X = 7|X > 5\}$.

2.24. **[A knockout game]**
Five distinct numbers are randomly distributed to players numbered 1 through 5. Whenever two players compare their numbers, the one with the higher number is declared the winner. Initially, players 1 and 2 compare their numbers; the winner then compares with player 3, and so on. Let X denote the number of times player 1 is a winner. Find $P\{X = 0\}$, $P\{X = 1\}$, and $P\{X = 2\}$. (Hint: Your answers should sum to $\frac{3}{4}$.)

2.25. **[Ultimate verdict]**
Suppose each time a certain defendant is given a jury trial for a particular charge (such as trying to sell a seat in the US Senate), an innocent verdict is given with probability q_I, a guilty verdict is given with probability q_G, and a mistrial occurs with probability q_M, where q_I, q_G, and q_M are positive numbers that sum to one. Suppose the prosecutors are determined to get a guilty or innocent verdict, so that after any number of consecutive mistrials, another trial is given. The process ends immediately after the first trial with a guilty or innocent verdict; appeals are not considered. Let T denote the total number of trials required, and let I denote the event that the verdict for the final trial is innocent.

(a) Find $P(I|T = 1)$. Express your answer in terms of q_I and q_G.

(b) Find the pmf of T.

(c) Find $P(I)$. Express your answer in terms of q_I and q_G.

(d) Compare your answers to parts (a) and (c). For example, is one always larger than the other?

2.26. **[ML parameter estimation for independent geometrically distributed rvs]**
A certain task needs to be completed n times, where each completion requires multiple attempts. Let L_i be the number of attempts that are needed to complete the task for the ith time. Suppose that $L_1, \ldots, L_n$ are independent and each is geometrically distributed with the same parameter p, to be estimated.

(a) Suppose it is observed that $(L_1, \ldots, L_n) = (k_1, \ldots, k_n)$ for some particular vector of positive integers, $(k_1, \ldots, k_n)$. Write the probability (i.e. likelihood) of this observation as simply as possible in terms of p and $(k_1, \ldots, k_n)$. (Hint: By the assumed independence, the likelihood factors: $P\{(L_1, \ldots, L_n) = (k_1, \ldots, k_n)\} = P\{L_1 = k_1\} \cdots P\{L_n = k_n\}$.)

(b) Find $\widehat{p}_{ML}$ if it is observed that $(L_1, \ldots, L_n) = (k_1, \ldots, k_n)$. Simplify your answer as much as possible.

2.27. **[Maximum likelihood estimation and the Poisson distribution]**
An auto insurance company wishes to charge monthly premiums based on an individual's risk factor. It defines the risk factor as the probability p that individual is involved in a auto accident during a trip. Assume that whether an accident occurred on one trip is independent of accidents occurring on others, i.e., the insurance company assumes that drivers are reckless and don't learn to be cautious after being in an accident. The insurance company assumes that each driver will be driving 120 trips a month.

(a) Determine the maximum likelihood estimate of the risk factor $\hat{p}_{ML}$ if no accidents are reported by a driver in a month. Repeat for the cases when the driver reports 1, 2 and 3.

(b) Assume that the actual value of $p = 0.01$. Compute the approximate values of $P\{X = k\}$ for $k = 0, 1, 2, 3$ using the Poisson approximation to the binomial distribution, and compare those approximations to the actual probabilities computed using the binomial distribution.

2.28. **[Scaling of a confidence interval]**
Suppose the fraction of people in Tokyo in favor of a certain referendum will be estimated by a poll. A confidence interval based on the Chebychev bound will be used (i.e. the interval is centered at $\widehat{p}$ with width $\frac{a}{\sqrt{n}}$ and confidence level $1 - \frac{1}{a^2}$, for some constant a, where $\widehat{p}$ is the fraction of the n people sampled that are in favor of the referendum.)

(a) Suppose the width of the confidence interval would be 0.1 for sample size $n = 300$ and some given confidence level. How many samples would be needed instead to yield a confidence interval that has only half the width, for the same level of confidence?

(b) What is the confidence level for the test of part (a)?

(c) Keeping the width of the confidence interval at 0.1 as in (a), how many samples would be required for a 96% confidence level?

2.29. **[Estimation of signal amplitude for Poisson observation]**
The number of photons X detected by a particular sensor over a particular time period is assumed to have the Poisson distribution with mean $1 + a^2$, where a is the amplitude of an incident field. It is assumed $a \geq 0$, but otherwise a is unknown.

(a) Find the maximum likelihood estimate, $\widehat{a}_{ML}$, of a for the observation $X = 6$.

(b) Find the maximum likelihood estimate, $\widehat{a}_{ML}$, of a given that it is observed $X = 0$.

2.30. **[Parameter estimation for the binomial distribution]**
Suppose X has the binomial distribution with parameters n and p.

(a) Suppose (for this part only) $p = 0.03$ and $n = 100$. Find the Poisson approximation to $P\{X \geq 2\}$. (You may leave one or more powers of e in your answer, but not an infinite number of terms.)

(b) Suppose (for this part only) p is unknown and $n = 10,000$, and based on observation of X we want to estimate p within 0.025. That is, we will use a confidence interval with half-width 0.025. What's the largest confidence level we can claim? (The confidence level is the probability, from the viewpoint before X is observed, that the confidence interval will contain the true value p.)

(c) Suppose (for this part only) it is known that $p = 0.03$, but n is unknown. The parameter n is to be estimated. Suppose it is observed that $X = 7$. Find the maximum likelihood estimate $\hat{n}_{ML}$. (Hint: It is difficult to differentiate with respect to the integer parameter n, so another approach is needed to identify the minimizing value. Think about how a function on the integers behaves near its maximum. How can you tell whether the function is increasing or decreasing from one integer value to the next?)

Bayes' Formula and binary hypothesis testing Sections 2.10 & 2.11

2.31. **[Explaining a sum]**
Suppose $S = X_1 + X_2 + X_3 + X_4$ where X_1, X_2, X_3, X_4 are mutually independent and X_i has the Bernoulli distribution with parameter $p_i = \frac{i}{5}$ for $1 \leq i \leq 4$.

(a) Find $P\{S = 1\}$.

(b) Find $P(X_1 = 1|S = 1)$.

2.32. **[The weight of a positive]**
(Based on G. Gigerenzer, *Calculated Risks*, Simon and Schuster, 2002 and S. Strogatz NYT article, April 25, 2010.) Women aged 40 to 49 years of age have a low incidence of breast cancer; the fraction is estimated at 0.8%. Given a woman with breast cancer has a mammogram, the probability of detection (i.e. a positive mammogram) is estimated to be 90%. Given a woman does not have breast cancer, the probability of a false positive (i.e a false alarm) is estimated to be 7%.

(a) Based on the above numbers, given a woman aged 40 to 49 has a mammogram, what is the probability the mammogram will be positive?

(b) Given a woman aged 40 to 49 has a positive mammogram, what is the conditional probability the woman has breast cancer?

(c) For 1000 women aged 40 to 49 getting mammograms for the first time, how many are expected to have breast cancer, for how many of those is the mammogram positive, and how many are expected to get a false positive?

2.33. **[Which airline was late?]**
Three airlines fly out of the Bloomington airport:

- American has five flights per day; 20% depart late,
- AirTrans has four flights per day; 5% depart late,
- Delta has nine flights per day; 10% depart late.

(a) What fraction of flights flying out of the Bloomington airport depart late?

(b) Given that a randomly selected flight departs late (with all flights over a long period of time being equally likely to be selected) what is the probability the flight is an American flight?

2.34. **[Conditional distribution of half-way point]**
Consider a robot taking a random walk on the integer line. The robot starts at zero at time zero. After that, between any two consecutive integer times, the robot takes a unit length left step or right step, with each possibility having probability one half. Let F denote the event that the robot is at zero at time eight, and let X denote the location of the robot at time four.

(a) Find $P(F)$.

(b) Find the pmf of X.

(c) Find $P(\{X = i\}F)$ for all integer values of i. (For what values of i is $P(\{X = i\}F) > 0$?)

(d) Find the conditional pmf of X given that F is true. It is natural to use the notation $p_X(i|F)$ for this, and it is defined by $p_X(i|F) = P(X = i|F)$ for all integers i. Is the conditional pmf more spread out than the unconditional pmf p_X, or more concentrated?

2.35. **[Matching Poisson means]**
Consider hypotheses H_0 and H_1 about a two dimensional observation vector $X = (X_1, X_2)$. Under H_0, X_1 and X_2 are mutually independent, and both have the Poisson distribution with mean 4. Under H_1, X_1 and X_2 are mutually independent, X_1 has the Poisson distribution with mean 2, and X_2 has the Poisson distribution with mean 6.

(a) Describe the ML rule for H_0 vs. H_1. Display your answer by indicating how to partition the set of possible observations, $\{(i, j) : i \geq 0, j \geq 0\}$, into two sets, Γ_0 and Γ_1, for which the decision is H_0 if $(X_1, X_2) \in \Gamma_0$ and H_1 if $(X_1, X_2) \in \Gamma_1$.

(b) Describe the MAP rule for H_0 vs. H_1, assuming the prior distribution with $\frac{\pi_0}{\pi_1} = 2$. Display your answer by indicating how to partition the set of possible observations, $\{(i, j) : i \geq 0, j \geq 0\}$, into two sets, Γ_0 and Γ_1, for which the decision is H_0 if $(X_1, X_2) \in \Gamma_0$ and H_1 if $(X_1, X_2) \in \Gamma_1$.

2.36. **[A simple hypothesis testing problem with discrete observations]**
Suppose there are two hypotheses about an observation X, with possible values in $\{-4, -3, \ldots, 3, 4\}$:

$$H_0 : X \text{ has pmf } p_0(i) = \tfrac{1}{9} \text{ for } -4 \leq i \leq 4 \qquad H_1 : X \text{ has pmf } p_1(i) = \tfrac{i^2}{60} \text{ for } -4 \leq i \leq 4.$$

(a) Describe the ML rule. Express your answer directly in terms of X in a simple way.

(b) Find $p_{false\ alarm}$ and p_{miss} for the ML rule.

(c) Find the MAP rule for priori distribution $\pi_0 = \frac{2}{3}$ and $\pi_1 = \frac{1}{3}$.

(d) Find p_e for the MAP rule found in part (c), assuming the prior used in part (c) is true.

(e) For what values of $\frac{\pi_0}{\pi_1}$ does the MAP rule always decide H_0? Assume ties are broken in favor of H_1.

2.37. [**Dissecting a vote**]
A panel of three judges has to make a yes-no decision. Each judge votes yes or no by secret ballot; each judge votes yes with some probability p, where $0 < p < 1$; the votes of the judges are mutually independent; the majority rules. Let M be the event that the decision of the panel is yes (i.e. a majority of the judges vote yes) and let A denote the event that the first judge votes yes.

(a) Express $P(M)$ in terms of p.

(b) Express $P(M|A)$ in terms of p.

(c) Express $P(A|M)$ in terms of p. Sketch your answer as a function of p. What are the limits as $p \to 0$ or as $p \to 1$? Explain why these limits make sense.

2.38. [**Field goal percentages – home vs. away**]
The Illini woman's basketball team plays some games at home and some games away. During games, the players make field goal attempts (i.e. throw the ball towards the hoop), and some of the attempts result in actual field goals (i.e. the ball goes through the hoop). Let p_h be the probability a field goal attempt at a home game is successful and p_a denote the probability a field goal attempt at an away game is successful. We assume (perhaps this is quite inaccurate) that attempts are successful independently of each other. We'd like to test the hypothesis $H_1 : p_h \neq p_a$ vs. hypothesis $H_0 : p_h = p_a$, based on actual data.[4] Specifically, the following statistics were collected from the team's website, for the games played in November and December 2011. There were five home games and nine away games.

	field goals	field goal attempts	shooting percentage
home games	119	281	42.35%
away games	212	521	40.70%

We take the two numbers of attempts, 281 and 521, as given, and not part of the random experiment.

(a) Using the methodology of Section 2.9, use the given data to calculate 95% confidence intervals for p_h and for p_a. (Note: If the two intervals you find intersect each other, then the accepted scientific methodology would be to say that there is not significant evidence in the data to reject the null hypothesis, H_0. In other words, the experiment is inconclusive.)

(b) Suppose an analysis of data similar to this one were conducted, and the two intervals calculated in part (a) did not intersect. What statement could we make in support of

[4]This hypothesis testing problem falls outside the scope of Section 2.11, because the hypotheses are *composite* hypotheses, meaning that they involve one or more unknown parameters. For example, H_0 specifies only that $p_h = p_a$, without specifying the numerical value of the probabilities. Problems like this are often faced in scientific experiments. A common methodology is based on the notion of p-value, and on certain functions of the data that are not sensitive to the parameters, such as in T tests or$F$ tests. While the details are beyond the scope of this course, this problem aims to give some insight into this common problem in scientific data analysis.

hypothesis H_1? (Hint: Refer to equation (2.11) of the notes, for which the true value p is fixed and arbitrary, and $\widehat{p} = \frac{X}{n}$ is random.)

2.39. **[Detection problem with the geometric distribution]**
The number of attempts, Y, required for a certain basketball player to make a 25 foot shot, is observed, in order to choose one of the following two hypotheses:

H_1 (outstanding player) : Y has the geometric distribution with parameter $p = 0.5$
H_0 (average player) : Y has the geometric distribution with parameter $p = 0.2$.

(a) Describe the ML decision rule. Express it as directly in terms of Y as possible.

(b) Find $p_{\text{false alarm}}$ and p_{miss} for the ML rule.

(c) Describe the MAP decision rule under the assumption that H_0 is a priori twice as likely as H_1. Express the rule as directly in terms of Y as possible.

(d) Find the average error probability, p_e, for both the ML rule and the MAP rule, using the same prior distribution given in part (c). For which rule is the average error probability smaller?

2.40. **[Hypothesis testing for independent geometrically distributed observations]**
Suppose $(L_1, \ldots, L_n)$ is a vector of independent, geometrically distributed random variables, all with the same parameter p, and there are two hypotheses about p, namely: Under hypothesis H_0, $p = 0.5$ and under hypothesis H_1, $p = 0.25$. Let $(k_1, \ldots, k_n)$ be a vector of positive integers, which represents a possible observed value for $(L_1, \ldots, L_n)$.

(a) Describe the maximum likelihood rule for deciding which hypothesis is true. Express it as simply and directly as possible for a given observed vector $(k_1, \ldots, k_n)$.

(b) Describe the MAP rule for deciding which hypothesis is true, under the assumption that the prior probabilities satisfy $\pi_0 = 8\pi_1$. Express it as simply and directly as possible for a given observed vector $(k_1, \ldots, k_n)$.

2.41. **[Testing hypotheses about a die]**
Consider a binary hypothesis testing problem based on observation of n independent rolls of a die. Let $X_1, \ldots, X_n$ denote the numbers rolled on the die. Let a denote a known constant that is slightly greater than one. The two hypotheses are:

H_0 : the die is fair

H_1 : for each roll of the die, i shows with probability $p_i = Ca^i$, where $C = \frac{1-a}{a-a^7}$, so that the probabilities sum to one.

(a) Find simple expressions for $p_i(u_1, \ldots, u_n) = P(X_1 = u_1, \ldots, X_n = u_n | H_i)$ for $i = 0$ and $i = 1$, where $u_i \in \{1, 2, 3, 4, 5, 6\}$ for each i. Express your answers using the variables t_k,

for $1 \leq k \leq 6$, where t_k is the number of the n rolls that show k. The vector $(t_1, \ldots, t_6)$ is called the type vector of $(u_1, \ldots, u_n)$. Intuitively, the order of the observations shouldn't matter, so decision rules will naturally only depend on the type vector of the observation sequence.

(b) Find a simple expression for the likelihood ratio, $\Lambda(u_1, \ldots, u_n) = \frac{p_1(u_1,\ldots,u_n)}{p_0(u_1,\ldots,u_n).}$ and describe, as simply as possible, the likelihood ratio test for H_1 vs. H_0 given the observations $(u_1, \ldots, u_n)$.

(c) In particular, suppose that $n = 100$, $(t_1, \ldots, t_6) = (18, 12, 13, 19, 18, 20)$, and $a = 1.1$. Which hypothesis does the maximum likelihood decision rule select?

Reliability Section 2.12

2.42. **[The reliability of a hierarchical backup system]**
Consider a parallel storage system composed of nine subsystems, each of which contains nine servers. Each subsystem can tolerate a single server failure, and the overall system can tolerate a single subsystem failure. Thus, in order for the overall system to fail, there has to be at least two subsystems that each have at least two server failures. Suppose servers fail independently with probability p.

(a) Find an expression for the exact probability, p_0, that a particular subsystem fails, in terms of p. Also, compute the numerical value of p_0 assuming that $p = 0.001$.

(b) Find an expression for the exact probability, p_1, that the overall system fails, in terms of p_0. Also, compute the numerical value of p_1 assuming that $p = 0.001$.

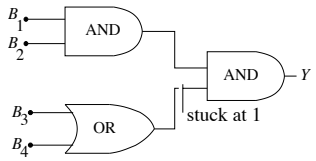
(c) Give an upper bound on p_0 and an upper bound on p_1 using the union bound, and compute their numerical values assuming that $p = 0.001$.

2.43. **[Failure of a cube]**
Suppose each corner of a three dimensional cube burns out with probability 0.001, independently of the other corners.

(a) Find a numerical upper bound, using the union bound, on the probability there exists two neighboring corners that both burn out. Explain your answer.

(b) Find a numerical upper bound, using the union bound, on the probability that two or more corners of the cube both burn out. It doesn't matter whether the burned out corners are neighbors. Explain your answer.

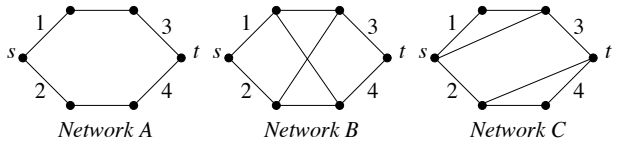
2.44. **[Fault detection in a Boolean circuit]**
Consider the Boolean circuit shown, with two AND gates, one OR gate, four binary input variables $B_1, \ldots, B_4$, and binary output variable Y.



(a) Suppose there is a *stuck at one* fault as shown, so that the value one is always fed into the second AND gate, instead of the output of the OR gate. Assuming that $B_1, \ldots, B_4$ are independent and equally likely to be zero or one, what is the probability that the output value Y is incorrect?

(b) Suppose that the circuit is working correctly with probability 0.5, or has the indicated stuck at one fault with probability 0.5. Suppose three distinct randomly generated test patterns are applied to the circuit. (Here, a test pattern is a binary sequence of length four. Assume all sets of three distinct test patterns are equally likely.) Given that the output is correct on all three of the patterns, what is the conditional probability the circuit is faulty?

2.45. **[Reliability of three $s-t$ networks]**
Consider $s-t$ networks A through C shown.



For each network, suppose that each link fails with probability p, independently of the other links, and suppose network outage occurs if at least one link fails on each path from s to t. Not all the links are labeled–for example, Network A has six links. For simplicity, we suppose that links can only be used in the forward direction, so that the paths with five links for Network B do not count;. Let F denote the event of network outage.

(a) Find $P(F)$ in terms of p for Network A, and give the numerical value of $P(F)$ for $p = 0.001$, accurate to within four significant digits.

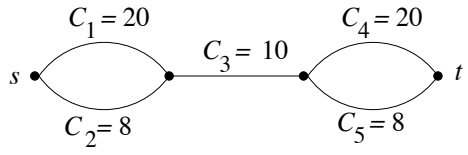
(b) This part aims to find $P(F)$ for Network B, using the law of total probability, based on the partition of the sample space into the events: $D_0, D_1, D_{2,s}, D_{2,d}, D_3, D_4$. Here, D_i is the event that exactly i links from among links $\{1,2,3,4\}$ fail, for $i = 0,1,3,4$; $D_{2,s}$ is the event that exactly two links from among links $\{1,2,3,4\}$ fail and they are on the same side (i.e. either links 1 and 2 fail or links 3 and 4 fail); $D_{2,d}$ is the event that exactly two links from among links $\{1,2,3,4\}$ fail and they are on different sides. Find the probability of each of these events, find the conditional probabilities of F given any one of these events, and finally, find $P(F)$. Express your answers as a function of p, and give the numerical value of $P(F)$ for $p = 0.001$, accurate to within four significant digits.

(c) Find the numerical value of $P(D_{2,d}|F)$ for $p = 0.001$.

(d) Find the limit of the ratio of the outage probability for Network B to the outage probability for Network A, as $p \to 0$. Explain the limit. Is it zero? Hint: For each network, $P(F)$ is a polynomial in p. As $p \to 0$, the term with the smallest power of p dominates.

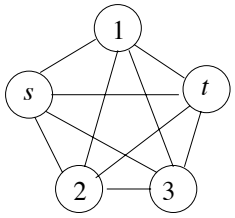
(e) Without doing any detailed calculations, based on the reasoning used in part (d), give the limit of the ratio of the outage probability for Network C to the outage probability for Network A, as $p \to 0$. Explain the limit. Is it zero?

2.46. **[Distribution of capacity of an $s-t$ flow network]**
Consider the following $s-t$ flow network. The link capacities, in units of some quantity per unit time, are shown for links that do not fail. Suppose each link fails with some probability p and if a link fails it can carry no flow. Let Y denote the $s-t$ capacity of the network.



Find the pmf of Y. Express your answer in terms of p. Te facilitate grading, express each nonzero term of the pmf as a polynomial in p, with terms arranged in increasing powers of p. (Hint: What is $P\{Y > 0\}$?)

2.47. **[Reliability of a fully connected $s-t$ network with five nodes]**
Consider the $s-t$ network shown.



Suppose each link fails independently with probability $p = 0.001$. The network is said to fail if at least one link fails along each path from s to t.

(a) Identify the minimum number of link failures, k^*, that can cause the network to fail.

(b) There are $\binom{10}{k^*}$ sets of k^* links. So by the union bound, the probability of network failure is bounded above: $P(F) \leq \binom{10}{k^*} p^{k^*}$. Compute the numerical value of this bound.

(c) A blocking set is a set of links such that if every link in the set fails then the network fails. By definition, k^* is the minimum number of links in a blocking set. Show that there are exactly two blocking sets with k^* links.

(d) Using the result of part (c), derive a tighter upper bound on the probability of system failure, and give its numerical value in case $p = 0.001$.

Chapter 3

Continuous-type random variables

Chapter 2 dealt largely with discrete-type random variables, and finite collections of events. Much of this chapter will involve continuous-type random variables, which have distributions described by density functions, rather than by mass functions. The relationship of mass functions to density functions is analogous to the relationship of peanuts to peanut butter. Whereas peanuts have mass in discrete quantities, peanut butter has similar mass that can be spread out. A general, although somewhat complicated, way to describe the distribution of any random variable is through the use of a cumulative distribution function, as described in the next section. Cumulative distribution functions form a natural bridge between discrete-type and continuous-type random variables.

3.1 Cumulative distribution functions

Let a probability space $(\Omega, \mathcal{F}, P)$ be given. Recall that in Chapter 2 we defined a random variable to be a function X from Ω to the real line $\mathbb{R}$. To be on mathematically firm ground, random variables are also required to have the property that sets of the form $\{\omega : X(\omega) \leq c\}$ should be events–meaning that they should be in $\mathcal{F}$. Since a probability measure P assigns a probability to every event, every random variable X has a *cumulative distribution function* (CDF), denoted by F_X. It is the function, with domain the real line $\mathbb{R}$, defined by

$$\begin{aligned} F_X(c) &= P\{\omega : X(\omega) \leq c\} \\ &= P\{X \leq c\} \text{ (for short).} \end{aligned}$$

Example 3.1.1 Let X denote the number showing for a roll of a fair die, so the pmf of X is $p_X(i) = \frac{1}{6}$ for integers i with $1 \leq i \leq 6$. The CDF F_X is shown in Figure 3.1.

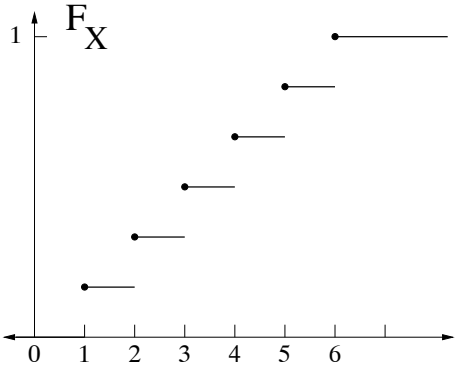


Figure 3.1: CDF for the roll of a fair die.

The notions of left limit and right limit are useful for discussing functions with jumps, such as CDFs. Given a function F on the real line and a value x, the *left limit* of F at x, denoted by $F(x-)$, is the limit of $F(y)$ as y converges to x from the left. Similarly, the *right limit* of F at x, denoted by $F(x+)$, is the limit of $F(y)$ as y converges to x from the right. That is,

$$F(x-) = \lim_{\substack{y \to x \\ y < x}} F(y) \qquad F(x+) = \lim_{\substack{y \to x \\ y > x}} F(y).$$

Note that the CDF in Example 3.1.1 has six jumps of size 1/6. The jumps are located at the six possible values of X, namely at the integers one through six, and the size of each of those jumps is 1/6, which is the probability assigned to each of the six possible values. The value of the CDF exactly at a jump point is equal to the right limit at that point. For example, $F_X(1) = F_X(1+) = 1/6$ and $F_X(1-) = 0$. The size of the jump at any x can be written as

$$\triangle F_X(x) = F_X(x) - F_X(x-).$$

The CDF of an arbitrary random variable X determines the probabilities of any events of the form $\{X \in A\}$. Of course, the CDF of a random variable X determines $P\{X \leq c\}$ for any real number c–by definition it is just $F_X(c)$. Similarly, $P\{X \in (a, b]\} = F_X(b) - F_X(a)$, whenever $a < b$. The next proposition explains how F_X also determines probabilities of the form $P\{X < c\}$ and $P\{X = c\}$.

Proposition 3.1.2 *Let X be a random variable and let c be any real number. Then $P\{X < c\} = F_X(c-)$ and $P\{X = c\} = \triangle F_X(c)$, where F_X is the CDF of X.*

Proof. Fix c and let $c_1, c_2, \ldots$ be a sequence with $c_1 < c_2 < \ldots$ such that $\lim_{j\to\infty} c_j = c$. Let G_1 be the event $G_1 = \{X \leq c_1\}$ and for $j \geq 2$ let $G_j = \{c_{j-1} < X \leq c_j\}$. Then for any $n \geq 1$, $\{X \leq c_n\} = G_1 \cup G_2 \cup \cdots \cup G_n$. Also, $\{X < c\} = G_1 \cup G_2 \cup \cdots$ and the events $G_1, G_2, \ldots$ are mutually exclusive. Therefore, by Axiom P.2,

$$P\{X < c\} = P(G_1) + P(G_2) + \cdots .$$

The sum of a series is, by definition, the limit of the sum of the first n terms as $n \to \infty$. Therefore,

$$P\{X < c\} = \lim_{n\to\infty} \sum_{k=1}^{n} P(G_k) = \lim_{n\to\infty} P\{G_1 \cup G_2 \cup \cdots \cup G_n\} = \lim_{n\to\infty} P\{X \le c_n\} = \lim_{n\to\infty} F_X(c_n).$$

That is, $P\{X < c\} = F_X(c-)$. The second conclusion of the proposition follows directly from the first: $P\{X = c\} = P\{X \le c\} - P\{X < c\} = F_X(c) - F_X(c-) = \triangle F_X(c)$. ■

Example 3.1.3 Let X have the CDF shown in Figure 3.2.
(a) Determine all values of u such that $P\{X = u\} > 0$. (b) Find $P\{X \le 0\}$. (c) Find $P\{X < 0\}$.

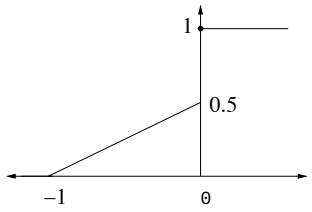


Figure 3.2: An example of a CDF.

Solution: (a) The CDF has only one jump, namely a jump of size 0.5 at $u = 0$. Thus, $P\{X = 0\} = 0.5$, and there are no values of u with $u \neq 0$ such that $P\{X = u\} > 0$.
(b) $P\{X \le 0\} = F_X(0) = 1$. (c) $P\{X < 0\} = F_X(0-) = 0.5$.

Example 3.1.4 Let X have the CDF shown in Figure 3.3. Find the numerical values of the following quantities:
(a) $P\{X \le 1\}$, (b) $P\{X \le 10\}$, (c) $P\{X \ge 10\}$, (d) $P\{X = 10\}$, (e) $P\{|X - 5| \le 0.1\}$.

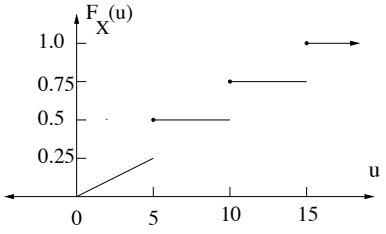


Figure 3.3: An example of a CDF.

Solution: (a) $P\{X \leq 1\} = F_X(1) = 0.05$.
(b) $P\{X \leq 10\} = F_X(10) = 0.75$.
(c) $P\{X \geq 10\} = 1 - P\{X < 10\} = 1 - F_X(10-) = 0.5$.
(d) $P\{X = 10\} = \triangle F_X(10) = 0.25$.
(e) $P\{|X - 5| \leq 0.1\} = P\{4.9 \leq X \leq 5.1\} = P\{X \leq 5.1\} - P\{X < 4.9\} = F_X(5.1) - F_X(4.9-) = 0.5 - 0.245 = 0.255$.

The following proposition follows from the axioms of probability and the definition of CDFs. The proof is omitted, but a proof of the only if part can be given along the lines of the proof of Proposition 3.1.2.

Proposition 3.1.5 *A function F is the CDF of some random variable if and only if it has the following three properties:*

F.1 *F is nondecreasing*

F.2 $\lim_{c\to+\infty} F(c) = 1$ *and* $\lim_{c\to-\infty} F(c) = 0$

F.3 *F is right continuous (i.e. $F_X(c) = F_X(c+)$ for all c).*

Example 3.1.6 Which of the six functions shown in Figure 3.4 are valid CDFs? For each one that is not valid, state a property from Proposition 3.1.5 that is violated.

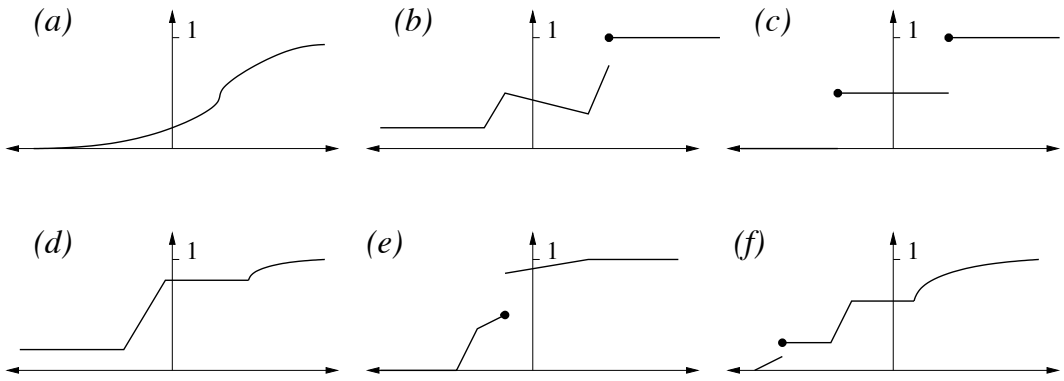


Figure 3.4: Six candidate CDFs.

Solution: The functions shown in plots (a), (c), and (f) are valid CDFs and the other three are not. The function in (b) is not nondecreasing, and it does not converge to zero at $-\infty$. The function in (d) does not converge to zero at $-\infty$. The function in (e) is not right continuous.

The vast majority of random variables described in applications are one of two types, to be described next. A random variable X is a *discrete-type* random variable if there is a finite or

countably infinite set of values $\{u_i : i \in I\}$ such that $P\{X \in \{u_i : i \in I\}\} = 1$. The probability mass function (pmf) of a discrete-type random variable X, denoted $p_X(u)$, is defined by $p_X(u) = P\{X = u\}$. Typically the pmf of a discrete random variable is much more useful than the CDF. However, the pmf and CDF of a discrete-type random variable are related by $p_X(u) = \triangle F_X(u)$ and conversely,

$$F_X(c) = \sum_{u:u\leq c} p_X(u), \tag{3.1}$$

where the sum in (3.1) is taken only over u such that $p_X(u) \neq 0$. If X is a discrete-type random variable with only finitely many mass points in any finite interval, then F_X is a piecewise constant function.

A random variable X is a *continuous-type* random variable if the CDF is the integral of a function:

$$F_X(c) = \int_{-\infty}^{c} f_X(u)du.$$

The function f_X is called the *probability density function*. Continuous-type random variables are the subject of the next section.

The relationship among CDFs, pmfs, and pdfs is summarized in Figure 3.5. Any random vari-

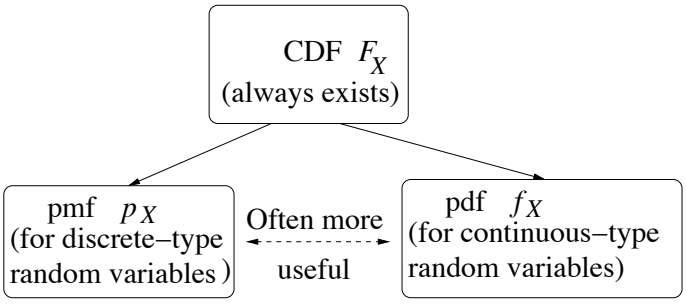


Figure 3.5: Ways to describe distributions of random variables

able X on an arbitrary probability space has a CDF F_X. The D in "CDF" stands for "distribution." But in common usage, the response to the question "What is the distribution of X?" is answered by giving one or more of F_X, p_X, or f_X, or possibly a transform of one of these, whichever is most convenient. Any of these have equivalent information and specify the probability of all events involving X alone, but in a given application often one way to express the distribution is more useful than another.

3.2 Continuous-type random variables

Definition 3.2.1 *A random variable X is a continuous-type random variable if there is a function f_X, called the probability density function (pdf) of X, such that*

$$F_X(c) = \int_{-\infty}^{c} f_X(u)du,$$

for all $c \in \mathbb{R}$. The support *of a pdf f_X is the set of u such that $f_X(u) > 0$.*

It follows by the fundamental theorem of calculus that if X is a continuous-type random variable and if the pdf f_X is continuous, then the pdf is the derivative of the CDF: $f_X = F_X'$. In particular, if X is a continuous-type random variable with a continuous pdf, F_X is differentiable, and therefore F_X is a continuous function.[1] That is, there are no jumps in F_X, so for any constant v, $P\{X = v\} = 0$.

It may seem strange at first that $P\{X = v\} = 0$ for all numbers v, because if we add these probabilities up over all v, we get zero. It would seem like X can't take on any real value. But remember that there are uncountably many real numbers, and the axiom that probability is additive, Axiom P.2, only holds for countably infinite sums.

If $a < b$ then

$$P\{a < X \leq b\} = F_X(b) - F_X(a) = \int_a^b f_X(u)du.$$

Since $P\{X = a\} = P\{X = b\} = 0$, it follows more generally that:

$$P\{a < X \leq b\} = P\{a < X < b\} = P\{a \leq X \leq b\} = P\{a \leq X < b\} = \int_a^b f_X(u)du.$$

So when we work with continuous-type random variables, we don't have to be precise about whether the endpoints of intervals are included when calculating probabilities.

It follows that the integral of f_X over every interval (a, b) is greater than or equal to zero, so f_X must be a nonnegative function. Also,

$$1 = \lim_{a \to -\infty} \lim_{b \to +\infty} F_X(b) - F_X(a) = \int_{-\infty}^{\infty} f_X(u)du.$$

Therefore, f_X integrates to one. In most applications, the density functions f_X are continuous, or piecewise continuous.

Although $P\{X = u\} = 0$ for any real value of u, there is still a fairly direct interpretation of f_X involving probabilities. Suppose u_o is a constant such that f_X is continuous at u_o. Since $f_X(u_o)$ is the derivative of F_X at u_o, it is also the symmetric derivative of F_X at u_o, because taking $h \to 0$ on each side of

$$\frac{F_X(u_o + h) - F_X(u_o - h)}{2h} = \frac{1}{2}\left(\frac{F_X(u_o + h) - F_X(u_o)}{h} + \frac{F_X(u_o - h) - F_X(u_o)}{-h}\right)$$

[1] In these notes we only consider pdf's f_X that are at least piecewise continuous, so that $f_X(u) = F_X'(u)$ except at discontinuity points of f_X. The CDF F_X is continuous in general for continuous-type random variables.

yields

$$\lim_{h\to 0} \frac{F_X(u_o+h) - F_X(u_o-h)}{2h} = \frac{1}{2}\left(f_X(u_o) + f_X(u_o)\right) = f_X(u_o). \tag{3.2}$$

Substituting $h = \epsilon/2$ and considering $\epsilon > 0$ only, (3.2) yields

$$f_X(u_o) = \lim_{\epsilon\to 0} \frac{P\{u_o - \frac{\epsilon}{2} < X < u_o + \frac{\epsilon}{2}\}}{\epsilon}, \tag{3.3}$$

or equivalently,

$$P\left\{u_o - \frac{\epsilon}{2} < X < u_o + \frac{\epsilon}{2}\right\} = \epsilon f_X(u_o) + o(\epsilon), \tag{3.4}$$

where $o(\epsilon)$ represents a term such that $\lim_{\epsilon\to 0} \frac{o(\epsilon)}{\epsilon} = 0$. The equivalent equations (3.3) and (3.4) show how the pdf f_X is directly related to probabilities of events involving X.

Many of the definitions and properties for discrete-type random variables carry over to continuous-type random variables, with summation replaced by integration. The mean (or expectation), $E[X]$, of a continuous-type random variable X is defined by:

$$\mu_X = E[X] = \int_{-\infty}^{\infty} u f_X(u) du.$$

The Law of the Unconscious Statistician (LOTUS) holds and is the fact that for a function g:

$$E[g(X)] = \int_{-\infty}^{\infty} g(u) f_X(u) du.$$

It follows from LOTUS, just as in the case of discrete-type random variables, that expectation is a linear operation. For example, $E[aX^2 + bX + c] = aE[X^2] + bE[X] + c$.

Variance is defined for continuous-type random variables exactly as it is for discrete-type random variables: $\mathrm{Var}(X) = E[(X-\mu_X)^2]$, and it has the same properties. It is a measure of how spread out the distribution of X is. As before, the variance of X is often denoted by σ_X^2, where $\sigma_X = \sqrt{\mathrm{Var}(X)}$ is called the standard deviation of X. If X is a measurement in some units, then σ_X is in the same units. For example. if X represents a measurement in feet, then $\mathrm{Var}(X)$ represents a number of feet^2 and σ_X represents a number of feet. Exactly as shown in Section 2.2 for discrete-type random variables, the variance for continuous-type random variables scales as $\mathrm{Var}(aX + b) = a^2\mathrm{Var}(X)$. The standardized random variable, $\frac{X-\mu_X}{\sigma_X}$, is a dimensionless random variable with mean zero and variance one.

Finally, as before, the variance of a random variable is equal to its second moment minus the square of its first moment:

$$\begin{aligned} \mathrm{Var}(X) &= E[X^2 - 2\mu_X X + \mu_X^2] \\ &= E[X^2] - 2\mu_X E[X] + \mu_X^2 \\ &= E[X^2] - \mu_X^2. \end{aligned}$$

Example 3.2.2 Suppose X has the following pdf, where A is a constant to be determined:

$$f_X(u) = \begin{cases} A(1-u^2) & -1 \le u \le 1 \\ 0 & \text{else.} \end{cases}$$

Find A, $P\{0.5 < X < 1.5\}$, F_X, μ_X, $\mathrm{Var}(X)$, and σ_X.

Solution: To find A we require that

$$\begin{aligned} 1 &= \int_{-\infty}^{\infty} f_X(u)du \\ &= \int_{-1}^{1} A(1-u^2)du \\ &= A\left(u - \frac{u^3}{3}\right)\Big|_{-1}^{1} = \frac{4A}{3}, \end{aligned}$$

so $A = \frac{3}{4}$.

The support of f_X (the set on which it is not zero) is the interval $[-1, 1]$. To find $P\{0.5 < X < 1.5\}$, we only need to integrate over the portion of the interval $[0.5, 1.5]$ in the support of f_X. That is, we only need to integrate over $[0.5, 1]$:

$$\begin{aligned} P\{0.5 < X < 1.5\} &= \int_{0.5}^{1.5} f_X(u)du \\ &= \int_{0.5}^{1} \frac{3(1-u^2)}{4}du \\ &= \frac{3}{4}\left(u - \frac{u^3}{3}\right)\Big|_{0.5}^{1} = \frac{3}{4}\left\{\frac{2}{3} - \frac{11}{24}\right\} = \frac{5}{32}. \end{aligned}$$

Because the support of f_X is the interval $[-1, 1]$, we can immediately write down the partial answer:

$$F_X(c) = \begin{cases} 0 & c \le -1 \\ ? & -1 < c < 1 \\ 1 & c \ge 1, \end{cases}$$

where the question mark represents what hasn't yet been determined, namely, the value of $F_X(c)$ for $-1 < c < 1$. So let $-1 < c < 1$. Then

$$\begin{aligned} F_X(c) &= P\{X \le c\} \\ &= \int_{-1}^{c} \frac{3(1-u^2)}{4}du \\ &= \frac{3}{4}\left(u - \frac{u^3}{3}\right)\Big|_{-1}^{c} \\ &= \frac{2 + 3c - c^3}{4}. \end{aligned}$$

This allows us to give the complete expression for F_X :

$$F_X(c) = \begin{cases} 0 & c \le -1 \\ \frac{2+3c-c^3}{4} & -1 < c < 1 \\ 1 & c \ge 1. \end{cases}$$

The mean, μ_X, is zero, because $\mu_X = \int_{-\infty}^{\infty} u f_X(u) du$, and $u f_X(u)$ is an odd function so its integral over $\mathbb{R}$ is zero. (For example, $u f_X(u)\Big|_{0.5} = \frac{9}{32}$ and $u f_X(u)\Big|_{-0.5} = -\frac{9}{32}$.)

Therefore, $\mathrm{Var}(X) = E[X^2] - \mu_X^2 = E[X^2]$, so we find

$$\begin{aligned} E[X^2] &= \int_{-\infty}^{\infty} u^2 f_X(u) du \\ &= \int_{-1}^{1} u^2 \frac{3}{4}(1-u^2) du \\ &= \frac{3}{4}\int_{-1}^{1} (u^2 - u^4) du = 0.2. \end{aligned}$$

Thus, $\mathrm{Var}(X) = 0.2$ and $\sigma_X = \sqrt{\mathrm{Var}(X)} = \sqrt{0.2} \approx 0.45$.

3.3 Uniform distribution

Let $a < b$. A random variable X is *uniformly distributed* over the interval $[a, b]$ if

$$f_X(u) = \begin{cases} \frac{1}{b-a} & a \le u \le b \\ 0 & \text{else.} \end{cases}$$

The mean of X is given by

$$E[X] = \frac{1}{b-a}\int_a^b u du = \frac{u^2}{2(b-a)}\Big|_a^b = \frac{b^2-a^2}{2(b-a)} = \frac{(b-a)(b+a)}{2(b-a)} = \frac{a+b}{2}.$$

Thus, the mean is the midpoint of the interval. The second moment of X is given by

$$E[X^2] = \frac{1}{b-a}\int_a^b u^2 du = \frac{u^3}{3(b-a)}\Big|_a^b = \frac{b^3-a^3}{3(b-a)} = \frac{(b-a)(b^2+ab+a^2)}{3(b-a)} = \frac{a^2+ab+b^2}{3}.$$

So

$$\mathrm{Var}(X) = E[X^2] - E[X]^2 = \frac{a^2+ab+b^2}{3} - \left(\frac{a+b}{2}\right)^2 = \frac{(a-b)^2}{12}.$$

Note that the variance is proportional to the length of the interval squared. A useful special case is when $a = 0$ and $b = 1$, in which case X is uniformly distributed over the unit interval $[0, 1]$. In that case, for any $k \ge 0$, the kth moment is given by

$$E[X^k] = \int_0^1 u^k du = \frac{u^{k+1}}{k+1}\Big|_0^1 = \frac{1}{k+1} \quad \text{(if } U \text{ is uniformly distributed over } [0,1]),$$

and the variance is $\frac{1}{3} - (\frac{1}{2})^2 = \frac{1}{12}$.

3.4 Exponential distribution

A random variable T has the exponential distribution with parameter $\lambda > 0$ if its pdf is given by

$$f_T(t) = \begin{cases} \lambda e^{-\lambda t} & t \geq 0 \\ 0 & \text{else.} \end{cases}$$

The pdfs for the exponential distributions with parameters $\lambda = 1, 2$, and 4 are shown in Figure 3.6. Note that the initial value of the pdf is given by $f_T(0) = \lambda$, and the rate of decay of $f_T(t)$ as

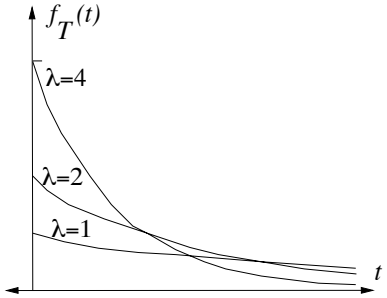


Figure 3.6: The pdfs for the exponential distributions with $\lambda = 1, 2$, and 4.

t increases is also λ. Of course, the area under the graph of the pdf is one for any value of λ. The mean is *decreasing* in λ, because the larger λ is, the more concentrated the pdf becomes towards the left.

The CDF evaluated at a $t \geq 0$, is given by

$$F_T(t) = \int_{-\infty}^{t} f_T(s)ds = \int_0^t \lambda e^{-\lambda s} ds = -e^{-\lambda s}\Big|_0^t = 1 - e^{-\lambda t}.$$

Therefore, in general,

$$F_T(t) = \begin{cases} 1 - e^{-\lambda t} & t \geq 0 \\ 0 & t < 0. \end{cases}$$

The complementary CDF, defined by $F_T^c(t) = P\{T > t\} = 1 - F_T(t)$, therefore satisfies:

$$F_T^c(t) = \begin{cases} e^{-\lambda t} & t \geq 0 \\ 1 & t < 0. \end{cases}$$

To find the first and second moments we use integration by parts:

$$\begin{aligned} E[T] &= \int_0^\infty t\lambda e^{-\lambda t} dt \\ &= -te^{-\lambda t}\Big|_0^\infty + \int_0^\infty e^{-\lambda t} dt \\ &= 0 + \frac{1}{\lambda} = \frac{1}{\lambda}. \end{aligned}$$

$$\begin{aligned} E[T^2] &= \int_0^\infty t^2 \lambda e^{-\lambda t} dt \\ &= -t^2 e^{-\lambda t}\Big|_0^\infty + \int_0^\infty 2te^{-\lambda t} dt \\ &= 0 + \frac{2}{\lambda^2} = \frac{2}{\lambda^2}. \end{aligned}$$

Therefore, $\text{Var}(T) = E[T^2] - E[T]^2 = \frac{2}{\lambda^2} - \frac{1}{\lambda^2} = \frac{1}{\lambda^2}$. The standard deviation of T is $\sigma_T = \frac{1}{\lambda}$. The standard deviation is equal to the mean. (Sound familiar? Recall that for p close to zero, the geometric distribution with parameter p has standard deviation nearly equal to the mean.)

Example 3.4.1 Let T be an exponentially distributed random variable with parameter $\lambda = \ln 2$. Find the simplest expression possible for $P\{T \geq t\}$ as a function of t for $t \geq 0$, and find $P(T \leq 1 | T \leq 2)$.

Solution. $P\{T \geq t\} = F_T^c(t) = e^{-\lambda t} = e^{-(\ln 2)t} = 2^{-t}$, and
$P(T \leq 1 | T \leq 2) = \frac{P\{T \leq 1, T \leq 2\}}{P\{T \leq 2\}} = \frac{P\{T \leq 1\}}{P\{T \leq 2\}} = \frac{1 - \frac{1}{2}}{1 - \frac{1}{4}} = \frac{2}{3}$.

Memoryless property of exponential distribution Suppose T is an exponentially distributed random variable with some parameter λ. Then $P\{T > t\} = e^{-\lambda t}$. It follows that

$$\begin{aligned} P(T > s + t | T > s) &= \frac{P\{T > s + t, T > s\}}{P\{T > s\}} \\ &= \frac{P\{T > s + t\}}{P\{T > s\}} \\ &= \frac{e^{-\lambda(s+t)}}{e^{-\lambda s}} \\ &= e^{-\lambda t} = P\{T > t\}. \end{aligned}$$

That is, $P(T > s + t | T > s) = P\{T > t\}$. This is called the *memoryless property* for continuous time. If T is the lifetime of a component installed in a system at time zero, the memoryless property of T has the following interpretation: Given that the component is still working after s time units, the probability it will continue to be working after t additional time units, is the same as the probability a new component would still be working after t time units. As discussed in Section 3.9, the memoryless property is equivalent to the failure rate being constant.

Connection between exponential and geometric distributions As just noted, the exponential distribution has the memoryless property in continuous time. Recall from Section 2.5 that the geometric distribution has the memoryless property in discrete time. We shall further illustrate the close connection between the exponential and geometric distributions by showing that the exponential distribution is the limit of scaled geometric distributions. In essence, the exponential

distribution is the continuous time analog of the geometric distribution. When systems are modeled or simulated, it is useful to be able to approximate continuous variables by discrete ones, which are easily represented in digital computers.

Fix $\lambda > 0$, which should be thought of as a failure rate, measured in inverse seconds. Let $h > 0$ represent a small duration of time, measured in seconds. Imagine there is a clock that ticks once every h time units. Thus, the clock ticks occur at times $h, 2h, 3h, \ldots$. If we measure how long a lightbulb lasts by the number of clock ticks, the smaller h is the larger the number of ticks. We can model the lifetime of a lightbulb as a discrete random variable if we assume the bulb can only fail at the times of the clock ticks. For small values of h this can give a good approximation to a continuous type random variable. For small values of h, the probability p that the lightbulb fails between consecutive clock ticks should be proportionally small. Specifically, set $p = \lambda h$. Consider a lightbulb which is new at time zero, and at each clock tick it fails with probability p, given it hasn't failed earlier. Let L_h be the number of ticks until the lightbulb fails. Then L_h has the geometric distribution with parameter p. The mean of L_h is given by $E[L_h] = \frac{1}{p} = \frac{1}{\lambda h}$, which converges to infinity as $h \to 0$. That's what happens when we measure the lifetime by the number of clock ticks using a clock with a very high tick rate. Let T_h be the amount of time until the lightbulb fails. Then $T_h = hL_h$. We shall show that the distribution of T_h is close to the exponential distribution with parameter λ for small h.

We have a handle on the distribution of T_h because we know the distribution of L_h. The complementary CDF of T_h can be found as follows. For any $c \geq 0$,

$$\begin{aligned} P\{T_h > c\} &= P\{L_h h > c\} \\ &= P\{L_h > \lfloor c/h \rfloor\} \\ &= (1 - h\lambda)^{\lfloor c/h \rfloor}, \end{aligned}$$

where the last equality follows from (2.8). Recall from (2.9) that $\left(1 - \frac{\lambda}{n}\right)^n \to e^{-\lambda}$ as $n \to \infty$. Replacing n by $1/h$ implies that $(1 - h\lambda)^{1/h} \to e^{-\lambda}$ as $h \to 0$, and therefore, $(1 - h\lambda)^{c/h} \to e^{-\lambda c}$ as $h \to 0$. Also, the difference between $\lfloor c/h \rfloor$ and c/h is not important, because $1 \geq (1 - h\lambda)^{c/h}/(1 - h\lambda)^{\lfloor c/h \rfloor} \geq (1 - h\lambda) \to 1$ as $h \to 0$. Therefore, if T has the exponential distribution with parameter λ,

$$P\{T_h > c\} = (1 - h\lambda)^{\lfloor c/h \rfloor} \to e^{-\lambda c} = P\{T > c\}$$

as $h \to 0$. So $1 - P\{T_h > c\} \to 1 - P\{T > c\}$ as well. That is, the CDF of T_h converges to the CDF of T. In summary, the CDF of h times a geometrically distributed random variable with parameter $p = \lambda h$ converges to the CDF of an exponential random variable with parameter λ, as $h \to 0$.

3.5 Poisson processes

Bernoulli processes are discussed in Section 2.6. Here we examine Poisson processes. Just as exponential random variables are limits of scaled geometric random variables (as seen in Section 3.4), Poisson processes are limits of scaled Bernoulli processes.

3.5.1 Time-scaled Bernoulli processes

Let $X_1, X_2, \ldots$ form a Bernoulli process with parameter p, with $0 \leq p \leq 1$. As discussed in Section 2.6, this means that the random variables are independent, and $P\{X_k = 1\} = p$ and $P\{X_k = 0\} = 1 - p$. We say that the kth trial results in a *count* if $X_k = 1$. Let $h > 0$, with h representing an amount of time. Suppose each trial takes h time units to perform. A time-scaled Bernoulli random process tracks the number of counts versus time. See Figure 3.7. Figure 3.7(a) shows a

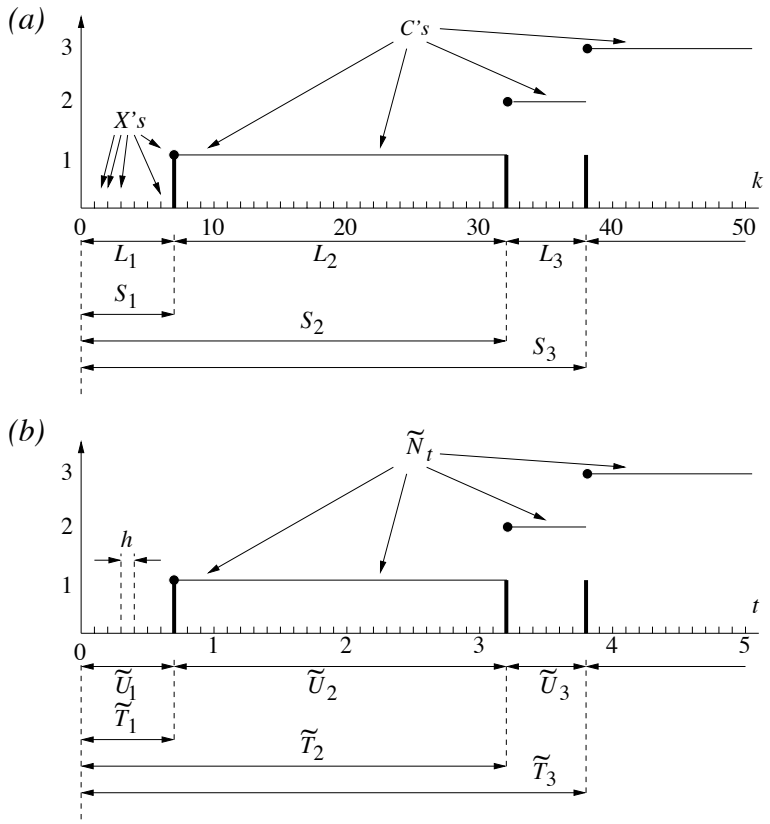


Figure 3.7: (a) A sample path of a Bernoulli process and (b) the associated time-scaled sample path of the time-scaled Bernoulli process, for $h = 0.1$.

Bernoulli process, along with the associated random variables defined in Section 2.6: the number of additional trials needed for each additional outcome of a one (the L's), the total number of trials needed for a given number of ones (the S's), and the cumulative number of ones for a given number of trials (the C's). Note that the index in the sketch is k, which indexes the trials, beginning with trial one. Figure 3.7(b) shows the corresponding time-scaled Bernoulli process for $h = 0.1$. The seventh trial is the first trial to result in a count, so the time of the first count is $7h$, as shown in 3.7(b) for $h = 0.1$. We define the following random variables to describe the time-scaled Bernoulli process with time step h:

- $\widetilde{U}_j = hL_j$: the amount of time between the $j-1^{th}$ count and the jth count
- $\widetilde{T}_j = hS_j$: the time the jth count occurs
- $\widetilde{N}_t = C_{\lfloor t/h \rfloor}$: the number of counts up to time t

The tilde's on the random variables here are used to distinguish the variables from the similar random variables for a Poisson process, defined below.

Suppose λ is fixed and that h is so small that $p = \lambda h$ is much smaller than one. Then the random variables describing the scaled Bernoulli process have simple approximate distributions. Each L_j is a geometrically distributed random variable with parameter p, so as explained in Section 3.4, the scaled version of L_j, namely $\widetilde{U}_j = hL_j$, approximately has the exponential distribution with parameter $\lambda = p/h$. For t fixed, $\widetilde{N}_t$ is the sum of $\lfloor t/h \rfloor$ Bernoulli random variables with parameter p. Therefore, $\widetilde{N}_t$ has the binomial distribution with parameters $\lfloor t/h \rfloor$ and $p = \lambda h$. So $E[\widetilde{N}_t] = \lfloor t/h \rfloor h\lambda \approx \lambda t$. Recall from Section 2.7 that the limit of a binomial distribution as $n \to \infty$ and $p \to 0$ with $np \to \lambda$ is the Poisson distribution with parameter λ. Therefore, as $h \to 0$, the limiting distribution of $\widetilde{N}_t$ is the Poisson distribution with mean λt. More generally, if $0 \le s < t$, the distribution of the increment $\widetilde{N}_t - \widetilde{N}_s$ converges to the Poisson distribution with parameter $(t-s)\lambda$. Also, the increments of $\widetilde{N}_t$ over disjoint intervals are independent random variables.

3.5.2 Definition and properties of Poisson processes

A Poisson process with rate $\lambda > 0$ is obtained as the limit of scaled Bernoulli random counting processes as $h \to 0$ and $p \to 0$ such that $p/h \to \lambda$. This limiting picture is just used to motivate the definition of Poisson processes, given below, and to explain why Poisson processes naturally arise in applications. A sample path of a Poisson process (i.e. the function of time the process yields for some particular ω in Ω) is shown in Figure 3.8. The variable N_t for each $t \ge 0$ is the cumulative

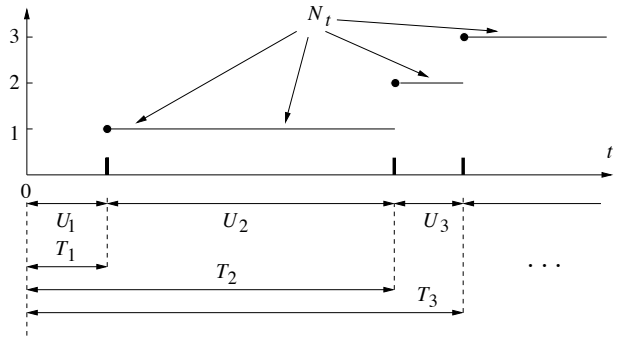


Figure 3.8: A sample path of a Poisson process.

number of counts up to time t. The random variables T_1, T_2, ... are called the *count times* and the

random variables U_1, U_2, ... are called the *intercount times*. The following equations clearly hold:

$$\begin{aligned} N_t &= \sum_{n=1}^{\infty} I_{\{t \geq T_n\}} \\ T_n &= \min\{t : N_t \geq n\} \\ T_n &= U_1 + \cdots + U_n. \end{aligned}$$

Definition 3.5.1 *Let $\lambda \geq 0$. A Poisson process with rate λ is a random counting process $N = (N_t : t \geq 0)$ such that*

N.1 N has independent increments: if $0 \leq t_0 \leq t_1 \leq \cdots \leq t_n$. the increments $N_{t_1} - N_{t_0}, N_{t_2} - N_{t_1}, \ldots, N_{t_n} - N_{t_{n-1}}$ are independent.

N.2 The increment $N_t - N_s$ has the $Poi(\lambda(t-s))$ distribution for $t \geq s$.

Proposition 3.5.2 *Let N be a random counting process and let $\lambda > 0$. The following are equivalent:*

(a) N is a Poisson process with rate λ.

(b) The intercount times U_1, $U_2, \ldots$ are mutually independent, exponentially distributed random variables with parameter λ.

Proof. Either (a) or (b) provides a specific probabilistic description of a random counting process. Furthermore, both descriptions carry over as limits from the Bernoulli random counting process, as $h, p \to 0$ with $p/h \to \lambda$. This provides a basis for a proof of the proposition. ■

Example 3.5.3 Consider a Poisson process on the interval $[0, T]$ with rate $\lambda > 0$, and let $0 < \tau < T$. Define X_1 to be the number of counts during $[0, \tau]$, X_2 to be the number of counts during $[\tau, T]$, and X to be the total number of counts during $[0, T]$. Let i, j, n be nonnegative integers such that $n = i + j$. Express the following probabilities in terms of n, i, j, τ, T, and λ, simplifying your answers as much as possible:
(a) $P\{X = n\}$, (b) $P\{X_1 = i\}$, (c) $P\{X_2 = j\}$, (d) $P(X_1 = i | X = n)$, (e) $P(X = n | X_1 = i)$.

Solution: (a) $P\{X = n\} = \frac{e^{-\lambda T}(\lambda T)^n}{n!}$.
(b) $P\{X_1 = i\} = \frac{e^{-\lambda \tau}(\lambda \tau)^i}{i!}$.
(c) $P\{X_2 = j\} = \frac{e^{-\lambda (T-\tau)}(\lambda (T-\tau))^j}{j!}$.
(d)

$$\begin{aligned} P(X_1 = i | X = n) &= \frac{P\{X_1 = i, X = n\}}{P\{X = n\}} = \frac{P\{X_1 = i, X_2 = j\}}{P\{X = n\}} \\ &= \frac{P\{X_1 = i\}P\{X_2 = j\}}{P\{X = n\}} = \frac{n!}{i!j!}\left(\frac{\tau}{T}\right)^i \left(\frac{T-\tau}{T}\right)^j \\ &= \binom{n}{i} p^i (1-p)^{n-i}, \end{aligned}$$

where $p = \frac{\tau}{T}$. As a function of i, the answer is thus the pmf of a binomial distribution. This result is indicative of the following stronger property that can be shown: given there are n counts during $[0, T]$, the times of the counts are independent and uniformly distributed over the interval $[0, T]$.

(e) Since X_1 and X_2 are numbers of counts in disjoint time intervals, they are independent. Therefore,

$$\begin{aligned} P(X = n|X_1 = i) &= P(X_2 = j|X_1 = i) \\ &= P\{X_2 = j\} = \frac{e^{-\lambda(T-\tau)}(\lambda(T-\tau))^j}{j!}. \end{aligned}$$

which can also be written as $P(X = n|X_1 = i) = \frac{e^{-\lambda(T-\tau)}(\lambda(T-\tau))^{n-i}}{(n-i)!}$. That is, given $X_1 = i$, the total number of counts is i plus a random number of counts. The random number of counts has the Poisson distribution with mean $\lambda(T - \tau)$.

Example 3.5.4 Calls arrive to a cell in a certain wireless communication system according to a Poisson process with arrival rate $\lambda = 2$ calls per minute. Measure time in minutes and consider an interval of time beginning at time $t = 0$. Let N_t denote the number of calls that arrive up until time t. For a fixed $t > 0$, the random variable N_t is a Poisson random variable with parameter $2t$, so it's pmf is given by $P\{N_t = i\} = \frac{e^{-2t}(2t)^i}{i!}$ for nonnegative integers i.
(a) Find the probability of each of the following six events:

E_1="No calls arrive in the first 3.5 minutes."

E_2="The first call arrives after time $t = 3.5$."

E_3="Two or fewer calls arrive in the first 3.5 minutes."

E_4="The third call arrives after time $t = 3.5$."

E_5="The third call arrives after time t." (for general $t > 0$)

E_6="The third call arrives before time t." (for general $t > 0$)

(b) Derive the pdf of the arrival time of the third call.
(c) Find the expected arrival time of the tenth call?

Solution: Since $\lambda = 2$, $N_{3.5}$ has the Poisson distribution with mean 7. Therefore, $P(E_1) = P\{N_{3.5} = 0\} = \frac{e^{-7}(7)^0}{0!} = e^{-7} = 0.00091$.
Event E_2 is the same as E_1, so $P(E_2) = 0.00091$.
Using the pmf of the Poisson distribution with mean 7 yields: $P(E_3) = P\{N_{3.5} \leq 2\} = P\{N_{3.5} = 0\} + P\{N_{3.5} = 1\} + P\{N_{3.5} = 2\} = e^{-7}(1 + 7 + \frac{7^2}{2}) = 0.0296$.
Event E_4 is the same as event E_3, so $P(E_4) = 0.0296$.
Event E_5 is the same as the event that the number of calls that arrive by time t is less than or

equal to two, so $P(E_5) = P\{N_t \leq 2\} = P\{N_t = 0\} + P\{N_t = 1\} + P\{N_t = 2\} = e^{-2t}(1 + 2t + \frac{(2t)^2}{2})$. Event E_6 is just the complement of E_5, so: $P(E_6) = 1 - P(E_5) = 1 - e^{-2t}(1 + 2t + \frac{(2t)^2}{2})$.

(b) As a function of t, $P(E_6)$ is the CDF for the time of the arrival time of the third call. To get the pdf, differentiate it to get

$$f(t) = e^{-2t}\left(2\left(1 + 2t + \frac{(2t)^2}{2}\right) - 2 - 4t\right) = e^{-2t}(2t)^2 = e^{-2t}\frac{2^3t^2}{2}.$$

This is the Erlang density with parameters $r = 3$ and $\lambda = 2$.

(c) The expected time between arrivals is $1/\lambda$, so the expected time until the tenth arrival is $10/\lambda = 5$.

Example 3.5.5 Consider a Poisson process with rate $\lambda > 0$.
(a) Find the probability there is exactly one count in each of the intervals (0,1], (1,2], and (2,3].
(b) Find the probability of the event A, that there are two counts in the interval $(0, 2]$ and two counts in the interval $(1, 3]$. Note that these intervals overlap.
(c) Find the conditional probability there are two counts in the interval (1,2], given that there are two counts in the interval (0,2] and two counts in the the interval (1,3].

Solution (a) The numbers of counts in the these disjoint intervals are independent, Poisson random variables with mean λ. Thus, the probability is $(\lambda e^{-\lambda})^3 = \lambda^3 e^{-3\lambda}$.
(b) The event A is the union of three disjoint events: $A = B_{020} \cup B_{111} \cup B_{202}$, where B_{ijk} is the event that there are i counts in the interval $(0, 1]$, j counts in the interval $(1, 2]$, and k counts in the interval $(2, 3]$. Since the events B_{ijk} involve numbers of counts in disjoint intervals, we can easily write down their probabilities. For example,

$$\begin{aligned} P(B_{020}) &= P(\text{no counts in (0,1])} \cdot P(\text{two counts in (1,2])} \cdot P(\text{no counts in (2,3])} \\ &= e^{-\lambda}\left(\frac{\lambda^2 e^{-\lambda}}{2!}\right)e^{-\lambda} = \frac{\lambda^2}{2}e^{-3\lambda}. \end{aligned}$$

So, by the law of total probability,

$$\begin{aligned} P(A) &= P(B_{020}) + P(B_{111}) + P(B_{202}) \\ &= \frac{\lambda^2}{2}e^{-3\lambda} + (\lambda e^{-\lambda})^3 + \left(\frac{\lambda^2 e^{-\lambda}}{2}\right)e^{-\lambda}\left(\frac{\lambda^2 e^{-\lambda}}{2}\right) = \left(\frac{\lambda^2}{2} + \lambda^3 + \frac{\lambda^4}{4}\right)e^{-3\lambda}. \end{aligned}$$

(c) By the definition of conditional probability,

$$\begin{aligned} P(B_{020}|A) &= \frac{P(B_{020}A)}{P(A)} = \frac{P(B_{020})}{P(A)} \\ &= \frac{\frac{\lambda^2}{2}}{\frac{\lambda^2}{2} + \lambda^3 + \frac{\lambda^4}{4}} = \frac{2}{2 + 4\lambda + \lambda^2}. \end{aligned}$$

Notice that in part (b) we applied the law of total probability to find A, and in part (c) we applied the definition of the conditional probability $P(B_{020}|A)$. Together, this amounts to application of Bayes rule for finding $P(B_{020}|A)$.

3.5.3 The Erlang distribution

Let T_r denote the time of the rth count of a Poisson process. Thus, $T_r = U_1 + \cdots + U_r$, where $U_1, \ldots, U_r$ are independent, exponentially distributed random variables with parameter λ. One way to derive the pdf of f_{T_r} is to use this characterization of T_r and the method of Section 4.5.2, showing how to find the pdf of the sum of independent continuous-type random variables. But the following method is less work. Notice that for a fixed time t, the event $\{T_r > t\}$ can be written as $\{N_t \leq r-1\}$, because the rth count happens after time t if and only if the number of counts that happened by time t is less than or equal to $r-1$. Therefore,

$$P\{T_r > t\} = \sum_{k=0}^{r-1} \frac{\exp(-\lambda t)(\lambda t)^k}{k!}.$$

The pdf is thus

$$\begin{aligned}
f_{T_r}(t) &= -\frac{dP\{T_r > t\}}{dt} \\
&= \exp(-\lambda t)\left(\sum_{k=0}^{r-1} \lambda \frac{(\lambda t)^k}{k!} - \sum_{k=1}^{r-1} \frac{k\lambda^k t^{k-1}}{k!}\right) \\
&= \exp(-\lambda t)\left(\sum_{k=0}^{r-1} \frac{\lambda^{k+1} t^k}{k!} - \sum_{k=1}^{r-1} \frac{\lambda^k t^{k-1}}{(k-1)!}\right) \\
&= \exp(-\lambda t)\left(\sum_{k=0}^{r-1} \frac{\lambda^{k+1} t^k}{k!} - \sum_{k=0}^{r-2} \frac{\lambda^{k+1} t^k}{k!}\right) \\
&= \frac{\exp(-\lambda t)\lambda^r t^{r-1}}{(r-1)!}.
\end{aligned}$$

The distribution of T_r is called the *Erlang distribution*[2] with parameters r and λ. The mean of T_r is $\frac{r}{\lambda}$, because T_r is the sum of r random variables, each with mean $1/\lambda$. It is shown in Example 4.8.1 that $\mathrm{Var}(T_r) = \frac{r}{\lambda^2}$.

Recall from Section 3.4 that the exponential distribution is the limit of a scaled geometric random variable. In the same way, the sum of r independent exponential random variables is the scaled limit of the sum of r independent geometric random variables. That is exactly what we

[2]The Erlang distribution can be generalized to the case r is any positive real number (not necessarily an integer) by replacing the term $(r-1)!$ by $\Gamma(r)$, where Γ is the gamma function defined by $\Gamma(r) = \int_0^\infty t^{r-1}e^{-t}dt$. Distributions in this more general family are called *gamma distributions*. Thus, Erlang distributions are special cases of Gamma distributions for the case that r is a positive integer, in which case $\Gamma(r) = (r-1)!$.

just showed: The Erlang distribution with parameters r and λ is the limit of a scaled negative binomial random variable hS_r, where S_r has the negative binomial distribution with parameters r and $p = \lambda h$.

3.6 Linear scaling of pdfs and the Gaussian distribution

3.6.1 Scaling rule for pdfs

Let X be a random variable with pdf f_X and let $Y = aX + b$ where $a > 0$.[3] The pdf of Y is given by the following *scaling rule*:

$$Y = aX + b \quad \Rightarrow \quad f_Y(v) = f_X\left(\frac{v-b}{a}\right)\frac{1}{a}. \tag{3.5}$$

We explain what the scaling rule means graphically, assuming $a \geq 1$. The situation for $0 < a \leq 1$ is similar. To obtain f_Y from f_X, first stretch the graph of f_X horizontally by a factor a and shrink it vertically by a factor a. That operation leaves the area under the graph equal to one, and produces the pdf of aX. Then shift the graph horizontally by b (to the right if $b > 0$ or to the left if $b < 0$.)

Here is a derivation of the scaling rule (3.5). Since $a > 0$, the event $\{aX + b \leq v\}$ is the same as $\left\{X \leq \frac{v-b}{a}\right\}$, so the CDF of Y can be expressed as follows:

$$F_Y(v) = P\{aX + b \leq v\} = P\left\{X \leq \frac{v-b}{a}\right\} = F_X\left(\frac{v-b}{a}\right).$$

Differentiate $F_Y(v)$ with respect to v, using the chain rule of calculus and the fact $F_X' = f_X$, to obtain (3.5):

$$f_Y(v) = F_Y'(v) = f_X\left(\frac{v-b}{a}\right)\frac{1}{a}.$$

Section 3.2 recounts how the mean, variance, and standard deviation of Y are related to the mean, variance, and standard deviation of X, in case $Y = aX + b$. These relations are the same ones discussed in Section 2.2 for discrete-type random variables, namely:

$$E[Y] = aE[X] + b \qquad \mathrm{Var(Y)} = a^2\mathrm{Var}(X) \qquad \sigma_Y = a\sigma_X.$$

In particular, the standardized version of a random variable X, $\frac{X-\mu_X}{\sigma_X}$, has mean zero and variance one.

Example 3.6.1 Let X denote the pdf of the high temperature, in degrees C (Celsius), for a certain day of the year in some city. Let Y denote the pdf of the same temperature, but in degrees F (Fahrenheit). The conversion formula is $Y = (1.8)X + 32$. This is the linear transformation that maps zero degrees C to 32 degrees F and 100 degrees C to 212 degrees F.
(a) Express f_Y in terms of f_X.
(b) Sketch f_Y in the case X is uniformly distributed over the interval $[15, 20]$.

[3] The case $a < 0$ is discussed in Example 3.8.4, included in the section on functions of a random variable.

Solution: (a) By the scaling formula with $a = 1.8$ and $b = 32$, $f_Y(c) = f_X(\frac{c-32}{1.8})/1.8$.
(b) The case when X is uniformly distributed over $[15, 20]$ leads to Y uniformly distributed over $[59, 68]$. This is illustrated in Figure 3.9. The pdf of X is shown at the top of the figure. The pdf

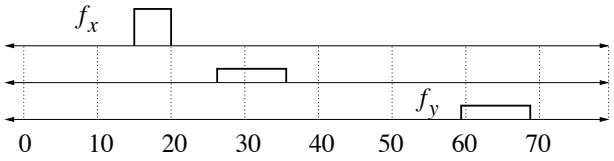


Figure 3.9: Rescaling from degrees C to degrees F.

of $(1.8)X$ is shown in the middle of the figure. It is obtained by stretching f_X out from the origin by a factor 1.8, and at the same time reducing the height by the same factor, so the area under the curve is still one. Finally, f_Y is obtained by sliding the pdf of $(1.8)X$ to the right by 32.

Example 3.6.2 Let T denote the duration of a waiting time in a service system, measured in seconds. Suppose T is exponentially distributed with parameter $\lambda = 0.01$. Let S denote the same waiting time, but measured in minutes. Find the mean and the pdf of S.

Solution: First we identify the pdf and mean of T. By definition of the exponential distribution with parameter $\lambda = 0.01$, the pdf of T is

$$f_T(u) = \begin{cases} (0.01)e^{-(0.01)u} & \text{for } u \geq 0 \\ 0 & \text{for } u < 0, \end{cases}$$

where the variable u is a measure of time in seconds. The mean of an exponentially distributed random variable with parameter λ is $\frac{1}{\lambda}$, so $E[T] = \frac{1}{0.01} = 100$.

The waiting time in minutes is given by $S = \frac{T}{60}$. By the linearity of expectation, $E[S] = \frac{E[T]}{60} = \frac{100}{60} = 1.66\ldots$. That is, the mean waiting time is 100 seconds, or $1.666\ldots$ minutes. By the scaling formula with $a = 1/60$ and $b = 0$,

$$f_S(v) = f_T(60v)60 = \begin{cases} (60)(0.01)e^{-(0.01)60v} = (0.6)e^{-(0.6)v} & \text{for } v \geq 0 \\ 0 & \text{for } v < 0, \end{cases}$$

where v is a measure of time in minutes. Examining this pdf shows that S is exponentially distributed with parameter 0.60. From this fact, we can find the mean of S a second way–it is one over the parameter in the exponential distribution for S, namely $\frac{1}{0.6} = 1.666\ldots$, as already noted.

Example 3.6.3 Let X be a uniformly distributed random variable on some interval $[a, b]$. Find the distribution of the standardized random variable $\frac{X-\mu_X}{\sigma_X}$.

Solution The mean, μ_X, is the midpoint of the interval $[a, b]$, and the standard deviation is $\sigma_X = \frac{(b-a)}{2\sqrt{3}}$. The pdf for $X - \mu_X$ is obtained by shifting the pdf of X to be centered at zero. Thus, $X - \mu_X$ is uniformly distributed over the interval $[-\frac{b-a}{2}, \frac{b-a}{2}]$. When this random variable is divided by σ_X, the resulting pdf is shrunk horizontally by the factor σ_X. This results in a uniform distribution over the interval $[-\frac{b-a}{2\sigma_X}, \frac{b-a}{2\sigma_X}] = [-\sqrt{3}, \sqrt{3}]$. This makes sense, because the uniform distribution over the interval $[-\sqrt{3}, \sqrt{3}]$ is the unique uniform distribution with mean zero and variance one.

3.6.2 The Gaussian (normal) distribution

The Gaussian distribution, also known as the normal distribution, has the pdf

$$f(u) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(u-\mu)^2}{2\sigma^2}\right),$$

where the parameters are μ and σ^2. The distribution is often denoted by "$N(\mu, \sigma^2)$ distribution," which is read aloud as "the normal μ, σ^2 distribution." Later in this section it is shown that the pdf integrates to one, the mean is μ, and the variance is σ^2. The pdf is pictured in Figure 3.10. The

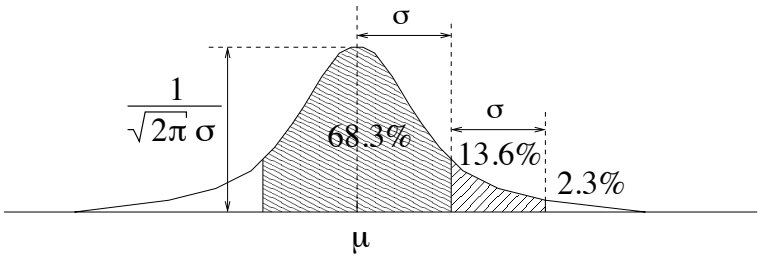


Figure 3.10: The Gaussian (or normal) pdf

parameter μ is sometimes called the location parameter, because the pdf is symmetric around the value μ. The effect of replacing μ by $\mu + 3$, for example, would be to slide the pdf to the right by three units. The parameter σ is sometimes called the scale parameter. As indicated in the figure, the width of the graph of the pdf is proportional to σ. For example, about 68.3% of the probability mass is in the interval $[\mu - \sigma, \mu + \sigma]$ and roughly 95% (more precisely, 95.44%) of the distribution is in the interval $[\mu - 2\sigma, \mu + 2\sigma]$. The peak height of the pdf is $\frac{1}{\sqrt{2\pi\sigma^2}}$. The height is inversely proportional to σ as expected, so for small σ the graph is tall and narrow, and for large σ it is short and spread out, but for all σ the area under the pdf is one.

The *standard normal distribution* is the normal distribution with $\mu = 0$ and $\sigma^2 = 1$. It is also called the $N(0, 1)$ distribution. The CDF of the $N(0, 1)$ distribution is traditionally denoted by the letter Φ (Phi):

$$\Phi(u) = \int_{-\infty}^{u} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{v^2}{2}\right) dv,$$

and the complementary CDF of the $N(0,1)$ distribution is traditionally denoted by the letter Q, (at least in much of the systems engineering literature). So

$$Q(u) = \int_u^\infty \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{v^2}{2}\right) dv = 1 - \Phi(u) = \Phi(-u).$$

Since $Q(u) = 1 - \Phi(u) = \Phi(-u)$, any probabilities we can express using the Q function we can also express using the Φ function. There is no hard and fast rule for whether to use the Φ function or the Q function, but typically we use $Q(u)$ for values of u that are larger than three or four, and $\Phi(u)$ for smaller positive values of u. When we deal with probabilities that are close to one it is usually more convenient to represent them as one minus something, for example writing $1 - 8.5 \times 10^{-6}$ instead of 0.9999915. The Φ and Q functions are available on many programable calculators and on Internet websites.[4] Some numerical values of these functions are given in Tables 6.1 and 6.2, in the appendix.

Let μ be any number and $\sigma > 0$. If X is a standard Gaussian random variable, and $Y = \sigma X + \mu$, then Y is a $N(\mu, \sigma^2)$ random variable. Indeed,

$$f_X(u) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u^2}{2}\right),$$

so by the scaling rule (3.5),

$$\begin{aligned} f_Y(v) &= \frac{1}{\sigma} f_X\left(\frac{v-\mu}{\sigma}\right) \\ &= \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(v-\mu)^2}{2\sigma^2}\right), \end{aligned}$$

so f_Y is indeed the $N(\mu, \sigma^2)$ pdf. Graphically, this means that the $N(\mu, \sigma^2)$ pdf can be obtained from the standard normal pdf by stretching it horizontally by a factor σ, shrinking it vertically by a factor σ, and sliding it over by μ.

Working in the other direction, if Y is a $N(\mu, \sigma^2)$ random variable, then the standardized version of Y, namely $X = \frac{Y-\mu}{\sigma}$, is a standard normal random variable. Graphically, this means that the standard normal pdf can be obtained from a $N(\mu, \sigma^2)$ pdf by sliding it over by μ (so it becomes centered at zero), shrinking it by a factor σ horizontally and stretching it by a factor σ vertically.

Let's check that the $N(\mu, \sigma^2)$ density indeed integrates to one, has mean μ, and variance σ^2. To show that the normal density integrates to one, it suffices to check that the standard normal density integrates to one, because the density for general μ and σ^2 is obtained from the standard normal pdf by the scaling rule, which preserves the total integral of the density. Let $I = \int_{-\infty}^{\infty} e^{-u^2/2} du$.

[4] http://www.stat.tamu.edu/~west/applets/normaldemo.html for example.

Then, switching to polar coordinates, we have:

$$\begin{aligned} I^2 &= \int_{-\infty}^{\infty} e^{-u^2/2}du \int_{-\infty}^{\infty} e^{-v^2/2}dv \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-(u^2+v^2)/2}dudv \\ &= \int_0^{2\pi}\int_0^{\infty} e^{-r^2/2}rdrd\theta \\ &= 2\pi\int_0^{\infty} e^{-r^2/2}rdr \\ &= -2\pi e^{-r^2/2}\Big|_0^{\infty} = 2\pi. \end{aligned}$$

Therefore, $I = \sqrt{2\pi}$, which means the standard normal density integrates to one, as claimed.

The fact that μ is the mean of the $N(\mu, \sigma^2)$ density follows from the fact that the density is symmetric about the point μ. To check that σ^2 is indeed the variance of the $N(\mu, \sigma^2)$ density, first note that if X is a standard normal random variable, then LOTUS and integration by parts yields

$$\begin{aligned} E[X^2] &= \int_{-\infty}^{\infty} \frac{u^2}{\sqrt{2\pi}} \exp\left(-\frac{u^2}{2}\right) du \\ &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} u \cdot u \exp\left(-\frac{u^2}{2}\right) du \\ &= -\frac{u}{\sqrt{2\pi}} \exp\left(-\frac{u^2}{2}\right)\Big|_{-\infty}^{\infty} + \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u^2}{2}\right) du = 0 + 1 = 1. \end{aligned}$$

Since X has mean zero, the variance, $\mathrm{Var}(X)$, for a standard normal random variable is one. Finally, a $N(\mu, \sigma^2)$ random variable Y can be written as $\sigma X + \mu$, where X is a standard normal random variable, so by the scaling formula for variances, $\mathrm{Var}(Y) = \sigma^2 \mathrm{Var}(X) = \sigma^2$, as claimed.

Example 3.6.4 Let X be a $N(10, 16)$ random variable (i.e. a Gaussian random variable with mean 10 and variance 16). Find the numerical values of the following probabilities:

$$P\{X \geq 15\},\ P\{X \leq 5\},\ P\{X^2 \geq 400\}, \text{ and } P\{X = 2\}.$$

Solution: The idea is to use the fact that $\frac{X-10}{4}$ is a standard normal random variable, and use either the Φ or Q function.

$$\begin{aligned} P\{X \geq 15\} &= P\left\{\frac{X-10}{4} \geq \frac{15-10}{4}\right\} = Q\left(\frac{15-10}{4}\right) = Q(1.25) = 1 - \Phi(1.25) = 0.1056. \\ P\{X \leq 5\} &= P\left\{\frac{X-10}{4} \leq \frac{5-10}{4}\right\} = \Phi\left(\frac{5-10}{4}\right) = Q(1.25) = 0.1056. \\ P\{X^2 \geq 400\} &= P\{X \geq 20\} + P\{X \leq -20\} = P\left\{\frac{X-10}{4} \geq 2.5\right\} + P\left\{\frac{X-10}{4} \leq -7.5\right\} \\ &= Q(2.5) + Q(7.5) \approx Q(2.5) = 1 - \Phi(2.5) = 0.0062. \quad (\text{Note: } Q(7.5) < 10^{-12}.) \end{aligned}$$

Finally, $P\{X = 2\} = 0$, because X is a continuous-type random variable.

Example 3.6.5 Suppose X is a random variable with mean 10 and variance 3. Find the numerical value of $P\{X < 10 - \sqrt{3}\}$ (or, nearly equivalently, $P\{X < 8.27\}$) for the following two choices of distribution type: (a) Assuming X is a Gaussian random variable, (b) Assuming X is a uniform random variable.

Solution: (a) If X is $N(10, 3)$,

$$P\{X < 10 - \sqrt{3}\} = P\left\{\frac{X - 10}{\sqrt{3}} \leq -1\right\} = \Phi(-1) = 1 - \Phi(1) = Q(1) \approx 0.1587.$$

(b) A random variable uniformly distributed on $[a, b]$ has mean $\frac{a+b}{2}$ and variance $\frac{(b-a)^2}{12}$. Hence, $a + b = 20$ and $b - a = 6$, giving $a = 7, b = 13$. That is, X is uniformly distributed over $[7, 13]$. Therefore, $P\{X < 8.27\} = \frac{8.27-7}{6} \approx 0.211$.

In some applications the distribution of a random variable seems nearly Gaussian, but the random variable is also known to take values in some interval. One way to model the situation is used in the following example.

Example 3.6.6 Suppose the random variable X has pdf

$$f_X(u) = \begin{cases} \dfrac{K}{2\sqrt{2\pi}} \exp\left(-\dfrac{(u-2)^2}{8}\right), & 0 \leq u \leq 4, \\ \\ 0, & \text{otherwise.} \end{cases}$$

where K is a constant to be determined. Determine K, the CDF F_X, and the mean $E[X]$.

Solution: Note that for $u \in [0, 4]$, $f_X(u) = K f_Z(u)$, where Z is a $N(\mu = 2, \sigma^2 = 4)$ random variable. That is, f_X is obtained by truncating f_Z to the interval $[0, 4]$ (i.e. setting it to zero outside the interval) and then multiplying it by a constant $K > 1$ so f_X integrates to one. Therefore,

$$\begin{aligned} 1 &= \int_{-\infty}^{\infty} f_X(u)du = \left(\int_0^4 f_Z(u)du\right) K = P\{0 \leq Z \leq 4\} K \\ &= P\left\{\frac{0-2}{2} \leq \frac{Z-2}{2} \leq \frac{4-2}{2}\right\} K \\ &= P\left\{-1 \leq \frac{Z-2}{2} \leq 1\right\} K \\ &= (\Phi(1) - \Phi(-1))\, K = (\Phi(1) - (1 - \Phi(1)))K \\ &= (2\Phi(1) - 1)K = (0.6826)K, \end{aligned}$$

so $K = 1/0.6826 \approx 1.46$. The same reasoning can be used to find F_X. For $0 \le v \le 4$,

$$\begin{aligned} F_X(v) &= P\{0 \le X \le v\} = P\{0 \le Z \le v\}K = P\left\{-1 \le \frac{Z-2}{2} \le \frac{v-2}{2}\right\}K \\ &= \left(\Phi\left(\frac{v-2}{2}\right) - \Phi(-1)\right)K = \left(\Phi\left(\frac{v-2}{2}\right) - 0.1587\right)K. \end{aligned}$$

Thus,

$$F_X(v) = \begin{cases} 0 & \text{if } v \le 0 \\ \left(\Phi\left(\frac{v-2}{2}\right) - 0.1587\right)K & \text{if } 0 < v \le 4 \\ 1 & \text{if } v \ge 4. \end{cases}$$

Finally, as for finding $E[X]$, since the pdf $f_X(u)$ is symmetric about the point $u = 2$, and $\int_0^\infty u f_X(u)du$ and $\int_{-\infty}^0 u f_X(u)du$ are both finite, it follows that $E[X] = 2$.

3.6.3 The central limit theorem and the Gaussian approximation

The Gaussian distribution arises frequently in practice, because of the phenomenon known as the central limit theorem (CLT), and the associated Gaussian approximation. There are many mathematical formulations of the CLT which differ in various details, but the main idea is the following: If many independent random variables are added together, and if each of them is small in magnitude compared to the sum, then the sum has an approximately Gaussian distribution. That is, if the sum is X, and if $\widetilde{X}$ is a Gaussian random variable with the same mean and variance as X, then X and $\widetilde{X}$ have approximately the same CDF:

$$P\{X \le v\} \approx P\{\widetilde{X} \le v\} \qquad \text{(Gaussian approximation)}.$$

An important special case is when X is the sum of n Bernoulli random variables, each having the same parameter p. In other words, when X has the binomial distribution, with parameters n and p. The approximation is most accurate if both np and $n(1-p)$ are at least moderately large, and the probabilities being approximated are not extremely close to zero or one. As an example, suppose X has the binomial distribution with parameters $n = 10$ and $p = 0.2$. Then $E[X] = np = 2$ and $\mathrm{Var}(X) = np(1-p) = 1.6$. Let $\widetilde{X}$ be a Gaussian random variable with the same mean and variance as X. The CDFs of X and $\widetilde{X}$ are shown in Figure 3.11. The two CDFs cannot be close everywhere because the CDF of X is piecewise constant with jumps at integer points, and the CDF of $\widetilde{X}$ is continuous. Notice, however, that the functions are particularly close when the argument is halfway between two consecutive integers. For example, $P\{X \le 2.5\} = 0.6778$ and $P\{\widetilde{X} \le 2.5\} = 0.6537$. Since X is an integer-valued random variable, $P\{X \le 2.5\} = P\{X \le 2\}$. Therefore, we have:

$$P\{X \le 2\} = P\{X \le 2.5\} \approx P\{\widetilde{X} \le 2.5\}.$$

So, $P\{\widetilde{X} \le 2.5\}$ is a fairly good approximation to $P\{X \le 2\}$. In particular, it is a better approximation than $P\{\widetilde{X} \le 2\}$ is–in this case $P\{\widetilde{X} \le 2\} = 0.5$. In general, when X is an integer-valued

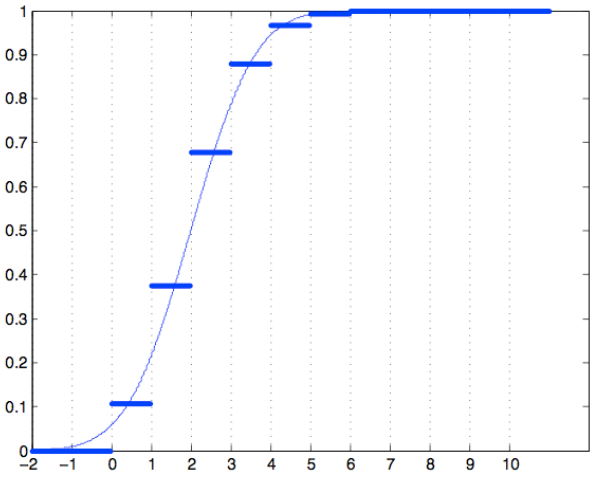


Figure 3.11: The CDF of a binomial random variable with parameters $n = 10$ and $p = 0.2$, and the Gaussian approximation of it.

random variable, the *Gaussian approximation with the continuity correction* is:

$$\begin{aligned} P\{X \le k\} &\approx P\{\widetilde{X} \le k + 0.5\} \\ P\{X \ge k\} &\approx P\{\widetilde{X} \ge k - 0.5\}. \end{aligned} \qquad \text{(Gaussian approximation with continuity correction)}$$

The example just given shows that, with the continuity correction, the Gaussian approximation is fairly accurate even for small values of n. The approximation improves as n increases. The CDF of X and $\widetilde{X}$ are shown in Figure 3.12 in case X has the binomial distribution with parameters $n = 30$ and $p = 0.2$.

As mentioned earlier, the Gaussian approximation is backed by various forms of the central limit theorem (CLT). Historically, the first version of the CLT proved is the following theorem, which pertains to the case of binomial distributions. Recall that if n is a positive integer and $0 < p < 1$, a binomial random variable $S_{n,p}$ with parameters n and p has mean np and variance $np(1-p)$. Therefore, the standardized version of $S_{n,p}$ is $\frac{S_{n,p}-np}{\sqrt{np(1-p)}}$.

Theorem 3.6.7 *(DeMoivre-Laplace limit theorem) Suppose $S_{n,p}$ is a binomial random variable with parameters n and p. For p fixed, with $0 < p < 1$, and any constant c,*

$$\lim_{n\to\infty} P\left\{\frac{S_{n,p} - np}{\sqrt{np(1-p)}} \le c\right\} = \Phi(c).$$

The practical implication of the DeMoivre-Laplace limit theorem is that, for large n, the standardized version of $S_{n,p}$ has approximately the standard normal distribution, or equivalently, that $S_{n,p}$ has approximately the same CDF as a Gaussian random variable with the same mean and variance. That is, the DeMoivre-Laplace limit theorem gives evidence that the Gaussian approximation to

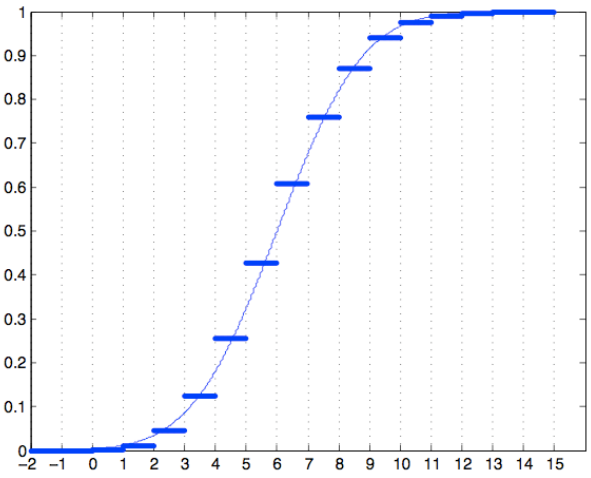


Figure 3.12: The CDF of a binomial random variable with parameters $n = 30$ and $p = 0.2$, and the Gaussian approximation of it.

the binomial distribution is a good one. A more general version of the CLT is stated in Section 4.10.2.

Example 3.6.8 (a) Suppose a fair coin is flipped a thousand times, and let X denote the number of times heads shows. Using the Gaussian approximation with the continuity correction,[5] find the approximate numerical value K so $P\{X \geq K\} \approx 0.01$. (b) Repeat, but now assume the coin is flipped a million times.

Solution: (a) The random variable X has the binomial distribution with parameters $n = 1000$ and $p = 0.5$. It thus has mean $\mu_X = np = 500$ and standard deviation $\sigma = \sqrt{np(1-p)} = \sqrt{250} \approx 15.8$. By the Gaussian approximation with the continuity correction,

$$P\{X \geq K\} = P\{X \geq K - 0.5\} = P\left\{\frac{X-\mu}{\sigma} \geq \frac{K-0.5-\mu}{\sigma}\right\} \approx Q\left(\frac{K-0.5-\mu}{\sigma}\right).$$

Since $Q(2.325) \approx 0.01$ we thus want to select K so $\frac{K-0.5-\mu}{\sigma} \approx 2.325$ or $K = \mu + 2.325\sigma + 0.5 = 537.26$. Thus, $K = 537$ or $K = 538$ should do. So, if the coin is flipped a thousand times, there is about a one percent chance that heads shows for more than 53.7% of the flips.
(b) By the same reasoning, we should take $K \approx \mu + 2.325\sigma + 0.5$, where, for $n = 10^6$, $\mu = 500000$ and $\sigma = \sqrt{250,000} = 500$. Thus, $K = 501163$ should do. So, if the coin is flipped a million times, there is about a one percent chance that heads shows for more than 50.12% of the flips.

---

[5] Use of the continuity correction is specified here to be definite, although n is so large that the continuity correction is not very important.

Example 3.6.9 Suppose a testing service is to administer a standardized test at a given location on a certain test date, and only students who have preregistered can take the test. Suppose 60 students preregistered, and each of those students actually shows up to take the test with probability $p = 5/6$. Let X denote the number of preregistered students showing up to take the test.
(a) Using the Gaussian approximation with the continuity correction, find the approximate numerical value of $P\{X \leq 52\}$.
(b) Similarly, find the approximate numerical value of $P\{|X - 50| \geq 6\}$.
(c) Find the Chebychev upper bound on $P\{|X - 50| \geq 6\}$. Compare it to your answer to part (b).

Solution: (a) The random variable X has the binomial distribution with parameters $n = 60$ and $p = 5/6$. It thus has mean $\mu_X = np = 50$ and standard deviation $\sigma_X = \sqrt{np(1-p)} \approx 2.887$. Thus, $\frac{X-50}{2.89}$ is approximately a standard normal random variable. Therefore,

$$\begin{aligned} P\{X \leq 52\} &= P\{X \leq 52.5\} = P\left\{\frac{X-50}{2.887} \leq \frac{52.5-50}{2.887}\right\} \\ &= P\left\{\frac{X-50}{2.887} \leq 0.866\right\} \approx \Phi(0.866) \approx 0.807. \end{aligned}$$

The Gaussian approximation in this case is fairly accurate: numerical calculation, using the binomial distribution, yields that $P\{X \leq 52\} \approx 0.8042$.
(b)

$$\begin{aligned} P\{|X - 50| \geq 6\} &= P\{X \leq 44 \text{ or } X \geq 56\} \\ &= P\{X \leq 44.5 \text{ or } X \geq 55.5\} \\ &= P\left\{\frac{X-50}{2.887} \leq -1.905 \text{ or } \frac{X-50}{2.887} \geq 1.905\right\} \approx 2Q(1.905) \approx 0.0567. \end{aligned}$$

The Gaussian approximation in this case is fairly accurate: numerical calculation, using the binomial distribution, yields that $P\{|X - 50| \geq 6\} \approx 0.0540$.
(c) The Chebychev inequality implies that $P\{|X - 50| \geq 6\} \leq \frac{\sigma_X^2}{36} = 0.231$, which is about four times larger than the value found in part (b).

Example 3.6.10 A campus network engineer would like to estimate the fraction p of packets going over the fiber optic link from the campus network to Bardeen Hall that are digital video disk (DVD) packets. The engineer writes a script to examine n packets, counts the number X that are DVD packets, and uses $\hat{p} = \frac{X}{n}$ to estimate p. The inspected packets are separated by hundreds of other packets, so it is reasonable to assume that each packet is a DVD packet with probability p, independently of the other packets.
(a) Using the Gaussian approximation to the binomial distribution, find an approximation to $P\{|\hat{p} - p| \leq \delta\}$ as a function of p, n, and δ. Evaluate it for $p = 0.5$ and for $p = 0.1$, with $\delta = 0.02$

and $n = 1000$.
(b) If $p = 0.5$ and $n = 1000$, find δ so $P\{|\hat{p}-p| \leq \delta\} \approx 0.99$. Equivalently, $P\{p \in [\hat{p}-\delta, \hat{p}+\delta]\} \approx 0.99$. Note that p is not random, but the *confidence interval* $[\hat{p}-\delta, \hat{p}+\delta]$ is random. So we want to find the half-width δ of the interval so we have 99% confidence that the interval will contain the true value of p.
(c) Repeat part (b), but for $p = 0.1$.
(d) However, the campus network engineer doesn't know p to begin with, so she can't select the halfwidth δ of the confidence interval as a function of p. A reasonable approach is to select δ so that, the Gaussian approximation to $P\{p \in [\hat{p}-\delta, \hat{p}+\delta]\}$ is greater than or equal to 0.99 for any value of p. Find such a δ for $n = 1000$.
(e) Using the same approach as in part (d), what n is needed (not depending on p) so that the random confidence interval $[\hat{p}-0.01, \hat{p}+0.01]$ contains p with probability at least 0.99 (according to the Gaussian approximation of the binomial)?

Solution: (a) Since X has the binomial distribution with parameters n and p, the Gaussian approximation yields

$$\begin{aligned} P[|\hat{p}-p| \leq \delta] &= P\left\{\left|\frac{X}{n}-p\right| \leq \delta\right\} = P\left\{\left|\frac{X-np}{\sqrt{np(1-p)}}\right| \leq \delta\sqrt{\frac{n}{p(1-p)}}\right\} \\ &\approx \Phi\left(\delta\sqrt{\frac{n}{p(1-p)}}\right) - \Phi\left(-\delta\sqrt{\frac{n}{p(1-p)}}\right) = 2\Phi\left(\delta\sqrt{\frac{n}{p(1-p)}}\right) - 1. \end{aligned}$$

For $n = 1000$, $\delta = 0.02$, and $p = 0.5$, this is equal to $2\Phi(1.265) = 0.794 = 79.4\%$, and
for $n = 1000$, $\delta = 0.02$, and $p = 0.1$, this is equal to $2\Phi(2.108) = 0.965 = 96.5\%$.
(b) Select δ so that $2\Phi\left(\delta\sqrt{\frac{1000}{p(1-p)}}\right) - 1 = 0.99$. Observing from Table 6.1 for the standard normal CDF that $2\Phi(2.58) - 1 \approx 0.99$, we select δ so that $\delta\sqrt{\frac{1000}{p(1-p)}} = 2.58$, or $\delta = 2.58\sqrt{\frac{p(1-p)}{1000}}$. The 99% confidence interval for $p = 0.5$ requires $\delta = 2.58\sqrt{\frac{0.5(1-0.5)}{1000}} = 0.04$.
(c) Similarly, the 99% confidence interval for $p = 0.1$ requires $\delta = 2.58\sqrt{\frac{0.1(0.9)}{1000}} = 0.025$.
(d) The product $p(1-p)$, and hence the required δ, is maximized by $p = 0.5$. Thus, if $\delta = 0.04$ as found in part (b), then the confidence interval contains p with probability at least 0.99 (no matter what the value of p is), at least up to the accuracy of the Gaussian approximation.
(e) The value of n needed for $p = 0.5$ works for any p (the situation is similar to that in part (d)) so n needs to be selected so that $0.01 = 2.58\sqrt{\frac{0.5(1-0.5)}{n}}$. This yields $n = (\frac{2.58}{0.01})^2(0.5)(1-0.5) \approx 16,641$.

3.7 ML parameter estimation for continuous-type variables

As discussed in Section 2.8 for discrete-type random variables, sometimes when we devise a probability model for some situation we have a reason to use a particular type of probability distribution,

but there may be a parameter that has to be selected. A common approach is to collect some data and then estimate the parameter using the observed data. For example, suppose the parameter is θ, and suppose that an observation is given with pdf f_θ that depends on θ. Section 2.8 suggests estimating θ by the value that maximizes the probability that the observed value is u. But for continuous-type observations, the probability of a specific observation u is zero, for any value of θ.

However, if f_θ is a continuous pdf for each value of θ, then recall from the interpretation, (3.4), of a pdf, that for ϵ sufficiently small, $f_\theta(u) \approx \frac{1}{\epsilon} P\left\{u - \frac{\epsilon}{2} < X < u + \frac{\epsilon}{2}\right\}$. That is, $f_\theta(u)$ is proportional to the probability that the observation is in an ϵ-width interval centered at u, where the constant of proportionality, namely $\frac{1}{\epsilon}$, is the same for all θ. Following tradition, in this context, we call $f_\theta(u)$ the likelihood of the observation u. The maximum likelihood estimate of θ for observation u, denoted by $\widehat{\theta}_{ML}(u)$, is defined to be the value of θ that maximizes the likelihood, $f_\theta(u)$, with respect to θ.

Example 3.7.1 Suppose a random variable T has the exponential distribution with parameter λ, and suppose it is observed that $T = t$, for some fixed value of t. Find the ML estimate, $\widehat{\lambda}_{ML}(t)$, of λ, based on the observation $T = t$.

Solution: The estimate, $\widehat{\lambda}_{ML}(t)$, is the value of $\lambda > 0$ that maximizes $\lambda e^{-\lambda t}$ with respect to λ, for t fixed. Since

$$\frac{d(\lambda e^{-\lambda t})}{d\lambda} = (1 - \lambda t)e^{-\lambda t},$$

the likelihood is increasing in λ for $0 \leq \lambda \leq \frac{1}{t}$, and it is decreasing in λ for $\lambda \geq \frac{1}{t}$, so the likelihood is maximized at $\frac{1}{t}$. That is, $\widehat{\lambda}_{ML}(t) = \frac{1}{t}$.

Example 3.7.2 Suppose it is assumed that X is drawn at random from the uniform distribution on the interval $[0, b]$, where b is a parameter to be estimated. Find the ML estimator of b given $X = u$ is observed. (This is the continuous-type distribution version of Example 2.8.2.)

Solution: The pdf can be written as $f_b(u) = \frac{1}{b} I_{\{0 \leq u \leq b\}}$. Recall that $I_{\{0 \leq u \leq b\}}$ is the indicator function of $\{0 \leq u \leq b\}$, equal to one on that set and equal to zero elsewhere. The whole idea now is to think of $f_b(u)$ not as a function of u (because u is the given observation), but rather, as a function of b. It is zero if $b < u$; it jumps up to $\frac{1}{u}$ at $b = u$; as b increases beyond u the function decreases. It is thus maximized at $b = u$. That is, $\widehat{b}_{ML}(u) = u$.

3.8 Functions of a random variable

3.8.1 The distribution of a function of a random variable

Often one random variable is a function of another. Suppose $Y = g(X)$ for some function g and a random variable X with a known pdf, and suppose we want to describe the distribution of Y. A general way to approach this problem is the following three step procedure:

Step 1: Scope the problem. Identify the support of X. Sketch the pdf of X and sketch g. Identify the support of Y. Determine whether Y is a continuous-type or discrete-type random variable. Then take a breath–you've done very important ground work here.

Step 2: Find the CDF of Y. Use the definition of the CDF: For any constant c, $F_Y(c) = P\{Y \leq c\} = P\{g(X) \leq c\}$. In order to find the probability of the event $\{g(X) \leq c\}$, try to describe it in a way that involves X in a simple way. In Step 2 most of the work is usually in finding $F_Y(c)$ for values of c that are in the support of Y.

Step 3: Differentiate F_Y **to find its derivative, which is** f_Y. Typically the pdf gives a more intuitive idea about the distribution of Y than the CDF.

The above three step procedure addresses the case that Y is continuous-type. If the function g is piecewise constant, such as a quantizer function, then Y is a discrete-type random variable. In that case, Steps 2 and 3 should be replaced by the following single step:

Step 2: (if Y **is discrete-type) Find the pmf of** Y. Work with the definition of the pmf: $p_Y(v) = P\{Y = v\} = P\{g(X) = v\}$. In order to find the probability of the event $\{g(X) = v\}$, try to describe it in a way that involves X in a simple way. Basically, $P\{g(X) = v\} = \int_{\{u:g(u)=v\}} f_X(u)du$ Here $p_Y(v)$ only needs to be identified for v in the support of Y. Usually the pmf is the desired answer and there is no need to find the CDF.

Example 3.8.1 Suppose $Y = X^2$, where X is a random variable with pdf $f_X(u) = \frac{e^{-|u|}}{2}$ for $u \in \mathbb{R}$. Find the pdf, mean, and variance of Y.

Solution: Note that $Y = g(X)$, where g is the function $g(u) = u^2$. Sketches of f_X and g are

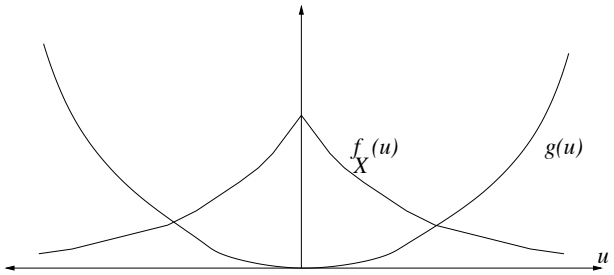


Figure 3.13: The pdf of X and the function g.

shown in Figure 3.13. The support of f_X is the entire real line, and g maps the real line onto the nonnegative reals, so the support of Y is the nonnegative reals. Also, there is no value v such that $P\{Y = v\} > 0$. So Y is a continuous-type random variable and we will find its CDF. At this point we have:

$$F_Y(c) = \begin{cases} 0 & c < 0 \\ ??? & c \geq 0; \end{cases}$$

we must find $F_Y(c)$ for $c \geq 0$. This completes Step 1, "scoping the problem," so we take a breath.

Continuing, for $c \geq 0$,

$$\begin{aligned} P\{X^2 \leq c\} &= P\{-\sqrt{c} \leq X \leq \sqrt{c}\} \\ &= \int_{-\sqrt{c}}^{\sqrt{c}} f_X(u)du = \int_{-\sqrt{c}}^{\sqrt{c}} \frac{\exp(-|u|)}{2} du \\ &= \int_0^{\sqrt{c}} \exp(-u)du = 1 - e^{-\sqrt{c}}, \end{aligned}$$

where the setup here is illustrated in Figure 3.14. So

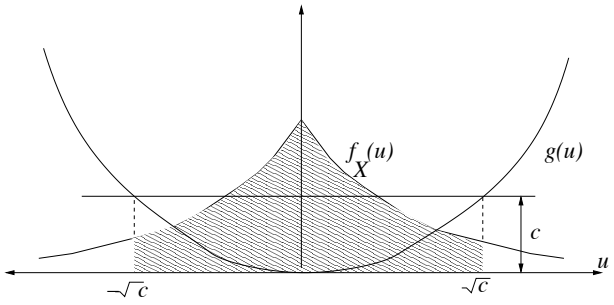


Figure 3.14: $F_X(c)$ is area of the shaded region.

$$F_Y(c) = \begin{cases} 0 & c < 0 \\ 1 - e^{-\sqrt{c}} & c \geq 0. \end{cases}$$

Finally, taking the derivative of F_Y (using the chain rule of calculus) yields

$$f_Y(c) = \begin{cases} 0 & c \leq 0 \\ \frac{e^{-\sqrt{c}}}{2\sqrt{c}} & c > 0, \end{cases}$$

Note that the value of a pdf at a single point can be changed without changing the integrals of the pdf. In this case we decided to let $f_Y(0) = 0$.

To find the mean and variance of Y we could use the pdf just found. However, instead of that, we will use LOTUS and the fact $\int_0^\infty u^n e^{-u} du = n!$ for all integers $n \geq 0$[6] to get

$$\begin{aligned} E[Y] &= \int_{-\infty}^{\infty} g(u) f_X(u) du \\ &= \int_{-\infty}^{\infty} \frac{u^2 e^{-|u|}}{2} du \\ &= \int_0^{\infty} u^2 e^{-u} du = 2! = 2. \end{aligned}$$

[6]This identity can be proved by induction on n, using integration by parts for the induction step.

and

$$\begin{aligned} E[Y^2] &= \int_{-\infty}^{\infty} (g(u))^2 f_X(u) du \\ &= \int_{-\infty}^{\infty} \frac{u^4 e^{-|u|}}{2} du \\ &= \int_0^{\infty} u^4 e^{-u} du = 4! = 24. \end{aligned}$$

Therefore, $\mathrm{Var}(Y) = E[Y^2] - E[Y]^2 = 20$.

Example 3.8.2 Suppose $Y = X^2$, where X is a $N(\mu, \sigma^2)$ random variable with $\mu = 2$ and $\sigma^2 = 3$. Find the pdf of Y.

Solution: Note that $Y = g(X)$ where $g(u) = u^2$. The support of the distribution of X is the whole real line, and the range of g over this support is $\mathbb{R}_+$. Next we find the CDF, F_Y. Since $P\{Y \geq 0\} = 1$, $F_Y(c) = 0$ for $c < 0$. For $c \geq 0$,

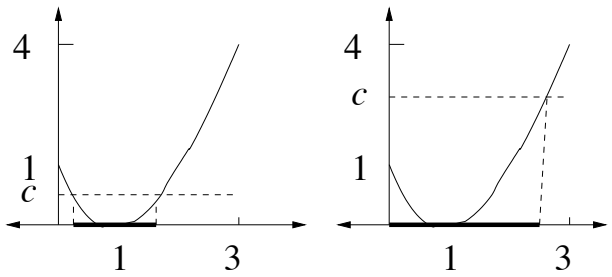
$$\begin{aligned} F_Y(c) = P\{X^2 \leq c\} &= P\{-\sqrt{c} \leq X \leq \sqrt{c}\} \\ &= P\left\{\frac{-\sqrt{c}-2}{\sqrt{3}} \leq \frac{X-2}{\sqrt{3}} \leq \frac{\sqrt{c}-2}{\sqrt{3}}\right\} \\ &= \Phi\left(\frac{\sqrt{c}-2}{\sqrt{3}}\right) - \Phi\left(\frac{-\sqrt{c}-2}{\sqrt{3}}\right). \end{aligned}$$

Differentiate with respect to c, using the chain rule and the fact: $\Phi'(s) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{s^2}{2})$, to obtain

$$f_Y(c) = \begin{cases} \frac{1}{\sqrt{24\pi c}} \left\{ \exp\left(-\frac{(\sqrt{c}-2)^2}{6}\right) + \exp\left(-\frac{(\sqrt{c}+2)^2}{6}\right) \right\} & \text{if } c \geq 0 \\ 0 & \text{if } c < 0. \end{cases} \tag{3.6}$$

Example 3.8.3 Suppose X is uniformly distributed over the interval [0, 3], and $Y = (X-1)^2$. Find the CDF, pdf, and expectation of Y.

Solution. Since X ranges over the interval $[0, 3]$, Y ranges over the interval $[0, 4]$. The expression for $F_Y(c)$ is qualitatively different for $0 \leq c \leq 1$ and $1 \leq c \leq 4$, as seen in the following sketch:



In each case, $F_Y(c)$ is equal to one third the length of the shaded interval. For $0 \le c \le 1$,

$$F_Y(c) = P\{(X-1)^2 \le c\} = P\{1-\sqrt{c} \le X \le 1+\sqrt{c}\} = \frac{2\sqrt{c}}{3}.$$

For $1 \le c \le 4$,

$$F_Y(c) = P\{(X-1)^2 \le c\} = P\{0 \le X \le 1+\sqrt{c}\} = \frac{1+\sqrt{c}}{3}.$$

Combining these observations yields:

$$F_Y(c) = \begin{cases} 0 & c < 0 \\ \frac{2\sqrt{c}}{3} & 0 \le c < 1 \\ \frac{1+\sqrt{c}}{3} & 1 \le c < 4 \\ 1 & c \ge 4. \end{cases}$$

Differentiating yields

$$f_Y(c) = \frac{dF_Y(c)}{dc} = \begin{cases} \frac{1}{3\sqrt{c}} & 0 \le c < 1 \\ \frac{1}{6\sqrt{c}} & 1 \le c < 4 \\ 0 & else. \end{cases}$$

By LOTUS,

$$E[Y] = E[(X-1)^2] = \int_0^3 (u-1)^2 \frac{1}{3} du = 1$$

Example 3.8.4 Suppose X is a continuous-type random variable. (a) Describe the distribution of $-X$ in terms of f_X. (b) More generally, describe the distribution of $aX+b$ in terms of f_X, for constants a and b with $a \neq 0$. (This generalizes Section 3.6.1, which covers only the case $a > 0$.)

Solution: (a) Let $Y = -X$, or equivalently, $Y = g(X)$ where $g(u) = -u$. We shall find the pdf of Y after first finding the CDF. For any constant c, $F_Y(c) = P\{Y \le c\} = P\{-X \le c\} = P\{X \ge -c\} = 1 - F_X(-c)$. Differentiating with respect to c yields $f_Y(c) = f_X(-c)$. Geometrically, the graph of f_Y is obtained by reflecting the graph of f_X about the vertical axis.
(b) Suppose now that $Y = aX + b$. The pdf of Y in case $a > 0$ is given in Section 3.6.1. So suppose $a < 0$. Then $F_Y(c) = P\{aX+b \le c\} = P\{aX \le c-b\} = P\left\{X \ge \frac{c-b}{a}\right\} = 1 - F_X\left(\frac{c-b}{a}\right)$. Differentiating with respect to c yields

$$f_Y(c) = f_X\left(\frac{c-b}{a}\right)\frac{1}{|a|}, \tag{3.7}$$

where we use the fact that $a = -|a|$ for $a < 0$. Actually, (3.7) is also true if $a > 0$, because in that case it is the same as (3.5). So (3.7) gives the pdf of $Y = aX + b$ for any $a \neq 0$.

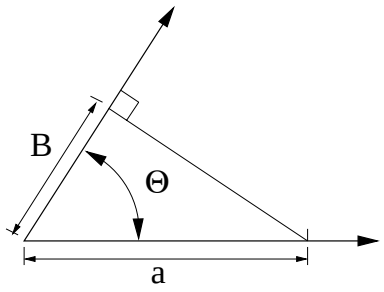


Figure 3.15: Direction of travel and a random direction.

Example 3.8.5 Suppose a vehicle is traveling in a straight line at constant speed a, and that a random direction is selected, subtending an angle Θ from the direction of travel. Suppose Θ is uniformly distributed over the interval $[0, \pi]$. See Figure 3.15. Then the effective speed of the vehicle in the random direction is $B = a\cos(\Theta)$. Find the pdf of B.

Solution: The range of $a\cos(\theta)$ as θ ranges over $[0, \pi]$ is the interval $[-a, a]$. Therefore, $F_B(c) = 0$ for $c \leq -a$ and $F_B(c) = 1$ for $c \geq a$. Let now $-a < c < a$. Then, because cos is monotone nonincreasing on the interval $[0, \pi]$,

$$\begin{aligned} F_B(c) &= P\{a\cos(\Theta) \leq c\} &&= P\left\{\cos(\Theta) \leq \frac{c}{a}\right\} \\ &&&= P\left\{\Theta \geq \cos^{-1}\left(\frac{c}{a}\right)\right\} \\ &&&= 1 - \frac{\cos^{-1}(\frac{c}{a})}{\pi}. \end{aligned}$$

Therefore, because $\cos^{-1}(y)$ has derivative, $-(1-y^2)^{-\frac{1}{2}}$,

$$f_B(c) = \begin{cases} \frac{1}{\pi\sqrt{a^2-c^2}} & |c| < a \\ 0 & |c| > a \end{cases}.$$

A sketch of the density is given in Figure 3.16.

Example 3.8.6 Suppose $Y = \tan(\Theta)$, as illustrated in Figure 3.17, where Θ is uniformly distributed over the interval $(-\frac{\pi}{2}, \frac{\pi}{2})$. Find the pdf of Y.

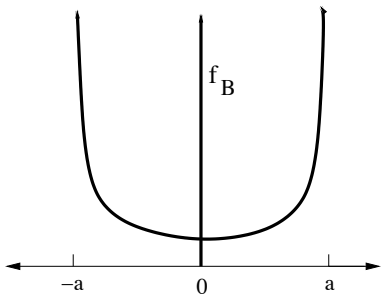


Figure 3.16: The pdf of the effective speed in a uniformly distributed direction.

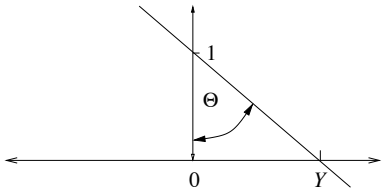


Figure 3.17: A horizontal line, a fixed point at unit distance, and a line through the point with random direction.

Solution: The function $\tan(\theta)$ increases from $-\infty$ to ∞ over the interval $(-\frac{\pi}{2}, \frac{\pi}{2})$, so the support of f_Y is the entire real line. For any real c,

$$\begin{aligned} F_Y(c) &= P\{Y \le c\} \\ &= P\{\tan(\Theta) \le c\} \\ &= P\{\Theta \le \tan^{-1}(c)\} \quad = \frac{\tan^{-1}(c) + \frac{\pi}{2}}{\pi}. \end{aligned}$$

Differentiating the CDF with respect to c yields that Y has the *Cauchy distribution*, with pdf:

$$f_Y(c) = \frac{1}{\pi(1+c^2)} \qquad -\infty < c < \infty.$$

Example 3.8.7 Given an angle θ expressed in radians, let $(\theta \mod 2\pi)$ denote the equivalent angle in the interval $[0, 2\pi]$. Thus, $(\theta \mod 2\pi)$ is equal to $\theta + 2\pi n$, where the integer n is such that $0 \le \theta + 2\pi n < 2\pi$.

Let Θ be uniformly distributed over $[0, 2\pi]$, let h be a constant, and let $\tilde{\Theta} = ((\Theta + h) \mod 2\pi)$. Find the distribution of $\widetilde{\Theta}$.

Solution: By its definition, $\tilde{\Theta}$ takes values in the interval $[0, 2\pi]$, so fix c with $0 \leq c \leq 2\pi$ and seek to find $P\{\tilde{\Theta} \leq c\}$. Since h can be replaced by (h mod 2π) if necessary, we can assume without loss of generality that $0 \leq h < 2\pi$. Then $\widetilde{\Theta} = g(\Theta)$, where the function $g(u) = ((u + h) \mod 2\pi)$ is graphed in Figure 3.18. Two cases are somewhat different: The first, shown in Figure 3.18(a), is

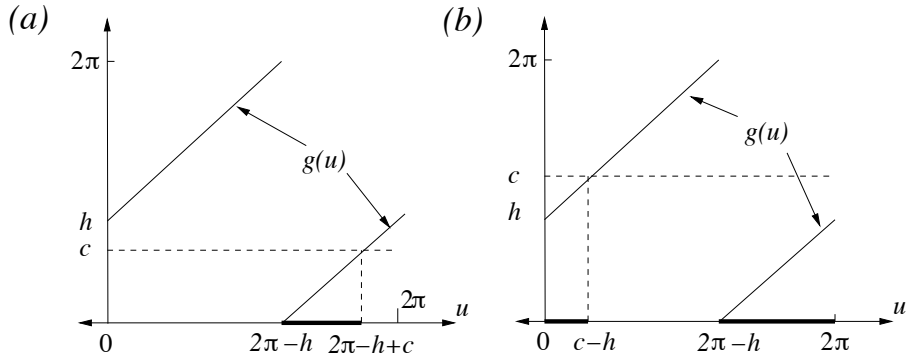


Figure 3.18: The function g such that $\widetilde{\Theta} = g(\Theta)$.

that $0 \leq c \leq h$. In this case, $\widetilde{\Theta} \leq c$ if Θ is in the interval $[2\pi - h, 2\pi - h + c]$, of length c. Therefore, in this case, $P\{\tilde{\Theta} \leq c\} = \frac{c}{2\pi}$. The other case is that $h < c \leq 2\pi$, shown in Figure 3.18(b). In this case, $\widetilde{\Theta} \leq c$ if Θ is in the union of intervals $[2\pi - h, 2\pi] \cup [0, c - h]$, which has total length c. So, again, $P\{\tilde{\Theta} \leq c\} = \frac{c}{2\pi}$. Therefore, in either case, $P\{\tilde{\Theta} \leq c\} = \frac{c}{2\pi}$, so that $\tilde{\Theta}$ is itself uniformly distributed over $[0, 2\pi]$.

Angles can be viewed as points on the unit circle in the plane. The result of this example is that, if an angle is uniformly distributed on the unit circle, then the angle plus a constant is also uniformly distributed over the unit circle.

Example 3.8.8 Express the pdf of $|X|$ in terms of the pdf of X, for an arbitrary continuous-type random variable X. Draw a sketch and give a geometric interpretation of the solution.

Solution: We seek the pdf of Y, where $Y = g(X)$ and $g(u) = |u|$. The variable Y takes nonnegative values, and for $c \geq 0$, $F_Y(c) = P\{Y \leq c\} = P\{-c \leq X \leq c\} = F_X(c) - F_X(-c)$. Thus,

$$F_Y(c) = \begin{cases} F_X(c) - F_X(-c) & c \geq 0 \\ 0 & c \leq 0; \end{cases}$$

Differentiating to get the pdf yields:

$$f_Y(c) = \begin{cases} f_X(c) + f_X(-c) & c \geq 0 \\ 0 & c < 0; \end{cases}$$

Basically, for each $c > 0$, there are two terms in the expression for $f_Y(c)$ because there are two ways for Y to be c–either $X = c$ or $X = -c$. A geometric interpretation is given in Figure 3.19. Figure

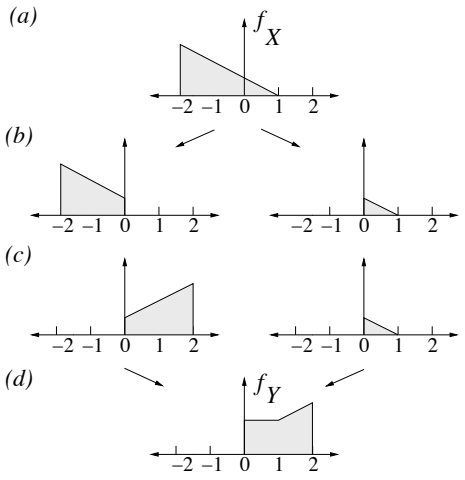


Figure 3.19: Geometric interpretation for pdf of $|X|$.

3.19(a) pictures a possible pdf for X. Figure 3.19(b) shows a decomposition of the probability mass into a part on the negative line and part on the positive line. Figure 3.19(c) shows the result of reflecting the probability mass on the negative line to the positive line. Figure 3.19(d) shows the pdf of $|Y|$, obtained by adding the two functions in Figure 3.19(c).

Another way to think of this geometrically would be to fold a picture of the pdf of X in half along the vertical axis, so that $f_X(c)$ and $f_X(-c)$ are lined up for each c, and then add these together to get $f_Y(c)$ for $c \geq 0$..

Example 3.8.9 Let X be an exponentially distributed random variable with parameter λ. Let $Y = \lfloor X \rfloor$, which is the integer part of X, and let $R = X - \lfloor X \rfloor$, which is the remainder. Describe the distributions of Y and R, and find the limit of the pdf of R as $\lambda \to 0$.

Solution: Clearly Y is a discrete-type random variable with possible values $0, 1, 2, \ldots$, so it is sufficient to find the pmf of Y. For integers $k \geq 0$,

$$p_Y(k) \quad = \quad P\{k \leq X < k+1\} \quad = \quad \int_k^{k+1} \lambda e^{-\lambda u} du = e^{-\lambda k}(1 - e^{-\lambda}),$$

and $p_Y(k) = 0$ for other k.

Turn next to the distribution of R. Note that $R = g(X)$, where g is the function sketched in Figure 3.20. Since R takes values in the interval $[0, 1]$, we shall let $0 < c < 1$ and find $F_R(c) = P\{R \leq c\}$. The event $\{R \leq c\}$ is equivalent to the event that X falls into the union of intervals,

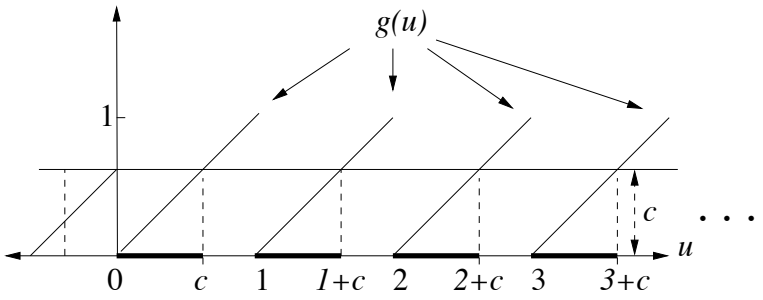


Figure 3.20: Function g such that $R = g(X)$.

$[0, c] \cup [1, 1 + c] \cup [2, 2 + c] \cup \cdots$, indicated in bold in the figure. Therefore,

$$\begin{aligned} F_R(c) &= P\{X - \lfloor X \rfloor \le c\} \\ &= \sum_{k=0}^{\infty} P\{k \le X \le k + c\} \\ &= \sum_{k=0}^{\infty} \int_k^{k+c} \lambda e^{-\lambda u} du = \sum_{k=0}^{\infty} \left(e^{-\lambda k} - e^{-\lambda(k+c)} \right) \\ &= \sum_{k=0}^{\infty} e^{-\lambda k}(1 - e^{-\lambda c}) = \frac{1 - e^{-\lambda c}}{1 - e^{-\lambda}}, \end{aligned}$$

where we used the formula $1 + \alpha + \alpha^2 + \cdots = \frac{1}{1-\alpha}$ for the sum of a geometric series, with $\alpha = e^{-\lambda}$. Differentiating F_R yields the pdf:

$$f_R(c) = \begin{cases} \frac{\lambda e^{-\lambda c}}{1 - e^{-\lambda}} & 0 \le c \le 1 \\ 0 & \text{otherwise.} \end{cases}$$

To find the limit of the pdf of R as $\lambda \to 0$, apply l'Hospital's rule to get

$$\lim_{\lambda \to 0} f_R(c) = \begin{cases} 1 & 0 \le c \le 1 \\ 0 & \text{otherwise.} \end{cases}$$

The limit of f_R is the pdf for the uniform distribution on the interval $[0, 1]$. Intuitively, the remainder R is nearly uniformly distributed over $[0, 1]$ for small λ because for such λ the density of X is spread out over a large range of integers.

Example 3.8.10 This example illustrates many possible particular cases. Suppose X has a pdf f_X which is supported on an interval $[a, b]$. Suppose $Y = g(X)$ where g is a strictly increasing function mapping the interval (a, b) onto the interval (A, B). The situation is illustrated in Figure

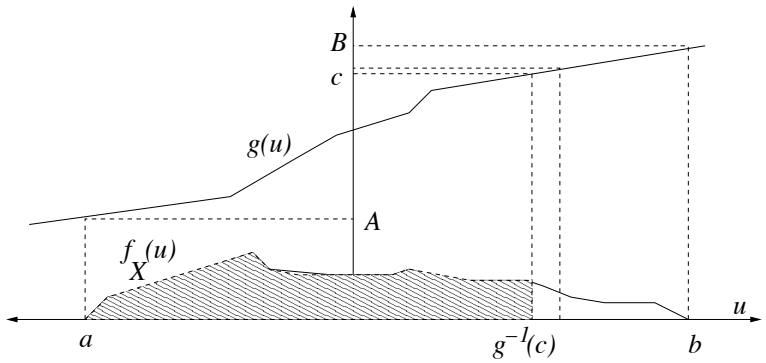


Figure 3.21: Monotone function of a continuous-type random variable.

3.21. The support of Y is the interval $[A, B]$. So let $A < c < B$. There is a value $g^{-1}(c)$ on the u axis such that $g(g^{-1}(c)) = c$, and:

$$F_Y(c) = P\{Y \leq c\} = P\{X \leq g^{-1}(c)\} = F_X(g^{-1}(c)).$$

The derivative of the inverse of a function is one over the derivative of the function itself[7]: $g^{-1}(c)' = \frac{1}{g'(g^{-1}(c))}$, where $g'(g^{-1}(c))$ denotes the derivative, $g'(u)$, evaluated at $u = g^{-1}(c)$. Thus, differentiating F_Y yields:

$$f_Y(c) = \begin{cases} f_X(g^{-1}(c))\frac{1}{g'(g^{-1}(c))} & A < c < B \\ 0 & \text{else.} \end{cases} \tag{3.8}$$

The expression for f_Y in (3.8) has an appealing form. Figure 3.21 shows a small interval with one endpoint c on the vertical axis, and the inverse image of the interval on the u axis, which is a small interval with one endpoint $g^{-1}(c)$. The probability Y falls into the vertical interval is equal to the probability that X falls into the horizontal interval on the u axis. The approximate ratio of the length of the vertical interval to the length of the horizontal interval is $g'(g^{-1}(c))$. The density $f_X(g^{-1}(c))$ is divided by this ratio in (3.8). Assuming Figure 3.21 is drawn to scale, the derivative of g at $g^{-1}(c)$ is about 1/4. Therefore the density of Y at c is about four times larger than the density of X at $g^{-1}(c)$.

A variation of this example would be to assume g is strictly decreasing. Then (3.8) holds with $g'(g^{-1}(c))$ replaced by $|g'(g^{-1}(c))|$. More generally, suppose g is continuously differentiable with no flat spots, but increasing on some intervals and decreasing on other intervals. For any real number c let $u_1, \ldots, u_n$ be the values of u such that $f(u_k) = c$ for all k. Here $n \geq 0$. Then (3.8) becomes $f_Y(c) = \sum_k f_X(u_k)\frac{1}{|g'(u_k)|}$.

Example 3.8.11 Suppose X is a continuous-type random variable with CDF F_X. Let Y be the result of applying F_X to X, that is, $Y = F_X(X)$. Find the distribution of Y.

[7]Prove this by differentiating both sides of the identity $g(g^{-1}(c)) = c$

Solution: Since Y takes values in the interval $[0,1]$, let $0 < v < 1$. Since F_X increases continuously from zero to one, there is a value c_v such that $F_X(c_v) = v$. Then $P\{F_X(X) \leq v\} = P\{X \leq c_v\} = F_X(c_v) = v$. That is, $F_X(X)$ is uniformly distributed over the interval $[0,1]$. This result may seem surprising at first, but it is natural if it is thought of in terms of percentiles. Consider, for example, the heights of all the people within a large group. Ignore ties. A particular person from the group is considered to be in the 90th percentile according to height, if the fraction of people in the group shorter than that person is 90%. So 90% of the people are in the 90th percentile or smaller. Similarly, 50% of the people are in the 50th percentile or smaller. So the percentile ranking of a randomly selected person from the group is uniformly distributed over the range from zero to one hundred percent. This result does not depend on the distribution of heights within the group. For this example, Y can be interpreted as the rank of X (expressed as a fraction rather than as a percentile) relative to the distribution assumed for X.

3.8.2 Generating a random variable with a specified distribution

An important step in many computer simulations of random systems is to generate a random variable with a specified CDF. This is often done by applying a function to a random variable that is uniformly distributed on the interval $[0,1]$. The method is basically to use Example 3.8.11 in reverse–if applying F_X to X produces a uniformly distributed random variable, applying F^{-1} to a uniform random variable should produce a random variable with CDF F. Let F be a function satisfying the three properties required of a CDF, as described in Proposition 3.1.5, and let U be uniformly distributed over the interval $[0,1]$. The problem is to find a function g so that F is the CDF of $g(U)$. An appropriate function g is given by the inverse function of F. Although F may not be strictly increasing, a suitable version of F^{-1} always exists, defined for $0 < u < 1$ by

$$F^{-1}(u) \quad = \quad \min\{c : F(c) \geq u\}. \tag{3.9}$$

If the graphs of F and F^{-1} are closed up by adding vertical lines at jump points, then the graphs are reflections of each other about the line through the origin of slope one, as illustrated in Figure 3.22. It is not hard to check that for any real c_o and u_o with $0 < u_o < 1$,

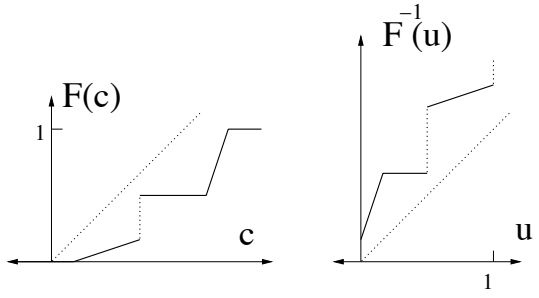


Figure 3.22: A CDF and its inverse.

$$F^{-1}(u_o) \le c_o \qquad \text{if and only if} \qquad u_o \le F(c_o).$$

Thus, if $X = F^{-1}(U)$ then

$$F_X(c) = P\{F^{-1}(U) \le c\} = P\{U \le F(c)\} = F(c),$$

so indeed F is the CDF of X.

Example 3.8.12 Find a function g so that, if U is uniformly distributed over the interval $[0, 1]$, $g(U)$ is exponentially distributed with parameter $\lambda = 1$.

Solution: The desired exponential distribution has support $\mathbb{R}_+$ and CDF F given by: $F(c) = 1 - e^{-c}$ for $c \ge 0$ and $F(c) = 0$ for $c < 0$. We'll let $g(u) = F^{-1}(u)$. Since F is strictly and continuously increasing over the support, if $0 < u < 1$ then the value c of $F^{-1}(u)$ is such that $F(c) = u$. That is, we would like $1 - e^{-c} = u$ which is equivalent to $e^{-c} = 1 - u$, or $c = -\ln(1 - u)$. Thus, $F^{-1}(u) = -\ln(1 - u)$. So we can take $g(u) = -\ln(1 - u)$ for $0 < u < 1$. To double check the answer, note that if $c \ge 0$, then

$$P\{-\ln(1 - U) \le c\} = P\{\ln(1 - U) \ge -c\} = P\{1 - U \ge e^{-c}\} = P\{U \le 1 - e^{-c}\} = F(c).$$

The choice of g is not unique in general. For example, $1 - U$ has the same distribution as U, so the CDF of $-\ln(U)$ is also F.

Example 3.8.13 Find a function g so that, if U is uniformly distributed over the interval $[0, 1]$, then $g(U)$ has the distribution of the number showing for the experiment of rolling a fair die.

Solution: The desired CDF of $g(U)$ is shown in Figure 3.1. Using $g = F^{-1}$ and using (3.9) or the graphical method illustrated in Figure 3.22 to find F^{-1}, we get that for $0 < u < 1$, $g(u) = i$ for $\frac{i-1}{6} < u \le \frac{i}{6}$ for $1 \le i \le 6$. To double check the answer, note that if $1 \le i \le 6$, then

$$P\{g(U) = i\} = P\left\{\frac{i-1}{6} < U \le \frac{i}{6}\right\} = \frac{1}{6},$$

so $g(U)$ has the correct pmf, and hence the correct CDF.

Example 3.8.14 (This example is a puzzle that doesn't use the theory of this section.) We've discussed how to generate a random variable with a specified distribution starting with a uniformly distributed random variable. Suppose instead that we would like to generate Bernoulli random variables with parameter $p = 0.5$ using flips of a *biased* coin. That is, suppose that the probability the coin shows H (for heads) is a, where a is some number that might not be precisely known. Explain how this coin can be used to generate independent Bernoulli random variables with $p = 0.5$.

Solution: Here is one solution. Flip the coin repeatedly, and look at the outcomes two at a time. If the first two outcomes are the same, ignore them. If they are different, then they are either HT (a head followed by a tail) or TH, and these two possibilities each have probability $a(1-a)$. In this case, give as output either H or T, whichever appeared first in those two flips. Then

$$\begin{aligned} & P(\text{output} = H | \text{there is output after two flips}) \\ = \; & P(\text{first two flips are } HT | \text{first two flips are } HT \text{ or } TH) \\ = \; & \frac{P(\text{first two flips are } HT)}{P(\text{first two flips are } HT) + P(\text{first two flips are } TH)} = \frac{a(1-a)}{2a(1-a)} = 0.5. \end{aligned}$$

Then repeat this procedure using the third and fourth flips, and so fourth. One might wonder how many times the biased coin must be flipped on average to produce one output. The required number of pairs of flips has the geometric distribution, with parameter $2a(1-a)$. Therefore, the mean number of pairs of coin flips required until an HT or TH is observed is $1/(2a(1-a))$. So the mean number of coin flips required to produce one Bernoulli random variable with parameter 0.5 using the biased coin by this method is $1/(a(1-a))$.

3.8.3 The area rule for expectation based on the CDF

There is a simple rule for determining the expectation, $E[X]$, of a random variable X directly from its CDF, called the *area rule for expectation.* See Figure 3.23, which shows an example of a CDF

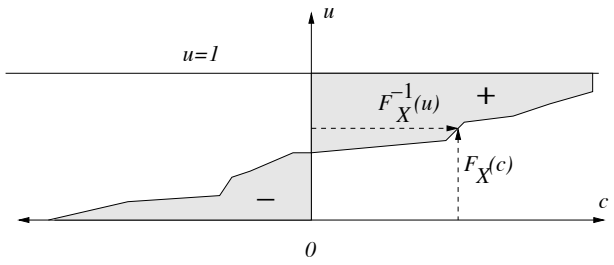


Figure 3.23: $E[X]$ is the area of the + region minus the area of the − region.

F_X plotted as a function of c in the $c-u$ plane. Consider the infinite strip bounded by the c axis and the horizontal line given by $u = 1$. Then $E[X]$ is the area of the region in the strip to the right of the u axis above the CDF, minus the area of the region in the strip to the left of the u axis below the CDF, as long as at least one of the two regions has finite area. The area rule can also be written as an equation, by integrating over the c axis:

$$E[X] = \int_0^\infty (1 - F_X(c))dc - \int_{-\infty}^0 F_X(c)dc, \tag{3.10}$$

or equivalently, by integrating over the u axis:

$$E[X] = \int_0^1 F_X^{-1}(u)du. \tag{3.11}$$

The area rule can be justified as follows. As noted in Section 3.8.2, if U is a random variable uniformly distributed on the interval $[0,1]$, then $F_X^{-1}(U)$ has the same distribution as X. Therefore, it also has the same expectation: $E[X] = E[F_X^{-1}(U)]$. Since $F_X^{-1}(U)$ is a function of U, its expectation can be found by LOTUS. Therefore, $E[X] = E[F_X^{-1}(U)] = \int_0^1 F_X^{-1}(u)du$, which proves (3.11), and hence, the area rule itself, because the right hand sides of both (3.10) and (3.11) are equal to the area of the region in the strip to the right of the u axis above the CDF, minus the area of the region in the strip to the left of the u axis below the CDF.

3.9 Failure rate functions

Eventually a system or a component of a particular system will fail. Let T be a random variable that denotes the lifetime of this item. Suppose T is a positive random variable with pdf f_T. The *failure rate function,* $h = (h(t) : t \geq 0)$, of T (and of the item itself) is defined by the following limit:

$$h(t) \stackrel{\triangle}{=} \lim_{\epsilon \to 0} \frac{P(t < T \leq t + \epsilon | T > t)}{\epsilon}.$$

That is, given the item is still working after t time units, the probability the item fails within the next ϵ time units is $h(t)\epsilon + o(\epsilon)$.

The failure rate function is determined by the distribution of T as follows:

$$\begin{aligned} h(t) &= \lim_{\epsilon \to 0} \frac{P\{t < T \leq t + \epsilon\}}{P\{T > t\}\epsilon} = \lim_{\epsilon \to 0} \frac{F_T(t+\epsilon) - F_T(t)}{(1 - F_T(t))\epsilon} \\ &= \frac{1}{(1 - F_T(t))} \left(\lim_{\epsilon \to 0} \frac{F_T(t+\epsilon) - F_T(t)}{\epsilon} \right) \\ &= \frac{f_T(t)}{1 - F_T(t)}, \end{aligned} \tag{3.12}$$

because the pdf f_T is the derivative of the CDF F_T.

Conversely, a nonnegative function $h = (h(t) : t \geq 0)$ with $\int_0^\infty h(t)dt = \infty$ determines a probability distribution with failure rate function h as follows. The CDF is given by

$$F(t) = 1 - e^{-\int_0^t h(s)ds}. \tag{3.13}$$

It is easy to check that F given by (3.13) has failure rate function h. To derive (3.13), and hence show it gives the unique distribution with failure rate function h, start with the fact that we would like $F'/(1-F) = h$. Equivalently, $(\ln(1-F))' = -h$, or, integrating over $[0,t]$, $\ln(1 - F(t)) = \ln(1 - F(0)) - \int_0^t h(s)ds$. Since F should be the CDF of a nonnegative continuous type random variable, $F(0) = 0$ or $\ln(1 - F(0)) = 0$. So $\ln(1 - F(t)) = -\int_0^t h(s)ds$, which is equivalent to (3.13).

Example 3.9.1 (a) Find the failure rate function for an exponentially distributed random variable with parameter λ. (b) Find the distribution with the linear failure rate function $h(t) = \frac{t}{\sigma^2}$ for $t \geq 0$. (c) Find the failure rate function of $T = \min\{T_1, T_2\}$, where T_1 and T_2 are independent random variables such that T_1 has failure rate function h_1 and T_2 has failure rate function h_2.

Solution: (a) If T has the exponential distribution with parameter λ, then for $t \geq 0$, $f_T(t) = \lambda e^{-\lambda t}$ and $1 - F_T(t) = e^{-\lambda t}$, so by (3.12), $h(t) = \lambda$ for all $t \geq 0$. That is, the exponential distribution with parameter λ has constant failure rate λ. The constant failure rate property is connected with the memoryless property of the exponential distribution; the memoryless property implies that $P(t < T \leq T + \epsilon | T > t) = P\{T \leq \epsilon\}$, which in view of the definition of h shows that h is constant.
(b) If $h(t) = \frac{t}{\sigma^2}$ for $t \geq 0$, then by (3.13), $F_T(t) = 1 - e^{-\frac{t^2}{2\sigma^2}}$. The corresponding pdf is given by

$$f_T(t) = \begin{cases} \frac{t}{\sigma^2} e^{-\frac{t^2}{2\sigma^2}} & t \geq 0 \\ 0 & \text{else.} \end{cases}$$

This is the pdf of the Rayleigh distribution with parameter σ^2.
(c) By the independence and (3.13) applied to T_1 and T_2,

$$P\{T > t\} = P\{T_1 > t \text{ and } T_2 > t\} = P\{T_1 > t\}P\{T_2 > t\} = e^{\int_0^t -h_1(s)ds} e^{\int_0^t -h_2(s)ds} = e^{-\int_0^t h(s)ds}$$

where $h = h_1 + h_2$. Therefore, the rate function for the minimum of two independent random variables is the sum of their failure rate functions. This makes intuitive sense; for a two-component system that fails when either components fails, the rate of system failure is the sum of the rates of component failure.

Two commonly appearing types of failure rate functions are shown in Figure 3.24. The failure

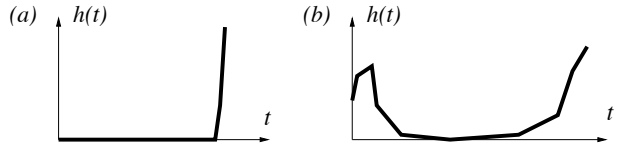


Figure 3.24: Failure rate functions for (a) nearly constant lifetime and (b) an item with a burn-in period (bath tub failure rate function)

rate function in Figure 3.24(a) corresponds to the lifetime T being nearly deterministic, because the failure rate is very small until some fixed time, and then it shoots up quickly at that time; with high probability, T is nearly equal to the time at which the failure rate function shoots up.

The failure rate function in Figure 3.24(b) corresponds to an item with a significant failure rate when the item is either new or sufficiently old. The large initial failure rate could be due to manufacturing defects or a critical burn-in requirement, and the large failure rate for sufficiently old items could be a result of wear or fatigue.

3.10 Binary hypothesis testing with continuous-type observations

Section 2.11 describes the binary hypothesis testing problem, with a focus on the case that the observation X is discrete-type. There are two hypotheses, H_1 and H_0, and if hypothesis H_i is true

then the observation is assumed to have a pmf p_i, for $i = 1$ or $i = 0$. The same framework works for continuous-type observations, as we describe in this section.

Suppose f_1 and f_0 are two known pdfs, and suppose that if H_i is true, then the observation X is a continuous-type random variable with pdf f_i. If a particular value of X is observed, the decision rule has to specify which hypothesis to declare. For continuous-type random variables, the probability that $X = u$ is zero for any value of u, for either hypothesis. But recall that if ϵ is small, then $f_i(u)\epsilon$ is the approximate probability that the observation is within a length ϵ interval centered at u, if H_i is the true hypothesis. For this reason, we still call $f_i(u)$ the likelihood of $X = u$ if H_i is the true hypothesis. We also define the likelihood ratio, $\Lambda(u)$, for u in the support of f_1 or f_0, by

$$\Lambda(u) = \frac{f_1(u)}{f_0(u)}.$$

A likelihood ratio test (LRT) with threshold τ is defined by:

$$\Lambda(X) \begin{cases} > \tau & \text{declare } H_1 \text{ is true} \\ < \tau & \text{declare } H_0 \text{ is true.} \end{cases}$$

Just as for the case of discrete-type observations, the maximum likelihood (ML) test is the LRT with threshold $\tau = 1$, and the maximum a posteriori probability (MAP) decision rule is the LRT with threshold $\tau = \frac{\pi_0}{\pi_1}$, where π_1 is the prior probability H_1 is true and π_0 is the prior probability H_0 is true. In particular, the ML rule is the special case of the MAP rule for the uniform prior: $\pi_1 = \pi_0$. The following definitions are the same as in the case of discrete-type observations:

$$\begin{aligned} p_{\text{false alarm}} &= P(\text{declare } H_1 \text{ true } | H_0) \\ p_{\text{miss}} &= P(\text{declare } H_0 \text{ true } | H_1) \\ p_e &= \pi_0 p_{\text{false alarm}} + \pi_1 p_{\text{miss}}. \end{aligned}$$

The MAP rule based on a given prior probability distribution (π_1, π_0) minimizes the average error probability, p_e, computed using the same prior.

Example 3.10.1 Suppose under hypothesis H_i, the observation X has the $N(m_i, \sigma^2)$ distribution, for $i = 0$ or $i = 1$, where the parameters are known and satisfy: $\sigma^2 > 0$ and $m_0 < m_1$. Identify the ML and MAP decision rules and their associated error probabilities, p_{false_alarm} and p_{miss}. Assume prior probabilities π_1 and π_0 are given where needed.

Solution: The pdfs are given by

$$f_i(u) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{ -\frac{(u - m_i)^2}{2\sigma^2} \right\},$$

so

$$\begin{aligned}\Lambda(u) &= \frac{f_1(u)}{f_0(u)} \\ &= \exp\left\{-\frac{(u-m_1)^2}{2\sigma^2}+\frac{(u-m_0)^2}{2\sigma^2}\right\} \\ &= \exp\left\{\left(u-\frac{m_0+m_1}{2}\right)\left(\frac{m_1-m_0}{\sigma^2}\right)\right\}.\end{aligned}$$

Observe that $\Lambda(X) > 1$ if and only if $X \geq \frac{m_0+m_1}{2}$, so the ML rule for this example is:

$$X \begin{cases} > \gamma_{ML} & \text{declare } H_1 \text{ is true} \\ < \gamma_{ML} & \text{declare } H_0 \text{ is true.} \end{cases}$$

where $\gamma_{ML} = \frac{m_0+m_1}{2}$. (The letter "$\gamma$" is used for the threshold here because it is the threshold for X directly, whereas the letter "τ" is used for the threshold applied to the likelihood ratio.)

The LRT for a general threshold τ and a general binary hypothesis testing problem with continuous-type observations is equivalent to

$$\ln \Lambda(X) \begin{cases} > \ln \tau & \text{declare } H_1 \text{ is true} \\ < \ln \tau & \text{declare } H_0 \text{ is true,} \end{cases}$$

which for this example can be expressed as a threshold test for X :

$$X \begin{cases} > \left(\frac{\sigma^2}{m_1-m_0}\right)\ln\tau + \frac{m_0+m_1}{2} & \text{declare } H_1 \text{ is true} \\ < \left(\frac{\sigma^2}{m_1-m_0}\right)\ln\tau + \frac{m_0+m_1}{2} & \text{declare } H_0 \text{ is true.} \end{cases}$$

In particular, the MAP rule is obtained by letting $\tau = \frac{\pi_0}{\pi_1}$, and it becomes:

$$X \begin{cases} > \gamma_{MAP} & \text{declare } H_1 \text{ is true} \\ < \gamma_{MAP} & \text{declare } H_0 \text{ is true.} \end{cases}$$

where $\gamma_{MAP} = \left(\frac{\sigma^2}{m_1-m_0}\right)\ln\left(\frac{\pi_0}{\pi_1}\right) + \frac{m_0+m_1}{2}$.

For this example, both the ML and MAP rules have the form

$$X \begin{cases} > \gamma & \text{declare } H_1 \text{ is true} \\ < \gamma & \text{declare } H_0 \text{ is true.} \end{cases}$$

Therefore, we shall examine the error probabilities for a test of that form. The error probabilities are given by the areas of the shaded regions shown in Figure 3.25.

$$\begin{aligned} p_{\text{false alarm}} &= P(X > \gamma | H_0) \\ &= P\left(\frac{X-m_0}{\sigma} > \frac{\gamma-m_0}{\sigma}\middle| H_0\right) \\ &= Q\left(\frac{\gamma-m_0}{\sigma}\right). \end{aligned}$$

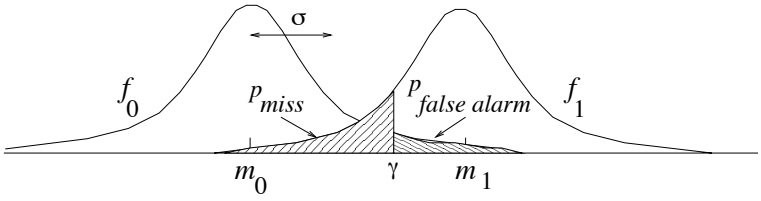


Figure 3.25: Error probabilities for direct threshold detection between two normal pdfs with the same variance.

$$\begin{aligned} p_{\text{miss}} &= P(X < \gamma | H_1) \\ &= P\left(\frac{X - m_1}{\sigma} < \frac{\gamma - m_1}{\sigma} \middle| H_1\right) \\ &= Q\left(\frac{m_1 - \gamma}{\sigma}\right). \end{aligned}$$

$$p_e = \pi_0 p_{\text{false alarm}} + \pi_1 p_{\text{miss}}.$$

Substituting in $\gamma = \gamma_{ML} = \frac{m_0+m_1}{2}$ yields that the error probabilities for the ML rule for this example satisfy:

$$p_{\text{false alarm}} = p_{\text{miss}} = p_e = Q\left(\frac{m_1 - m_0}{2\sigma}\right).$$

Note that $\frac{m_1-m_0}{\sigma}$ can be interpreted as a signal-to-noise ratio. The difference in the means, m_1-m_0, can be thought of as the difference between the hypotheses due to the signal, and σ is the standard deviation of the noise. The error probabilities for the MAP rule can be obtained by substituting in $\gamma = \gamma_{MAP}$ in the above expressions.

Example 3.10.2 Based on a sensor measurement X, it has to be decided which hypothesis about a remotely monitored machine is true: H_0: the machine is working vs. H_1: the machine is broken. Suppose if H_0 is true X is normal with mean 0 and variance a^2, and if H_1 is true X is normal with mean 0 and variance b^2. Suppose a and b are known and that $0 < a < b$. Find the ML decision rule, the MAP decision rule (for a given choice of π_0 and π_1) and the error probabilities, p_{false_alarm} and p_{miss}, for both rules.

Solution: To begin, we get as simple expression for the likelihood ratio as we can:

$$\Lambda(u) = \frac{\frac{1}{b\sqrt{2\pi}} e^{-\frac{u^2}{2b^2}}}{\frac{1}{a\sqrt{2\pi}} e^{-\frac{u^2}{2a^2}}} = \frac{a}{b} e^{-\frac{u^2}{2b^2} + \frac{u^2}{2a^2}} = \frac{a}{b} e^{\frac{u^2}{2}(\frac{1}{a^2} - \frac{1}{b^2})}$$

The ML rule is to choose H_1 when $\Lambda(X) > 1$. Thus, by taking the natural logarithm of both sides of this inequality we obtain the rule: If $(\ln \frac{a}{b}) + \frac{X^2}{2}(\frac{1}{a^2} - \frac{1}{b^2}) > 0$ choose H_1. Equivalently, after a bit of algebra, we find that the ML rule selects H_1 when

$$\ln \frac{b}{a} < \frac{X^2}{2}\frac{b^2 - a^2}{a^2 b^2} \quad \text{or} \quad \left(\frac{2a^2b^2 \ln(b/a)}{b^2 - a^2}\right) < X^2.$$

Thus, the ML rule can be expressed as a threshold test on the magnitude $|X|$ of X:

$$|X| \begin{cases} > K & \text{declare } H_1 \text{ is true} \\ < K & \text{declare } H_0 \text{ is true.} \end{cases} \tag{3.14}$$

where $K = K_{ML} = ab\sqrt{\frac{2\ln(b/a)}{b^2 - a^2}}$. Figure 3.10.2 plots the pdfs for the case $a^2 = 1$ and $b^2 = 4$.

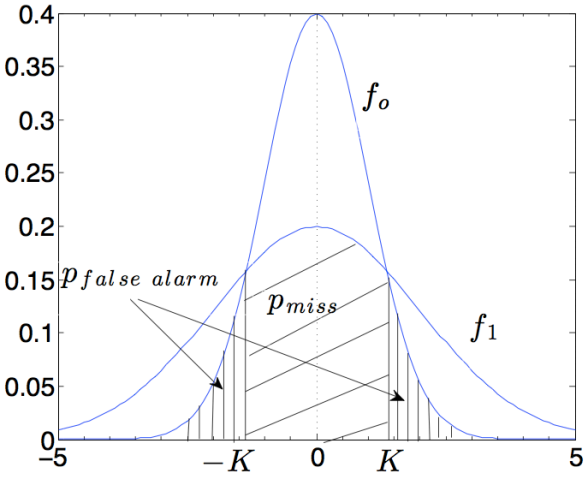


Figure 3.26: $N(0,1)$ and $N(0,4)$ pdfs and ML threshold K.

The MAP rule is to choose H_1 when $\Lambda(X) > \frac{\pi_0}{\pi_1}$. After a bit of algebra, we derive the rule that H_1 should be chosen when

$$\ln \frac{b\pi_0}{a\pi_1} < \frac{X^2}{2}\frac{b^2 - a^2}{a^2 b^2}$$

Or, equivalently, the MAP rule is given by the magnitude threshold test (3.14) with the threshold $K = K_{MAP} = ab\sqrt{\frac{2\ln(b\pi_0/a\pi_1)}{b^2 - a^2}}$.

Finally, we find the error probabilities for the magnitude threshold test (3.14) with an arbitrary threshold $K > 0$. Substituting in K_{ML} or K_{MAP} for K gives the error probabilities for the ML and MAP tests:

$$p_{false_alarm} = P\{|X| > K \mid H_0\} = \int_{-\infty}^{-K} f_0(u)du + \int_{K}^{\infty} f_0(u)du = \Phi(-K/a) + 1 - \Phi(K/a) = 2Q(K/a).$$

$$p_{miss} = P\{|X| < K \mid H_1\} = \int_{-K}^{K} f_1(u)du = \Phi(K/b) - \Phi(-K/b) = 1 - 2Q(K/b).$$

Example 3.10.3 Based on a sensor measurement X, it has to be decided which hypothesis about a remotely monitored room is true: H_0: the room is empty vs. H_1: a person is present in the room. Suppose if H_0 is true then X has pdf $f_0(x) = \frac{1}{2}e^{-|x+1|}$ and if H_1 is true then X has pdf $f_1(x) = \frac{1}{2}e^{-|x-1|}$. Both densities are defined on the whole real line. These distributions are examples of the *Laplace distribution*. Find the ML decision rule, the MAP decision rule for prior probability distribution $(\pi_0, \pi_1) = (2/3, 1/3)$, and the associated error probabilities, including the average error probability based on the given prior.

Solution: To help with the computations, we express the pdfs without using absolute value signs:

$$f_1(u) = \begin{cases} \frac{1}{2}e^{u-1} & : \quad u < 1 \\ \frac{1}{2}e^{-u+1} & : \quad u \geq 1 \end{cases} \qquad f_0(u) = \begin{cases} \frac{1}{2}e^{u+1} & : \quad u < -1 \\ \frac{1}{2}e^{-u-1} & : \quad u \geq -1 \end{cases}$$

Therefore,

$$\Lambda(u) = \frac{e^{-|u-1|}}{e^{-|u+1|}} = \begin{cases} \frac{e^{u-1}}{e^{u+1}} = e^{-2} & : \quad u < -1 \\ \frac{e^{u-1}}{e^{-u-1}} = e^{2u} & : \quad -1 \leq u < 1 \\ \frac{e^{-u+1}}{e^{-u-1}} = e^{2} & : \quad 1 < u \end{cases}$$

The likelihood ratio $\Lambda(u)$ is nondecreasing and it crosses 1 at $u = 0$. Thus, the ML decision rule is to decide H_1 is true if $X > 0$ and decide H_0 otherwise. The error probabilities for the ML decision rule are:

$$p_{false_alarm} = \int_0^\infty f_0(u)du = \int_0^\infty \frac{e^{-u-1}}{2}du = \frac{1}{2e} \approx 0.1839.$$

By symmetry, or by a similar computation, we see that p_{miss} is also given by $p_{miss} = \frac{1}{2e}$. Of course the average error probability for the ML rule is also $\frac{1}{2e}$ for any prior distribution.

The MAP decision rule is to choose H_1 if $\Lambda(X) > \frac{\pi_0}{\pi_1} = 2$, and choose H_0 otherwise. Note that the solution of $e^{2u} = 2$ is $u = \frac{\ln 2}{2}$, which is between -1 and 1. Thus, the MAP decision rule is to choose H_1 if $X \geq \gamma_{MAP}$ and choose H_0 otherwise, where $\gamma_{MAP} = \frac{\ln 2}{2} = \ln\sqrt{2}$. The MAP decision rule for this problem is illustrated in Figure 3.10.3. For the MAP decision rule:

$$\begin{aligned} p_{false_alarm} &= \int_{\ln\sqrt{2}}^\infty f_0(u)du = \frac{1}{2}e^{-1}\int_{\ln\sqrt{2}}^\infty e^{-u}du = \frac{1}{2e\sqrt{2}} \approx 0.1301 \\ p_{miss} &= \int_{-\infty}^{\ln\sqrt{2}} f_1(u)du = \frac{1}{2}e^{-1}\int_{-\infty}^{\ln\sqrt{2}} e^{u}du = \frac{1}{e\sqrt{2}} \approx 0.2601. \end{aligned}$$

The average error probability for the MAP rule using the given prior distribution is $p_e = (2/3)p_{false_alarm} + (1/3)p_{miss} = \frac{2}{3e\sqrt{2}} = 0.1734$.

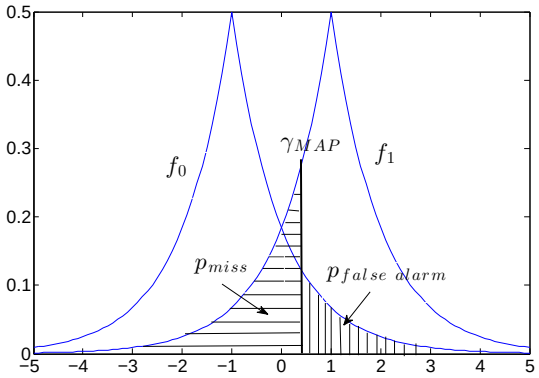


Figure 3.27: Laplace densities centered at ± 1, and γ_{MAP} for $\frac{\pi_o}{\pi_1} = 2$.

3.11 Short Answer Questions

Section 3.1[video]

1. Suppose X is a random variable with CDF $F_X(u) = \frac{1}{1+e^{-u}}$. Find $P\{(X-3)^2 \geq 4\}$.
2. Suppose X is a random variable with CDF $F_X(u) = \frac{1}{1+\exp(-\lfloor u \rfloor)}$ (where $\lfloor u \rfloor$ is the greatest integer less than or equal to u. Find $P\{|X| \leq 3\}$.

Section 3.2[video]

1. Find $\mathrm{Var}(X)$ if $f_X(u) = 2u$ for $0 \leq u \leq 1$ and $f_X(u) = 0$ elsewhere.
2. Find $E[\cos(X)]$ if $f_X(u) = \frac{\cos(u)}{2}$ for $|u| \leq \frac{\pi}{2}$, and $f_X(u) = 0$ elsewhere.

Section 3.3[video]

1. Find $\mathrm{Var}(U^2)$ if U is uniformly distributed over the interval $[0, 1]$.
2. Find $P\{U \leq 4\}$ if U is a uniformly distributed random variable with mean 5 and variance 3.

Section 3.4[video]

1. Suppose T is exponentially distributed with parameter λ. Find λ so that $P\{T \geq 5\} = \frac{1}{3}$.

2. Find the median of the exponential distribution with parameter $\lambda = .02$.

Section 3.5[video]

1. Server outages at a certain cloud computing center are well modeled by a Poisson process with an average of two outages per hour. What is the probability of four or more outages happening within each of three consecutive hour long time intervals?

2. Spurious counts in a particular photon detector occur according to a Poisson process. If the probability of no counts in a one second time interval is 0.5, what is the mean and standard deviation of the number of counts that appear in a one minute time interval?

3. Given three counts of a Poisson process occur during the interval $[0, 2]$, what is the conditional probability that no counts occur during $[0, 1]$? (Hint: The answer does not depend on the rate parameter.)

Section 3.6[video]

1. Suppose the graph of the pdf f_Y of Y has a triangular shape, symmetric about zero, and Y is standardized (mean zero and variance one). Find $P\{Y \geq 1\}$.

2. Suppose X has the $N(3, 4)$ distribution. Find $P\{X \geq 5\}$.

3. Let Y have the $N(1, 9)$ distribution. Find $P\{Y^2 \geq 4\}$.

4. Let X have the binomial distribution with parameters 100 and 0.4. Find the approximation to $P\{X \geq 45\}$ yielded by the Gaussian approximation with the continuity correction.

Section 3.7[video]

1. Let T have the pdf $f_{\sigma^2}(t) = I_{\{t \geq 0\}} \frac{t}{\sigma^2} \exp\left(-\frac{t}{2\sigma^2}\right)$ (i.e. T has the *Rayleigh* pdf with parameter σ^2) where σ^2 is unknown. Find the ML estimate of σ^2 for the observation $T = 10$.

2. Suppose X has the $N(\theta, \theta)$ distribution. (This is a Gaussian approximation of the Poisson distribution, for which the mean and variance are equal.) Find the ML estimate of θ for the observation $X = 5$.

Section 3.8[video]

1. Find the pdf of X^2 assuming X has the $N(0, \sigma^2)$ distribution for some $\sigma^2 > 0$.

2. Find the increasing function g so that if U is uniformly distributed on $[0, 1]$, then $g(U)$ has pdf $f(t) = I_{\{0 \leq t \leq 1\}} 2t$.

Section 3.9[video]

1. Find $E[T]$ assuming T is a random variable with failure rate function $h(t) = \frac{a}{1-t}$ for $0 \leq t < 1$, where $a > 0$.

2. The Pareto distribution with minimum value one and shape parameter $\alpha > 0$ has complementary CDF $P\{T \geq t\} = t^{-\alpha}$ for $t \geq 1$. Find the failure rate function, $h(t)$ for $t \geq 0$. (Hint: It is zero for $0 \leq t < 1$.)

Section 3.10[video]

1. Suppose the observation X is exponentially distributed with parameter one if H_0 is true, and is uniformly distributed over the interval $[0, 1]$ if H_1 is true. Find p_{miss} and $p_{false\ alarm}$ for the ML decision rule.

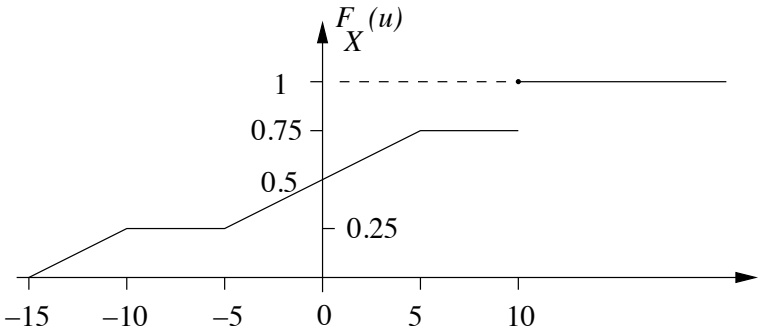
2. Suppose the observation X is exponentially distributed with parameter one if H_0 is true, and is uniformly distributed over the interval $[0, 1]$ if H_1 is true. Suppose H_0 is a priori true with probability $\pi_0 = \frac{2}{3}$. Find the average error probability for the MAP decision rule.

3. Suppose the observation X is has pdf $f_0(u) = \exp(-2|u|)$ if H_0 is true and pdf $f_1(u) = \frac{1}{2}\exp(-|u|)$ if H_1 is true. Find the largest values of prior probability for H_0, π_0, so that the MAP rule always declares that H_1 is true.

3.12 Problems

Cumulative distribution functions (CDFs) Section 3.1

3.1. **[Using a CDF I]**
Let X be a random variable with the CDF shown.



Compute the following probabilities:

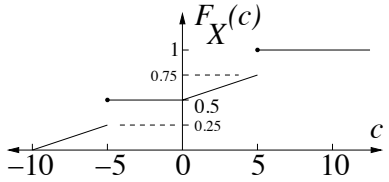
(a) $P\{X \leq 10\}$

(b) $P\{X \geq -7\}$

(c) $P\{|X| < 10\}$

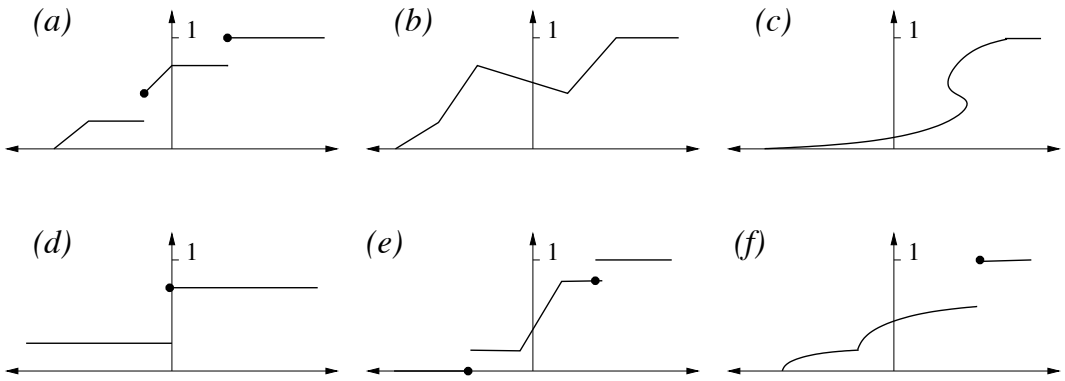
(d) $P\{X^2 \leq 16\}$

3.2. **[Using a CDF II]**
The CDF of a random variable X is shown below. Find the numerical values of the following.



(a) $P\{X = 5\}$
(b) $P\{X = 0\}$
(c) $P\{|X| \leq 5\}$
(d) $P\{X^2 \leq 4\}$

3.3. **[Recognizing a valid CDF]**
Which of the six plots below show valid CDFs? For each one that is not valid, state a property of CDFs that is violated.



Continuous-type random variables, uniform and exponential distributions, and Poisson processes Sections 3.2-3.5

3.4. **[Continuous-type random variables I]**
Consider a pdf of the following form:

$$f(u) = \begin{cases} A & u \leq 0 \\ u & 0 < u \leq 1 \\ B(2-u) & 1 < u \leq 2 \\ C & u \geq 2 \end{cases}$$

(a) Find the values of A, B, and C that make f a valid pdf.

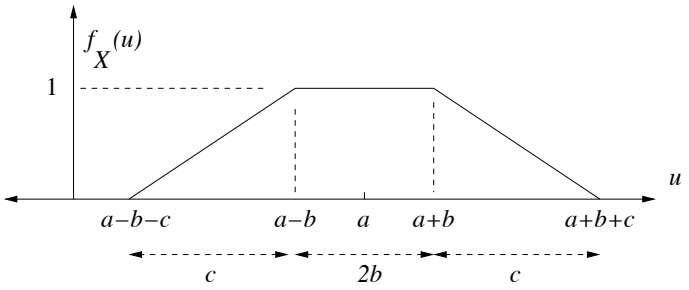
(b) Derive the CDF F(c) corresponding to f.

3.5. **[Continuous-type random variables II]**
The pdf of a random variable X is given by:

$$f_X(u) = \begin{cases} a + bu^2, & 0 \le u \le 1 \\ 0, & \text{else} \end{cases}$$

If E[X]=5/8, find a and b.

3.6. **[Selecting pdf parameters to match a mean and CDF]**
Let X be a continuous-type random variable with the pdf shown, where a, b, and c are (strictly) positive constants.



Suppose it is known $\mu_X = 2$, and the CDF of X satisfies $F_X(u) = \frac{5}{6}(u - (a - b - c))^2$ in the interval $[a - b - c, a - b]$. Find the values of a, b, and c. Show/explain your reasoning.

3.7. **[Using an exponential distribution]**
Let T be an exponentially distributed random variable with parameter $\lambda = \ln 2$.

(a) Find the simplest expression possible for $P\{T \ge t\}$ as a function of t.

(b) Find $P(T \le 1 | T \le 2)$.

3.8. **[A continuous approximation of the Zipf distribution]**
Let M be a positive integer and let $\alpha > 0$. Let Y be a random variable with the pdf

$$f_Y(u) = \begin{cases} \frac{u^{-\alpha}}{C} & 0.5 \le u \le M + 0.5 \\ 0 & \text{else.} \end{cases}$$

(Note that for integer values of u, the pdf f_Y is the same as the Zipf pmf p for X appearing in problem set 2, and for $1 \le k \le M$, $p(k) \approx \int_{k-0.5}^{k+0.5} f_Y(u)du$. An advantage of using f_Y is that it can be analytically integrated, whereas there is no closed form expression for the CDF of the Zipf pmf.)

(a) Express the constant C in terms of M and α. (For simplicity, assume $\alpha \neq 1$.)

(b) If $M = 2000$ and $\alpha = 0.8$, what is $P\{Y \leq 500.5\}$? (This approximates $P\{X \leq 500\}$ from problem set 2.)

3.9. **[Uniform and exponential distribution I]**
A fire station is to be build along a street of length L.

(a) If fires will occur at points uniformly chosen on (0, L), where should the station be located so as to minimize the expected distance from the fire? That is, choose a so as to minimize $E[|X - a|]$, when X is uniformly distributed over $(0, L)$.

(b) Now suppose the street is of infinite length- stretching from point 0 outward to ∞. If the distance of a fire from point 0 is exponentially distributed with rate λ, where should the fire station now be located?

3.10. **[Uniform and exponential distribution II]**
A factory produced two equal size batches of radios. All the radios look alike, but the lifetime of a radio in the first batch is uniformly distributed from zero to two years, while the lifetime of a radio in the second batch is exponentially distributed with parameter $\lambda = 0.1(\text{years})^{-1}$.

(a) Suppose Alicia bought a radio and after five years it is still working. What is the conditional probability it will still work for at least three more years?

(b) Suppose Venkatesh bought a radio and after one year it is still working. What is the conditional probability it will work for at least three more years?

3.11. **[Poisson process]**
A certain application in a cloud computing system is accessed on average by 15 customers per minute. Find the probability that in a one minute period, three customers access the application in the first ten seconds and two customers access the application in the last fifteen seconds. (Any number could access the system in between these two time intervals.)

3.12. **[Disk crashes modeled by a Poisson process]**
Suppose disk crashes in a particular cloud computing center occur according to a Poisson process with mean rate λ per hour, 24 hours per day, seven days a week (i.e. 24/7). Express each of the following three quantities in terms of λ, and *explain or show your work.*

(a) The expected number of disk crashes in a 24 hour period.

(b) The probability there are exactly three disk crashes in a five hour period.

(c) The mean number of hours from the beginning of a particular day, until three disks have crashed.

(d) The pdf of the time the third disk crashes.

3.13. **[Poisson buses]**
Buses arrive at a station starting at time zero according to a Poisson process $(N_t : t \geq 0)$ with rate λ, where N_t denotes the number of buses arriving up to time t.

(a) Given exactly one bus arrives during the interval $[0, 3]$, find the probability it arrives before $t = 1$. Show your work or explain your reasoning.

(b) Let T_2 be the arrival time of the second bus. What is the distribution of T_2? Find the pdf.

(c) What is the probability that there is a bus arriving exactly at time $t = 1$?

3.14. **[Poisson tweets]**
Suppose tweets from @313 follow a Poisson process with rate 1/7 per day.

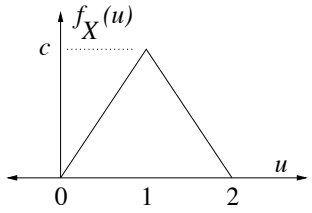
(a) What is the probability there is exactly one tweet every week for the next four weeks?

(b) What is the probability it takes more than two weeks to receive three tweets?

(c) If there have been no tweets during two weeks of waiting, what is the mean amount of additional time until five tweets have been sent?

Linear Scaling, Gaussian Distribution, ML Parameter Estimation
Sections 3.6 & 3.7

3.15. **[Standardizing a random variable with a triangular pdf]**
Suppose X has the pdf shown:



(a) Find the constant c.

(b) Let $\widetilde{X}$ denote the standardized version of X. Thus, $\widetilde{X} = \frac{X-a}{b}$ for some constants a and b so that $\widetilde{X}$ has mean zero and variance one. Carefully **sketch** the pdf of $\widetilde{X}$. Be sure to indicate both the horizontal and vertical scales of your sketch by labeling at least one nonzero point on each of the axes. (Hint: $\text{Var}(X) = \frac{1}{6}$.)

3.16. **[Gaussian distribution]**
The random variable X has a $\mathcal{N}(-4, 9)$ distribution. Express the following probabilities in terms of the Q function:

(a) $P\{X = 0\}$

(b) $P\{|X + 4| \geq 2\}$

(c) $P\{0 < X < 2\}$

(d) $P\{X^2 < 9\}$

3.17. **[A mixture of Gaussian distributions]**
Suppose X, Y and Z are three mutually independent random variables such that $X \sim N(2, 4)$, $Y \sim N(7, 9)$, and $B \sim$ Bernoulli(2/3). Let $Z = XB + Y(1 - B)$. The distribution of Z is a mixture of the distributions of X and Y.

(a) Find the pdf of Z. (Hint: First find the CDF. Use the law of total probability, taking into account the two possible cases $B = 1$ and $B = 0$.)

(b) Find $P\{Z \geq 4|B = 1\}$.

(c) Find $P\{Z = 4|B = 1\}$.

(d) Find $P\{Z \leq 4|B = 0\}$.

(e) Find $P\{Z \geq 4\}$.

3.18. **[Blind guessing answers on an exam]**
A particular exam has ten True/False questions. A student gets 3 points for each correct answer, -3 points for each wrong answer, and zero points for any unanswered question. Assume the student has no idea what the correct answers are, but attempts to answer all ten questions by randomly guessing True or False on each one. Let S be the students total score from these questions.

(a) Find the upper bound on $P\{S \geq 12\}$ yielded by the Markov inequality. Since the Markov inequality is for nonnegative random variables, apply it to the random variable X, where X is the number of correct answers.

(b) Find the upper bound on $P\{S \geq 12\}$ yielded by the Chebychev inequality and the observation that, by symmetry, $P\{S \geq 12\} = P\{S \leq -12\}$. (Use this observation to make the bound tighter.)

(c) Find the approximate value of $P\{S \geq 12\}$ using the Gaussian approximation suggested by the central limit theorem. To be definite, *use the continuity correction.*

3.19. **[Heads minus tails]**
Suppose a fair coin is flipped 100 times, and A is the event:
$A = \{$ |(number of heads) - (number of tails)| ≥ 10 $\}$.

(a) Let S denote the number of heads. Express A in terms of S. Specifically, identify which values of S make A true.

(b) Using the Gaussian approximation with the continuity correction, express the approximate value of $P(A)$ in terms of the Q function.

3.20. **[Betting with doubling or halving]**
Suppose Bob plays a series of card games at a casino. He declares prior to each game how much money he bets on the game. If he wins the game, he wins as much money as he

bet. Otherwise, he loses as much money as he bet. Assume that he wins each game with probability 0.5, and the outcomes of the games are mutually independent. Suppose he initially has \$1,048,576 ($= 2^{20}$) and, for each game, he bets one half of the money he has prior to the game. (Assume that his bet can be any positive real, i.e., \$2/3.)

(a) Let X be the number of wins in the first 196 games, and let S be the money remaining after 196 games. Express S in terms of X. Simplify your answer as much as possible.

(b) Using the Gaussian approximation, express the probability Bob has \$1 or less after 196 games in terms of the Q-function or Φ-function. Use of the continuity correction is optional. (Hint: Approximate $\log_2(3)$ by 1.6.)

3.21. **[Approximations to a Binomial Distribution]**
A communication receiver recovers a block of $n = 10^5$ bits. It is known that each bit in the block can be in error with probability 10^{-4}, independently of whether other bits are in error.

(a) Write down an exact expression for the probability of observing $k = 15$ errors in the block. A numerical value isn't required to be calculated.

(b) Determine an approximate value of $P\{X = 15\}$ via the Gaussian approximation with continuity correction.

(c) Solve part (b) using the Poisson approximation of a binomial distribution.

3.22. **[ML parameter estimation for Rayleigh and uniform distributions]**

(a) Suppose X has the pdf: $f_\theta(u) = \begin{cases} \left(\frac{u}{\theta}\right) e^{-\frac{u^2}{2\theta}} & u \geq 0 \\ 0 & u < 0 \end{cases}$ with parameter $\theta > 0$. Suppose the value of θ is unknown and it is observed that $X = 10$. Find the maximum likelihood estimate, $\widehat{\theta}_{ML}$, of θ.

(b) Suppose Y is uniformly distributed over the interval $\left[\frac{1}{a}, \frac{2}{a}\right]$, where a is an unknown positive parameter. Find the maximum likelihood estimate $\hat{a}_{ML}$ for the observation $Y = 3$.

3.23. **[ML parameter estimation of the rate of a Poisson process]**
Calls arrive in a call center according to a Poisson process with arrival rate λ (calls/minute). Derive the maximum likelihood estimate $\widehat{\lambda}_{ML}$ if k calls are received in a T minute interval.

3.24. **[ML parameter estimation for independent exponentially distributed rvs]**
The failure rate λ for a new model of high power laser is to be estimated. Suppose n lasers are tested, and let U_i be the observed lifetime of the ith laser operating at high power. Suppose that $U_1, \ldots, U_n$ are independent and each is exponentially distributed with the same parameter λ, to be estimated.

(a) Suppose it is observed that $(U_1, \ldots, U_n) = (u_1, \ldots, u_n)$ for some particular vector of positive numbers, $(u_1, \ldots, u_n)$. Write the likelihood of this observation as simply as possible in terms of λ and $(u_1, \ldots, u_n)$. (Hint: By the assumed independence, the likelihood is the product of the pdfs of the individual lasers: $f_\lambda(u_1) \cdots f_\lambda(u_n)$, where f_λ is the pdf for the exponential distribution with parameter λ.

(b) Find $\widehat{\lambda}_{ML}$ if it is observed that $(U_1, \ldots, U_n) = (u_1, \ldots, u_n)$. Simplify your answer as much as possible.

3.25. [**A hypothesis testing problem for the mean of a binomial distribution**]
Consider a binary hypothesis testing problem with observation X. Under H_0, X has the binomial distribution with parameters $n = 72$ and $p = \frac{1}{3}$. Under H_1, X has the binomial distribution with parameters $n = 72$ and $p = \frac{2}{3}$.

(a) Describe the maximum likelihood decision rule for an observation k, where k is an arbitrary integer with $0 \leq k \leq 72$. Express the rule in terms of k as simply as possible. To be definite, in case of a tie in likelihoods, declare H_1 to be the hypothesis.

(b) Suppose a particular decision rule declares that H_1 is the true hypothesis if and only if $X \geq 34$. Find the approximate value of $p_{false\ alarm}$ for this rule by using the Gaussian approximation to the binomial distribution. (To be definite, don't use the continuity correction.)

(c) Describe the MAP decision rule for an observation k, where k is an arbitrary integer with $0 \leq k \leq 72$, for the prior distribution $\pi_0 = 0.9$ and $\pi_1 = 0.1$. Express the rule in terms of k as simply as possible.

(d) Assuming the same prior distribution as in part (c), find $P(H_0|X = 38)$.

Functions of a random variable, failure rate functions, and binary hypothesis testing for continuous-type observations Sections 3.8-3.10

3.26. [**Some simple questions about a uniformly distributed random variable**]
Suppose X is uniformly distributed on the interval [0,3].

(a) Find $E[X^2]$.

(b) Find $P\{\lfloor X^2 \rfloor = 3\}$, where $\lfloor v \rfloor$ is the greatest integer less than or equal to v.

(c) Find the cumulative distribution function (CDF) of $Y = \ln X$. Be sure to specify it over the entire real line.

3.27. [**Some simple questions about an exponentially distributed random variable**]
Suppose X has the exponential distribution with parameter $\lambda > 0$. Express the answers below in terms of λ.

(a) Find $E[X^2]$.

(b) Find $P\{\lfloor X^2 \rfloor = 3\}$, where $\lfloor v \rfloor$ is the greatest integer less than or equal to v.

(c) Find the cumulative distribution function (CDF) of $Y = e^{-X}$. Be sure to specify it over the entire real line.

3.28. **[A binary quantizer with Laplacian input]**
Suppose X has pdf $f_X(u) = \frac{e^{-|u|}}{2}$ and $Y = g(X)$ where $g(u) = \alpha(\text{sign}(u)) = \begin{cases} \alpha & \text{if } u \geq 0 \\ -\alpha & \text{if } u < 0 \end{cases}$ for some positive constant α. So Y is the output of a binary quantizer with input X.

(a) Describe the pdf or pmf of Y.

(b) Find the mean square quantization error, $E[(X - Y)^2]$. Your answer should depend on α.
(Hint: $\int_0^\infty u^k e^{-u} du = k!$ for nonnegative integers k.)

(c) Find α to minimize the mean square quantization error.

3.29. **[The exponential of an exponentially distributed random variable]**
Suppose $Y = e^X$, where X is an exponentially distributed random variable with parameter λ.

(a) What is $E[Y]$? Give a simple answer depending on λ that is valid for $\lambda > 1$.

(b) What is the support of the pdf $f_Y(v)$ (the set of v for which $f_Y(v) \neq 0$)?

(c) For the set of v that you specified in part (b), find the CDF $F_Y(v)$.

3.30. **[Log uniform and log normal random variables]**

(a) Let $Z = e^U$, where U is uniformly distributed over an interval $[a, b]$. Find the pdf, f_Z. Be sure to specify it over the entire real line. (The random variable Z is said to have a log uniform distribution because $\ln(Z)$ has a uniform distribution.)

(b) Find $E[Z]$. (Hint: Use LOTUS.)

(c) Let $Y = e^X$, where X has a normal distribution. For simplicity, suppose X has mean zero and variance one. Find the pdf, f_Y. (The random variable Y is said to have a log normal distribution because $\ln(Y)$ has a normal distribution. This distribution arises as the amplitude of a signal after propagation through a heterogeneous media such as in tomography or atmospheric propagation, where attenuation factors in different parts of the media are mutually independent.)

(d) Find $E[Y]$. (Hint: Use LOTUS. Rewrite the integrand as a constant times a Gaussian pdf by completing the square in the exponent, and integrate out the pdf to get a simple answer.)

3.31. **[A simple hypothesis testing problem with continuous-type observations]**
Consider the hypothesis testing problem in which the pdf's of the observation X under hypotheses H_0 and H_1 are given, respectively, by:

$$f_0(u) = \begin{cases} \frac{1}{2} & \text{if } -1 \le u \le 1 \\ 0 & \text{otherwise} \end{cases}$$

and

$$f_1(u) = \begin{cases} |u| & \text{if } -1 \le u \le 1 \\ 0 & \text{otherwise} \end{cases}$$

Assume the priors on the hypotheses satisfy $\pi_1 = 2\pi_0$.

(a) Find the MAP rule.

(b) Find $p_{\text{false alarm}}$, p_{miss} and the average probability of error, p_e, for the MAP rule.

3.32. **[Function of a random variable]**
Let X have pdf $f_X(u) = \frac{1}{2u^2}$ for $|u| \ge 1$ and $f_X(u) = 0$ for $|u| < 1$. Let $Y = \sqrt{|X|}$.

(a) Using LOTUS, find $E[Y]$.

(b) Find the pdf of Y.

(c) (This part does not involve Y.) Find the nondecreasing function h so that $h(X)$ is uniformly distributed over $[0, 1]$. Be as explicit as possible.

3.33. **[Linearization of a quadratic function of a random variable]**
Suppose $Y = g(X)$ where $g(u) = 8u^2$ and X is uniformly distributed over $[9.9, 10.1]$. (For example, Y could be the total energy stored in a capacitor if X is the voltage across the capacitor.)

(a) Using LOTUS and the fact $\text{Var}(Y) = E[Y^2] - E[Y]^2$, find the mean and variance of Y.

(b) Find and sketch the pdf of Y.

(c) Note that X is always relatively close to 10. The first order Taylor approximation yields that $g(u) \approx g(10) + g'(10)(u - 10) = 800 + 160(u - 10)$ for u near 10. Let $Z = 800+160(X-10)$. We expect Z to be a good approximation to Y. Identify the probability distribution of Z.

(d) Find the mean and variance of Z.

(e) Your answers to (a)-(d) should show that Y and Z have nearly the same pdfs, means, and variances. To get another idea of how close Y and Z are, compute $E[(Y - Z)^2]$. (Hint: Express $(Y - Z)^2$ as a simple function of X and use LOTUS.)

3.34. **[Generation of a random variable with a given failure rate function]**
Suppose $(r(t), t \ge 0)$ is a positive, continuous function with $\int_0^\infty r(t)dt = \infty$. Let X be an exponentially distributed random variable with parameter one. Let T be implicitly determined

by X through the equation $\int_0^T r(s)ds = X$. For example, if X is the amount of water in a well at time zero, and if water is scheduled to be drawn out with a time-varying flow rate $r(t)$, then the well becomes dry at time T.

(a) Express the CDF of T in terms of the function r. (Hint: For any $t \geq 0$, the event $\{T \leq t\}$ is equivalent to $\left\{X \leq \int_0^t r(s)ds\right\}$. Using the analogy above, it is because the well is dry at time t if and only if X is less than or equal to the amount of water scheduled to be drawn out by time t.

(b) Express the failure rate function h of T in terms of the function r.

3.35. [**A simple hypothesis testing problem with continuous observations**]
On the basis of a sensor output X, it is to be decided which hypothesis is true: H_0 or H_1. Suppose that if H_0 is true then X has density f_0 and if H_1 is true then X has density f_1, where the densities are given by

$$f_0(u) = \begin{cases} \frac{1}{2} & |u| \leq 1 \\ 0 & |u| > 1 \end{cases} \qquad f_1(u) = \begin{cases} |u| & |u| \leq 1 \\ 0 & |u| > 1 \end{cases}$$

(a) Describe the ML decision rule for deciding which hypothesis is true for observation X.

(b) Find $p_{false\ alarm}$ and p_{miss} for the ML rule.

(c) Suppose it is assumed *a priori* that H_0 is true with probability π_0 and H_1 is true with probability $\pi_1 = 1 - \pi_0$. For what values of π_0 does the MAP decision rule declare H_1 with probability one, no matter which hypothesis is really true?

(d) Suppose it is assumed *a priori* that H_0 is true with probability π_0 and H_1 is true with probability $\pi_1 = 1 - \pi_0$. For what values of π_0 does the MAP decision rule declare H_0 with probability one, no matter which hypothesis is really true?

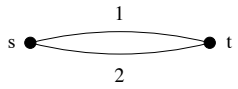
3.36. [**(COMPUTER EXERCISE) Running averages of independent, identically distributed random variables**]
Consider the following experiment, for an integer $N \geq 1$. (a) Suppose $U_1, U_2, \ldots, U_N$ are mutually independent, uniformly distributed random variables on the interval $[-0.5, 0.5]$ Let $S_n = U_1 + \cdots + U_n$ denote the cumulative sum for $1 \leq n \leq N$. Simulate this experiment on a computer and make two plots, the first showing $\frac{S_n}{n}$ for $1 \leq n \leq 100$ and the second showing $\frac{S_n}{n}$ for $1 \leq n \leq 10000$. (b) Repeat part (a), but change S_n to $S_n = Y_1 + \cdots + Y_n$ where $Y_k = \tan(\pi U_k)$. (This choice makes each Y_k have the Cauchy distribution, $f_Y(v) = \frac{1}{\pi(1+v^2)}$; see Example 3.8.6. Since the pdf f_Y is symmetric about zero, it is tempting to think that $E[Y_k] = 0$, but that is *false*; $E[Y_k]$ is not well defined because $\int_0^\infty v f_Y(v)dv = +\infty$ and $\int_{-\infty}^0 v f_Y(v)dv = -\infty$. Thus, for any function $g(n)$ defined for $n \geq 1$, it is possible to select $a_n \to -\infty$ and $b_n \to \infty$ so that $\int_{a_n}^{b_n} v f_Y(v)dv = g(n)$ for all n. It is said, therefore, that the integral $\int_{-\infty}^\infty v f_Y(v)dv$ is *indeterminate*.)

3.37. **[Generation of random variables with specified probability density function]**
Find a function g so that, if U is uniformly distributed over the interval $[0,1]$, and $X = g(U)$, then X has the pdf:

$$f_X(v) = \begin{cases} 2v & \text{if } 0 \le u \le 1 \\ 0 & \text{else.} \end{cases}$$

(Hint: Begin by finding the cumulative distribution function F_X.)

3.38. **[Failure rate of a network with two parallel links]**
Consider the $s-t$ network with two parallel links, as shown:



Suppose that for each i, link i fails at time T_i, where T_1 and T_2 are independent, exponentially distributed with some parameter $\lambda > 0$. The network fails at time T, where $T = \max\{T_1, T_2\}$.

(a) Express $F_T^c(t) = P\{T > t\}$ for $t \ge 0$ in terms of t and λ.

(b) Find the pdf of T.

(c) Find the failure rate function, $h(t)$, for the network. Simplify your answer as much as possible. (Hint: Check that your expression for h satisfies $h(0) = 0$ and $\lim_{t\to\infty} h(t) = \lambda$.)

(d) Find $P(\min\{T_1, T_2\} < t | T > t)$ and verify that $h(t) = \lambda P\{\min\{T_1, T_2\} < t | T > t)$. That is, the network failure rate at time t is λ times the conditional probability that at least one of the links has already failed by time t, given the network has not failed by time t.

3.39. **[Cauchy vs. Gaussian detection problem]**
On the basis of a sensor output X, it is to be decided which hypothesis is true: H_0 or H_1. Under H_1, X is a Gaussian random variable with mean zero and variance σ^2=0.5: $f_1(u) = \frac{1}{\sqrt{\pi}} e^{-u^2}$. Under H_0, X has the Cauchy density, $f_0(u) = \frac{1}{\pi(1+u^2)}$.

(a) Find the ML decision rule. Express it as simply as possible.

(b) Find $p_{\text{false alarm}}$ and p_{miss}. Hint: the CDFs of X under the hypotheses are $F_0(c) = 0.5 + \frac{\arctan(c)}{\pi}$ and $F_1(c) = \Phi(c\sqrt{2})$, respectively.

(c) Consider the decision rule that decides H_1 is true if $|X| \le 1.4$ and decides H_0 otherwise. This rule is the MAP rule for some prior distribution (π_1, π_0). Find the ratio $\tau = \pi_0/\pi_1$ for that distribution.

Chapter 4

Jointly Distributed Random Variables

This chapter focuses on dealing with multiple random variables that may not be independent. Earlier in these notes, we've sometimes worked with multiple random variables. For example, a Bernoulli process, as discussed in Section 2.6, is composed of independent Bernoulli random variables, and other random variables are also associated with the processes, such as the number of trials required for each one, and the cumulative number of ones up to some time n. Similarly, Poisson processes, discussed in Section 2.7, involve multiple random variables defined on the same space. But typically we considered independent random variables. In many applications, there are multiple dependent (i.e. not independent) random variables. The strength of the dependence is sometimes of primary importance. For example, researchers in medicine are often interested in the correlation between some behavior and some aspect of health, such as between smoking and heart attacks. Sometimes one random variable is observed and we wish to estimate another. For example, we may be interested in predicting the performance of a financial market based on observation of an employment statistic. Some questions that arise in this context are the following: What does it mean for a predictor or estimator to be the best? What do we need to know to compute an estimator? What computations are required and how complex are they? What if we restrict attention to linear estimators? Is there a generalization of the Gaussian distribution, central limit theorem, and Gaussian approximation for multiple dependent random variables?

4.1 Joint cumulative distribution functions

Recall that any random variable has a CDF, but pmfs (for discrete-type random variables) and pdfs (for continuous-type random variables) are usually simpler. This situation is illustrated in Figure 3.5. In this chapter we'll see that the same scheme holds for joint distributions of multiple random variables. We begin in this section with a brief look at joint CDFs.

Let X and Y be random variables on a single probability space $(\Omega, \mathcal{F}, P)$. The *joint cumulative distribution function* (CDF) is the function of two variables defined by

$$F_{X,Y}(u_o, v_o) \quad = \quad P\{X \leq u_o, Y \leq v_o\}.$$

for any $(u_o, v_o) \in \mathbb{R}^2$. (Here, $\mathbb{R}^2$ denotes the set of all pairs of real numbers, or, equivalently, the two-dimensional Euclidean plane.) That is, $F_{X,Y}(u_o, v_o)$ is the probability that the random point (X, Y) falls into the shaded region in the $u - v$ plane, shown in Figure 4.1.

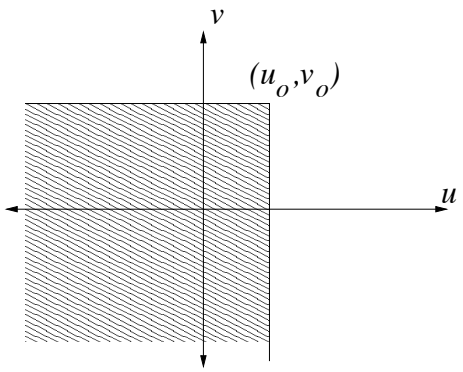


Figure 4.1: Region defining $F_{X,Y}(u_o, v_o)$.

The joint CDF determines the probabilities of all events concerning X and Y. For example, if R is the rectangular region $(a, b] \times (c, d]$ in the plane, then

$$P\{(X, Y) \in R\} = F_{X,Y}(b, d) - F_{X,Y}(b, c) - F_{X,Y}(a, d) + F_{X,Y}(a, c), \tag{4.1}$$

as illustrated in Figure 4.2. The joint CDF of X and Y also determines the probabilities of any

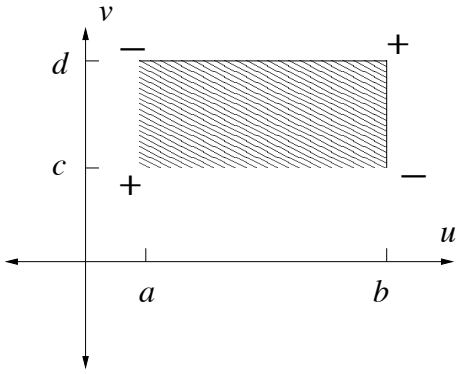


Figure 4.2: $P\{(X, Y) \in$ shaded region$\}$ is equal to $F_{X,Y}$ evaluated at the corners with signs shown.

event concerning X alone, or Y alone. To show this, we show that the CDF of X alone is determined by the joint CDF of X and Y.

Proposition 4.1.1 *Suppose X and Y have joint CDF $F_{X,Y}$. For each u fixed the limit, $\lim_{v\to\infty} F_{X,Y}(u, v)$ exists; call it $F_{X,Y}(u, \infty)$. Furthermore, $F_X(u) = F_{X,Y}(u, \infty)$. Similarly, $F_Y(v) = F_{X,Y}(\infty, v)$.*

Proof. Fix $u \in \mathbb{R}$. Let $G_0 = \{X \leq u, Y \leq 0\}$ and for $n \geq 1$ let $G_n = \{X \leq u, n-1 < Y \leq n\}$. The events $G_0, G_1, \ldots$ are mutually exclusive, and $\bigcup_{n=0}^{\infty} G_n = \{X \leq u\}$. Therefore, by Axiom P.2, about the additivity of probabilities measures,

$$\begin{aligned} F_X(u) &= \lim_{n\to\infty} P(G_0) + P(G_1) + \cdots + P(G_n) \\ &= \lim_{n\to\infty} P(G_0 \cup G_1 \cup \cdots \cup G_n) \\ &= \lim_{n\to\infty} P\{X \leq u, Y \leq n\} = \lim_{n\to\infty} F_{X,Y}(u,n) \end{aligned}$$

Also, $F_{X,Y}(u,v)$ is nondecreasing in v. Thus, $\lim_{v\to\infty} F_{X,Y}(u,v)$ exits and is equal to $F_X(u)$, as claimed. ■

Properties of CDFs The joint CDF, $F_{X,Y}$, for a pair of random variables X and Y, has the following properties. For brevity, we drop the subscripts on $F_{X,Y}$ and write it simply as F:

- $0 \leq F(u,v) \leq 1$ for all $(u,v) \in \mathbb{R}^2$
- $F(u,v)$ is nondecreasing in u and is nondecreasing in v
- $F(u,v)$ is right-continuous in u and right-continuous in v
- If $a < b$ and $c < d$, then $F(b,d) - F(b,c) - F(a,d) + F(a,c) \geq 0$
- $\lim_{u\to-\infty} F(u,v) = 0$ for each v, and $\lim_{v\to-\infty} F(u,v) = 0$ for each u
- $\lim_{u\to\infty} \lim_{v\to\infty} F(u,v) = 1$

It can be shown that any function F satisfying the above properties is the joint CDF for some pair of random variables (X,Y).

4.2 Joint probability mass functions

If X and Y are each discrete-type random variables on the same probability space, they have a *joint probability mass function* (joint pmf), denoted by $p_{X,Y}$. It is defined by $p_{X,Y}(u,v) = P\{X = u, Y = v\}$. If the numbers of possible values are small, the joint pmf can be easily described in a table. The joint pmf determines the probabilities of any events that can be expressed as conditions on X and Y. In particular, the pmfs of X and Y can be recovered from the joint pmf, using the law of total probability, as follows. By definition, since X is a discrete-type random variable, there is a finite or countably infinite set of possible values of X; denote them by $u_1, u_2, \ldots$. Similarly, there is a finite or countably infinite set of possible values of Y; denote them by $v_1, v_2, \ldots$

The events of the form $\{Y = v_j\}$ are mutually exclusive, there are at most countably infinitely many such events, and together with the zero probability event $F = \{Y \notin \{v_1, v_2, \ldots\}\}$, they

partition Ω. So for any $u \in \mathbb{R}$, the event $\{X = u\}$ can be written as a union of mutually exclusive events:

$$\{X = u\} = \left(\bigcup_j \{X = u, Y = v_j\} \right) \cup (\{X = u\}F).$$

Therefore, by the additivity axiom of probability, Axiom P.2, and the fact $P(\{X = u\}F) = 0$ (because $P(F) = 0$):

$$P\{X = u\} = \sum_j P\{X = u, Y = v_j\} \tag{4.2}$$

or equivalently,

$$p_X(u) = \sum_j p_{X,Y}(u, v_j). \tag{4.3}$$

It is useful to consider (4.2), or equivalently, (4.3), to be an instance of the law of total probability based on the partition $F, \{Y = v_1\}, \{Y = v_2\}, \cdots$ Similarly,

$$p_Y(v) = \sum_i p_{X,Y}(u_i, v)$$

In this case, p_X and p_Y are called the *marginal* pmfs of the joint pmf, $p_{X,Y}$. The *conditional pmfs* are also determined by the joint pmf. For example, the conditional pmf of Y given X, $p_{Y|X}(v|u_o)$, is defined for all u_o such that $p_X(u_o) > 0$ by:

$$p_{Y|X}(v|u_o) = P(Y = v|X = u_o) = \frac{p_{X,Y}(u_o, v)}{p_X(u_o).}$$

If $p_X(u_o) = 0$ then $p_{Y|X}(v|u_o)$ is undefined.

Example 4.2.1 Let (X, Y) have the joint pmf given by Table 4.1. The graph of the pmf is shown

Table 4.1: A simple joint pmf.

$Y = 3$	0.1	0.1	
$Y = 2$		0.2	0.2
$Y = 1$		0.3	0.1
	$X = 1$	$X = 2$	$X = 3$

in Figure 4.3. Find (a) the pmf of X, (b) the pmf of Y, (c) $P\{X = Y\}$, (d) $P\{X > Y\}$, (e) $p_{Y|X}(v|2)$, which is a function of v.

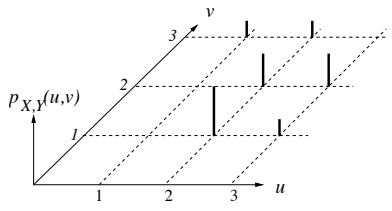


Figure 4.3: The graph of $p_{X,Y}$.

Solution: (a) The pmf of X is given by the column sums:
$p_X(1) = 0.1$, $p_X(2) = 0.3 + 0.2 + 0.1 = 0.6$, $p_X(3) = 0.1 + 0.2 = 0.3$.
(b) The pmf of Y is given by the row sums:
$p_Y(1) = 0.3 + 0.1 = 0.4$, $p_Y(2) = 0.2 + 0.2 = 0.4$, and $p_Y(3) = 0.1 + 0.1 = 0.2$.
(c) $P\{X = Y\} = p_{X,Y}(1,1) + p_{X,Y}(2,2) + p_{X,Y}(3,3) = 0 + 0.2 + 0 = 0.2$.
(d) $P\{X > Y\} = p_{X,Y}(2,1) + p_{X,Y}(3,1) + p_{X,Y}(3,2) = 0.3 + 0.1 + 0.2 = 0.6$.
(e) $p_{Y|X}(v|2) = \begin{cases} 3/6 & v = 1 \\ 2/6 & v = 2 \\ 1/6 & v = 3. \end{cases}$

The following three properties are necessary and sufficient for p to be a valid joint pmf:

Property pmf.1: p is nonnegative,

Property pmf.2 There are finite or countably infinite sets $\{u_1, u_2, \ldots\}$ and $\{v_1, v_2, \ldots\}$ such that $p(u, v) = 0$ if $u \notin \{u_1, u_2, \ldots\}$ or if $v \notin \{v_1, v_2, \ldots\}$.

Property pmf.3: $\sum_i \sum_j p(u_i, v_j) = 1$.

4.3 Joint probability density functions

The random variables X and Y are *jointly continuous-type* if there exists a function $f_{X,Y}$, called the *joint probability density function* (pdf), such that $F_{X,Y}(u_o, v_o)$ is obtained for any $(u_o, v_o) \in \mathbb{R}^2$ by integration over the shaded region in Figure 4.1:

$$F_{X,Y}(u_o, v_o) \quad = \quad \int_{-\infty}^{u_o} \int_{-\infty}^{v_o} f_{X,Y}(u, v) dv du.$$

It follows from (4.1) that if R is the rectangular region $(a, b] \times (c, d]$ in the plane, shown in Figure 4.2, then

$$P\{(X, Y) \in R\} = \int \int_R f_{X,Y}(u, v) du dv. \tag{4.4}$$

More generally, (4.4) holds for any set R in the plane that has a piecewise differentiable boundary, because such sets can be approximated by a finite or countably infinite union of disjoint rectangular

regions. If g is a function on the plane then the expectation of the random variable $g(X,Y)$ can be computed using the law of the unconscious statistician (LOTUS), just as for functions of a single random variable:

$$E[g(X,Y)] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(u,v) f_{X,Y}(u,v)\,du\,dv.$$

The following two properties are necessary and sufficient for f to be a valid joint pdf:

Property pdf.1: For any $(u,v) \in \mathbb{R}^2$, $f(u,v) \geq 0$.

Property pdf.2: The integral of f over $\mathbb{R}^2$ is one, i.e. $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(u,v)\,du\,dv = 1$.

Property pdf.1 has to be true for any pdf f because, by (4.4), the integral of f over any set is the probability of an event, which must be nonnegative. Property pdf.2 has to be true for any pdf because the integral of f over $\mathbb{R}^2$ is, by (4.4), equal to $P\{(X,Y) \in \mathbb{R}^2\}$, which is equal to one because, by definition, the pair (X,Y) always takes a value in $\mathbb{R}^2$.

Note that pdfs are defined over the entire plane $\mathbb{R}^2$, although often the pdfs are zero over large regions. The *support* of a joint pdf is the set over which it is nonzero.

The pdf of X alone or of Y alone can be obtained from the joint pdf, as follows. Given any real number u_o, let $R(u_o)$ be the region to the left of the vertical line $\{u = u_o\}$ in the (u,v) plane: $R(u_o) = \{(u,v) : -\infty < u \leq u_o, -\infty < v < \infty\}$. Then by (4.4), if X and Y are jointly continuous-type,

$$\begin{aligned} F_X(u_o) &= P\{X \leq u_o\} = P\{(X,Y) \in R(u_o)\} \\ &= \int_{-\infty}^{u_o}\left[\int_{-\infty}^{\infty} f_{X,Y}(u,v)\,dv\right] du. \end{aligned} \tag{4.5}$$

Equation (4.5) expresses $F_X(u_o)$ as the integral over $(-\infty, u_o]$, of a quantity in square brackets, for any real number u_o. So by the definition of pdfs (i.e. Definition 3.2.1), the quantity in square brackets is the pdf of X :

$$f_X(u) = \int_{-\infty}^{\infty} f_{X,Y}(u,v)\,dv. \tag{4.6}$$

Similarly,

$$f_Y(v) = \int_{-\infty}^{\infty} f_{X,Y}(u,v)\,du. \tag{4.7}$$

The pdfs f_X and f_Y are called the *marginal pdfs* of the joint distribution of X and Y. Since X is trivially a function of X and Y, the mean of X can be computed directly from the joint pdf by LOTUS:

$$E[X] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} u f_{X,Y}(u,v)\,du\,dv.$$

If the integration over v is performed first it yields the definition of $E[X]$ in terms of the marginal pdf, f_X :

$$E[X] = \int_{-\infty}^{\infty} u\left\{\int_{-\infty}^{\infty} f_{X,Y}(u,v)\,dv\right\} du = \int_{-\infty}^{\infty} u f_X(u)\,du.$$

The *conditional pdf* of Y given X, denoted by $f_{Y|X}(v|u_o)$, is undefined if $f_X(u_o) = 0$. It is defined for u_o such that $f_X(u_o) > 0$ by

$$f_{Y|X}(v|u_o) = \frac{f_{X,Y}(u_o, v)}{f_X(u_o)} \qquad -\infty < v < +\infty. \tag{4.8}$$

Graphically, the connection between $f_{X,Y}$ and $f_{Y|X}(v|u_o)$ for u_o fixed, is quite simple. For u_o fixed, the right hand side of (4.8) depends on v in only one place; the denominator, $f_X(u_o)$, is just a constant. So $f_{Y|X}(v|u_o)$ as a function of v is proportional to $f_{X,Y}(u_o, v)$, and the graph of $f_{X,Y}(u_o, v)$ with respect to v is obtained by slicing through the graph of the joint pdf along the line $u \equiv u_o$ in the $u - v$ plane. The choice of the constant $f_X(u_o)$ in the denominator in (4.8) makes $f_{Y|X}(v|u_o)$, as a function of v for u_o fixed, itself a pdf. Indeed, it is nonnegative, and

$$\int_{-\infty}^{\infty} f_{Y|X}(v|u_o)dv = \int_{-\infty}^{\infty} \frac{f_{X,Y}(u_o, v)}{f_X(u_o)}dv = \frac{1}{f_X(u_o)} \int_{-\infty}^{\infty} f_{X,Y}(u_o, v)dv = \frac{f_X(u_o)}{f_X(u_o)} = 1.$$

In practice, given a value u_o for X, we think of $f_{Y|X}(v|u_o)$ as a new pdf for Y, based on our change of view due to observing the event $X = u_o$. There is a bit of difficulty in this interpretation, because $P\{X = u_o\} = 0$ for any u_o. But if the joint density function is sufficiently regular, then the conditional density is a limit of conditional probabilities of events:

$$f_{Y|X}(v|u_o) = \lim_{\epsilon \to 0} \frac{P\left(Y \in \left(v - \frac{\epsilon}{2}, v + \frac{\epsilon}{2}\right) \middle| X \in \left(u_o - \frac{\epsilon}{2}, u_o + \frac{\epsilon}{2}\right)\right)}{\epsilon}.$$

Also, the definition of conditional pdf has the intuitive property that $f_{X,Y}(u, v) = f_{Y|X}(v|u)f_X(u)$. So (4.7) yields a version of the law of total probability for the pdf of Y :

$$f_Y(v) = \int_{-\infty}^{\infty} f_{Y|X}(v|u)f_X(u)du. \tag{4.9}$$

The expectation computed using the conditional pdf is called the *conditional expectation* (or *conditional mean*) of Y given $X = u$, written as

$$E[Y|X = u] \;\; = \;\; \int_{-\infty}^{\infty} v f_{Y|X}(v|u)dv.$$

So $E[Y|X = u]$ is a deterministic function of u. To emphasize that fact, we could denote it by g, so $g(u) = E[Y|X = u]$ for all u. If that function is applied to the random variable X, the result is a random variable denoted by $E[Y|X] = g(X)$. In summary, $E[Y|X = u]$ is a deterministic function of u. In contrast, $E[Y|X]$ is obtained by applying that same deterministic function, but to X instead of to u, so $E[Y|X]$ is a random variable. Conditional expectations play an important role in estimation of one random variable from another, as shown in Section 4.9.

Example 4.3.1 Suppose X and Y have the joint pdf:

$$f_{X,Y}(u, v) = \begin{cases} c(1 - u - v) & \text{if } u \geq 0, v \geq 0, u + v \leq 1 \\ 0 & \text{else,} \end{cases} \tag{4.10}$$

where c is a constant to be determined. The pdf and its support are shown in Figure 4.4. Find the constant c, the marginal pdf of X, and the conditional pdf of Y given X.

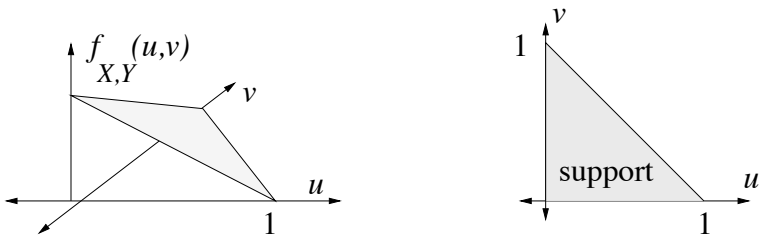


Figure 4.4: The pdf (4.10) and its support.

Solution: We first derive c by a geometric argument. The value of c is such that the integral of $f_{X,Y}$ over the plane is one. Equivalently, the volume of the region between the $u-v$ plane and the graph of $f_{X,Y}$ should be one. That region is similar to a cone or pyramid, in that it is the space between a two-dimensional base and a vertex. The volume of such regions is one third the area of the base times the height of the vertex above the plane of the base. The base in this case is the support of $f_{X,Y}$, which is a triangular region with area 1/2. The height is c. So the volume is $c/6$, which should be one, so $c = 6$.

The marginal pdf of X is given by

$$\begin{aligned} f_X(u_o) &= \int_{-\infty}^{\infty} f_{X,Y}(u_o, v)dv \\ &= \begin{cases} \int_0^{1-u_o} c(1-u_o-v)dv = \frac{c(1-u_o)^2}{2} & 0 \le u_o \le 1 \\ 0 & \text{otherwise.} \end{cases} \end{aligned}$$

Note that if we hadn't already found that $c = 1/6$, we could find c using the expression for f_X just found, because f_X must integrate to one. That yields:

$$1 = \int_{-\infty}^{\infty} f_X(u)du = \int_0^1 \frac{c(1-u)^2}{2}du = -\frac{c(1-u)^3}{6}\Big|_0^1 = \frac{c}{6},$$

or, again, $c = 6$. Thus, f_X has support $[0, 1)$ with $f_X(u_o) = 3(1-u_o)^2$ for $0 \le u_o \le 1$.

The conditional density $f_{Y|X}(v|u_o)$ is defined only if u_o is in the support of f_X: $0 \le u_o < 1$. For such u_o,

$$f_{Y|X}(v|u_o) = \begin{cases} \frac{2(1-u_o-v)}{(1-u_o)^2} & 0 \le v \le 1-u_o \\ 0 & \text{else.} \end{cases}$$

That is, the conditional pdf of Y given $X = u_o$ has a triangular shape, over the interval $[0, 1-u_o]$, as shown in Figure 4.5. This makes sense geometrically–a slice through the three dimensional region bounded by the pdf along $u \equiv u_o$ is a two-dimensional triangular region.

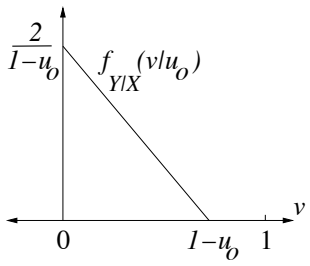


Figure 4.5: The conditional pdf of Y given $X = u_o$, where $0 \leq u_o < 1$.

Uniform joint pdfs A simple class of joint distributions are the distributions that are uniform over some set. Let S be a subset of the plane with finite area. Then the *uniform distribution over* S is the one with the following pdf:

$$f_{X,Y}(u,v) = \begin{cases} \frac{1}{\text{area of } S} & \text{if } (u,v) \in S \\ 0 & \text{else.} \end{cases}$$

If A is a subset of the plane, then

$$P\{(X,Y) \in A\} = \frac{\text{area of } A \cap S}{\text{area of } S}. \tag{4.11}$$

The support of the uniform pdf over S is simply S itself. The next two examples concern uniform distributions over two different subsets of $\mathbb{R}^2$.

Example 4.3.2 Suppose (X,Y) is uniformly distributed over the unit circle. That is, take $S = \{(u,v) : u^2 + v^2 = 1\}$. Since the area of S is π, the joint pdf is given by

$$f_{X,Y}(u,v) = \begin{cases} \frac{1}{\pi} & \text{if } u^2 + v^2 \leq 1 \\ 0 & \text{else.} \end{cases}$$

The pdf and its support are pictured in Figure 4.6. The three dimensional region under the pdf (and above the $u-v$ plane) is a cylinder centered at the origin with height $\frac{1}{\pi}$. Note that the volume of the region is 1, as required.
(a) Find $P\{(X,Y) \in A\}$ for $A = \{(u,v) : u \geq 0, v \geq 0\}$.
(b) Find $P\{X^2 + Y^2 \leq r^2\}$ for $r \geq 0$.
(c) Find the pdf of X.
(d) Find the conditional pdf of Y given X.

Solution: (a) The region $A \cap S$ is a quarter of the support, S, so by symmetry, $P\{(X,Y) \in A\} = 1/4$.
(b) If $0 \leq r \leq 1$, the region $\{(u,v) : u^2 + v^2 \leq r^2\}$ is a disk of radius r contained in S. The area of

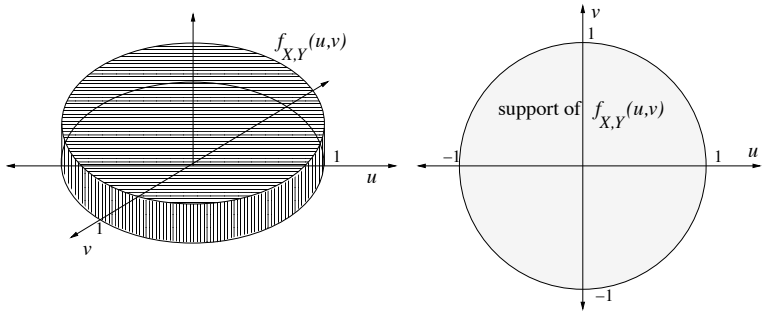


Figure 4.6: The pdf for X, Y uniformly distributed over the unit circle in $\mathbb{R}^2$

this region intersect S is thus the area of this region, which is πr^2. Dividing this by the area of S yields: $P\{X^2+Y^2 \le r\} = r^2$, for $0 \le r \le 1$. If $r > 1$, the region $\{(u,v) : u^2+v^2 \le r^2\}$ contains S, so $P\{X^2+Y^2 \le r\} = 1$ for $r > 1$.
(c) The marginal, f_X, is given by

$$f_X(u_o) \;=\; \int_{-\infty}^{\infty} f_{X,Y}(u_o,v)dv = \begin{cases} \int_{-\sqrt{1-u_o^2}}^{\sqrt{1-u_o^2}} \frac{1}{\pi} dv = \frac{2\sqrt{1-u_o^2}}{\pi} & \text{if } |u_o| \le 1 \\ 0 & \text{else.} \end{cases}$$

(d) The conditional density $f_{X|Y}(v|u_o)$ is undefined if $|u_o| \ge 1$, so suppose $|u_o| < 1$. Then

$$f_{X|Y}(v|u_o) = \begin{cases} \frac{\frac{1}{\pi}}{\frac{2\sqrt{1-u_o^2}}{\pi}} = \frac{1}{2\sqrt{1-u_o^2}} & \text{if } -\sqrt{1-u_o^2} \le v \le \sqrt{1-u_o^2} \\ 0 & \text{else.} \end{cases}$$

That is, if $|u_o| < 1$, then given $X = u_o$, Y is uniformly distributed over the interval $\left[-\sqrt{1-u_o^2}, \sqrt{1-u_o^2}\right]$. This makes sense geometrically–a slice through the cylindrically shaped region under the joint pdf is a rectangle.

Example 4.3.3 Here's a second example of a uniform distribution over a set in the plane. Suppose (X,Y) is uniformly distributed over the set $S = \{(u,v) : 0 \le u \le 1, 0 \le v \le 1, \max\{u,v\} \ge 0.5\}$. The pdf and its support set S are shown in Figure 4.7. Since the area of S is $3/4$, the pdf is:

$$f_{X,Y}(u,v) = \begin{cases} \frac{4}{3} & \text{if } (u,v) \in S \\ 0 & \text{else.} \end{cases}$$

Find the marginal and conditional pdfs.

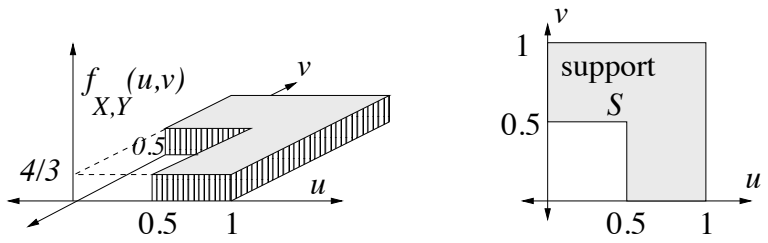


Figure 4.7: The pdf $f_{X,Y}$ and its support, S, for (X, Y) uniformly distributed over S.

Solution:

$$f_X(u_o) = \int_{-\infty}^{\infty} f_{X,Y}(u_o, v)dv = \begin{cases} \int_{0.5}^{1} \frac{4}{3}dv = \frac{2}{3} & 0 \le u_o < 0.5 \\ \int_{0}^{1} \frac{4}{3}dv = \frac{4}{3} & 0.5 \le u_o \le 1 \\ 0 & \text{else.} \end{cases}$$

The graph of f_X is shown in Figure 4.8. By symmetry, f_Y is equal to f_X.

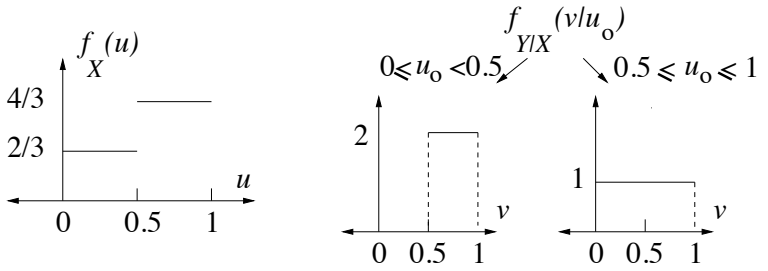


Figure 4.8: The marginal and conditional pdfs for (X, Y) uniformly distributed over S.

The conditional density $f_{Y|X}(v|u_o)$ is undefined if $u_o < 0$ or $u_o > 1$. It is well defined for $0 \le u_o \le 1$. Since $f_X(u_o)$ has different values for $u_o < 0.5$ and $u_o \ge 0.5$, we consider these two cases separately. The result is

$$f_{Y|X}(v|u_o) = \begin{cases} 2 & 0.5 \le v \le 1 \\ 0 & \text{else} \end{cases} \qquad \text{for } 0 \le u_o < 0.5,$$

$$f_{Y|X}(v|u_o) = \begin{cases} 1 & 0 \le v \le 1 \\ 0 & \text{else} \end{cases} \qquad \text{for } 0.5 \le u_o \le 1.$$

That is, if $0 \le u_o < 0.5$, the conditional distribution of Y given $X = u_o$ is the uniform distribution over the interval $[0.5, 1]$, And if $0.5 \le u_o \le 1$, the conditional distribution of Y given $X = u_o$ is the uniform distribution over the interval $[0, 1]$. The conditional distribution $f_{Y|X}(v|u_o)$ is not defined for other values of u_o. The conditional pdfs are shown in Figure 4.8.

Example 4.3.4 Suppose X and Y have the joint pdf:

$$f_{X,Y}(u,v) = \begin{cases} e^{-u} & \text{if } 0 \le v \le u \\ 0 & \text{else.} \end{cases} \tag{4.12}$$

The pdf and its support are shown in Figure 4.9. Find the marginal and conditional pdfs.

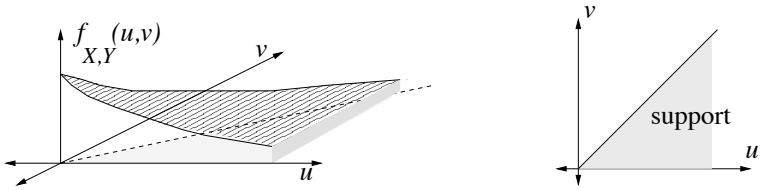


Figure 4.9: The pdf (4.12) and its support.

Solution:

$$f_X(u_o) = \begin{cases} \int_0^{u_o} e^{-u_o} dv = u_o e^{-u_o} & u_o \ge 0 \\ 0 & \text{else.} \end{cases}$$

$$f_Y(v_o) = \begin{cases} \int_{v_o}^{\infty} e^{-u} du = e^{-v_o} & v_o \ge 0 \\ 0 & \text{else.} \end{cases}$$

The pdfs are shown in Figure 4.10.

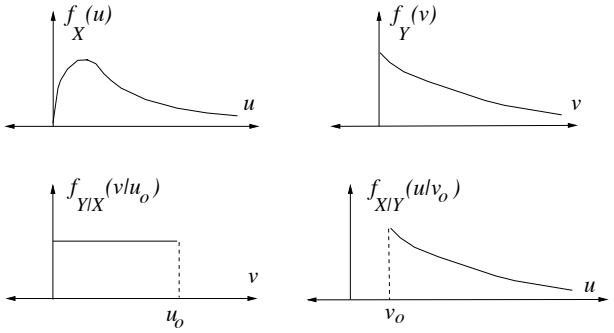


Figure 4.10: The marginal and conditional pdfs for the joint pdf in (4.12).

The conditional density $f_{Y|X}(v|u_o)$ is undefined if $u_o \le 0$. For $u_o > 0$:

$$f_{Y|X}(v|u_o) = \begin{cases} \frac{e^{-u_o}}{u_o e^{-u_o}} = \frac{1}{u_o} & 0 \le v \le u_o \\ 0 & \text{else.} \end{cases}$$

That is, the conditional distribution of Y given $X = u_o$ is the uniform distribution over the interval $[0, u_o]$. The conditional density $f_{X|Y}(u|v_o)$ is undefined if $v_o \leq 0$. For $v_o > 0$:

$$f_{X|Y}(u|v_o) = \begin{cases} \frac{e^{-u}}{e^{-v_o}} = e^{-(u-v_o)} & u \geq v_o \\ 0 & \text{else.} \end{cases}$$

That is, the conditional distribution of X given $Y = v_o$ is the exponential distribution with parameter one, shifted to the right by v_o. The conditional pdfs are shown in Figure 4.10.

Example 4.3.5 Suppose the unit interval $[0, 1]$ is randomly divided into three subintervals by two points, located as follows. The point X is uniformly distributed over the interval $[0, 1]$. Given $X = u$, a second point is uniformly distributed over the interval $[u, 1]$. Let Y denote the length of the middle subinterval. If the unit interval represents a stick, and the two points represent locations of breaks, then X and Y are the lengths of the left and center sticks formed. See Figure 4.11. Find

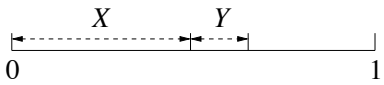


Figure 4.11: Unit interval divided into intervals of length X, Y, and $1 - X - Y$.

the pdf of Y.

Solution: We shall first determine the joint pdf of X and Y, and then integrate it to find the marginal pdf, f_Y. Basically we will be using the version of the law of total probability for pdfs in (4.9). The pdf of X is simple–X is uniformly distributed over the interval $[0, 1]$. The support of $f_{X,Y}$ is the triangular region, $T = \{(u, v) : u \geq 0, v \geq 0, \text{ and } u + v \leq 1\}$. In particular, $f_{X,Y}(u, v) = 0$ if $u \leq 0$ or if $u \geq 1$. For $0 < u < 1$, given $X = u$, the set of possible values of Y is the interval $[0, 1-u]$. Since the second random point is uniformly distributed over $[u, 1]$, and Y is the difference between that point and the constant u, the conditional distribution of Y is uniform over the interval $[0, 1 - u]$. That is, if $0 < u < 1$:

$$f_{Y|X}(v|u) = \begin{cases} \frac{1}{1-u} & 0 \leq v \leq 1 - u \\ 0 & \text{else.} \end{cases}$$

For $0 < u < 1$, $f_{X,Y}(u, v) = f_X(u) f_{Y|X}(v|u) = f_{X|Y}(v|u)$, and $f_{X,Y}(u, v) = 0$ otherwise. Therefore,

$$f_{X,Y}(u, v) = \begin{cases} \frac{1}{1-u} & (u, v) \in T \\ 0 & \text{else.} \end{cases}$$

The support of the pdf of Y is $[0, 1]$ so let $v_o \in [0, 1]$. Then

$$f_Y(v_o) = \int_{-\infty}^{\infty} f_{X,Y}(u, v_o) du = \int_0^{1-v_o} \frac{1}{1-u} du = -\ln(1-u)\Big|_0^{1-v_o} = -\ln(v_o).$$

This is infinite if $v_o = 0$, but the density at a single point doesn't make a difference, so we set $f_Y(0) = 0$. Thus, we have:

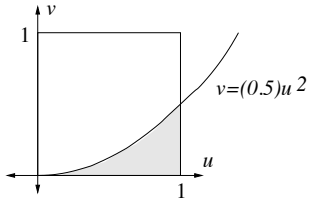
$$f_Y(v) = \begin{cases} -\ln(v) & 0 < v < 1 \\ 0 & \text{else.} \end{cases}$$

Example 4.3.6 Let $Z = \frac{Y}{X^2}$, such that X and Y have joint pdf given by

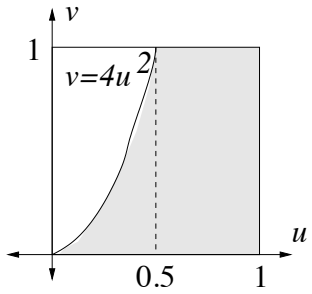
$$f_{X,Y}(u,v) = \begin{cases} 1 & \text{if } 0 \le u \le 1, 0 \le v \le 1 \\ 0 & \text{else.} \end{cases}$$

(a) Find the numerical value of $P\{Z \le 0.5\}$. Also, sketch a region in the plane so that the area of the region is $P\{Z \le 0.5\}$. (b) Find the numerical value of $P\{Z \le 4\}$. Also, sketch a region in the plane so that the area of the region is $P\{Z \le 4\}$.

Solution. (a) $P\{Z \le 0.5\} = P\{Y \le (0.5)X^2\} = \int_0^1 (0.5)u^2 du = \frac{1}{6}$.



(b) $P\{Z \le 4\} = P\{Y \le 4X^2\} = \int_0^{0.5} 4u^2 du + 0.5 = \frac{2}{3}$.



4.4 Independence of random variables

Independence of events is discussed in Section 2.4. Recall that two events, A and B, are independent, if and only if $P(AB) = P(A)P(B)$. Events A, B, and C are mutually independent if they are pairwise independent, and $P(ABC) = P(A)P(B)P(C)$. The general definition of mutual independence for n random variables was given in Section 2.4.2. Namely, $X_1, X_2, \ldots, X_n$ are mutually independent if any set of events of the form $\{X_1 \in A_1\}, \{X_2 \in A_2\}, \ldots, \{X_n \in A_n\}$ are mutually independent. In this section we cover in more detail the special case of independence for two

random variables, although factorization results are given which can be extended to the case of n mutually independent random variables.

4.4.1 Definition of independence for two random variables

Definition 4.4.1 *Random variables X and Y are defined to be independent if any pair of events of the form $\{X \in A\}$ and $\{Y \in B\}$, are independent. That is:*

$$P\{X \in A, Y \in B\} = P\{X \in A\}P\{Y \in B\}. \tag{4.13}$$

Taking A and B to be sets of the form $A = \{u : u \leq u_o\}$ and $B = \{v : v \leq v_o\}$ shows that if X and Y are independent, the joint CDF factors:

$$F_{X,Y}(u_o, v_o) = F_X(u_o)F_Y(v_o). \tag{4.14}$$

Since the CDF completely determines probabilities of the form $P\{X \in A, Y \in B\}$, it turns out that that the converse is also true: If the CDF factors (i.e. (4.14) holds for all u_o, v_o). then X and Y are independent (i.e. (4.13) holds for all A, B.

In practice, we like to use the generality of (4.13) when doing calculations, but the condition (4.14) is easier to check. To illustrate that (4.14) is stronger than it might appear, suppose that (4.14) holds for all values of u_o, v_o, and suppose $a < b$ and $c < d$. By the four point difference formula for CDFs, illustrated in Figure 4.2,

$$\begin{aligned} P\{a < X \leq b, c < Y \leq d\} &= F_X(b)F_Y(d) - F_X(b)F_Y(c) - F_X(a)F_Y(d) + F_X(a)F_Y(c) \\ &= (F_X(b) - F_X(a))(F_Y(d) - F_Y(c)) = P\{a < X \leq b\}P\{c < Y \leq d\}. \end{aligned}$$

Therefore, if (4.14) holds for all values of u_o, v_o, then (4.13) holds whenever $A = (a, b]$ and $B = (c, d]$ for some a, b, c, d. It can be shown that (4.13) also holds when A and B are finite unions of intervals, and then for all choices of A and B.

Recall that for discrete-type random variables it is usually easier to work with pmfs, and for jointly continuous-type random variables it is usually easier to work with pdfs, than with CDFs. Fortunately, in those instances, independence is also equivalent to a factorization property for a joint pmf or pdf. Therefore, discrete-type random variables X and Y are independent if and only if the joint pmf factors:

$$p_{X,Y}(u, v) = p_X(u)p_Y(v),$$

for all u, v. And for jointly continuous-type random variables, X and Y are independent if and only if the joint pdf factors:

$$f_{X,Y}(u, v) = f_X(u)f_Y(v).$$

4.4.2 Determining from a pdf whether independence holds

Suppose X and Y are jointly continuous with joint pdf $f_{X,Y}$. So X and Y are independent if and only if $f_{X,Y}(u.v) = f_X(u)f_Y(v)$ for all u and v. It takes a little practice to be able to tell, given a choice of $f_{X,Y}$, whether independence holds. Some propositions are given in this section that can often be used to help determine whether X and Y are independent. To begin, the following proposition gives a condition equivalent to independence, based on conditional pdfs.

Proposition 4.4.2 *X and Y are independent if and only if the following condition holds: For all $u \in \mathbb{R}$, either $f_X(u) = 0$ or $f_{Y|X}(v|u) = f_Y(v)$ for all $v \in \mathbb{R}$.*

Proof. (if part) If the condition holds, then for any $(u, v) \in \mathbb{R}^2$ there are two cases: (i) $f_X(u) = 0$ so $f_{X,Y}(u, v) = 0$ so $f_{X,Y}(u, v) = f_X(u)f_Y(v)$. (ii) $f_X(u) > 0$ and then $f_{X,Y}(u, v) = f_X(u)f_{Y|X}(v|u) = f_X(u)f_Y(v)$. So in either case, $f_{X,Y}(u, v) = f_X(u)f_Y(v)$. Since this is true for all $(u, v) \in \mathbb{R}^2$, X and Y are independent.
(only if part) If X and Y are independent, then $f_{X,Y}(u, v) = f_X(u)f_Y(v)$ for all $(u, v) \in \mathbb{R}^2$, so if $f_X(u) > 0$, then $f_{Y|X}(v|u) = \frac{f_{X,Y}(u,v)}{f_X(u)} = f_Y(v)$, for all $v \in \mathbb{R}$. ■

The remaining propositions have to do with the notion of product sets. Suppose A and B each consist of a finite union of disjoint finite intervals of the real line. Let $|A|$ denote the sum of the lengths of all intervals making up A, and $|B|$ denote the sum of the lengths of all intervals making up B. The *product set* of A and B, denoted by $A \times B$, is defined by $A \times B = \{(u, v) : u \in A, v \in B\}$, as illustrated in Figure 4.12. The total area of the product set, $|A \times B|$, is equal to $|A| \times |B|$.

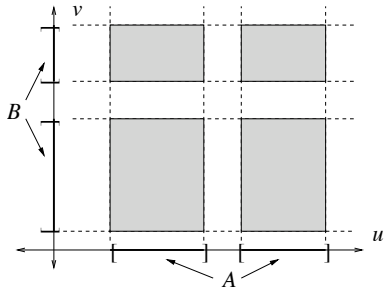


Figure 4.12: The product set $A \times B$ for sets A and B.

Definition 4.4.3 *A subset S in $\mathbb{R}^2$ has the* swap property *if for any two points $(a, b) \in S$ and $(c, d) \in S$, the points (a, d) and (b, c) are also in S.*

The swap property holds for a product set $A \times B$ because if $(a, b) \in A \times B$ and $(c, d) \in A \times B$, then both a and c are in A and both c and d are in B, so (a, d) and (b, c) must be in $A \times B$. It is not difficult to show that the converse is true too, so following proposition holds.

Proposition 4.4.4 *(swap test for product sets) Let $S \in \mathbb{R}^2$. Then S is a product set if and only if it has the swap property.*

The next proposition gives a necessary condition for X and Y to be independent. While the condition is not sufficient for independence, it is sufficient for an important special case, covered in the corollary to the proposition.

Proposition 4.4.5 *If X and Y are independent, jointly continuous-type random variables, then the support of $f_{X,Y}$ is a product set.*

Proof. By assumption, $f_{X,Y}(u,v) = f_X(u)f_Y(v)$. So if $f_{X,Y}(a,b) > 0$ and $f_{X,Y}(c,d) > 0$, it must be that $f_X(a), f_Y(b), f_X(c)$, and $f_Y(d)$ are all strictly positive, so $f_{X,Y}(a,d) > 0$ and $f_{X,Y}(b,d) > 0$. That is, the support of $f_{X,Y}$ has the swap property; it is a product set by Proposition 4.4.4. ■

Corollary 4.4.6 *Suppose (X,Y) is uniformly distributed over a set S in the plane. Then X and Y are independent if and only if S is a product set.*

Proof.(*if*) If X and Y are independent, the set S must be a product set by Proposition 4.4.5.
(only if) Conversely, if S is a product set, then $S = A \times B$ for some sets A and B, and $|S| = |A||B|$. Thus, the joint density of X and Y is given by

$$f_{X,Y}(u,v) = \begin{cases} \frac{1}{|A||B|} & u \in A, v \in B \\ 0 & \text{else.} \end{cases}$$

The standard integral formulas for marginal pdfs, (4.6) and (4.7), imply that X is uniformly distributed over A, Y is uniformly distributed over B. So $f_{X,Y}(u.v) = f_X(u)f_Y(v)$ for all u, v. Therefore, X is independent of Y. ■

Example 4.4.7 Decide whether X and Y are independent for each of the following three pdfs:
(a) $f_{X,Y}(u,v) = \begin{cases} Cu^2v^2 & u,v \geq 0, u+v \leq 1 \\ 0 & \text{else,} \end{cases}$ for an appropriate choice of C.
(b) $f_{X,Y}(u,v) = \begin{cases} u+v & u,v \in [0,1] \\ 0, & \text{else} \end{cases}$
(c) $f_{X,Y}(u,v) = \begin{cases} 9u^2v^2 & u,v \in [0,1] \\ 0, & \text{else} \end{cases}$

Solution: (a) No, X and Y are not independent. This one is a little tricky because the function Cu^2v^2 does factor into a function of u times a function of v. However, $f_{X,Y}(u,v)$ is only equal to Cu^2v^2 over a portion of the plane. One reason (only one reason is needed!) X and Y are not independent is because the support of $f_{X,Y}$ is not a product set. For example, $f_{X,Y}(0.3, 0.6) > 0$ and $f_{X,Y}(0.6, 0.3) > 0$, but $f_{X,Y}(0.6, 0.6) \not> 0$. That is, $(0.3, 0.6)$ and $(0.6, 0.3)$ are both in the suppert, but $(0.6, 0.6)$ is not, so the support fails the swap test and is not a product set by Proposition 4.4.4. Another reason X and Y are not independent is that the conditional distribution of Y given $X = u$ depends on u: for u fixed the conditional density of Y given $X = u$ has support equal to the interval $[0, 1-u]$. Since the support of the conditional pdf of Y given $X = u$ depends on u, the conditional pdf itself depends on u. So X and Y are not independent by Proposition 4.4.2.
(b) No, X and Y are not independent. One reason is that the function $u+v$ on $[0,1]^2$ does not factor into a function of u times a function of v.[1] Another reason is that the conditional distributions of Y given $X = u$ depend on u: $f_{Y|X}(v|u) = \begin{cases} \frac{u+v}{u+0.5} & v \in [0,1] \\ 0, & \text{else} \end{cases}$ for $0 \leq u \leq 1$. So X and Y are not

[1]If the density did factor, then for $0 < u_1 < u_2 < 1$ and $0 < v_1 < v_2 < 1$ the following would have to hold: $(u_1+v_1)(u_2+v_2) = (u_1+v_2)(u_2+v_1)$, or equivalently, $(u_2-u_1)(v_2-v_1) = 0$, which is a contradiction.

independent by Proposition 4.4.2.

(c) Yes, because $f_{X,Y}(u,v) = f_X(u)f_Y(v)$ where $f_X(u) = \begin{cases} 3u^2 & u \in [0,1] \\ 0, & \text{else,} \end{cases}$ and $f_Y \equiv f_X$.

4.5 Distribution of sums of random variables

Section 3.8 describes a two or three step procedure for finding the distribution of a random variable that is is a function, $g(X)$, of another random variable. If $g(X)$ is a discrete random variable, Step 1 is to scope the problem, and Step 2 is to find the pmf of $g(X)$. If $g(X)$ is a continuous-type random variable, Step 1 is to scope the problem, Step 2 is to find the CDF, and Step 3 is to differentiate the CDF to find the pdf. The same procedures work to find the pmf or pdf of a random variable that is a function of two random variables, having the form $g(X,Y)$. An important function of (X,Y) is the sum, $X+Y$. This section focuses on determining the distribution of the sum, $X+Y$, under various assumptions on the joint distribution of X and Y.

4.5.1 Sums of integer-valued random variables

Suppose $S = X+Y$, where X and Y are integer-valued random variables. We shall derive the pmf of S in terms of the joint pmf of X and Y. For a fixed value k, the possible ways to get $S = k$ can be indexed according to the value of X. That is, for $S = k$ to happen, it must be that $X = j$ and $Y = k - j$ for some value of j. Therefore, by the law of total probability,

$$\begin{aligned} p_S(k) &= P\{X+Y=k\} \\ &= \sum_j P\{X=j, Y=k-j\} \\ &= \sum_j p_{X,Y}(j,k-j). \end{aligned} \tag{4.15}$$

If X and Y are independent, then $p_{X,Y}(j,k-j) = p_X(j)p_Y(k-j)$, and (4.15) becomes:

$$p_S(k) = \sum_j p_X(j)p_Y(k-j). \tag{4.16}$$

By the definition of the convolution operation "$*$", (4.16) is equivalent to:

$$p_S = p_X * p_Y \qquad \text{(if } S = X+Y\text{, where } X, Y \text{ are independent).}$$

Example 4.5.1 Suppose X has the binomial distribution with parameters m and p, Y has the binomial distribution with parameters n and p (the same p as for X), and X and Y are independent. Describe the distribution of $S = X+Y$.

Solution: This problem can be solved with a little thought and no calculation, as follows. Recall that the distribution of X arises as the number of times heads shows if a coin with bias p is flipped m times. Then Y could be thought of as the number of times heads shows in n additional flips of the same coin. Thus, $S = X + Y$ is the number of times heads shows in $m + n$ flips of the coin. So S has the binomial distribution with parameters $m + n$ and p.

That was easy, or, at least it didn't require tedious computation. Let's try doing it another way, and at the same time get practice using the convolution formula (4.16). Since the support of X is $\{0, 1, \ldots, m\}$ and the support of Y is $\{0, 1, \ldots, n\}$, the support of S is $\{0, 1, \ldots, m + n\}$. So select an integer k with $0 \le k \le m + n$. Then (4.16) becomes

$$\begin{aligned} p_S(k) &= \sum_{j=0}^{\min(m,k)} \binom{m}{j} p^j (1-p)^{m-j} \binom{n}{k-j} p^{k-j} (1-p)^{n-k+j} \\ &= \left(\sum_{j=0}^{\min(m,k)} \binom{m}{j} \binom{n}{k-j} \right) p^k (1-p)^{m+n-k} \\ &= \binom{m+n}{k} p^k (1-p)^{m+n-k}, \end{aligned}$$

where the last step can be justified as follows. Recall that $\binom{m+n}{k}$ is the number of subsets of size k that can be formed by selecting from among $m + n$ distinct objects. Suppose the first m objects are orange and the other n objects are blue. Then the number of subsets of size k that can be made using the $m + n$ objects, such that j of the objects in the set are orange, is $\binom{m}{j}\binom{n}{k-j}$, because the j orange objects in the set can be chosen separately from the $k - j$ blue objects. Then summing over j gives $\binom{m+n}{k}$. This verifies the equation, and completes the second derivation of the fact that S has the binomial distribution with parameters $m + n$ and p.

Example 4.5.2 Suppose X and Y are independent random variables such that X has the Poisson distribution with parameter λ_1 and Y has the Poisson distribution with parameter λ_2. Describe the distribution of $S = X + Y$.

Solution: This problem can also be solved with a little thought and no calculation, as follows. Recall that the Poisson distribution is a limiting form of the binomial distribution with large n and small p. So let p be a very small positive number, and let $m = \lambda_1/p$ and $n = \lambda_2/p$. (Round to the nearest integer if necessary so that m and n are integers.) Then the distribution of X is well approximated by the binomial distribution with parameters m and p, and the distribution of Y is well approximated by the binomial distribution with parameters n and p. So by Example 4.5.1, the distribution of $X + Y$ is well approximated by the binomial distribution with parameters $m + n$ and p. But $m + n$ is large and p is small, with the product $(m + n)p = mp + np = \lambda_1 + \lambda_2$. Therefore, the distribution of $X + Y$ is well approximated by the Poisson distribution with parameter $\lambda_1 + \lambda_2$. The approximations become better and better as $p \to 0$, so we conclude that $X + Y$ has the Poisson distribution with parameter $\lambda_1 + \lambda_2$.

Let's derive the solution again, this time using the convolution formula (4.16). The support of $X+Y$ is the set of nonnegative integers, so select an integer $k \geq 0$. Then (4.16) becomes

$$\begin{aligned} p_{X+Y}(k) &= \sum_{j=0}^{k} \frac{\lambda_1^j e^{-\lambda_1}}{j!} \frac{\lambda_2^{k-j} e^{-\lambda_2}}{(k-j)!} \\ &= \left(\sum_{j=0}^{k} \frac{\lambda_1^j \lambda_2^{k-j}}{j!(k-j)!} \right) e^{-(\lambda_1+\lambda_2)} \\ &= \left(\sum_{j=0}^{k} \binom{k}{j} p^j (1-p)^{k-j} \right) \frac{\lambda^k e^{-\lambda}}{k!} \\ &= \frac{\lambda^k e^{-\lambda}}{k!}, \end{aligned}$$

where $\lambda = \lambda_1 + \lambda_2$, and $p = \frac{\lambda_1}{\lambda}$. In the last step we used the fact that the sum over the binomial pmf with parameters k and p is one. Thus, we see again that $X+Y$ has the Poisson distribution with parameter $\lambda_1 + \lambda_2$.

Example 4.5.3 Suppose X and Y represent the numbers showing for rolls of two fair dice. Thus, each is uniformly distributed over the set $\{1,2,3,4,5,6\}$ and they are independent. The convolution formula for the pmf of $S = X+Y$ is an instance of the fact that the convolution of two identical rectangle functions results in a triangle shaped function. See Fig. 4.13.

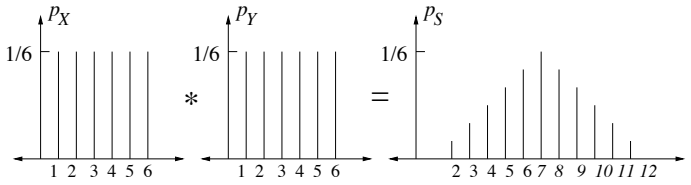


Figure 4.13: The pmf for the sum of the numbers showing for rolls of two fair dice.

4.5.2 Sums of jointly continuous-type random variables

Suppose $S = X+Y$ where X and Y are jointly continuous-type. We will express the pdf f_S in terms of the joint pdf, $f_{X,Y}$. The method will be to first find the CDF of Y and then differentiate it to get the pdf. For any $c \in \mathbb{R}$, the event $\{S \leq c\}$ is the same as the event that the random point (X,Y) in the plane falls into the shaded region of Figure 4.14. The shaded region can be integrated

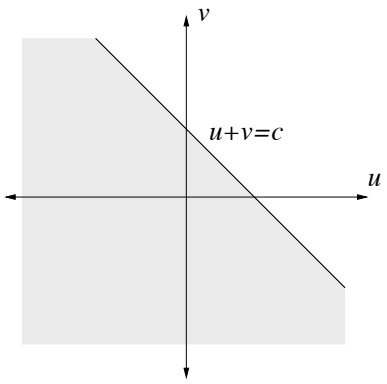


Figure 4.14: Shaded region for computation of $F_S(c)$, where $S = X + Y$.

over by integrating over all u, and for each u fixed, integrate over v from $-\infty$ to $c - u$, so

$$F_S(c) = P\{S \le c\} = \int_{-\infty}^{\infty} \left(\int_{-\infty}^{c-u} f_{X,Y}(u,v)dv \right) du.$$

Therefore,

$$\begin{aligned} f_S(c) &= \frac{dF_S(c)}{dc} = \int_{-\infty}^{\infty} \frac{d}{dc} \left(\int_{-\infty}^{c-u} f_{X,Y}(u,v)dv \right) du \\ &= \int_{-\infty}^{\infty} f_{X,Y}(u, c-u)du. \end{aligned} \tag{4.17}$$

The integral in (4.17) can be viewed as the integral of $f_{X,Y}$ over the line $u + v = c$ shown in Figure 4.14. This is an integral form of the law of total probability, because in order for $X + Y = c$, it is necessary that there is some value u such that $X = u$ and $Y = c - u$. The integral in (4.17) integrates (and integration is a type of summation) over all possible values of u.

If X and Y are independent, so that $f_{X,Y}(u,v) = f_X(u)f_Y(v)$, then (4.17) becomes

$$f_S(c) = \int_{-\infty}^{\infty} f_X(u)f_Y(c-u)du. \tag{4.18}$$

which, by the definition of the convolution operation "$*$", means

$$f_S = f_X * f_Y \qquad \text{(if } S = X + Y\text{, where } X, Y \text{ are independent)}.$$

Note the strong similarity between (4.15) and (4.16), derived for sums of integer-valued random variables, and (4.17) and (4.18), derived for sums of continuous-type random variables.

Example 4.5.4 Suppose X and Y are independent, with each being uniformly distributed over the interval $[0, 1]$. Find the pdf of $S = X + Y$.

Solution: We have $f_S = f_X * f_Y$, and we will use Figure 4.15 to help work out the integral in (4.18). As shown, the graph of the function $f_Y(c-u)$ is a rectangle over the interval $[c-1, c]$.

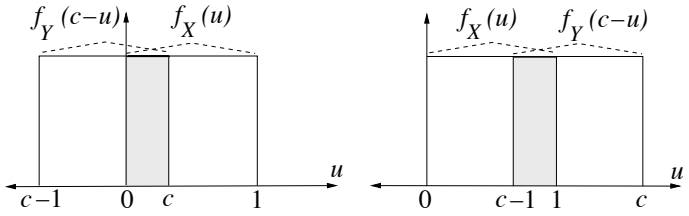


Figure 4.15: Calculating the convolution of two rectangle functions.

To check this, note that if $u = c$ then $f_Y(c-u) = f_Y(0)$, and as u decreases from c to c-1, $c-u$ increases from zero to c, so $f_Y(c-u) = 1$ for this range of u. The product $f_X(u)f_Y(c-u)$ is equal to one on $[0,1] \cup [c-1, c]$ and is equal to zero elsewhere. As shown in Figure 4.15, if $0 < c \leq 1$, then the overlap is the interval $[0, c]$, and therefore $f_X * f_Y(c) = c$. If $1 < c < 2$, the overlap is the interval $[c-1, 1]$ which has length $2-c$, and therefore $f_X * f_Y(c) = 2-c$. Otherwise, the value of the convolution is zero. Summarizing,

$$f_S(c) = \begin{cases} c & 0 < c \leq 1 \\ 2-c & 1 < c \leq 2 \\ 0 & \text{else.} \end{cases}$$

so the graph of f_S has the triangular shape shown in Figure 4.16.

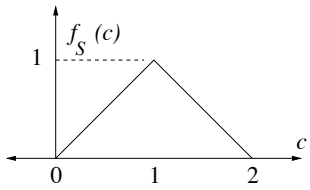


Figure 4.16: The convolution of two identical rectangle functions.

Example 4.5.5 Suppose X and Y are independent, with X having the $N(0, \sigma_1^2)$ distribution and Y having the $N(0, \sigma_2^2)$ distribution. Find the pdf of $X + Y$.

Solution:

$$\begin{aligned} f_{X+Y}(c) &= f_X * f_Y(c) = \int_{-\infty}^{\infty} f_X(u) f_Y(c-u) du \\ &= \int_{-\infty}^{\infty} f_X(u) f_Y(c-u) du \\ &= \int_{-\infty}^{\infty} \frac{1}{\sigma_1\sqrt{2\pi}} e^{-\frac{u^2}{2\sigma_1^2}} \frac{1}{\sigma_2\sqrt{2\pi}} e^{-\frac{(c-u)^2}{2\sigma_2^2}} du \\ &= \frac{1}{2\pi\sigma_1\sigma_2} \int_{-\infty}^{\infty} e^{-\frac{u^2}{2\sigma_1^2} - \frac{(c-u)^2}{2\sigma_2^2}} du. \end{aligned} \tag{4.19}$$

The integrand in (4.19) is equal to $e^{-(Au^2+2Bu+C)/2}$, where $A = \frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}$, $B = -\frac{c}{\sigma_2^2}$, and $C = \frac{c^2}{\sigma_2^2}$. But by completing the square,

$$e^{-(Au^2+2Bu+C)/2} = \left[\frac{1}{\sqrt{2\pi/A}} e^{-\frac{(u+B/A)^2}{2/A}} \right] \sqrt{\frac{2\pi}{A}} e^{B^2/2A - C/2}. \tag{4.20}$$

and the quantity in square brackets in (4.20) is a normal pdf that integrates to one, yielding the identity:

$$\int_{-\infty}^{\infty} e^{-(Au^2+2Bu+C)/2} du = \sqrt{\frac{2\pi}{A}} e^{B^2/2A - C/2}.$$

Therefore,

$$f_{X+Y}(c) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{c^2}{2\sigma^2}},$$

where $\sigma^2 = \sigma_1^2 + \sigma_2^2$. That is, the sum of two independent, mean zero Gaussian random variables is again Gaussian, and the variance of the sum, σ^2, is equal to the sum of the variances. The result can be extended to the case that X_1 and X_2 have nonzero means μ_1 and μ_2. In that case, the sum is again Gaussian, with mean $\mu = \mu_1 + \mu_2$ and variance $\sigma^2 = \sigma_1^2 + \sigma_2^2$.

The above derivation is somewhat tedious. There are two more elegant but more advanced ways to show that the convolution of two Gaussian pdfs is again a Gaussian pdf. (1) One way is to use Fourier transforms–convolution of signals is equivalent to multiplication of their Fourier transforms in the Fourier domain. The Fourier transform of the $N(0, \sigma^2)$ pdf is $\exp(-\sigma^2(2\pi f)^2/2)$, so the result is equivalent to the fact:

$$\exp(-\sigma_1^2(2\pi f)^2/2) \exp(-\sigma_2^2(2\pi f)^2/2) = \exp(-(\sigma_1^2 + \sigma_2^2)(2\pi f)^2/2).$$

(2) The other way is to use the fact that the the sum of two independent binomial random variables with the same p parameter is again a binomial random variable, and then appeal to the DeMoivre-Laplace limit theorem, stating Gaussian distributions are limits of binomial distributions. A similar approach was used in Examples 4.5.1 and 4.5.2.

4.6 Additional examples using joint distributions

Three examples are presented in this section. The first two examples continue the theme from the previous section; they illustrate, for functions of two random variables, use of the three step procedure of Section 3.8 for finding the pdf of a function $g(X, Y)$ of two random variables. A key step in the procedure is to calculate the probabilities of events involving two random variables in terms of their joint pdf. The third example also involves calculating the probability of an event determined by two random variables.

Example 4.6.1 Suppose $W = \max(X, Y)$, where X and Y are independent, continuous-type random variables. Express f_W in terms of f_X and f_Y. (Note: This example complements the analysis of the minimum of two random variables, discussed in Section 3.9 in connection with failure rate functions.)

Solution: We compute F_W first. A nice observation is that, for any constant t, $\max(X, Y) \le t$ if and only if $X \le t$ and $Y \le t$. Equivalently, the following equality holds for events: $\{\max(X, Y) \le t\} = \{X \le t\} \cap \{Y \le t\}$. By the independence of X and Y,

$$F_W(t) = P\{\max(X, Y) \le t\} = P\{X \le t\}P\{Y \le t\} = F_X(t)F_Y(t).$$

Differentiating with respect to t yields

$$f_W(t) = f_X(t)F_Y(t) + f_Y(t)F_X(t). \tag{4.21}$$

There is a nice interpretation of (4.21). If h is a small positive number, the probability that W is in the interval $(t, t + h]$ is $f_W(t)h + o(h)$, where $o(h)/h \to 0$ as $h \to 0$. There are three mutually exclusive ways that W can be in $(t, t + h]$:

$Y \le t$ and $X \in (t, t + h]$: has probability $F_Y(t)f_X(t)h + o(h)$

$X \le t$ and $Y \in (t, t + h]$: has probability $F_X(t)f_Y(t)h + o(h)$

$X \in (t, t + h]$ and $Y \in (t, t + h]$: has probability $f_X(t)hf_Y(t)h + o(h)$

So $f_W(t)h = F_Y(t)f_X(t)h + F_X(t)f_Y(t)h + f_X(t)f_Y(t)h^2 + o(h)$. Dividing this by h and letting $h \to 0$ yields (4.21).

Example 4.6.2 Suppose X and Y are jointly continuous-type random variables. Let $R = \sqrt{X^2 + Y^2}$, so R is the distance of the random point (X, Y) from the origin. Express f_R in terms of $f_{X,Y}$.

Solution: Clearly R is a nonnegative random variable. We proceed to find its CDF. Let $c > 0$, and let $D(c)$ denote the disk of radius c centered at the origin. Using polar coordinates,

$$\begin{aligned} F_R(c) &= P\{R \le c\} = P\{(X,Y) \in D(c)\} \\ &= \int\int_{D(c)} f_{X,Y}(u,v)dudv \\ &= \int_0^{2\pi}\int_0^c f_{X,Y}(r\cos(\theta), r\sin(\theta))rdrd\theta. \end{aligned}$$

Differentiating F_R to obtain f_R yields

$$\begin{aligned} f_R(c) &= \int_0^{2\pi} \frac{d}{dc}\left(\int_0^c f_{X,Y}(r\cos(\theta), r\sin(\theta))rdr\right)d\theta \\ &= \int_0^{2\pi} f_{X,Y}(c\cos(\theta), c\sin(\theta))\ c\ d\theta. \end{aligned} \tag{4.22}$$

The integral in (4.22) is just the path integral of $f_{X,Y}$ over the circle of radius c. This makes sense, because the only way R can be close to c is if (X,Y) is close to the circle, and so (4.22) is a continuous-type example of the law of total probability.

A special case is when $f_{X,Y}$ is circularly symmetric, which by definition means that $f_{X,Y}(u,v)$ depends on (u,v) only through the value $r = \sqrt{u^2+v^2}$. Equivalently, circularly symmetric means that $f_{X,Y}(u,v) = f_{X,Y}(r,0)$. So, if $f_{X,Y}$ is circularly symmetric, (4.22) simplifies to:

$$f_R(c) = (2\pi c)f_{X,Y}(c,0) \quad \text{(if } f_{X,Y} \text{ is circularly symmetric).}$$

Example 4.6.3 Buffon's needle problem Suppose a needle of unit length is thrown at random onto a large grid of lines with unit spacing. Find the probability the needle, after it comes to rest, intersects a grid line.

Solution: An important aspect of solving this problem is how to model it, and that depends on how we think about it. Here is one possible way to think about it–but there are others. Imagine that the needle lands in a plane with horizontal and vertical directions, and that the grid lines are horizontal. When the needle lands, let Θ, with $0 \le \Theta \le \pi$, denote the angle between the line going through the needle and a grid line, measured counter-clockwise from the grid line, as shown in Figure 4.17. Assume that Θ is uniformly distributed over the interval $[0,\pi]$. The vertical displacement of the needle is $\sin(\Theta)$. Let U denote the vertical distance from the lower endpoint of the needle to the first grid line above that endpoint. It is reasonable to assume that U is uniformly distributed from 0 to 1, and that U is independent of Θ. The needle intersects the grid if and only if $U \le \sin(\Theta)$. The joint density of (Θ, U), is equal to $\frac{1}{\pi}$ over the rectangular region $[0,\pi] \times [0,1]$. Integrating that density over the region $\{(\theta,u) : u \le \sin(\theta)\}$ shown in Figure 4.18 yields

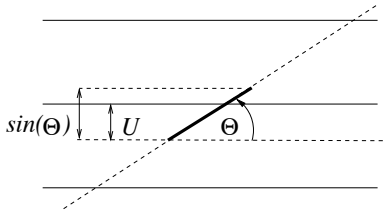


Figure 4.17: Solving the Buffon needle problem.

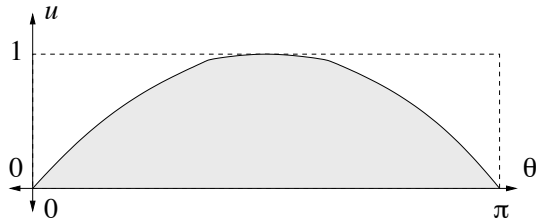


Figure 4.18: Region of integration for Buffon needle problem.

$$\begin{aligned} P\{\text{needle intersects grid}\} &= P\{U \le \sin(\Theta)\} \\ &= \int_0^{\pi}\int_0^{\sin\theta} \frac{1}{\pi} du d\theta \\ &= \int_0^{\pi} \sin(\theta)\frac{d\theta}{\pi} = -\frac{\cos(\theta)}{\pi}\Big|_0^{\pi} = \frac{2}{\pi} \approx 0.6366. \end{aligned}$$

The answer can also be written as $E[\sin(\Theta)]$. A way to interpret this is that, given Θ, the probability the needle intersects a line is $\sin(\Theta)$, and averaging over all Θ (in a continuous-type version of the law of total probability) gives the overall answer.

Letting $\alpha = P\{\text{needle intersects grid}\}$, the above shows that $\alpha = \frac{2}{\pi}$. Here is another derivation you can show your friends or students. Begin by asking "What is the mean number of times the needle crosses the grid?" Of course, the answer is α because the number of crossings is either one (with probability α) or zero (with probability $1-\alpha$). Then bend the needle to about a right angle in the middle, throw it on the grid and ask again, "What is the mean number of times the needle crosses the grid?" Convince yourself and the audience that the number is still α–bending the needle does not change the mean; each small segment of the needle has the same probability of intersecting the grid no matter the shape of the needle, and the sum of expectations is the expectation of the sum. Next ask, what if a longer bent needle, of length π, is randomly dropped onto the grid. Then what is the mean number of intersections of the needle with the grid? Convince yourself and the audience that the expected number is proportional to the length of the needle, so the expected number is $\pi\alpha$. Now, bend the length π needle into a ring of diameter one; the circumference is π. By the discussion so far, the mean number of intersections of the ring with the grid when the ring is randomly tossed onto the grid is $\pi\alpha$. But no matter where the ring lands, it intersects the grid

exactly twice. Thus, the mean number of intersections is two, which is equal to $\pi\alpha$. So, $\alpha = \frac{2}{\pi}$.

Example 4.6.4 Consider the following variation of the Buffon's needle problem (Example 4.6.3). Suppose a needle of unit length is thrown at random onto a plane with both a vertical grid and a horizontal grid, each with unit spacing. Find the probability the needle, after it comes to rest, does NOT intersect any grid line.

Solution: Let M_h be the event that the needle misses the horizontal grid (i.e. does not intersect a horizontal grid line) and let M_v denote the event that the needle misses the vertical grid. We seek to find $P(M_h M_v)$. By the solution to Buffon's needle problem, $P(M_h) = P(M_v) = 1 - \frac{2}{\pi}$. If M_h and M_v were independent, we would have that $P(M_h M_v) = (1 - \frac{2}{\pi})^2 \approx (0.363)^2 \approx 0.132$. But these events are not independent.

Let Θ be defined relative to the horizontal grid as in the solution of Buffon's needle problem. Then the vertical displacement of the needle is $\sin(\Theta)$ and the horizontal displacement is $|\cos(\Theta)|$. Assume that the position of the needle relative to the horizontal grid is independent of its position relative to the vertical grid. Let U be as in the solution to Buffon's needle problem, and let V similarly denote the distance from the leftmost endpoint of the needle to the first vertical grid line to the right of that point, as shown in Figure 4.19. Then U and V are independent, and the

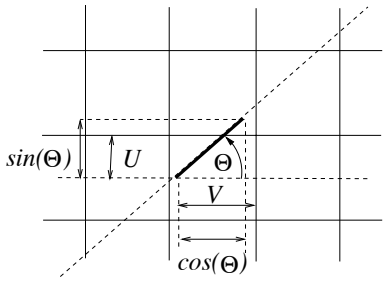


Figure 4.19: Variation of the Buffon needle problem, with horizontal and vertical grids.

needle misses both grids if and only if $U \geq \sin(\Theta)$ and $V \geq |\cos(\Theta)|$. Therefore, $P(M_h M_v | \Theta = \theta) = P\{U \geq \sin(\theta), V \geq |\cos(\theta)|\} = P\{U \geq \sin(\theta)\}P\{V \geq |\cos(\theta)|\} = (1 - \sin(\theta))(1 - |\cos(\theta)|)$.

Averaging over Θ using its pdf yields (using the trigometric identity $2\sin(\theta)\cos(\theta) = \sin(2\theta)$)

$$\begin{aligned} P(M_h M_h) &= \frac{1}{\pi}\int_0^\pi (1-\sin(\theta))(1-|\cos(\theta)|)d\theta \\ &= \frac{2}{\pi}\int_0^{\pi/2} (1-\sin(\theta))(1-\cos(\theta))d\theta \\ &= \frac{2}{\pi}\int_0^{\pi/2} 1-\sin(\theta)-\cos(\theta)-\sin(\theta)\cos(\theta)d\theta \\ &= \frac{2}{\pi}\int_0^{\pi/2} 1-\sin(\theta)-\cos(\theta)-\frac{\sin(2\theta)}{2}d\theta \\ &= \frac{2}{\pi}\left(\frac{\pi}{2}-1-1+\frac{1}{2}\right) = 1-\frac{3}{\pi} \approx 0.045. \end{aligned}$$

The true probability of missing both grids is almost three times smaller than what it would be if M_h and M_v were independent. Intuitively, the events are negatively correlated. That is because if M_h is true, it is likely that the position of the needle is more horizontal than vertical, and that makes it less likely that M_v is true. A way to express the negative correlation is to note that $P(M_v|M_h) = \frac{P(M_h M_h)}{P(M_h)} \approx \frac{0.045}{0.363} \approx 0.124$, which is nearly three times smaller than the unconditional probability, $P(M_v) \approx 0.363$.

4.7 Joint pdfs of functions of random variables

The previous two sections present examples with one random variable that is a function of two other random variables. For example, $X+Y$ is a function of (X,Y). In this section we consider the case that there are two random variables, W and Z, that are both functions of (X,Y), and we see how to determine the joint pdf of W and Z from the joint pdf of X and Y. For example, (X,Y) could represent a random point in the plane in the usual rectangular coordinates, and we may be interested in determining the joint distribution of the polar coordinates of the same point.

4.7.1 Transformation of pdfs under a linear mapping

To get started, first consider the case that W and Z are both linear functions of X and Y. It is much simpler to work with matrix notation, and following the usual convention, we represent points in the plane as column vectors. So sometimes we write $\binom{X}{Y}$ instead of (X,Y), and we write $f_{X,Y}(\binom{u}{v})$ instead of $f_{X,Y}(u,v)$.

Suppose X and Y have a joint pdf $f_{X,Y}$, and suppose $W = aX+bY$ and $Z = cX+dY$ for some constants a, b, c, and d. Equivalently, in matrix notation, suppose $\binom{X}{Y}$ has a joint pdf $f_{X,Y}$, and suppose

$$\begin{pmatrix} W \\ Z \end{pmatrix} = A\begin{pmatrix} X \\ Y \end{pmatrix} \quad \text{where} \quad A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}.$$

Thus, we begin with a random point $\binom{X}{Y}$ and get another random point $\binom{W}{Z}$. For ease of analysis, we can suppose that $\binom{X}{Y}$ is in the $u-v$ plane and $\binom{W}{Z}$ is in the $\alpha-\beta$ plane. That is, $\binom{W}{Z}$ is the image of $\binom{X}{Y}$ under the linear mapping:

$$\begin{pmatrix}\alpha\\ \beta\end{pmatrix} = A\begin{pmatrix}u\\ v\end{pmatrix}.$$

The *determinant* of A is defined by $\det(A) = ad - bc$. If $\det A \neq 0$ then the mapping has an inverse, given by

$$\begin{pmatrix}u\\ v\end{pmatrix} = A^{-1}\begin{pmatrix}\alpha\\ \beta\end{pmatrix} \text{ where } A^{-1} = \frac{1}{\det A}\begin{pmatrix} d & -b\\ -c & a\end{pmatrix}.$$

An important property of such linear transformations is that if R is a set in the $u-v$ plane and if S is the image of the set under the mapping, then $\text{area}(S) = |\det(A)|\text{area}(R)$, where $\det(A)$ is the determinant of A. Consider the problem of finding $f_{W,Z}(\alpha,\beta)$ for some fixed (α,β). If there is a small rectangle S with a corner at (α,β), then $f_{W,Z}(\alpha,\beta) \approx \frac{P\{(W,Z)\in S\}}{\text{area}(S)}$. Now $\{(W,Z)\in S\}$ is the same as $\{(X,Y)\in R\}$, where R is the preimage of S under the linear transformation. Since S is a small set, R is also a small set. So $P\{(W,Z)\in S\} = P\{(X,Y)\in R\} \approx f_{X,Y}(u,v)\text{area}(R)$. Thus, $f_{W,Z}(\alpha.\beta) \approx f_{X,Y}(u,v)\frac{\text{area}(R)}{\text{area}(S)} = \frac{1}{|\det A|}f_{X,Y}(u,v)$. This observation leads to the following proposition:

Proposition 4.7.1 *Suppose* $\binom{W}{Z} = A\binom{X}{Y}$, *where* $\binom{X}{Y}$ *has pdf* $f_{X,Y}$, *and* A *is a matrix with* $\det(A)\neq 0$. *Then* $\binom{W}{Z}$ *has joint pdf given by*

$$f_{W,Z}(\alpha,\beta) = \frac{1}{|\det A|}f_{X,Y}\left(A^{-1}\begin{pmatrix}\alpha\\ \beta\end{pmatrix}\right).$$

Example 4.7.2 Suppose X and Y have joint pdf $f_{X,Y}$, and $W = X - Y$ and $Z = X + Y$. Express the joint pdf of W and Z in terms of $f_{X,Y}$.

Solution: We apply Proposition 4.7.1, using

$$A = \begin{pmatrix}1 & -1\\ 1 & 1\end{pmatrix} \qquad \det(A) = 2 \qquad A^{-1} = \begin{pmatrix}\frac{1}{2} & \frac{1}{2}\\ -\frac{1}{2} & \frac{1}{2}\end{pmatrix}.$$

That is, the linear transformation used in this example maps $\binom{u}{v}$ to $\binom{\alpha}{\beta}$ such that $\alpha = u - v$ and $\beta = u + v$. The inverse mapping is given by $u = \frac{\alpha+\beta}{2}$ and $v = \frac{-\alpha+\beta}{2}$, or equivalently, $\binom{u}{v} = A^{-1}\binom{\alpha}{\beta}$. Proposition 4.7.1 yields:

$$f_{W,Z}(\alpha,\beta) = \frac{1}{2}f_{X,Y}\left(\frac{\alpha+\beta}{2}, \frac{-\alpha+\beta}{2}\right),$$

for all $(\alpha,\beta)\in\mathbb{R}^2$.

Example 4.7.3 Suppose X and Y are independent, continuous-type random variables. Find the joint pdf of W and Z, where $W = X + Y$ and $Z = Y$. Also, find the pdf of W.

Solution: We again apply Proposition 4.7.1, this time using

$$A = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \qquad \det(A) = 1 \qquad A^{-1} = \begin{pmatrix} 1 & -1 \\ 0 & 1 \end{pmatrix}.$$

That is, the linear transformation used in this example maps $\binom{u}{v}$ to $\binom{\alpha}{\beta}$ such that $\alpha = u + v$ and $\beta = v$. The inverse mapping is given by $u = \alpha - \beta$ and $v = \beta$. or equivalently, $\binom{u}{v} = A^{-1}\binom{\alpha}{\beta}$. Proposition 4.7.1 yields:

$$f_{W,Z}(\alpha, \beta) = f_{X,Y}(\alpha - \beta, \beta) = f_X(\alpha - \beta) f_Y(\beta),$$

for all $(\alpha, \beta) \in \mathbb{R}^2$. The marginal pdf of W is obtained by integrating out β :

$$f_W(\alpha) = \int_{\infty}^{-\infty} f_X(\alpha - \beta) f_Y(\beta) d\beta.$$

Equivalently, f_W is the convolution: $f_W = f_X * f_Y$. This expression for the pdf of the sum of two independent continuous-type random variables was found by a different method in Section 4.5.2.

4.7.2 Transformation of pdfs under a one-to-one mapping

We shall discuss next how joint pdfs are transformed for possibly nonlinear functions. Specifically, we assume that $\binom{W}{Z} = g(\binom{X}{Y})$, where g is a mapping from $\mathbb{R}^2$ to $\mathbb{R}^2$. As in the special case of linear transformations, think of the mapping as going from the $u - v$ plane to the $\alpha - \beta$ plane. Therefore, for each (u, v), there corresponds a point (α, β). This can be written as: $\alpha = g_1(u, v)$ and $\beta = g_2(u, v)$, where g_1 and g_2 are the two coordinate functions of g, each mapping $\mathbb{R}^2$ to $\mathbb{R}$. In vector notation this can be expressed by $\binom{\alpha}{\beta} = \binom{g_1(u,v)}{g_2(u,v)}$, or, for short, by $\binom{\alpha}{\beta} = g(\binom{u}{v})$.

The Jacobian of g, which we denote be J, is the matrix-valued function defined by:

$$J = J(u, v) = \begin{pmatrix} \frac{\partial g_1(u,v)}{\partial u} & \frac{\partial g_1(u,v)}{\partial v} \\ \frac{\partial g_2(u,v)}{\partial u} & \frac{\partial g_2(u,v)}{\partial v} \end{pmatrix}.$$

The Jacobian is also called the matrix derivative of g. Just as for functions of one variable, the function g near a fixed point $\binom{u_o}{v_o}$ can be approximated by a linear function:

$$g\left(\begin{pmatrix} u \\ v \end{pmatrix}\right) \approx g\left(\begin{pmatrix} u_o \\ v_o \end{pmatrix}\right) + A\left(\begin{pmatrix} u \\ v \end{pmatrix} - \begin{pmatrix} u_o \\ v_o \end{pmatrix}\right),$$

where the matrix A is given by $A = J(u_o, v_o)$. More relevant, for our purposes, is the related fact that for a small set R near a point (u, v), if S is the image of R under the mapping, then $\frac{\text{area}(S)}{\text{area}(R)} \approx |\det(J)|$.

If $\binom{W}{Z} = g(\binom{X}{Y})$, and we wish to find the pdf of $\binom{W}{Z}$ at a particular point (α, β), we need to consider values of (u, v) such that $g(u, v) = (\alpha, \beta)$. The simplest case is if there is at most one value of (u, v) so that $g(u, v) = (\alpha, \beta)$. If that is the case for all (α, β), g is said to be a *one-to-one* function. If g is one-to-one, then $g^{-1}(\alpha, \beta)$ is well-defined for all (α, β) in the range of g.

These observations lead to the following proposition:

Proposition 4.7.4 *Suppose* $\binom{W}{Z} = g(\binom{X}{Y})$, *where* $\binom{X}{Y}$ *has pdf* $f_{X,Y}$, *and* g *is a one-to-one mapping from the support of* $f_{X,Y}$ *to* $\mathbb{R}^2$. *Suppose the Jacobian* J *of* g *exists, is continuous, and has nonzero determinant everywhere. Then* $\binom{W}{Z}$ *has joint pdf given by*

$$f_{W,Z}(\alpha,\beta) = \frac{1}{|\det J|} f_{X,Y}\left(g^{-1}\left(\binom{\alpha}{\beta}\right)\right).$$

for (α,β) *in the support of* $f_{W,Z}$.

Proposition 4.7.4 generalizes Proposition 4.7.1. Indeed, suppose A is a two-by-two matrix with $\det(A) \neq 0$ and let g be the linear mapping defined by $g(u,v) = A\binom{u}{v}$. Then g satisfies the hypotheses of Proposition 4.7.4, the Jacobian function of g is constant and equal to A everywhere, and $g^{-1}(\binom{\alpha}{\beta}) = A^{-1}\binom{\alpha}{\beta}$. Therefore, the conclusion of Proposition 4.7.4 reduces to the conclusion of Proposition 4.7.1. Proposition 4.7.4 can also be viewed as the two dimensional generalization of (3.8), where g is a function of one variable and $g'(g^{-1}(c))$ plays the role that $|\det(J)|$ does in Proposition 4.7.4.

Example 4.7.5 Let X, Y have the joint pdf:

$$f_{X,Y}(u,v) = \begin{cases} u+v & (u,v) \in [0,1]^2 \\ 0 & \text{else .} \end{cases}$$

and let $W = X^2$ and $Z = X(1+Y)$. Find the pdf, $f_{W,Z}$.

Solution: The vector (X,Y) in the $u-v$ plane is transformed into the vector (W,Z) in the $\alpha-\beta$ plane under a mapping g that maps u,v to $\alpha = u^2$ and $\beta = u(1+v)$. The image in the $\alpha-\beta$ plane of the square $[0,1]^2$ (with its left side removed) in the $u-v$ plane is the set A given by

$$A = \{(\alpha,\beta) : 0 < \alpha \leq 1, \text{ and } \sqrt{\alpha} \leq \beta \leq 2\sqrt{\alpha}\}.$$

See Figure 4.20. The mapping from the square (with its left side removed) is one-to-one, because if $(\alpha,\beta) \in A$ then (u,v) can be recovered by $u = \sqrt{\alpha}$ and $v = \frac{\beta}{\sqrt{\alpha}} - 1$. The determinant of the Jacobian is

$$\det(J) = \det\begin{pmatrix} 2u & 0 \\ 1+v & u \end{pmatrix} = 2u^2 = 2\alpha.$$

Therefore, Proposition 4.7.4 yields

$$f_{W,Z}(\alpha,\beta) = \begin{cases} \frac{\sqrt{\alpha}+\frac{\beta}{\sqrt{\alpha}}-1}{2\alpha} & \text{if } (\alpha,\beta) \in A \\ 0 & \text{else.} \end{cases}$$

Example 4.7.6 Suppose X and Y are independent $N(0,\sigma^2)$ random variables. View $\binom{X}{Y}$ as a random point in the $u-v$ plane, and let (R,Θ) denote the polar coordinates of that point. Find the joint pdf of R and Θ.

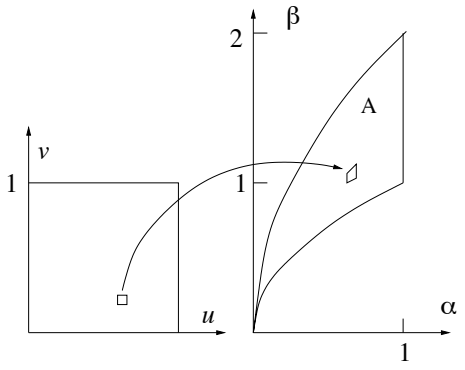


Figure 4.20: Transformation from the $u - v$ plane to the $x - y$ plane.

Solution: Changing from rectangular coordinates to polar coordinates can be viewed as a mapping g from the $u - v$ plane to the set $[0, \infty) \times [0, 2\pi)$ in the $r - \theta$ plane, where $r = (u^2 + v^2)^{\frac{1}{2}}$ and $\theta = \tan^{-1}(\frac{v}{u})$. The inverse of this mapping is given by

$$\begin{aligned} u &= r\cos(\theta) \\ v &= r\sin(\theta). \end{aligned}$$

The joint pdf of X and Y is given by

$$f_{X,Y}(u, v) = f_X(u)f_Y(v) = \frac{1}{2\pi\sigma^2} e^{-\frac{u^2+v^2}{2\sigma^2}} = \frac{1}{2\pi\sigma^2} e^{-\frac{r^2}{2\sigma^2}}.$$

The Jacobian of the mapping from (u, v) to (r, θ) is given by

$$J = \begin{pmatrix} \frac{\partial r}{\partial u} & \frac{\partial r}{\partial v} \\ \\ \frac{\partial \theta}{\partial u} & \frac{\partial \theta}{\partial v} \end{pmatrix} = \begin{pmatrix} \frac{u}{r} & \frac{v}{r} \\ \\ -\frac{v}{r^2} & \frac{u}{r^2} \end{pmatrix}.$$

and so $\det(J) = \frac{1}{r}$. (This corresponds to the well known fact from calculus that $dudv = rdrd\theta$ for changing integrals in rectangular coordinates to integrals in polar coordinates.) Therefore, Proposition 4.7.4 yields that for (r, θ) in the support $[0, \infty) \times [0, 2\pi)$ of (R, Θ),

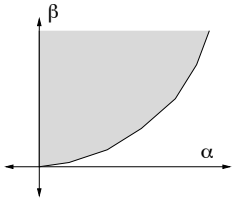
$$f_{R,\Theta}(r, \theta) = r f_{X,Y}(r\cos(\theta), r\sin(\theta)) = \frac{r}{2\pi\sigma^2} e^{-\frac{r^2}{2\sigma^2}}.$$

Of course $f_{R,\Theta}(r, \theta) = 0$ off the range of the mapping. The joint density factors into a function of r and a function of θ, so R and Θ are independent. Moreover, R has the Rayleigh distribution with parameter σ^2, and Θ is uniformly distributed on $[0, 2\pi]$. The distribution of R here could also be found using the result of Example 4.6.2, but the analysis here shows that, in addition, for the case X and Y are independent and both $N(0, \sigma^2)$, the distance R from the origin is independent of the angle θ.

The result of this example can be used to generate two independent $N(0, \sigma^2)$ random variables, X and Y, beginning with two independent random variables, A and B, that are uniform on the interval $[0, 1]$, as follows. Let $R = \sigma\sqrt{-\ln(A)}$, which can be shown to have he Rayleigh distribution with parameter σ, and let $\Theta = 2\pi B$, which is uniformly distributed on $[0, 2\pi]$. Then let $X = R\cos(\Theta)$ and $Y = R\sin(\Theta)$.

Example 4.7.7 Suppose a ball is thrown upward at time $t = 0$ with initial height X and initial upward velocity Y, such that X and Y are independent, exponentially distributed with parameter λ. The height at time t is thus $H(t) = X + Yt - \frac{ct^2}{2}$, where c is the gravitational constant. Let T denote the time the ball reaches its maximum height and M denote the maximum height. Find the joint pdf of T and M, $f_{T,M}(\alpha, \beta)$.

Solution Solving $H'(t) = 0$ to find T yields that $T = \frac{Y}{c}$ and $M = X + \frac{Y^2}{2c}$. Or $\binom{T}{M} = g\left(\binom{X}{Y}\right)$, where g is the mapping from the $u - v$ plane to the $\alpha - \beta$ plane given by $\alpha = g_1(u, v) = \frac{v}{c}$ and $\beta = g_2(u, v) = u + \frac{v^2}{2c}$. The support of $f_{X,Y}$ is the positive quadrant of the plane, and $f_{X,Y}(u, v) = f_X(u)f_Y(v) = \lambda^2 e^{-\lambda(u+v)}$ for (u, v) in the positive quadrant. The support of $f_{T,M}$ is the range of $g(u, v)$ as (u, v) ranges over the support of $f_{X,Y}$. For (α, β) to be in the support of $f_{T,M}$, it must be that $\alpha = \frac{v}{c}$ for some $v \geq 0$, or $v = \alpha c$, and then $\beta = u + \frac{v^2}{2c} = u + \frac{\alpha^2 c}{2}$. Therefore, for $\alpha \geq 0$ fixed, the point (α, β) is in the support if $\beta \geq \frac{\alpha^2 c}{2}$. That is, the support of $f_{T,M}(\alpha, \beta)$ is $\{(\alpha, \beta) : \alpha \geq 0, \beta \geq \frac{\alpha^2 c}{2}\}$, pictured:



.

The mapping g is one-to-one, because for any (α, β) in the range set, $u = \beta - \frac{\alpha^2 c}{2}$ and $v = \alpha c$. The Jacobian of g is given by

$$J = J(u, v) = \begin{pmatrix} 0 & \frac{1}{c} \\ 1 & \frac{v}{c} \end{pmatrix}.$$

so that $|\det(J)| = \frac{1}{c}$ (which happens to be constant). Therefore,

$$f_{T,M}(\alpha, \beta) = \frac{1}{|\det J|} f_{X,Y}\left(g^{-1}\left(\begin{pmatrix} \alpha \\ \beta \end{pmatrix}\right)\right) = \begin{cases} \lambda^2 c \exp(-\lambda(\beta - \frac{\alpha^2 c}{2} + \alpha c)) & \alpha \geq 0, \beta \geq \frac{\alpha^2 c}{2} \\ 0 & \text{else.} \end{cases}$$

4.7.3 Transformation of pdfs under a many-to-one mapping

In the previous subsection we considered one-to-one mappings g. If a mapping g from $\mathbb{R}^2$ to $\mathbb{R}^2$ has a continuous Jacobian with a nonzero determinant, but there are multiple points in the $u-v$ plane mapping to a single point in the $\alpha-\beta$ plane, then Proposition 4.7.4 requires a modification in order to apply. Namely, the pdf of $f_{W,Z}(\alpha,\beta)$ is given by a sum of terms, with the sum running over points of the form (u_i, v_i) such that $g(u_i, v_i) = (\alpha, \beta)$. This modification is illustrated by an example for which the mapping is two-to-one.

Example 4.7.8 Suppose $W = \min\{X, Y\}$ and $Z = \max\{X, Y\}$, where X and Y are jointly continuous-type random variables. Express $f_{W,Z}$ in terms of $f_{X,Y}$.

Solution: Note that (W, Z) is the image of (X, Y) under the mapping from the $u-v$ plane to the $\alpha-\beta$ plane defined by $\alpha = \min\{u, v\}$ and $\beta = \max\{u, v\}$, shown in Figure 4.21. This mapping

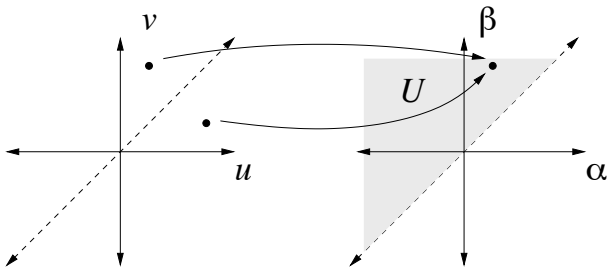


Figure 4.21: The mapping of (u, v) to $(\min\{u, v\}, \max\{u, v\})$.

maps $\mathbb{R}^2$ into the set $\{(\alpha,\beta) : \alpha \le \beta\}$, which is the set of points on or above the diagonal (i.e. the line $\alpha = \beta$) in the $\alpha-\beta$ plane. Since X and Y are jointly continuous, $P\{W = Z\} = P\{X = Y\} = \int_{-\infty}^{\infty}\int_{v}^{v} f_{X,Y}(u,v)dudv = 0$, so it does not matter how the joint density of (W, Z) is defined exactly on the diagonal; we will set it to zero there. Let $U = \{(\alpha,\beta) : \alpha < \beta\}$, which is the region that lies strictly above the diagonal. For any subset $A \subset U$, $\{(W, Z) \in A\} = \{(X, Y) \in A\} \cup \{(Y, X) \in A\}$, where the two sets in this union are disjoint. Therefore, using the fact $f_{Y,X}(u, v) = f_{X,Y}(v, u)$,

$$\begin{aligned} P\{(W,Z) \in A\} &= P\{(X,Y) \in A\} + P\{(Y,X) \in A\} \\ &= \int\int_A f_{X,Y}(u,v) + f_{Y,X}(u,v)dudv \\ &= \int\int_A f_{X,Y}(u,v) + f_{X,Y}(v,u)dudv \\ &= \int\int_A f_{X,Y}(\alpha,\beta) + f_{X,Y}(\beta,\alpha)d\alpha d\beta, \end{aligned}$$

where in the last step we simply changed the variables of integration. Consequently, the joint pdf of $f_{W,Z}$ is given by

$$f_{W,Z}(\alpha,\beta) = \begin{cases} f_{X,Y}(\alpha,\beta) + f_{X,Y}(\beta,\alpha) & \alpha < \beta \\ 0 & \alpha \geq \beta. \end{cases}$$

There are two terms on the right hand side because for each (α,β) with $\alpha < \beta$ there are two points in the (u,v) plane that map into that point: (α,β) and (β,α). No Jacobian factors such as those in Section 4.7.2 appear for this example because the mappings $(u,v) \to (u,v)$ and $(u,v) \to (v,u)$ both have Jacobians with determinant equal to one. Geometrically, to get $f_{W,Z}$ from $f_{X,Y}$ imagine spreading the probability mass for (X,Y) on the plane, and then folding the plane at the diagonal by swinging the part of the plane below the diagonal to above the diagonal, and then adding together the two masses above the diagonal. This interpretation is similar to the one found for the pdf of $|X|$, in Example 3.8.8.

4.8 Correlation and covariance

The first and second moments, or equivalently, the mean and variance, of a single random variable convey important information about the distribution of the variable, and the moments are often simpler to deal with than pmfs, pdfs, or CDFs. Use of moments is even more important when considering more than one random variable at a time. That is because joint distributions are much more complex than distributions for individual random variables.

Let X and Y be random variables with finite second moments. Three important related quantities are:

$$\begin{aligned} &\text{the correlation: } E[XY] \\ &\text{the covariance: } \mathrm{Cov}(X,Y) = E[(X - E[X])(Y - E[Y])] \\ &\text{the correlation coefficient: } \rho_{X,Y} = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}} = \frac{\mathrm{Cov}(X,Y)}{\sigma_X \sigma_Y}. \end{aligned}$$

Covariance generalizes variance, in the sense that $\mathrm{Var}(X) = \mathrm{Cov}(X,X)$. Recall that there are useful shortcuts for computing variance: $\mathrm{Var}(X) = E[X(X - E[X])] = E[X^2] - E[X]^2$. Similar shortcuts exist for computing covariances:

$$\mathrm{Cov}(X,Y) = E[X(Y - E[Y])] = E[(X - E[X])Y] = E[XY] - E[X]E[Y].$$

In particular, if either X or Y has mean zero, then $E[XY] = \mathrm{Cov}(X,Y)$.

Random variables X and Y are called *uncorrelated* if $\mathrm{Cov}(X,Y) = 0$. (If $\mathrm{Var}(X) > 0$ and $\mathrm{Var}(Y) > 0$, so that $\rho_{X,Y}$ is well defined, then X and Y being uncorrelated is equivalent to $\rho_{X,Y} = 0$.) If $\mathrm{Cov}(X,Y) > 0$, or equivalently, $\rho_{X,Y} > 0$, the variables are said to be *positively correlated*, and if $\mathrm{Cov}(X,Y) < 0$, or equivalently, $\rho_{X,Y} < 0$, the variables are said to be *negatively correlated.* If X and Y are independent, then $E[XY] = E[X]E[Y]$, which implies that X and Y are uncorrelated. The converse is false–uncorrelated does not imply independence–and in fact,

independence is a much stronger condition than being uncorrelated. Specifically, independence requires a large number of equations to hold, namely $F_{XY}(u,v) = F_X(u)F_Y(v)$ for every real value of u and v. The condition of being uncorrelated requires only a single equation to hold.

Three or more random variables are said to be uncorrelated if they are pairwise uncorrelated. That is, there is no difference between a set of random variables being uncorrelated or being pairwise uncorrelated. Recall from Section 2.4.1 that, in contrast, independence of three or more events is a stronger property than pairwise independence. Therefore, mutual independence of n random variables is a stronger property than pairwise independence. Pairwise independence of n random variables implies that they are uncorrelated.

Covariance is linear in each of its two arguments, and adding a constant to a random variable does not change the covariance of that random variable with other random variables:

$$\begin{aligned} \mathrm{Cov}(X+Y,U+V) &= \mathrm{Cov}(X,U)+\mathrm{Cov}(X,V)+\mathrm{Cov}(Y,U)+\mathrm{Cov}(Y,V) \\ \mathrm{Cov}(aX+b,cY+d) &= ac\mathrm{Cov}(X,Y), \end{aligned}$$

for constants a, b, c, d. The variance of the sum of uncorrelated random variables is equal to the sum of the variances of the random variables. For example, if X and Y are uncorrelated,

$$\mathrm{Var}(X+Y) = \mathrm{Cov}(X+Y,X+Y) = \mathrm{Cov}(X,X)+\mathrm{Cov}(Y,Y)+2\mathrm{Cov}(X,Y) = \mathrm{Var}(X)+\mathrm{Var}(Y),$$

and this calculation extends to three or more random variables

For example, consider the sum $S_n = X_1 + \cdots + X_n$, such that $X_1, \cdots, X_n$ are uncorrelated (so $\mathrm{Cov}(X_i, X_j) = 0$ if $i \neq j$) with $E[X_i] = \mu$ and $\mathrm{Var}(X_i) = \sigma^2$ for $1 \leq i \leq n$. Then

$$E[S_n] = n\mu \tag{4.23}$$

and

$$\begin{aligned} \mathrm{Var}(S_n) &= \mathrm{Cov}(S_n,S_n) = \mathrm{Cov}\left(\sum_{i=1}^n X_i, \sum_{j=1}^n X_j\right) \\ &= \sum_{i=1}^n \sum_{j=1}^n \mathrm{Cov}(X_i,X_j) \\ &= \sum_{i=1}^n \mathrm{Cov}(X_i,X_i) + \sum_{i,j:i\neq j} \mathrm{Cov}(X_i,X_j) \\ &= \sum_{i=1}^n \mathrm{Var}(X_i) + 0 = n\sigma^2. \end{aligned} \tag{4.24}$$

Therefore, the standardized version of S_n is the random variable $\frac{S_n - n\mu}{\sqrt{n\sigma^2}}$.

Example 4.8.1 Identify the mean and variance of (a) a binomial random variable with parameters n and p, (b) a negative binomial random variable with parameters r and p, and (c) an Erlang random variable with parameters r and λ.

Solution: (a) A binomial random variable with parameters n and p has the form $S = X_1 + \ldots + X_n$, where $X_1, \ldots, X_n$ are independent Bernoulli random variables with parameter p. So $E[X_i] = p$ for each i, $\text{Var}(X_i) = p(1-p)$, and, since the X_i's are independent, they are uncorrelated. Thus, by (4.23) and (4.24), $E[S] = np$ and $\text{Var}(S) = np(1-p)$.
(b) Similarly, as seen in Section 2.6, a negative binomial random variable, S_r, with parameters r and p, arises as the sum of r independent geometrically distributed random variables with parameter p. Each such geometric random variable has mean $1/p$ and variance $(1-p)/p^2$. So, $E[S_r] = \frac{r}{p}$ and $\text{Var}(S_r) = \frac{r(1-p)}{p^2}$.
(c) Likewise, as seen in Section 3.5, an Erlang random variable, T_r, with parameters r and λ, arises as the sum of r independent exponentially distributed random variables with parameter λ. An exponentially distributed random variable with parameter λ has mean $1/\lambda$ and variance $1/\lambda^2$. Therefore, $E[T_r] = \frac{r}{\lambda}$ and $\text{Var}(T_r) = \frac{r}{\lambda^2}$.

Example 4.8.2 Simplify the following expressions:
(a) $\text{Cov}(8X+3, 5Y-2)$, (b) $\text{Cov}(10X-5, -3X+15)$, (c) Cov(X+2,10X-3Y), (d) $\rho_{10X,Y+4}$.

Solution (a) $\text{Cov}(8X+3, 5Y-2) = \text{Cov}(8X, 5Y) = 40\text{Cov}(X,Y)$.
(b) $\text{Cov}(10X-5, -3X+15) = \text{Cov}(10X, -3X) = -30\text{Cov}(X,X) = -30\text{Var}(X)$.
(c) Cov(X+2,10X-3Y)=Cov(X,10X-3Y)=10Cov(X,X)-3Cov(X,Y)=10Var(X)-3Cov(X,Y).
(d) Since $\text{Cov}(10X, Y+4) = 10\text{Cov}(X,Y)$, the standard deviation of $10X$ is $10\sigma_X$, and the standard deviation of $Y+4$ is σ_Y, $\rho_{10X,Y+4} = \frac{10\text{Cov}(X,Y)}{(10\sigma_X)\sigma_Y} = \frac{\text{Cov}(X,Y)}{\sigma_X\sigma_Y} = \rho_{X,Y}$.

It is clear from the definition that the correlation coefficient $\rho_{X,Y}$ is a scaled version of $\text{Cov}(X,Y)$. The units that $E[XY]$ or $\text{Cov}(X,Y)$ are measured in are the product of the units that X is measured in times the units that Y is measured in. For example, if X is in kilometers and Y is in seconds, then $\text{Cov}(X,Y)$ is in kilometer-seconds. If we were to change units of the first variable to meters, then X in kilometers would be changed to $1000X$ in meters, and the covariance between the new measurement in meters and Y would be $\text{Cov}(1000X, Y) = 1000\text{Cov}(X,Y)$, which would be measured in meter-seconds. In contrast, the correlation coefficient $\rho_{X,Y}$ is dimensionless–it carries no units. That is because the units of the denominator, $\sigma_X\sigma_Y$, in the definition of $\rho_{X,Y}$, are the units of X times the units of Y, which are also the units of the numerator, $\text{Cov}(X,Y)$. The situation is similar to the use of the standardized versions of random variables X and Y, namely $\frac{X-E[X]}{\sigma_X}$ and $\frac{Y-E[Y]}{\sigma_Y}$. These standardized versions have mean zero, variance one, and are dimensionless. In fact, the covariance between the standardized versions of X and Y is $\rho_{X,Y}$:

$$\text{Cov}\left(\frac{X-E[X]}{\sigma_X}, \frac{Y-E[Y]}{\sigma_Y}\right) = \text{Cov}\left(\frac{X}{\sigma_X}, \frac{Y}{\sigma_Y}\right) = \frac{\text{Cov}(X,Y)}{\sigma_X\sigma_Y} = \rho_{X,Y}.$$

If the units of X or Y are changed (by linear or affine scaling, such as changing from kilometers to meters, or degrees C to degrees F) the correlation coefficient does not change:

$$\rho_{aX+b,cY+d} = \rho_{X,Y} \quad \text{for } a, c > 0.$$

In a sense, therefore, the correlation coefficient $\rho_{X,Y}$ is the standardized version of the covariance, $\mathrm{Cov}(X,Y)$, or of the correlation, $E[XY]$. As shown in the corollary of the following proposition, correlation coefficients are always in the interval $[-1,1]$. As shown in Section 4.9, covariance or correlation coefficients play a central role for estimating Y by a linear function of X. Positive correlation (i.e. $\rho_{X,Y} > 0$) means that X and Y both tend to be large or both tend to be small, whereas a negative correlation (i.e. $\rho_{X,Y} < 0$) means that X and Y tend to be opposites: if X is larger than average it tends to indicate that Y is smaller than average. The extreme case $\rho_{X,Y} = 1$ means Y can be perfectly predicted by a linear function $aX + b$ with $a > 0$, and the extreme case $\rho_{X,Y} = -1$ means Y can be perfectly predicted by a linear function $aX + b$ with $a < 0$. As mentioned earlier in this section, X and Y are said to be uncorrelated if $\mathrm{Cov}(X,Y) = 0$, and being uncorrelated does not imply independence.

Proposition 4.8.3 (Schwarz's inequality) *For two random variables X and Y :*

$$|E[XY]| \leq \sqrt{E[X^2]E[Y^2]}. \tag{4.25}$$

Furthermore, if $E[X^2] \neq 0$, equality holds in (4.25) *(i.e. $|E[XY]| = \sqrt{E[X^2]E[Y^2]}$) if and only if $P\{Y = cX\} = 1$ for some constant c.*

Proof. Take $\lambda = E[XY]/E[X^2]$ and note that

$$\begin{aligned} 0 \leq E[(Y - \lambda X)^2] &= E[Y^2] - 2\lambda E[XY] + \lambda^2 E[X^2] \\ &= E[Y^2] - \frac{E[XY]^2}{E[X^2]}, \end{aligned} \tag{4.26}$$

which implies that $E[XY]^2 \leq E[X^2]E[Y^2]$, which is equivalent to the Schwarz inequality. If $P\{Y = cX\} = 1$ for some c then equality holds in (4.25), and conversely, if equality holds in (4.25) then equality also holds in (4.26), so $E[(Y - \lambda X)^2] = 0$, and therefore $P\{Y = cX\} = 1$ for $c = \lambda$. ■

Corollary 4.8.4 *For two random variables X and Y,*

$$|Cov(X,Y)| \leq \sqrt{Var(X)\,Var(Y)}.$$

Furthermore, if $Var(X) \neq 0$ then equality holds if and only if $Y = aX + b$ for some constants a and b. Consequently, if $Var(X)$ and $Var(Y)$ are not zero, so the correlation coefficient $\rho_{X,Y}$ is well defined, then

- $|\rho_{X,Y}| \leq 1$,
- $\rho_{X,Y} = 1$ *if and only if $Y = aX + b$ for some a, b with $a > 0$, and*
- $\rho_{X,Y} = -1$ *if and only if $Y = aX + b$ for some a, b with $a < 0$.*

Proof. The corollary follows by applying the Schwarz inequality to the random variables $X - E[X]$ and $Y - E[Y]$. ■

Example 4.8.5 Suppose n fair dice are independently rolled. Let

$$X_k = \begin{cases} 1 & \text{if one shows on the } k^{th} \text{ die} \\ 0 & \text{else} \end{cases} \qquad Y_k = \begin{cases} 1 & \text{if two shows on the } k^{th} \text{ die} \\ 0 & \text{else.} \end{cases}$$

Let $X = \sum_{k=1}^{n} X_k$, which is the number of one's showing, and $Y = \sum_{k=1}^{n} Y_k$, which is the number of two's showing. Note that if a histogram is made recording the number of occurrences of each of the six numbers, then X and Y are the heights of the first two entries in the histogram.
(a) Find $E[X_1]$ and $\text{Var}(X_1)$.
(b) Find $E[X]$ and $\text{Var}(X)$.
(c) Find $\text{Cov}(X_i, Y_j)$ if $1 \leq i \leq n$ and $1 \leq j \leq n$ (Hint: Does it make a difference if $i = j$?)
(d) Find $\text{Cov}(X, Y)$.
(e) Find the correlation coefficient $\rho_{X,Y}$. Are X and Y positively correlated, uncorrelated, or negatively correlated?

Solution (a) Each X_k is a Bernoulli random variable with parameter $p = \frac{1}{6}$, so $E[X_k] = \frac{1}{6}$ and $\text{Var}(X_k) = E[X_k^2] - E[X_k]^2 = p - p^2 = p(1-p) = \frac{5}{36}$.
(b) $E[X] = nE[X_1] = \frac{n}{6}$, and $\text{Var}(X) = n\text{Var}(X_1) = \frac{5n}{36}$.
(c) If $i \neq j$ then $\text{Cov}(X_i, X_j) = 0$, because X_i and Y_j are associated with different, independent, dice rolls. If $i = j$ the situation is different. The joint pmf of X_i and Y_i for any i is:

$$p_{X_i,Y_i}(1,0) = \frac{1}{6} \qquad p_{X_i,Y_i}(0,1) = \frac{1}{6} \qquad p_{X_i,Y_i}(0,0) = \frac{4}{6}.$$

In particular, $X_i Y_i$ is always equal to zero, because it is not possible for both a one and a two to show on a single roll of the die. Thus, $E[X_i Y_i] = 0$. Therefore, $\text{Cov}(X_i, Y_i) = E[X_i Y_i] - E[X_i]E[Y_i] = 0 - \frac{1}{6}\frac{1}{6} = -\frac{1}{36}$. Not surprisingly, X_i and Y_i are negatively correlated.
(d) Using the answer to part (c) yields

$$\begin{aligned} \text{Cov}(X,Y) &= \text{Cov}\left(\sum_{i=1}^{n} X_i, \sum_{j=1}^{n} Y_j\right) \\ &= \sum_{i=1}^{n}\sum_{j=1}^{n} \text{Cov}(X_i, Y_j) \\ &= \sum_{i=1}^{n} \text{Cov}(X_i, Y_i) + \sum_{i,j:i\neq j} \text{Cov}(X_i, Y_j) \\ &= \sum_{i=1}^{n} \left(-\frac{1}{36}\right) \quad + \quad 0 = -\frac{n}{36}. \end{aligned}$$

(e) Using the definition of correlation coefficient and answers to (b) and (d) yields:

$$\rho_{X,Y} = \frac{-\frac{n}{36}}{\sqrt{\frac{5n}{36}\frac{5n}{36}}} = -\frac{1}{5}.$$

Since $\mathrm{Cov}(X,Y) < 0$, or, equivalently, $\rho_{X,Y} < 0$, X and Y are negatively correlated. This makes sense; if X is larger than average, it means that there were more one's showing than average, which would imply that there should be somewhat fewer two's showing than average.

Example 4.8.6 Suppose $X_1, \ldots, X_n$ are independent and identically distributed random variables, with mean μ and variance σ^2. It might be that the mean and variance are unknown, and that the distribution is not even known to be a particular type, so maximum likelihood estimation is not appropriate. In this case it is reasonable to estimate μ and σ^2 by the *sample mean* and *sample variance* defined as follows:

$$\widehat{X} = \frac{1}{n}\sum_{k=1}^{n} X_k \qquad \widehat{\sigma^2} = \frac{1}{n-1}\sum_{k=1}^{n}(X_k - \widehat{X})^2.$$

Note the perhaps unexpected appearance of $n-1$ in the sample variance. Of course, we should have $n \geq 2$ to estimate the variance (assuming we don't know the mean) so it is not surprising that the formula is not defined if $n = 1$. An estimator is called *unbiased* if the mean of the estimator is equal to the parameter that is being estimated. (a) Is the sample mean an unbiased estimator of μ? (b) Find the mean square error, $E[(\mu - \widehat{X})^2]$, for estimation of the mean by the sample mean. (c) Is the sample variance an unbiased estimator of σ^2?

Solution (a) By the linearity of expectation,

$$E[\widehat{X}] = \frac{1}{n}\sum_{k=1}^{n} E[X_k] = \frac{1}{n}\sum_{k=1}^{n} \mu = \mu,$$

so $\widehat{X}$ is an unbiased estimator of μ.
(b) The mean square error for estimating μ by $\widehat{X}$ is given by

$$E[(\mu - \widehat{X})^2] = \mathrm{Var}(\widehat{X}) = \frac{1}{n^2}\mathrm{Var}\left(\sum_{k=1}^{n} X_k\right) = \frac{1}{n^2}\sum_{k=1}^{n}\mathrm{Var}(X_k) = \frac{1}{n^2}\sum_{k=1}^{n}\sigma^2 = \frac{\sigma^2}{n}.$$

(c)

$$E\left[\widehat{\sigma^2}\right] = \frac{1}{n-1}\sum_{k=1}^{n} E[(X_k - \widehat{X})^2] = \frac{n}{n-1}E[(X_1 - \widehat{X})^2],$$

because, by symmetry, $E[(X_k - \widehat{X})^2] = E[(X_1 - \widehat{X})^2]$ for all k. Now, $E[X_1 - \widehat{X}] = \mu - \mu = 0$, so

$$\begin{aligned} E\left[(X_1 - \widehat{X})^2\right] &= \mathrm{Var}(X_1 - \widehat{X}) \\ &= \mathrm{Var}\left(\frac{(n-1)X_1}{n} - \sum_{k=2}^{n} \frac{X_k}{n}\right) \\ &= \left(\left(\frac{n-1}{n}\right)^2 + \sum_{k=2}^{n} \frac{1}{n^2}\right)\sigma^2 = \frac{(n-1)\sigma^2}{n}. \end{aligned}$$

Therefore,

$$E\left[\widehat{\sigma^2}\right] = \frac{n}{n-1}\frac{(n-1)\sigma^2}{n} = \sigma^2,$$

so, $\widehat{\sigma^2}$ is an unbiased estimator of σ^2.

Example 4.8.7 A portfolio selection problem Suppose you are an investment fund manager with three financial instruments to invest your funds in for a one year period. Assume that, based on past performance, the returns on investment for the instruments have the following means and standard deviations.

Instrument	Expected value after one year	Standard deviation of value after one year
Stock fund (S)	$\mu_S = 1.10$ i.e., 10% expected gain	σ_S=0.15
Bond fund (B)	$\mu_B = 1.00$ i.e., expected gain is zero	σ_B=0.15
T-bills (T)	$\mu_T = 1.02$ i.e. 2% gain	σ_T=0

(So $T \equiv 1.02$.) Also assume the correlation coefficient between the stocks and bonds is $\rho_{S,B} = -0.8$. Some fraction of the funds is to be invested in stocks, some fraction in bonds, and the rest in T-bills, and at the end of the year the return per unit of funds is R. There is no single optimal choice of what values to use for these fractions; there is a tradeoff between the mean, μ_R, (larger is better) and the standard deviation, σ_R (smaller is better). Plot your answers to the parts below using a horizontal axis for mean return ranging from 1.0 to 1.1, and a vertical axis for standard deviation ranging from 0 to 0.15. Label the points $P_S = (1.1, 0.15)$, $P_B = (1, 0.15)$ and $P_T = (1.02, 0)$ on the plot corresponding to the three possibilities for putting all the funds into one instrument.

(a) Let $R_\lambda = \lambda S + (1-\lambda)T$, so R_λ is the random return resulting when a fraction λ of the funds is put into stocks and a fraction $1 - \lambda$ is put into T-bills. Determine and plot the set of $(\mu_{R_\lambda}, \sigma_{R_\lambda})$ pairs as λ ranges from zero to one.

(b) Let $R_\alpha = \alpha S + (1-\alpha)B$, so R_α is the random return resulting when a fraction α of the funds is put into stocks and a fraction $1 - \alpha$ is put into bonds. Determine and plot the set of $(\mu_{R_\alpha}, \sigma_{R_\alpha})$ pairs as α ranges from zero to one. (Hint: Use the fact $\rho_{S,B} = -0.8$.).

(c) Combining parts (a) and (b), let $R_{\lambda,\alpha} = \lambda R_\alpha + (1-\lambda)T$, so $R_{\lambda,\alpha}$ is the random return resulting when a fraction $1 - \lambda$ of the funds is invested in T-bills as in part (a), and a fraction λ of the funds is invested in the same mixture of stock and bond funds considered in part (b). For each

$\alpha \in \{0.2,\ 0.4,\ 0.6,\ 0.8\}$, determine and plot the set of $(\mu_{R_{\lambda,\alpha}}, \sigma_{R_{\lambda,\alpha}})$ pairs as λ ranges from zero to one. (Hint: You may express your answers in terms of $(\mu_{R_\alpha}, \sigma_{R_\alpha})$ found in part (b).)

(d) As α and λ both vary over the interval $[0, 1]$, the corresponding point $(\mu_{R_{\lambda,\alpha}}, \sigma_{R_{\lambda,\alpha}})$ sweeps out the set of *achievable* (mean, standard deviation) pairs. An achievable pair $(\widetilde{\mu}_R, \widetilde{\sigma}_R)$ is said to be strictly better than an achievable pair (μ_R, σ_R) if either ($\mu_R < \widetilde{\mu}_R$ and $\sigma_R \geq \widetilde{\sigma}_R$) or ($\mu_R \leq \widetilde{\mu}_R$ and $\sigma_R > \widetilde{\sigma}_R$.). An achievable pair $(\widetilde{\mu}_R, \widetilde{\sigma}_R)$ is *undominated* if there is no other achievable pair strictly better than it. Identify the set of *undominated* achievable pairs.

Solution (a) By linearity of expectation, $\mu_R = \lambda\mu_S + (1-\lambda)\mu_T$. Since $T \equiv 1.02$, $\sigma^2_{R_\lambda} = \mathrm{Var}(\lambda S) = \lambda^2\sigma_S^2$, and so $\sigma_{R_\lambda} = \lambda\sigma S$. As λ ranges from 0 to 1, the point $(\mu_{R_\lambda}, \sigma_{R_\lambda})$ sweeps out the line segment from P_T to P_S. See Figure 4.22.

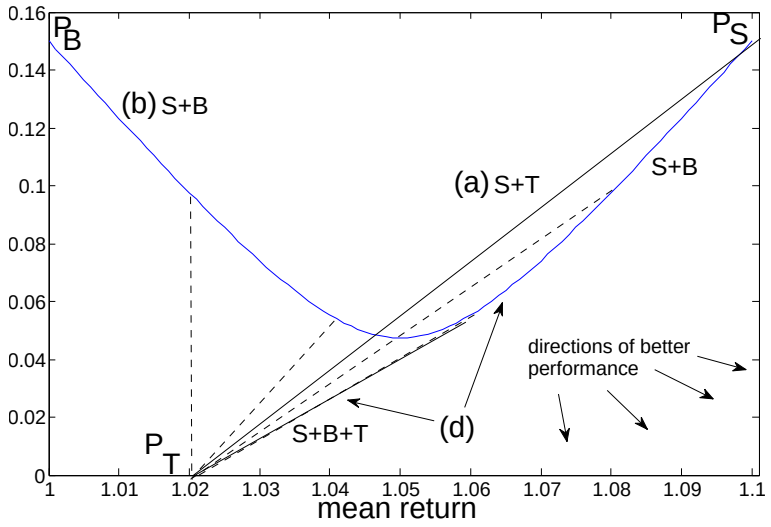


Figure 4.22: Mean and standard deviation of returns for various portfolios

(b) By linearity of expectation, $\mu_{R_\alpha} = \alpha\mu_S + (1-\alpha)\mu_B$. By the fact $\mathrm{Cov}(S, B) = \sigma_S\sigma_B\rho_{S,B}$, and

properties of covariance, $\sigma^2_{R_\alpha} = \text{Cov}(\alpha S + (1-\alpha)B) = \alpha^2\sigma_S^2 + 2\alpha(1-\alpha)\rho_{S,B}\sigma_S\sigma_B + (1-\alpha)^2\sigma_B^2 = (0.15)^2(\alpha^2 + 2\rho_{S,B}\alpha(1-\alpha) + (1-\alpha)^2)$. Thus, $\sigma_{R_\alpha} = (0.15)\sqrt{\alpha^2 + 2\rho_{S,B}\alpha(1-\alpha) + (1-\alpha)^2}$. As α ranges from 0 to 1, the point $(\mu_{R_\alpha}, \sigma_{R_\alpha})$ sweeps out a convex curve from P_B to P_S, as shown in Figure 4.22.

(c) By linearity of expectation, $\mu_{R_{\lambda,\alpha}} = \lambda\mu_{R_\alpha} + (1-\lambda)\mu_T$. Since $T \equiv 1.02$, $\sigma^2_R = \lambda^2\text{Var}(R_\alpha)$. Thus $\sigma_{R_{\lambda,\alpha}} = \lambda\sigma_{R_\alpha}$. As λ ranges from 0 to 1, the point $(\mu_{R_{\lambda,\alpha}}, \sigma_{R_{\lambda,\alpha}})$ sweeps out the line segment from P_T to the point found in part (b) corresponding to α. Those lines for $\alpha \in \{0.2,\ 0.4,\ 0.6,\ 0.8\}$ are shown as dashed lines in Figure 4.22.

(d) By part (c), the set of all achievable points is the union of all line segments from P_T to points on the curve found in part(b). Consider the line through P_T that is tangent to the curve found in part (b). The set of undominated points consists of the segment of that line from P_T up to the point of tangency, and then continues along the curve found in part (b) up to the point P_S. See Figure 4.22. Thus, the optimal (and only) portfolio with zero standard deviation is the one investing all funds in T-bills. For small, positive standard deviation, the optimal portfolio is to use a mixture of all three instruments. For larger standard deviation less than 0.15, the optimal portfolio is to use a mixture of stock and bond funds only. Finally, the largest mean return of 0.10 is uniquely achieved by investing only in the stock fund, for which the standard deviation is 0.15.

4.9 Minimum mean square error estimation

4.9.1 Constant estimators

Let Y be a random variable with some known distribution. Suppose Y is not observed but that we wish to estimate Y. If we use a constant δ to estimate Y, the estimation error will be $Y - \delta$. The mean square error (MSE) for estimating Y by δ is defined by $E[(Y-\delta)^2]$. By LOTUS, if Y is a continuous-type random variable,

$$\text{MSE (for estimation of } Y \text{ by a constant } \delta) = \int_{-\infty}^{\infty} (y-\delta)^2 f_Y(y)dy. \tag{4.27}$$

We seek to find δ to minimize the MSE. Since $(Y-\delta)^2 = Y^2 - 2\delta Y + \delta^2$, we can use linearity of expectation to get $E[(Y-\delta)^2] = E[Y^2] - 2\delta E[Y] + \delta^2$. This is quadratic in δ, and the derivative with resect to δ is $-2E[Y] + 2\delta$. Therefore the minimum occurs at $\delta^* = E[Y]$. For this value of δ, the MSE is $E[(Y-\delta^*)^2] = \text{Var}(Y)$. In summary, the constant δ^* that minimizes the mean square error for estimation of a random variable Y by a constant is the mean, and the minimum possible value of the mean square error for estimating Y by a constant is $\text{Var}(Y)$.

Another way to derive this result is to use the fact that $E[Y - EY] = 0$ and $EY - \delta$ is constant, to get

$$\begin{aligned} E[(Y-\delta)^2] &= E[((Y-EY) + (EY-\delta))^2] \\ &= E[(Y-EY)^2 + 2(Y-EY)(EY-\delta) + (EY-\delta)^2] \\ &= \text{Var}(Y) + (EY-\delta)^2. \end{aligned}$$

From this expression it is easy to see that the mean square error is minimized with respect to δ if and only if $\delta = EY$, and the minimum possible value is $\mathrm{Var}(Y)$.

4.9.2 Unconstrained estimators

Suppose instead that we wish to estimate Y based on an observation X. If we use the estimator $g(X)$ for some function g, the resulting mean square error (MSE) is $E[(Y - g(X))^2]$. We want to find g to minimize the MSE. The resulting estimator $g^*(X)$ is called the unconstrained optimal estimator of Y based on X because no constraints are placed on the function g.

Suppose you observe $X = 10$. What do you know about Y? Well, if you know the joint pdf of X and Y, you also know or can derive the conditional pdf of Y given $X = 10$, denoted by $f_{Y|X}(v|10)$. Based on the fact, discussed above, that the minimum MSE constant estimator for a random variable is its mean, it makes sense to estimate Y by the conditional mean:

$$E[Y|X = 10] = \int_{-\infty}^{\infty} v f_{Y|X}(v|10)dv.$$

The resulting conditional MSE is the variance of Y, computed using the conditional distribution of Y given $X = 10$.

$$E[(Y - E[Y|X = 10])^2|X = 10] = E[Y^2|X = 10] - (E[Y|X = 10]^2).$$

Conditional expectation indeed gives the optimal estimator, as we show now. Recall that $f_{X,Y}(u, v) = f_X(u)f_{Y|X}(v|u)$. So

$$\begin{aligned} \text{MSE} &= E[(Y - g(X))^2] \\ &= \int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} (v - g(u))^2 f_{Y|X}(v|u)dv \right) f_X(u)du. \end{aligned} \tag{4.28}$$

For each u fixed, the integral in parentheses in (4.28) has the same form as the integral (4.27). Therefore, for each u, the integral in parentheses in (4.28) is minimized by using $g(u) = g^*(u)$, where

$$g^*(u) = E[Y|X = u] = \int_{-\infty}^{\infty} v f_{Y|X}(v|u)dv. \tag{4.29}$$

We write $E[Y|X]$ for $g^*(X)$. The minimum MSE is

$$\begin{aligned} MSE &= E[(Y - E[Y|X])^2] \\ &= \int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} (v - g^*(u))^2 f_{Y|X}(v|u)dv \right) f_X(u)du \qquad (4.30) \\ &= \int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} (v^2 - (g^*(u))^2) f_{Y|X}(v|u)dv \right) f_X(u)du \qquad (4.31) \\ &= E[Y^2] - E[(E[Y|X])^2]. \qquad (4.32) \end{aligned}$$

In summary, the minimum MSE unconstrained estimator of Y given X is $E[Y|X] = g^*(X)$ where $g^*(u) = E[Y|X = u]$, and expressions for the MSE are given by (4.30)-(4.32).

4.9.3 Linear estimators

In practice it is not always possible to compute $g^*(u)$. Either the integral in (4.29) may not have a closed form solution, or the conditional density $f_{Y|X}(v|u)$ may not be available or might be difficult to compute. The problems might be more than computational. There might not even be a good way to decide what joint pdf $f_{X,Y}$ to use in the first place. A reasonable alternative to using g^* is to consider linear estimators of Y given X. A linear estimator has the form $L(X) = aX + b$, and to specify L we only need to find the two constants a and b, rather than finding a whole function g^*. The MSE for the linear estimator $aX + b$ is

$$MSE = E[(Y - (aX + b))^2].$$

Next we identify the linear estimator that minimizes the MSE. One approach is to multiply out $(Y - (aX + b))^2$, take the expectation, and set the derivative with respect to a equal to zero and the derivative with respect to b equal to zero. That would yield two equations for the unknowns a and b. We will take a slightly different approach, first finding the optimal value of b as a function of a, substituting that in, and then minimizing over a. The MSE can be written as follows:

$$E[((Y - aX) - b)^2].$$

Therefore, we see that for a given value of a, the constant b should be the minimum MSE constant estimator of $Y - aX$, which is given by $b = E[Y - aX] = \mu_Y - a\mu_X$. Therefore, the optimal linear estimator has the form $aX + \mu_Y - a\mu_X$ or, equivalently, $\mu_Y + a(X - \mu_X)$, and the corresponding MSE is given by

$$\begin{aligned} MSE &= E[(Y - \mu_Y - a(X - \mu_X))^2] \\ &= \mathrm{Var}(Y - aX) \\ &= \mathrm{Cov}(Y - aX, Y - aX) \\ &= \mathrm{Var}(Y) - 2a\mathrm{Cov}(Y, X) + a^2\mathrm{Var}(X). \end{aligned} \tag{4.33}$$

It remains to find the constant a. The MSE is quadratic in a, so taking the derivative with respect to a and setting it equal to zero yields that the optimal choice of a is $a^* = \frac{\mathrm{Cov}(Y,X)}{\mathrm{Var}(X)}$. Therefore, the minimum MSE linear estimator is given by $L^*(X) = \widehat{E}[Y|X]$, where

$$\widehat{E}[Y|X] = \mu_Y + \left(\frac{\mathrm{Cov}(Y, X)}{\mathrm{Var}(X)}\right)(X - \mu_X) \tag{4.34}$$

$$= \mu_Y + \sigma_Y \rho_{X,Y} \left(\frac{X - \mu_X}{\sigma_X}\right). \tag{4.35}$$

Setting a in (4.33) to a^* gives the following expression for the minimum possible MSE:

$$\text{minimum MSE for linear estimation} = \sigma_Y^2 - \frac{(\mathrm{Cov}(X, Y))^2}{\mathrm{Var}(X)} = \sigma_Y^2(1 - \rho_{X,Y}^2). \tag{4.36}$$

Since $\text{Var}\left(\widehat{E}[Y|X]\right) = \left(\frac{\sigma_Y \rho_{X,Y}}{\sigma_X}\right)^2 \text{Var}(X) = \sigma_Y^2 \rho_{X,Y}^2$, the following alternative expressions for the minimum MSE hold

$$\text{minimum MSE for linear estimation} = \sigma_Y^2 - \text{Var}(\widehat{E}[Y|X]) = E[Y^2] - E[\widehat{E}[Y|X]^2]. \tag{4.37}$$

In summary, the minimum mean square error linear estimator is given by (4.34) or (4.35), and the resulting minimum mean square error is given by (4.36) or (4.37). The minimum MSE linear estimator $\widehat{E}[Y|X]$ is called the *wide sense conditional expectation* of Y given X. In analogy with regular conditional expectation, we write

$$L^*(u) = \widehat{E}[Y|X = u] = \mu_Y + \left(\frac{\text{Cov}(Y,X)}{\text{Var}(X)}\right)(u - \mu_X).$$

so that $\widehat{E}[Y|X]$ is equal to the linear function $\widehat{E}[Y|X = u]$ applied to the random variable X. The notations $L^*(X)$ and $\widehat{E}[Y|X]$ are both used for the minimum MSE linear estimator of Y given X. The notation $L^*(X)$ emphasizes that it is linear in X, and the notation $\widehat{E}[Y|X]$ explicitly references both random variables X and Y. For example, if W and Z are random variables, $\widehat{E}[W|Z] = \mu_W + \left(\frac{\text{Cov}(W,Z)}{\text{Var}(Z)}\right)(Z - \mu_Z)$ is the minimum MSE linear estimator of W for observation Z.

Four of the five constants in the formula (4.35) for $\widehat{E}[Y|X]$ are the means and standard deviations. If the random variables X and Y are standard (i.e. have mean zero and variance one), then $\widehat{E}[Y|X] = \rho_{X,Y} X$ and the MSE is $1 - \rho_{X,Y}^2$. This fact highlights the central role of the correlation coefficient, $\rho_{X,Y}$, in the linear estimation of Y from X.

This section covers three types of estimators of Y: unconstrained estimators of the form $g(X)$, linear estimators of the form $L(X) = aX + b$, and constant estimators, of the form δ. Of the three, the unconstrained estimators are the most general, a subset of them consists of the linear estimators, and a subset of the linear estimators is the set of constant estimators. The larger the class of estimators optimized over, the smaller the resulting MSE is. Thus, the following ordering among the three MSEs holds:

$$\underbrace{E[(Y - g^*(X))^2]}_{\text{MSE for } g^*(X)=E[Y|X]} \leq \underbrace{\sigma_Y^2(1 - \rho_{X,Y}^2)}_{\text{MSE for } L^*(X)=\widehat{E}[Y|X]} \leq \underbrace{\sigma_Y^2}_{\text{MSE for } \delta^*=E[Y].} \tag{4.38}$$

Note that all three estimators are linear as functions of the variable to be estimated:

$$\begin{aligned} E[aY + bZ + c] &= aE[Y] + bE[Z] + c \\ E[aY + bZ + c|X] &= aE[Y|X] + bE[Z|X] + c \\ \widehat{E}[aY + bZ + c|X] &= a\widehat{E}[Y|X] + b\widehat{E}Z|X] + c \end{aligned}$$

Example 4.9.1 Let $X = Y + N$, where Y has the exponential distribution with parameter λ, and N is Gaussian with mean 0 and variance σ_N^2. Suppose the variables Y and N are independent, and the parameters λ and σ_N^2 are known and strictly positive. (Recall that $E[Y] = \frac{1}{\lambda}$ and $\text{Var}(Y) = \sigma_Y^2 = \frac{1}{\lambda^2}$.)
(a) Find $\widehat{E}[Y|X]$, the MSE linear estimator of Y given X, and also find the resulting MSE.
(b) Find an unconstrained estimator of Y yielding a strictly smaller MSE than $\widehat{E}[Y|X]$ does.

Solution: (a) Since Y and N are independent, $\mathrm{Cov}(Y,N)=0$. Therefore,

$$\begin{aligned}\mathrm{Cov}(Y,X) &= \mathrm{Cov}(Y,Y+N)=\mathrm{Cov}(Y,Y)+\mathrm{Cov}(Y,N)=\mathrm{Var}(Y)\\ \mathrm{Var}(X) &= \mathrm{Var}(Y+N)=\mathrm{Var}(Y)+\mathrm{Var}(N).\end{aligned}$$

So, by (4.34),

$$\widehat{E}[Y|X]=\frac{1}{\lambda}+\frac{1/\lambda^2}{1/\lambda^2+\sigma_N^2}\left(X-\frac{1}{\lambda}\right)=\frac{1}{\lambda}+\frac{1}{1+\lambda^2\sigma_N^2}\left(X-\frac{1}{\lambda}\right)=\frac{X+\lambda\sigma_N^2}{1+\lambda^2\sigma_N^2},$$

and by (4.36),

$$\text{MSE for } \widehat{E}[Y|X]=\sigma_Y^2-\frac{\sigma_Y^4}{\sigma_Y^2+\sigma_N^2}=\frac{\sigma_Y^2\sigma_N^2}{\sigma_Y^2+\sigma_N^2}=\frac{\sigma_N^2}{1+\lambda^2\sigma_N^2}.$$

(b) Although Y is always nonnegative, the estimator $L^*(X)$ can be negative. An estimator with smaller MSE is $\widehat{Y}=\max\{0,\widehat{E}[Y|X]\}$, because $(Y-\widehat{Y})^2\le(Y-\widehat{E}[Y|X])^2$ with probability one, and the inequality is strict whenever $\widehat{E}[Y|X]<0$.

Example 4.9.2 Suppose (X,Y) is uniformly distributed over the triangular region with vertices at $(-1,0)$, $(0,1)$, and $(1,1)$, shown in Figure 4.23. (a) Find and sketch the minimum MSE estimator

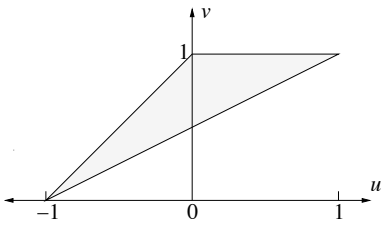


Figure 4.23: Support of $f_{X,Y}$.

of Y given $X=u$, $g^*(u)=E[Y|X=u]$, for all u such that it is well defined, and find the resulting minimum MSE for using $g^*(X)=E[Y|X]$ to estimate Y. (b) Find and sketch the function $\widehat{E}[Y|X=u]$, used for minimum MSE linear estimation of Y from X, and find the resulting MSE for using $\widehat{E}[Y|X]$ to estimate Y.

Solution: (a) We will first solve this by going through all the usual steps, first finding $f_{X,Y}$ and f_X in order to identify the ratio of the two, which is $f_{Y|X}(v|u)$. Then $E[Y|X=u]$ is the mean of the conditional density $f_{Y|X}(v|u)$ for u fixed. It will then be explained how the answer could be deduced by inspection.

The support of $f_{X,Y}$ is the triangular region, which has area 0.5. (Use the formula one half base times height, and take the base to be the line segment at the top of the triangle.) So the joint pdf

of $f_{X,Y}$ is 2 inside the triangle. The marginal pdf, f_X, is given by

$$f_X(u) = \begin{cases} \int_{(1+u)/2}^{1+u} 2dv = 1+u & -1 \le u \le 0 \\ \int_{(1+u)/2}^{1} 2dv = 1-u & 0 \le u \le 1 \\ 0 & \text{else,} \end{cases}$$

so the graph of f_X is the triangle shown if Figure 4.24(a). This is intuitively clear because the

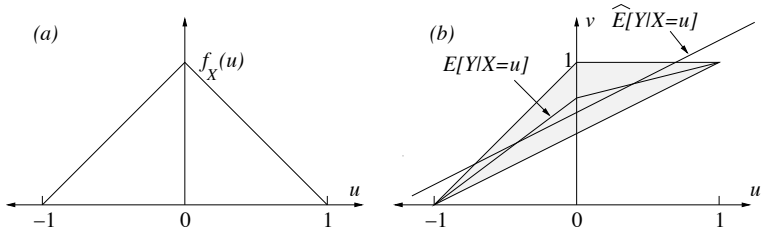


Figure 4.24: (a) The pdf of X. (b) The functions used for the minimum MSE error unconstrained estimator and linear estimator.

lengths of the vertical cross-sections of the support of $f_{X,Y}$ increase linearly from 0 over the interval $-1 < u \le 0$, and then decrease linearly back down to zero over the interval $0 \le u \le 1$. Since $f_X(u) > 0$ only for $-1 < u < 1$, the conditional density $f_{Y|X}(v|u)$ is defined only for u in that interval. The cases $-1 < u \le 0$ and $0 < u < 1$ will be considered separately, because f_X has a different form over those two intervals. So the conditional pdf of Y given $X = u$ is

$$-1 \le u < 0 \quad \leftrightarrow \quad f_{Y|X}(v|u) = \begin{cases} \frac{2}{1+u} & \frac{1+u}{2} \le v \le 1+u \\ 0 & else \end{cases} \quad \leftrightarrow \text{uniform on } \left[\frac{1+u}{2}, 1+u\right]$$

$$0 \le u < 1 \quad \leftrightarrow \quad f_{Y|X}(v|u) = \begin{cases} \frac{2}{1-u} & \frac{1+u}{2} \le v \le 1 \\ 0 & else \end{cases} \quad \leftrightarrow \text{uniform on } \left[\frac{1+u}{2}, 1\right],$$

As indicated, the distribution of Y given $X = u$ for u fixed, is the uniform distribution over an interval depending on u, as long as $-1 < u < 1$. The mean of a uniform distribution over an interval is the midpoint of the interval, and thus:

$$g^*(u) = E[Y|X=u] = \begin{cases} \frac{3(1+u)}{4} & -1 < u \le 0 \\ \frac{3+u}{4} & 0 < u < 1 \\ \text{undefined} & \text{else.} \end{cases}$$

This estimator is shown in Figure 4.24(b). Note that this estimator could have been drawn by inspection. For each value of u in the interval $-1 < u < 1$, $E[Y|X=u]$ is just the center of mass of the cross section of the support of $f_{X,Y}$ along the vertical line determined by u. That is true whenever (X,Y) is uniformly distributed over some region. The mean square error given $X = u$ is the variance of a uniform distribution on an interval of length $\frac{1-|u|}{2}$, which is $\frac{(1-|u|)^2}{48}$. Averaging

over u using the pdf f_X yields that

$$\begin{aligned}\text{MSE for } g^*(X) &= \int_{-1}^{0}(1+u)\frac{(1-|u|)^2}{48}du + \int_{0}^{1}(1-u)\frac{(1-|u|)^2}{48}du \\ &= 2\int_{0}^{1}(1-u)\frac{(1-|u|)^2}{48}du = \frac{1}{96}.\end{aligned}$$

(b) Finding $\widehat{E}[Y|X=u]$ requires calculating some moments. By LOTUS:

$$\begin{aligned}E[Y] &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} v f_{X,Y}(u,v)dudv \\ &= \int_{0}^{1}\int_{v-1}^{2v-1} 2v\,dudv \\ &= \int_{0}^{1} 2v^2 dv = \frac{2}{3},\end{aligned}$$

$$\begin{aligned}E[XY] &= \int_{0}^{1}\int_{v-1}^{2v-1} 2uv\,dudv \\ &= \int_{0}^{1} v[(2v-1)^2-(v-1)^2]dv = \int_{0}^{1} 3v^3-2v^2dv = \frac{1}{12},\end{aligned}$$

$$E[X^2] = 2\int_{0}^{1} u^2(1-u)du = \frac{1}{6}.$$

A glance at f_X shows $E[X]=0$, so $\text{Var}(X)=E[X^2]=\frac{1}{6}$ and $\text{Cov}(X,Y)=E[XY]=\frac{1}{12}$. Therefore, by (4.34),

$$\widehat{E}[Y|X=u] = L^*(u) = \frac{2}{3}+\frac{1/12}{1/6}u = \frac{2}{3}+\frac{u}{2}.$$

This estimator is shown in Figure 4.24(b). While the exact calculation of $\widehat{E}[Y|X=u]$ was tedious, its graph could have been drawn approximately by inspection. It is a straight line that tries to be close to $E[Y|X=u]$ for all u. To find the MSE we shall use (4.37). By LOTUS,

$$\begin{aligned}E[Y^2] &= \int_{0}^{1}\int_{v-1}^{2v-1} 2v^2 dudv \\ &= \int_{0}^{1} 2v^3 dv = \frac{1}{2}.\end{aligned}$$

and $E\left[\widehat{E}[Y|X=u]^2\right] = E\left[\left(\frac{2}{3}+\frac{X}{2}\right)^2\right] = \frac{4}{9}+\frac{E[X^2]}{4} = \frac{35}{72}$. Thus,

$$\text{MSE for } \widehat{E}[Y|X] = \frac{1}{2}-\frac{35}{72} = \frac{1}{72}.$$

Note that (MSE using $E[Y|X]$) $=\frac{1}{96} \le$ (MSE using $\widehat{E}[Y|X]$) $=\frac{1}{72} \le \text{Var}(Y)=\frac{1}{18}$, so the three MSEs are ordered in accordance with (4.38).

4.10 Law of large numbers and central limit theorem

The law of large numbers, in practical applications, has to do with approximating sums of random variables by a constant. The Gaussian approximation, backed by the central limit theorem, in practical applications, has to do with a more refined approximation: approximating sums of random variables by a single Gaussian random variable.

4.10.1 Law of large numbers

There are many forms of the law of large numbers (LLN). The law of large numbers is about the sample average of n random variables: $\frac{S_n}{n}$, where $S_n = X_1 + \ldots + X_n$. The random variables have the same mean, μ. The random variables are assumed to be independent, or weakly dependent, and some condition is placed on the sizes of the individual random variables. The conclusion is that as $n \to \infty$, $\frac{S_n}{n}$ converges in some sense to the mean, μ. The following version of the LLN has a simple proof.

Proposition 4.10.1 *(Law of large numbers) Suppose $X_1, X_2, \ldots$ is a sequence of uncorrelated random variables such that each X_k has finite mean μ and variance less than or equal to C. Then for any $\delta > 0$,*

$$P\left\{\left|\frac{S_n}{n} - \mu\right| \geq \delta\right\} \leq \frac{C}{n\delta^2} \overset{n\to\infty}{\to} 0.$$

Proof. The mean of $\frac{S_n}{n}$ is given by

$$E\left[\frac{S_n}{n}\right] = E\left[\frac{\sum_{k=1}^n X_k}{n}\right] = \frac{\sum_{k=1}^n E[X_k]}{n} = \frac{n\mu}{n} = \mu.$$

The variance of $\frac{S_n}{n}$ is bounded above by:

$$\mathrm{Var}\left(\frac{S_n}{n}\right) = \mathrm{Var}\left(\frac{\sum_{k=1}^n X_k}{n}\right) = \frac{\sum_{k=1}^n \mathrm{Var}(X_k)}{n^2} \leq \frac{nC}{n^2} = \frac{C}{n}.$$

Therefore, the proposition follows from the Chebychev inequality, (2.12), applied to the random variable $\frac{S_n}{n}$. ∎

The law of large numbers is illustrated in Figure 4.25, which was made using a random number generator on a computer. For each $n \geq 1$, S_n is the sum of the first n terms of a sequence of independent random variables, each uniformly distributed on the interval $[0, 1]$. Figure 4.25(a) illustrates the statement of the LLN, indicating convergence of the averages, $\frac{S_n}{n}$, towards the mean, 0.5, of the individual uniform random variables. The same sequence S_n is shown in Figure 4.25(b), except the S_n's are not divided by n. The sequence of partial sums S_n converges to $+\infty$. The LLN tells us that the asymptotic slope is equal to 0.5. The sequence S_n is not expected to get closer to $\frac{n}{2}$ as n increases–just to have the same asymptotic slope. In fact, the central limit theorem, given in the next section, implies that for large n, the difference $S_n - \frac{n}{2}$ has approximately the Gaussian distribution with mean zero, variance $\frac{n}{12}$, and standard deviation $\sqrt{\frac{n}{12}}$.

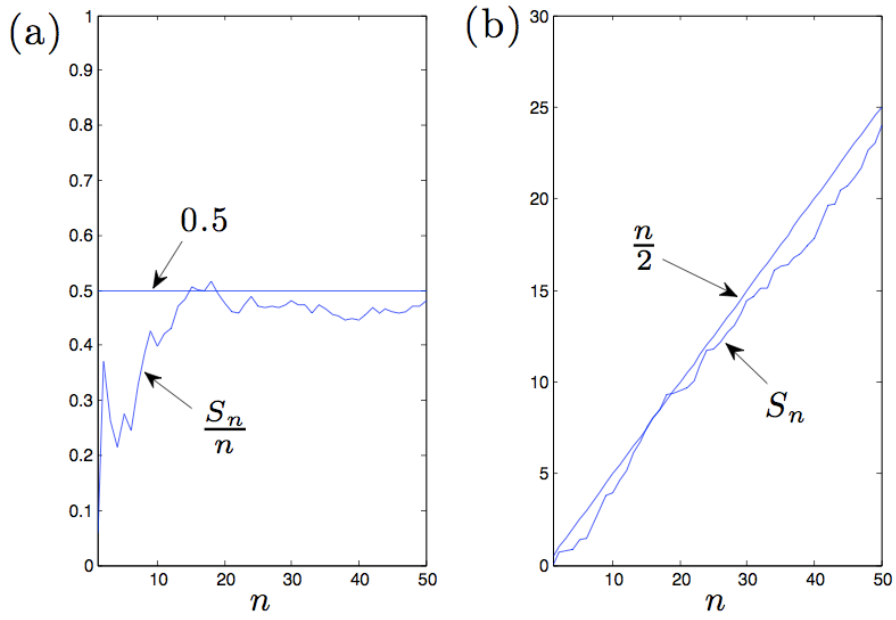


Figure 4.25: Results of simulation of sums of independent random variables, uniformly distributed on $[0, 1]$. (a) $\frac{S}{n}$ vs. n. (b) S_n vs. n.

Example 4.10.2 Suppose a fair die is rolled 1000 times. What is a rough approximation to the sum of the numbers showing, based on the law of large numbers?

Solution: The sum is $S_{1000} = X_1 + X_2 + \cdots + X_{1000}$, where X_k denotes the number showing on the kth roll. Since $E[X_k] = \frac{1+2+3+4+5+6}{6} = 3.5$, the expected value of S_{1000} is $(1000)(3.5) = 3500$. By the law of large numbers, we expect the sum to be near 3500.

Example 4.10.3 Suppose $X_1, \ldots, X_{100}$ are random variables, each with mean $\mu = 5$ and variance $\sigma^2 = 1$. Suppose also that $|\text{Cov}(X_i, X_j)| \leq 0.1$ if $i = j \pm 1$, and $\text{Cov}(X_i, X_j) = 0$ if $|i - j| \geq 2$. Let $S_{100} = X_1 + \cdots + X_{100}$.
(a) Show that $\text{Var}(S_{100}) \leq 120$.
(b) Use part (a) and Chebychev's inequality to find an upper bound on $P(|\frac{S_{100}}{100} - 5| \geq 0.5)$.

Solution: (a) For any n

$$\begin{aligned}
\mathrm{Var}(S_n) &= \mathrm{Cov}(X_1 + \cdots + X_n, X_1 + \cdots + X_n) \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n} \mathrm{Cov}(X_i, X_j) \\
&= \sum_{i=1}^{n-1} \mathrm{Cov}(X_i, X_{i+1}) \ + \ \sum_{i=1}^{n} \mathrm{Cov}(X_i, X_i) \ + \ \sum_{i=2}^{n} \mathrm{Cov}(X_i, X_{i-1}) \\
&\leq (n-1)(0.1) + n + (n-1)(0.1) < (1.2)n.
\end{aligned}$$

(b) Therefore, $\mathrm{Var}(\frac{S_{100}}{100}) = \frac{1}{(100)^2}\mathrm{Var}(S_{100}) \leq 0.012$. Also, $E\left[\frac{S_{100}}{100}\right] = \mu = 5$. Thus, by Chebychev's inequality, $P\left\{\left|\frac{S_{100}}{100} - 5\right| \geq 0.5\right\} \leq \frac{0.012}{(0.5)^2} = 0.048$.

4.10.2 Central limit theorem

The following version of the central limit theorem (CLT) generalizes the DeMoivre-Laplace limit theorem, discussed in Section 3.6.3. While the DeMoivre-Laplace limit theorem pertains to sums of independent, identically distributed Bernoulli random variables, the version here pertains to sums of independent, identically distributed random variables with any distribution, so long as their mean and variance are finite.

Theorem 4.10.4 *(Central limit theorem) Suppose $X_1, X_2, \ldots$ are independent, identically distributed random variables, each with mean μ and variance σ^2. Let $S_n = X_1 + \cdots + X_n$. Then:*

$$\lim_{n\to\infty} P\left\{\frac{S_n - n\mu}{\sqrt{n\sigma^2}} \leq c\right\} = \Phi(c).$$

The practical implication of the central limit theorem is the same as that of the DeMoivre-Laplace limit theorem. That is, the CLT gives evidence that the Gaussian approximation, discussed in Section 3.6.3, is a good one, for sums of independent, identically distributed random variables. Figure 4.26 illustrates the CLT for the case the X_k's are uniformly distributed on the interval $[0, 1]$. The approximation in this case is so good that some simulation programs generate Gaussian random variables by generating six uniformly distributed random variables and adding them together to produce one approximately Gaussian random variable.

Example 4.10.5 Let S denote the sum of the numbers showing in 1000 rolls of a fair die. By the law of large numbers, S is close to 3500 with high probability. Find the number L so that $P\{|S - 3500| \leq L\} \approx 0.9$. To be definite, use the continuity correction.

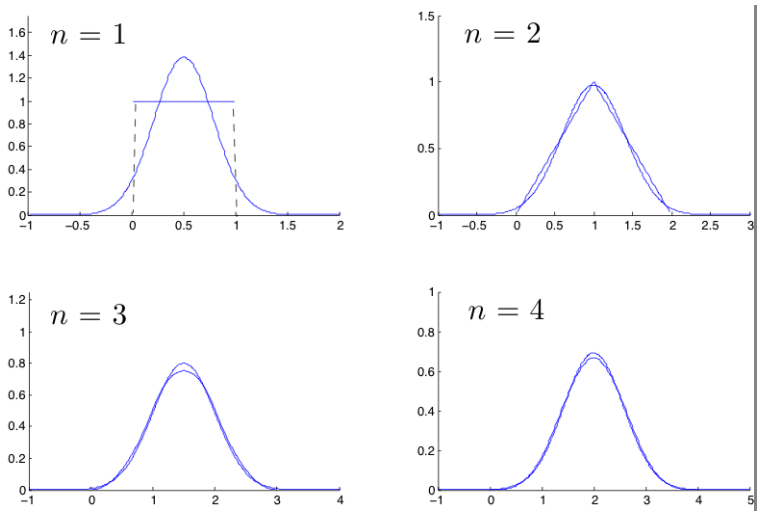


Figure 4.26: Comparison of the pdf of S_n to the Gaussian pdf with the same mean and variance, for sums of uniformly distributed random variables on $[0, 1]$. For $n = 3$ and $n = 4$, the Gaussian pdf is the one with a slightly higher peak.

Solution: As noted in Example 2.2.7, the number showing for one roll of a die has mean $\mu = 3.5$ and variance $\sigma^2 = 2.9167$. Therefore, S has mean 3500, variance 2916.7, and standard deviation $\sqrt{2916.7} = 54.01$. By the Gaussian approximation with the continuity correction,

$$\begin{aligned} P\{|S - 3500| \le L\} &= P\{|S - 3500| \le L + 0.5\} \\ &= P\left\{\left|\frac{S - 3500}{54.01}\right| \le \frac{L + 0.5}{54.01}\right\} \\ &\approx 1 - 2Q\left(\frac{L + 0.5}{54.01}\right). \end{aligned}$$

Now $Q(1.645) = 0.05$, so the desired value of L solves

$$\frac{L + 0.5}{54.01} \approx 1.645,$$

or $L \approx 88.34$. Thus, $L = 88$ should give the best approximation.

Example 4.10.6 Suppose each of 100 real numbers are rounded to the nearest integer and then added. Assume the individual roundoff errors are independent and uniformly distributed over the interval $[-0.5, 0.5]$. Using the Gaussian approximation suggested by the CLT, find the approximate probability that the absolute value of the sum of the errors is greater than 5.

Solution: The mean of each roundoff error is zero and the variance is $\int_{-0.5}^{0.5} u^2 du = \frac{1}{12}$. Thus, $E[S] = 0$ and $\text{Var}(S) = \frac{100}{12} = 8.333$. Thus, $P\{|S| \geq 5\} = P\left\{|\frac{S}{\sqrt{8.333}}| \geq \frac{5}{\sqrt{8.333}}\right\} \approx 2Q\left(\frac{5}{\sqrt{8.333}}\right) = 2Q(1.73) = 2(1 - \Phi(1.732)) = 0.083$.

Example 4.10.7 Suppose each day of the year, the value of a particular stock: increases by one percent with probability 0.5, remains the same with probability 0.4, and decreases by one percent with probability 0.1. Changes on different days are independent. Consider the value after one year (365 days), beginning with one unit of stock. Find (a) the probability the stock at least triples in value, (b) the probability the stock at least quadruples in value, (c) the median value after one year.

Solution: The value, Y, after one year, is given by $Y = D_1 D_2 \cdots D_{365}$, where D_k for each k is the growth factor for day k, with pmf $p_D(1.01) = 0.5$, $p_D(1) = 0.4$, $p_D(0.99) = 0.1$. The CLT pertains to sums of random variables, so we will apply the Gaussian approximation to $\ln(Y)$, which is a sum of a large number of random variables:

$$\ln Y = \sum_{k=1}^{365} \ln D_k.$$

By LOTUS, the mean and variance of $\ln(D_k)$ are given by

$$\mu = E[\ln(D_k)] = 0.5\ln(1.01) + 0.4\ln(1) + 0.1\ln(0.99) = 0.00397.$$

$\sigma^2 = \text{Var}(\ln(D_k)) = E[(\ln(D_k)^2] - \mu^2 = 0.5(\ln(1.01))^2 + 0.4(\ln(1))^2 + 0.1(\ln(0.99))^2 - \mu^2 = 0.00004384$.
Thus, $\ln Y$ is approximately Gaussian with mean $365\mu = 1.450$ and standard deviation $\sqrt{365\sigma^2} = 0.127$. Therefore,

$$\begin{aligned} P\{Y \geq c\} &= P\{\ln(Y) \geq \ln(c)\} \\ &= P\left\{\frac{\ln(Y) - 1.450}{0.127} \geq \frac{\ln(c) - 1.450}{0.127}\right\} \\ &\approx Q\left(\frac{\ln(c) - 1.450}{0.127}\right). \end{aligned}$$

In particular:
(a) $P\{Y \geq 3\} \approx Q(-2.77) \approx 0.997$.
(b) $P\{Y \geq 4\} \approx Q(-0.4965) = 0.69$.
(c) The median is the value c such that $P\{Y \geq c\} = 0.5$, which by the Gaussian approximation is $e^{1.450} = 4.26$. (This is the same result one would get by the following argument, based on the law of large numbers. We expect, during the year, the stock to increase by one percent about 365*(0.5) times, and to decrease by one percent about 365*(0.1) times. That leads to a year end value of $(1.01)^{182.5}(0.99)^{36.5} = 4.26$.)

4.11 Joint Gaussian distribution

Recall that Gaussian distributions often arise in practice; this fact is explained by the CLT. The CLT can be extended to two, or even more, correlated random variables. For example, suppose $(U_1, V_1), (U_2, V_2), \cdots$ are independent, identically distributed pairs of random variables. For example, U_i might be the height, and V_i the weight, of the ith student to enroll at a university. Suppose for convenience that they have mean zero. Then as $n \to \infty$, the pair $\left(\frac{U_1+\cdots+U_n}{\sqrt{n}}, \frac{V_1+\cdots+V_n}{\sqrt{n}}\right)$ has a limiting bivariate distribution, where "bivariate" means the limit distribution is a joint distribution of two random variables. Suppose X and Y have such a limit distribution. Then X and Y must each be Gaussian random variables, by the CLT. But also for any constants a and b, $aX + bY$ must be Gaussian, because $aX + bY$ has the limiting distribution of $\frac{(aU_1+bV_1)+\cdots+(aU_n+bV_n)}{\sqrt{n}}$, which is Gaussian by the CLT. This observation motivates the following definition:

Definition 4.11.1 *Random variables X and Y are said to be* jointly Gaussian *if every linear combination $aX + bY$ is a Gaussian random variable. (For the purposes of this definition, a constant is considered to be a Gaussian random variable with variance zero.)*

Being jointly Gaussian includes the case that X and Y are each Gaussian and linearly related: $X = aY + b$ for some a, b or $Y = aX + b$ for some a, b. In these cases, X and Y do not have a joint pdf. Aside from those two degenerate cases, a pair of jointly Gaussian random variables has a bivariate normal (or Gaussian) pdf, given by

$$f_{X,Y}(u,v) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \exp\left(-\frac{\left(\frac{u-\mu_X}{\sigma_X}\right)^2 + \left(\frac{v-\mu_Y}{\sigma_Y}\right)^2 - 2\rho\left(\frac{u-\mu_X}{\sigma_X}\right)\left(\frac{v-\mu_Y}{\sigma_Y}\right)}{2(1-\rho^2)}\right), \tag{4.39}$$

where the five parameters $\mu_X, \mu_Y, \sigma_X, \sigma_Y, \rho$ satisfy $\sigma_X > 0$, $\sigma_Y > 0$ and $-1 < \rho < 1$. As shown below, μ_X and μ_Y are the means of X and Y, respectively, σ_X and σ_Y are the standard deviations, respectively, and ρ is the correlation coefficient. We shall describe some properties of the bivariate normal pdf in the remainder of this section.

There is a simple way to recognize whether a pdf is a bivariate normal. Namely, such a pdf has the form: $f_{X,Y}(u,v) = C\exp(-P(u,v))$, where P is a second order polynomial of two variables:

$$P(u,v) = au^2 + buv + cv^2 + du + ev + f.$$

The constant C is selected so that the pdf integrates to one. Such a constant exists if and only if $P(u,v) \to +\infty$ as $|u| + |v| \to +\infty$, which requires $a > 0$, $c > 0$ and $b^2 - 4ac < 0$. Without loss of generality, we can take $f = 0$, because it can be incorporated into the constant C. Thus, the set of bivariate normal pdfs can be parameterized by the five parameters: a, b, c, d, e.

4.11.1 From the standard 2-d normal to the general

Suppose W and Z are independent, standard normal random variables. Their joint pdf is the product of their individual pdfs:

$$f_{W,Z}(\alpha,\beta) = \left(\frac{e^{-\frac{\alpha^2}{2}}}{\sqrt{2\pi}}\right)\left(\frac{e^{-\frac{\beta^2}{2}}}{\sqrt{2\pi}}\right) = \frac{e^{-\frac{\alpha^2+\beta^2}{2}}}{2\pi}.$$

This joint pdf is called the *standard bivariate normal pdf,* and it is the special case of the general bivariate normal pdf obtained by setting the means to zero, the standard deviations to one, and the correlation coefficient to zero. The general bivariate normal pdf can be obtained from $f_{W,Z}$ by a linear transformation. Specifically, if X and Y are related to W and Z by

$$\begin{pmatrix} X \\ Y \end{pmatrix} = A\begin{pmatrix} W \\ Z \end{pmatrix} + \begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix},$$

where A is the matrix

$$A = \begin{pmatrix} \sqrt{\frac{\sigma_X^2(1+\rho)}{2}} & -\sqrt{\frac{\sigma_X^2(1-\rho)}{2}} \\ \sqrt{\frac{\sigma_Y^2(1+\rho)}{2}} & \sqrt{\frac{\sigma_Y^2(1-\rho)}{2}} \end{pmatrix},$$

then $f_{X,Y}$ is given by (4.39), as can be shown by the method of Section 4.7.1.

Figure 4.27 illustrates the geometry of the bivariate normal pdf. The graph of the standard

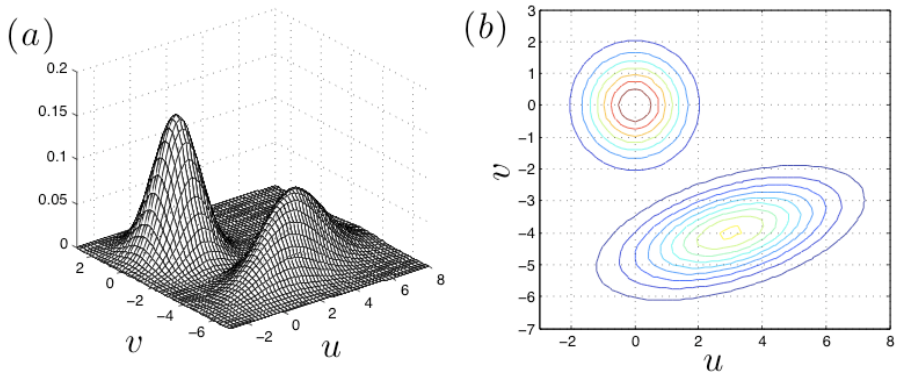


Figure 4.27: (a) Mesh plots of both the standard bivariate normal, and the bivariate normal with $\mu_X = 3, \mu_Y = -4, \sigma_X = 2, \sigma_Y = 1, \rho = 0.5$, shown on the same axes. (b) Contour plots of the same pdfs.

bivariate normal, $f_{W,Z}$, has a bell shape that is rotationally symmetric about the origin. The level sets are thus circles centered at the origin. The peak value of $f_{W,Z}$ is $\frac{1}{2\pi}$. The *half-peak level set* is

the set of (α, β) pairs such that $f_{W,Z}(\alpha, \beta)$ is equal to half of the peak value, or $\{(\alpha, \beta) : \alpha^2 + \beta^2 = 2 \ln 2 \approx (1.18)^2\}$, which is a circle centered at the origin with radius approximately 1.18. Since all the bivariate normal pdfs are obtained from $f_{W,Z}$ by linear scaling and translation, the half-peak level set of a general bivariate normal pdf completely determines the pdf. The half-peak level set for the general bivariate normal $f_{X,Y}$ given by (4.39) is

$$\left\{ (u, v) : \frac{\left(\frac{u-\mu_X}{\sigma_X}\right)^2 + \left(\frac{v-\mu_Y}{\sigma_Y}\right)^2 - 2\rho\left(\frac{u-\mu_X}{\sigma_X}\right)\left(\frac{v-\mu_Y}{\sigma_Y}\right)}{1-\rho^2} = 2\ln 2 \approx (1.18)^2 \right\},$$

which is an ellipse. The space of ellipses in the plane is five dimensional–two coordinates specify the center of an ellipse, two numbers give the lengths of the major and minor axes, and a final number gives the angle of the major axis from the horizontal. This gives another way to parameterize the set of bivariate normal pdfs using five parameters.

4.11.2 Key properties of the bivariate normal distribution

Proposition 4.11.2 *Suppose X and Y have the bivariate normal pdf with parameters $\mu_X, \mu_Y, \sigma_X, \sigma_Y$, and ρ. Then*

(a) X has the $N(\mu_X, \sigma_X^2)$ distribution, and Y has the $N(\mu_Y, \sigma_Y^2)$ distribution.

(b) Any linear combination of the form $aX + bY$ is a Gaussian random variable (i.e., X and Y are jointly Gaussian).

(c) ρ is the correlation coefficient between X and Y (i.e. $\rho_{X,Y} = \rho$).

(d) X and Y are independent if and only if $\rho = 0$.

(e) For estimation of Y from X, $L^*(X) = g^*(X)$. Equivalently, $E[Y|X] = \widehat{E}[Y|X]$. That is, the best unconstrained estimator $g^*(X)$ is linear.

(f) The conditional distribution of Y given $X = u$ is $N(\widehat{E}[Y|X = u], \sigma_e^2)$, where σ_e^2 is the MSE for $\widehat{E}[Y|X]$, given by (4.36) or (4.37).

Proof. It suffices to prove the proposition in case $\mu_X = \mu_Y = 0$ and $\sigma_X^2 = \sigma_Y^2 = 1$, because these parameters simply involve centering and scaling of X and Y separately, or, equivalently, translation and scaling of the joint pdf parallel to the u-axis or parallel to the v-axis. Such centering and scaling of X and Y separately does not change the correlation coefficient, as shown in Section 4.8.

The joint pdf in this case can be written as the product of two factors, as follows:

$$\begin{aligned} f_{X,Y}(u, v) &= \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left(-\frac{u^2 + v^2 - 2\rho uv}{2(1-\rho^2)}\right) \\ &= \left[\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u^2}{2}\right)\right] \left[\frac{1}{\sqrt{2\pi(1-\rho^2)}} \exp\left(-\frac{(v-\rho u)^2}{2(1-\rho^2)}\right)\right]. \end{aligned} \tag{4.40}$$

The first factor is a function of u alone, and is the standard normal pdf. The second factor, as a function of v for u fixed, is a Gaussian pdf with mean ρu and variance $1-\rho^2$. In particular, the integral of the second factor with respect to v is one. Therefore, the first factor is the marginal pdf of X and the second factor is the conditional pdf of Y given $X = u$:

$$\begin{aligned} f_X(u) &= \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u^2}{2}\right) \\ f_{Y|X}(v|u) &= \frac{1}{\sqrt{2\pi(1-\rho^2)}} \exp\left(-\frac{(v-\rho u)^2}{2(1-\rho^2)}\right). \end{aligned} \tag{4.41}$$

Thus, X is a standard normal random variable. By symmetry, Y is also a standard normal random variable. This proves (a).

The class of bivariate normal pdfs is preserved under linear transformations corresponding to multiplication of $\binom{X}{Y}$ by a matrix A if $\det A \neq 0$. Given a and b, we can select c and d so that the matrix $A = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)$ has $\det(A) = ad - bc \neq 0$. Then the random vector $A\binom{X}{Y}$ has a bivariate normal pdf, so by part (a) already proven, both of its coordinates are Gaussian random variables. In particular, its first coordinate, $aX + bY$, is a Gaussian random variable. This proves (b).

By (4.41), given $X = u$, the conditional distribution of Y is Gaussian with mean ρu and variance $1-\rho^2$. Therefore, $g^*(u) = E[Y|X = u] = \rho u$. Since X and Y are both standard (i.e. they have mean zero and variance one), $\rho_{X,Y} = E[XY]$, so

$$\begin{aligned} \rho_{X,Y} = E[XY] &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} uv f_X(u) f_{Y|X}(v|u)\,dv\,du \\ &= \int_{-\infty}^{\infty} u f_X(u) \left(\int_{-\infty}^{\infty} v f_{Y|X}(v|u)\,dv\right) du \\ &= \rho \int_{-\infty}^{\infty} u^2 f_X(u)\,du = \rho. \end{aligned}$$

This proves (c).

If $\rho \neq 0$ then $\rho_{X,Y} \neq 0$ so that X and Y are not independent. If $\rho = 0$ then $f_{X,Y}$ factors into the product of two single-variable normal pdfs, so X and Y are independent. This proves (d).

By (4.35), $L^*(u) = \rho u$. Therefore, $L^*(u) = g^*(u)$, as claimed, proving (e). By (4.36), the MSE for using L^* is $\sigma_e^2 = 1-\rho^2$, so (f) follows from (4.41). ■

Example 4.11.3 Let X and Y be jointly Gaussian random variables with mean zero, $\sigma_X^2 = 5$, $\sigma_Y^2 = 2$, and $\mathrm{Cov}(X,Y) = -1$. Find $P\{X + 2Y \geq 1\}$.

Solution: Let $Z = X + 2Y$. Then Z is a linear combination of jointly Gaussian random variables, so Z itself is a Gaussian random variable. Also, $E[Z] = E[X] + 2E[Y] = 0$ and

$$\begin{aligned} \sigma_Z^2 &= \mathrm{Cov}(X+2Y, X+2Y) = \mathrm{Cov}(X,X) + \mathrm{Cov}(X,2Y) + \mathrm{Cov}(2Y,X) + \mathrm{Cov}(2Y,2Y) \\ &= \sigma_X^2 + 4\mathrm{Cov}(X,Y) + 4\sigma_Y^2 = 5 - 4 + 8 = 9. \end{aligned}$$

Thus, $P\{X + 2Y \geq 1\} = P\{Z \geq 1\} = P\left\{\frac{Z}{3} \geq \frac{1}{3}\right\} = Q\left(\frac{1}{3}\right) = 1 - \Phi(\frac{1}{3}) \approx 0.3694$.

Example 4.11.4 Let X and Y be jointly Gaussian random variables with mean zero, variance one, and $\text{Cov}(X,Y)=\rho$. Find $E[Y^2|X]$, the best estimator of Y^2 given X. (Hint: X and Y^2 are not jointly Gaussian. But you know the conditional distribution of Y given $X=u$ and can use it to find the conditional second moment of Y given $X=u$.)

Solution: Recall the fact that $E[Z^2]=E[Z]^2+\text{Var}(Z)$ for a random variable Z. The idea is to apply the fact to the conditional distribution of Y given X. Given $X=u$, the conditional distribution of Y is Gaussian with mean ρu and variance $1-\rho^2$. Thus, $E[Y^2|X=u]=(\rho u)^2+1-\rho^2$. Therefore, $E[Y^2|X]=(\rho X)^2+1-\rho^2$.

Example 4.11.5 Suppose X and Y are zero-mean unit-variance jointly Gaussian random variables with correlation coefficient $\rho=0.5$.
(a) Find $\text{Var}(3X-2Y)$.
(b) Find the numerical value of $P\{(3X-2Y)^2\le 28\}$.
(c) Find the numerical value of $E[Y|X=3]$.

Solution: (a)

$$\begin{aligned}\text{Var}(3X-2Y) &= \text{Cov}(3X-2Y,3X-2Y)\\ &= 9\text{Var}(X)-12\text{Cov}(X,Y)+4\text{Var}(Y)=9-6+4=7.\end{aligned}$$

(b) Also, the random variable $3X-2Y$ has mean zero and is Gaussian. Therefore, $P\{(3X-2Y)^2\le 28\}=P\{-\sqrt{28}\le 3X-2Y\le\sqrt{28}\}=2P\left\{0\le\frac{3X-2Y}{\sqrt{7}}\le\sqrt{\frac{28}{7}}\right\}=2(\Phi(2)-0.5)\approx 0.9545$.
(c) Since X and Y are jointly Gaussian, $E[Y|X=3]=\widehat{E}[Y|X=3]$, so plugging numbers into (4.35) yields:

$$E[Y|X=3]=\mu_Y+\sigma_Y\rho_{X,Y}\left(\frac{3-\mu_X}{\sigma_X}\right)=3\rho_{X,Y}=1.5.$$

4.12 Short Answer Questions

Section 4.1[video]

1. Suppose X and Y are random variables with values in $[0,1]$ and joint CDF that satisfies $F_{X,Y}(u,v)=\frac{u^2v^2+uv^3}{2}$ for $(u,v)\in[0,1]^2$. Find $F_{X,Y}(2,0.5)$.
2. Suppose X and Y are random variables with values in $[0,1]$ and joint CDF that satisfies $F_{X,Y}(u,v)=\frac{u^2v^2+uv^3}{2}$ for $(u,v)\in[0,1]^2$. Find $P\left\{(X,Y)\in\left[\frac{1}{3},\frac{2}{3}\right]\times\left[\frac{1}{2},1\right]\right\}$.

Section 4.2[video]

1. Find $P\{X = Y\}$ if $p_{X,Y}(i, j) = 2^{-(i+j+2)}$ for nonnegative integers i and j.

Section 4.3[video]

1. Find the constant c so that f defined by $f(u, v) = c(u^2 + uv + v^2)I_{\{(u,v)\in[0,1]^2\}}$ is a valid pdf.
2. Find $E[X]$ if (X, Y) has joint pdf $f_{X,Y}(u, v) = \frac{4}{5}(1 + uv)I_{\{(u,v)\in[0,1]^2\}}$.
3. Consider an equilateral triangle inscribed in a circle. If a random point is uniformly distributed over the region bounded by the circle, what is the probability the point is inside the triangle?
4. Find $E[|X - Y|]$ assuming (X, Y) is uniformly distributed over the square region $[0, 1]^2$.
5. Suppose (X, Y) has joint pdf $f_{X,Y}(u, v) = c(1 + uv)I_{\{u\geq 0,v\geq 0,u+v\leq 1\}}$, where the value of the constant c is not needed to answer this question. Find $E[Y|X = 0.5]$.

Section 4.4[video]

1. Suppose X and Y are independent random variables such that X is uniformly distributed over the interval $[0, 2]$ and Y is uniformly distributed over the interval $[1, 3]$. Find $P\{X \geq Y\}$.
2. Suppose X and Y are independent random variables such that X has the exponential distribution with parameter $\lambda = 2$ and Y is uniformly distributed over the interval $[0, 1]$. Find $P\{X \geq Y\}$.
3. Find $E\left[\frac{1}{X+Y}\right]$, assuming that X and Y are independent random variables, each uniformly distributed over the interval $[0, 1]$.

Section 4.5[video]

1. Find the maximum value of the pmf of S (i.e. $\max_k p_S(k)$) where $S = X + Y$, and X and Y are independent with $p_X(i) = \frac{i}{10}$ for $1 \leq i \leq 4$, and $p_Y(j) = \frac{j^2}{30}$ for $1 \leq j \leq 4$.
2. Find the maximum value of the pdf of S (i.e. $\max_c f_S(c)$) where $S = X + Y$ and (X, Y) is uniformly distributed over $\{(u, v) : u^2 + v^2 \leq 1\}$.

Section 4.6[video]

1. Suppose an ant is positioned at a random point that is uniformly distributed over a square region with unit area. Let D be the minimum distance the ant would need to walk to exit the region. Find $E[D]$.
2. Suppose X and Y are independent random variables, with each being uniformly distributed over the interval $[0.9, 1.1]$. Find $P\{XY \geq 1.1\}$.

Section 4.7[video]

1. Suppose (X,Y) is uniformly distributed over $[0,1]^2$. Let $W = \frac{X}{X+Y}$ and $Z = X+Y$. Find $f_{W,Z}(0.3, 0.8)$.

Section 4.8[video]

1. Find $\text{Cov}(X,Y)$ if $\text{Var}(X+Y) = 9$ and $\text{Var}(X-Y) = 6$.

2. Find $\text{Var}(X+2Y)$ under the assumptions $E[X] = E[Y] = 1$, $E[X^2] = E[XY] = 5$ and $E[Y^2] = 10$.

3. Find the maximum possible value of $\text{Cov}(X,Y)$ subject to $\text{Var}(X) = 12$ and $\text{Var}(Y) = 3$.

4. Find $\text{Var}\left(\frac{S_{30}}{\sqrt{30}}\right)$ where $S_{30} = X_1 + \cdots + X_{30}$ such that the X_i's are uncorrelated, $E[X_i] = 1$ and $\text{Var}(X_i) = 4$.

5. Find $\text{Var}(X_1 + \cdots + X_{100})$ where $\text{Cov}(X_i, X_j) = \max\{3 - |i-j|, 0\}$ for $i, j \in \{1, \cdots, 100\}$.

Section 4.9[video]

1. Suppose $X = Y + N$ where Y and N are uncorrelated with mean zero, $\text{Var}(Y) = 4$, and $\text{Var}(N) = 8$. Find the minimum MSE for linear estimation, $E[(Y - \widehat{E}(Y|X])^2]$, and find $\widehat{E}[Y|X = 6]$.

2. Suppose X has pdf $f_X(u) = \frac{e^{-|u|}}{2}$. Find $\widehat{E}[X^3|X]$. (Hint: $E[X^n] = n!$ for n even.)

3. Suppose $(N_t : t \geq 0)$ is a Poisson random process with rate $\lambda = 10$. Find $E[N_7|N_5 = 42]$.

Section 4.10[video]

1. A gambler repeatedly plays a game, each time winning two units of money with probability 0.3 and losing one unit of money with probability 0.7. Let A be the event that after 1000 plays, the gambler has at least as much money as at the beginning. Find an upper bound on $P(A)$ using the Chebychev inequality as in Proposition 4.10.1.

2. A gambler repeatedly plays a game, each time winning two units of money with probability 0.3 and losing one unit of money with probability 0.7. Let A be the event that after 1000 plays, the gambler has at least as much money as at the beginning. Find the approximate value of $P(A)$ based on the Gaussian approximation. (To be definite, use the continuity correction.)

Section 4.11[video]

1. Suppose X and Y are jointly Gaussian with equal variances. Find $a \geq 0$ so that $X + aY$ is independent of $X - aY$.

2. Suppose X and Y are jointly Gausian with mean zero, $\sigma_X^2 = 3$, $\sigma_Y^2 = 12$, and $\rho = 0.5$. Find $P\{X - 2Y \geq 10\}$.

4.13 Problems

Jointly distributed random variables including independent random variables
Sections 4.1-4.4

4.1. **[A joint pmf]**
The joint pmf $p_{X,Y}(u,v)$ of X and Y is shown in the table below.

	u=0	u=1	u=2	u=3
v=4	0	0.1	0.1	0.2
v=5	0.2	0	0	0
v=6	0	0.2	0.1	0.1

(a) Find the marginal pmfs $p_X(u)$ and $p_Y(v)$.

(b) Let $Z = X + Y$. Find p_Z, the pmf of Z.

(c) Are X and Y independent random variables? Justify your answer.

(d) Find $p_{Y|X}(v|3)$ for all v and find $E[Y|X = 3]$.

4.2. **[A joint distribution]**
Suppose X and Y are jointly continuous with joint pdf

$$f_{X,Y}(u,v) = \begin{cases} ve^{-(1+u)v} & u \geq 0, v \geq 0 \\ 0 & \text{else.} \end{cases}$$

(a) Find the marginal pdfs, f_X and f_Y. Note that these functions are defined on the entire real line. (Hint: To find the marginal of X it might help to review the gamma density.)

(b) Find the conditional pdfs, $f_{Y|X}$ and $f_{X|Y}$. Be sure to indicate where these functions are well defined, and where they are zero, as well as giving the nonzero values.

(c) Find the joint CDF, $F_{X,Y}(u_o, v_o)$.

(d) Are X and Y independent? Justify your answer.

4.3. **[A joint distribution on a quarter disk]**
Let X and Y have joint pdf $f_{X,Y}(u,v) = \begin{cases} 8uv & \text{if } u \geq 0, v \geq 0,\ u^2 + v^2 \leq 1 \\ 0 & \text{else.} \end{cases}$

(a) Are X and Y independent random variables? Justify your answer.

(b) Find $P\{X \leq Y\}$.

(c) Find $P\{X \leq 0.5, Y \leq 0.5\}$.

(d) Find $E\left[\frac{1}{XY}\right]$.

(e) Find $P\{X + Y \leq 1\}$.

4.4. **[Working with the joint pdf of two independent variables]**
Suppose X and Y are independent random variables such that X is uniformly distributed over the interval $[0, 1]$ and Y is exponentially distributed with parameter $\lambda > 0$.

(a) What is the joint pdf, $f_{X,Y}$?

(b) Express $P\{Y \geq X\}$ in terms of λ. Also, identify the limits of your answer as $\lambda \to 0$ or $\lambda \to +\infty$.

(c) Express $P\{Xe^Y \geq 1\}$ in terms of λ. Also, identify the limits of your answer as $\lambda \to 0$ or $\lambda \to +\infty$.

4.5. **[Working with a joint pdf]**
Suppose random variables X and Y have the joint probability density function (pdf):

$$f_{XY}(u, v) = \begin{cases} \frac{3}{2}, & u > 0,\ u^2 < v < 1 \\ 0, & \text{elsewhere} \end{cases}$$

(a) Are X and Y independent? Explain your answer.

(b) Determine the marginal pdf of X, $f_X(u)$.

(c) For what values of u is the conditional pdf of Y given $X = u$, $f_{Y|X}(v|u)$, well defined?

(d) Determine $f_{Y|X}(v|u)$ for the values of u for which it is well defined. Be sure to indicate where its value is zero.

(e) Determine $P\{Y > X\}$.

4.6. **[Working with a simple joint pdf]**
Let X and Y be two random variables with joint pdf

$$f_{X,Y}(u, v) = \begin{cases} 2ue^{-u-2v} & u \geq 0, v \geq 0, \\ 0 & else. \end{cases}$$

(a) Find the marginal pdf of X.

(b) Find the conditional pdf $f_{Y|X}(v|u)$. Be sure to include specifying what values of u, v it is well defined for and what values it is zero for.

(c) Find $P\{X + 2Y \leq 2\}$.

(d) Find $E\left[\frac{Y}{X}\right]$.

4.7. **[Recognizing independence]**
Decide whether X and Y are independent for each of the following three joint pdfs. If they are independent, identify the marginal pdfs f_X and f_Y. If they are not, give a reason why.

(a) $f_{X,Y}(u,v) = \begin{cases} \frac{4}{\pi}e^{-(u^2+v^2)} & u,v \geq 0 \\ 0 & \text{else.} \end{cases}$

(b) $f_{X,Y}(u,v) = \begin{cases} -\frac{ln(u)v^2}{21} & 0 \leq u \leq 1, 1 \leq v \leq 4 \\ 0 & \text{else.} \end{cases}$

(c) $f_{X,Y}(u,v) = \begin{cases} \frac{(96)u^2v^2}{\pi} & u^2+v^2 \leq 1 \\ 0 & \text{else.} \end{cases}$

4.8. **[Independent or not?]**
Decide whether X and Y are independent for each of the following pdfs. The constant C in each case represents the value making the pdf integrate to one. Justify your answer.

(a) $f_{X,Y}(u,v) = \begin{cases} C(u^2+v^2) & \text{if } 0 \leq u \leq 1 \text{ and } 2 \leq v \leq 3 \\ 0 & \text{else} \end{cases}$

(b) $f_{X,Y}(u,v) = \begin{cases} Cuv(1+\cos(\pi u)) & \text{if } 0 \leq u \leq 1 \text{ and } 2 \leq v \leq 3 \\ 0 & \text{else} \end{cases}$

(c) $f_{X,Y}(u,v) = \begin{cases} C\exp(-u-2v) & \text{if } 0 \leq v \leq u \\ 0 & \text{else} \end{cases}$

4.9. **[Working with a joint pdf I]**
Suppose two random variables X and Y have the following joint pdf:

$$f_{X,Y}(u,v) = \begin{cases} \frac{uv+1}{C} & \text{if } -1 \leq u \leq 1 \text{ and } -1 \leq v \leq 1, \\ 0 & \text{else.} \end{cases}$$

(a) Find the pdf of f_X (you need not find the constant C at this point).

(b) Find the constant C.

(c) For $-1 \leq u_o \leq 1$, find the conditional pdf $f_{Y|X}(v|u_o)$. Specify it for all real values of v.

(d) Find $E[X^mY^n]$ for integers $m,n \geq 0$.

(e) Find $P\{X+Y \geq 1\}$.

4.10. **[Working with a joint pdf II]**
Suppose X and Y are jointly continuous-type random variables with joint pdf:

$$f_{X,Y}(u,v) = \begin{cases} \alpha(u+v^2) & \text{if } 0 \leq u \leq 1 \text{ and } 0 \leq v \leq 2u, \\ 0 & \text{else.} \end{cases}$$

where α is a positive real that you will find.

(a) Determine whether X and Y are independent. Justify your answer.

(b) Find the marginal pdf $f_X(u)$. (You can express your answer using α. Specify the pdf over the entire line, including where it is zero.)

(c) Find α.

(d) For what values of u is the conditional pdf $f_{Y|X}(v|u)$ well defined for all v?

(e) Find $P\{X < Y\}$.

4.11. **[Near simultaneous failure times]**
Suppose S and T represent the lifetimes of two computers. Suppose the lifetimes are mutually independent, and each has the exponential distribution with some parameter $\lambda = 1$.

(a) Find $P\{|S - T| \leq 1\}$ by setting up an integral over $\{(u, v) : u \geq 0, v \geq 0, |u - v| \leq 1\}$ and evaluating the integral.

(b) Use the memoryless property of the exponential distribution and the fact the computer lifetimes have the same distribution, to explain why $P\{|S-T| \leq 1\}$ is equal to $P\{S \leq 1\}$, or equivalently, to $P\{T \leq 1\}$.

4.12. **[The volume of a random cylinder]**
Consider a cylinder with height H and radius R. Its volume V is given by $V = \pi H R^2$. Suppose H and R are mutually independent, and each is uniformly distributed over the interval $[0, 1]$.

(a) Using LOTUS, find the mean of V.

(b) What is the support of the distribution of V?

(c) Find the CDF of V.

(d) Find the pdf of V.

Joint pdfs and functions of two random variables Sections 4.5&4.6

4.13. **[Joint densities]**
X and Y are two random variables with the following joint pdf:

$$f_{X,Y}(u, v) = \begin{cases} A(1 - |u - v|) & 0 < u < 1, 0 < v < 1 \\ 0 & else \end{cases}$$

(a) Find A.

(b) Find marginal pdfs for X and Y.

(c) Find $P\{X > Y\}$.

(d) Find $P(X + Y < 1|X > 1/2)$.

4.14. **[Functions of random variables]**
Two random variables X and Y have the following joint pdf:

$$f_{X,Y}(u, v) = \begin{cases} 2\exp(-(u + v)) & 0 < u < v < \infty \\ 0 & \text{else.} \end{cases}$$

A new random variable Z is defined as: Z=X+Y. Find the CDF and pdf of Z.

4.15. **[Joint Densities and functions of two random variables]**
Let X and Y have joint pdf

$$f_{X,Y}(u,v) = \begin{cases} A\left(1-\sqrt{u^2+v^2}\right) & u^2+v^2<1 \\ 0 & \text{else.} \end{cases}$$

Hint: Use of polar coordinates is useful for all parts of this problem.

(a) Find the value of A.

(b) Let $Z = X^2 + Y^2$. Find the pdf of the random variable Z.

(c) Find $E[Z^5]$ using LOTUS for joint pdfs: $E[g(X,Y)] = \int\int_{\mathbb{R}^2} g(u,v) f_{X,Y}(u,v) du dv$.

4.16. **[A function of two random variables]**
Two resistors are connected in series to a one volt voltage source. Suppose that the resistance values R_1 and R_2 (measured in ohms) are independent random variables, each uniformly distributed on the interval (0,1). Find the pdf $f_I(a)$ of the current I (measured in amperes) in the circuit.

Moments of jointly distributed random variables, minimum mean square error estimation Sections 4.8-4.9.

4.17. **[Deducing a covariance from variances]**
Consider random variables X and Y on the same probability space.

(a) If $\text{Var}(X+2Y) = 40$ and $\text{Var}(X-2Y) = 20$, what is $\text{Cov}(X,Y)$?

(b) In part (a), determine $\rho_{X,Y}$ if $\text{Var}(X) = 2 \cdot \text{Var}(Y)$.

4.18. **[Variance and correlation]**
Consider two random variables X and Y.

(a) If $\text{Var}(X+3Y) = \text{Var}(X-3Y)$, must X and Y be uncorrelated?

(b) If $\text{Var}(X) = \text{Var}(Y)$, must X and Y be uncorrelated?

(c) If $\text{Var}(X+Y) = \text{Var}(X) + \text{Var}(Y)$, must X and Y be uncorrelated?

(d) If $\text{Cov}(X^2, Y^2) = 0$, must X and Y be uncorrelated?

4.19. **[Working with covariances]**
Suppose X, Y, and Z are random variables, each with mean zero and variance 20, such that $\text{Cov}(X,Y) = \text{Cov}(X,Z) = 10$ and $\text{Cov}(Y,Z) = 5$. Be sure to show your work, as usual, for all parts below.

(a) Find $\text{Cov}(X+Y, X-Y)$.

(b) Find Cov(3X+Z,3X+Y).

(c) Find $E[(X+Y)^2]$.

(d) Find $\widehat{E}[Y+Z|X=3]$.

4.20. **[Variances and covariances of sums]**
Rewrite the expressions below in terms of $\text{Var}(X)$, $\text{Var}(Y)$, $\text{Var}(Z)$, and $\text{Cov}(X,Y)$.

(a) $\text{Cov}(3X+2, 5Y-1)$

(b) $\text{Cov}(2X+1, X+5Y-1)$.

(c) $\text{Cov}(2X+3Z, Y+2Z)$ where Z is uncorrelated with both X and Y.

4.21. **[Covariance for sampling without replacement]**
Five balls numbered one through five are in a bag. Two balls are drawn at random, without replacement, with all possible outcomes having equal probability. Let X be the number on the first ball drawn and Y be the number on the second ball drawn.

(a) Find $E[X]$.

(b) Find $\text{Var}(X)$.

(c) Find $E[XY]$.

(d) Find the correlation coefficient, $\rho_{X,Y}$.

4.22. **[Signal to noise ratio with correlated observations]**
Random variables X_1 and X_2 represent two observations of a signal corrupted by noise. They have the same mean μ and variance σ^2. The *signal-to-noise-ratio* (SNR) of the observation X_1 or X_2 is defined as the ratio $SNR_X = \frac{\mu^2}{\sigma^2}$. A system designer chooses the averaging strategy, whereby she constructs a new random variable $S = \frac{X_1+X_2}{2}$.

(a) Show that the SNR of S is twice that of the individual observations, if X_1 and X_2 are uncorrelated.

(b) The system designer notices that the averaging strategy is giving $SNR_S = (1.5)SNR_X$. She correctly assumes that the observations X_1 and X_2 are correlated. Determine the value of the correlation coefficient $\rho_{X_1X_2}$.

(c) Under what condition on $\rho_{X,Y}$ can the averaging strategy result in an SNR_S that is arbitrarily high?

4.23. **[The covariance of sums of correlated random variables]**
Suppose $X_1, \ldots X_n$ and $Y_1, \ldots, Y_n$ are random variables on a common probability space such that $\text{Var}(X_i) = \text{Var}(Y_i) = 4$ for all i, and

$$\rho_{X_i,Y_j} = \begin{cases} 3/4 & \text{if } i = j \\ -1/4 & \text{if } |i-j| = 1 \\ 0 & \text{else.} \end{cases}$$

Let $W = \sum_{i=1}^n X_i$ and $Z = \sum_{i=1}^n Y_i$. Express $\text{Cov}(W,Z)$ as a function of n.

4.24. **[Linear minimum MSE estimation from uncorrelated observations]**
Suppose Y is estimated by a linear estimator, $L(X_1, X_2) = a + bX_1 + cX_2$, such that X_1 and X_2 have mean zero and are uncorrelated with each other.

(a) Determine a, b and c to minimize the MSE, $E[(Y - (a + bX_1 + cX_2))^2]$. Express your answer in terms of $E[Y]$, the variances of X_1 and X_2, and the covariances $\mathrm{Cov}(Y, X_1)$ and $\mathrm{Cov}(Y, X_2)$.

(b) Express the MSE for the estimator found in part (a) in terms of the variances of X_1, X_2, and Y and the covariances $\mathrm{Cov}(Y, X_1)$ and $\mathrm{Cov}(Y, X_2)$.

4.25. **[An MMSE estimation problem]**
Suppose two random variables X and Y have the following joint pdf:

$$f_{X,Y}(u, v) = \begin{cases} \frac{uv+1}{4} & \text{if } -1 \le u \le 1 \text{ and } -1 \le v \le 1, \\ 0 & \text{else} \end{cases}$$

(a) Find the numerical value of $P\{Y \le -\frac{1}{3}\}$.

(b) Find the constant estimator δ^* of Y and the corresponding mean square error (MSE).

(c) Find the unconstrained estimator $g^*(X)$ for observation X and the corresponding MSE. (Hint: The MSE must be deterministic, not random.)

(d) Find the linear estimator $L^*(X)$ for observation X and the corresponding MSE.

4.26. **[What's new? or the innovations method]**
The problem above on linear MSE estimation from uncorrelated observations shows that if Y is to be estimated by a linear combination of X_1 and X_2, the solution is particularly simple if $\mathrm{Cov}(X_1, X_2) = 0$. Intuitively, that is because, even if X_1 is already known, X_2 offers completely new information. If $\mathrm{Cov}(X_1, X_2) \neq 0$, then by preprocessing, we can reduce the problem to the previous case. The idea is to let $\widetilde{X_2} = X_2 - hX_1$, where the constant h is chosen so that X_1 and $\widetilde{X_2}$ are uncorrelated. Intuitively, $\widetilde{X_2}$ is the part of X_2 that is new, or innovative, to someone who already knows X_1. Then, the best linear estimator of Y given X_1 and $\widetilde{X_2}$ can be easily found by the method of the previous problem. But a linear estimator based on X_1 and $\widetilde{X_2}$ can also be expressed as a linear estimator based on X_1 and X_2, or vice versa. So the estimator would also be the best linear estimator of Y given X_1 and X_2.

This recipe is illustrated by an example as follows. Suppose Y, X_1, and X_2 have mean zero and variance one, and suppose $\mathrm{Cov}(Y, X_1) = \mathrm{Cov}(Y, X_2) = 0.8$ and $\mathrm{Cov}(X_1, X_2) = 0.5$.

(a) Find h so that X_1 and $\widetilde{X_2} = X_2 - hX_1$ are uncorrelated.

(b) Find $\mathrm{Var}(\widetilde{X}_2)$ and $\mathrm{Cov}(Y, \widetilde{X}_2)$.

(c) Appealing to the previous problem, find the values of a, b, and c so that the MSE for estimating Y by the linear estimator $a + bX_1 + c\widetilde{X}_2$ is minimized, and find the resulting MSE.

(d) Express the linear estimator found in part (c) as a linear combination of X_1 and X_2.

4.27. **[A singular estimation problem]**
Suppose Y represents a signal, and $P\{Y = 1\} = P\{Y = -1\} = 0.5$. Suppose N represents noise, and N is uniformly distributed over the interval $[-1, 1]$. Assume S and N are mutually independent, and let $X = Y + N$.

(a) Find $g^*(X)$, the MMSE estimator of Y given X. (Hint: Consider some typical values of (X, Y) and think about how Y can be estimated from X.)

(b) Find $L^*(X)$, the MMSE linear estimator of Y given X.

4.28. **[Simple estimation problems]**
Suppose (X, Y) is uniformly distributed over the set $\{(u, v) : 0 \leq v \leq u \leq 1\}$.

(a) Find $E[Y^2|X = u]$ for $0 \leq u \leq 1$.

(b) Find $\widehat{E}[Y|X = u]$. (Hint: There is a Y^2 in part (a) but only a Y in this part. This problem can be done with little to no calculation.)

4.29. **[An estimation problem]**
Suppose X and Y have the following joint pdf:

$$f_{X,Y}(u, v) = \begin{cases} \frac{8uv}{(15)^4} & u \geq 0, v \geq 0, u^2 + v^2 \leq (15)^2 \\ 0 & \text{else} \end{cases}$$

(a) Find the constant estimator, δ^*, of Y, with the smallest mean square error (MSE), and find the MSE.

(b) Find the unconstrained estimator, $g^*(X)$, of Y based on observing X, with the smallest MSE, and find the MSE.

(c) Find the linear estimator, $L^*(X)$, of Y based on observing X, with the smallest MSE, and find the MSE. (Hint: You may use the fact $E[XY] = \frac{75\pi}{4} \approx 58.904$, which can be derived using integration in polar coordinates.)

LLN, CLT, and joint Gaussian distribution Sections 4.10 & 4.11

4.30. **[Law of Large Numbers and Central Limit Theorem]**
A fair die is rolled n times. Let $S_n = X_1 + X_2 + \ldots + X_n$, where X_i is the number showing on the ith roll. Determine a condition on n so the probability the sample average $\frac{S_n}{n}$ is within 1% of the mean μ_X, is greater than 0.95. (Note: This problem is related to Example 4.10.5.)

(a) Solve the problem using the form of the law of large numbers based on the Chebychev inequality (i.e. Proposition 4.10.1 in the notes).

(b) Solve the problem using the Gaussian approximation for S_n, which is suggested by the CLT. (Do not use the continuity correction, because, unless $3.5n \pm (0.01)n\mu_X$ are integers, inserting the term 0.5 is not applicable).

4.31. **[Rate of convergence in law of large numbers for independent Gaussians]**
By the law of large numbers, if $\delta > 0$ and $S_n = X_1 + \cdots + X_n$, where $X_1, X_2, \ldots$ are uncorrelated with mean zero and bounded variance, $\lim_{n\to\infty} P\left\{\left|\frac{S_n}{n}\right| \geq \delta\right\} = 0$.

(a) Express $P\left\{\left|\frac{S_n}{n}\right| \geq \delta\right\}$ in terms of n, δ and the Q function, for the special case $X_1, X_2, \ldots$ are independent, $N(0,1)$ random variables.

(b) An upper bound for $Q(x)$ for $x > 0$, which is also a good approximation for x at least moderately large, is given by

$$Q(x) = \int_x^\infty \frac{1}{\sqrt{2\pi}} e^{-u^2/2} du \leq \int_x^\infty \frac{1}{\sqrt{2\pi}} \left(\frac{u}{x}\right) e^{-u^2/2} du = \frac{1}{x\sqrt{2\pi}} e^{-x^2/2}.$$

Use this bound to obtain an upper bound on the probability in part (a) in terms of δ and n but not using the Q function.

4.32. **[Marathon blackjack]**
In a particular version of the card game *blackjack* offered by gambling casinos[2], if a player uses a particular optimized strategy, then in one game with one unit of money initial bet, the net return is -0.0029 and the standard deviation of the net return is 1.1418 (which can be squared to get the variance). Suppose a player uses this strategy and bets \$100 on each game, regardless of how much the player won or lost in previous games.

(a) What is the expected net gain of the player after 1000 games? (Answer should be a negative dollar amount.)

(b) What is the probability the player is ahead after 1000 games? (Use the Gaussian approximation suggested by the central limit theorem for this and other parts below.)

(c) What is the probability the player is ahead by at least \$1000 after 1000 games?

(d) What value of n is such that after playing n games (with the same initial bet per game), the probability the player is ahead after n games is about 0.4?

4.33. **[Achieving potential in a class]**
(The following is roughly based on the ECE 313 grading scheme, ignoring homework scores and the effect of partial credit.) Consider a class in which grades are based entirely on midterm and final exams. In all, the exams have 100 separate parts, each worth 5 points. A student scoring at least 85%, or 425 points in total, is guaranteed an A score. Throughout this problem, consider a particular student, who, based on performance in other courses and amount of effort put into the class, is estimated to have a 90% chance to complete any particular part correctly. Problem parts not completed correctly receive zero credit.

[2] See http://wizardofodds.com/games/blackjack/appendix/4/

(a) Assume that the scores on different parts are independent. Based on the LLN, about what total score for the semester are we likely to see?

(b) Under the same assumptions in part (a), using the CLT (without the continuity correction, to be definite) calculate the approximate probability the student scores enough points for a guaranteed A score.

(c) Consider the following variation of the assumptions in part (a). The problem parts for exams during the semester are grouped into problems of four parts each, and if X_i and X_j are the scores received on two different parts of the same problem, then the correlation coefficient is $\rho_{X_i,X_j} = 0.8$. The scores for different problems are independent. The total score for problem one, for example, could be written as $Y_1 = X_1 + X_2 + X_3 + X_4$, where X_i is the score for the ith part of the problem. Find the mean and variance of Y_1.

(d) Continuing with the assumptions of part (c), the student takes a total of 25 problems in the exams during the semester (with four parts per exam). Using the CLT (without the continuity correction, to be definite) calculate the approximate probability the student scores enough points for a guaranteed A score.

4.34. **[The CLT and the Poisson distribution]**
Let X be a Poisson random variable with parameter $\lambda = 10$, and let $\widetilde{X}$ be a Gaussian random variable with the same mean and variance as X.

(a) Explain why X can be written as the sum of ten independent, identically distributed random variables. What is the distribution of those random variables? (Hint: Think about Poisson processes. By the CLT, we thus expect that the CDFs of X and $\widetilde{X}$ are approximately equal.)

(b) Compute $p_X(12)$, the pmf of X evaluated at 12.

(c) Compute $P\{11.5 \leq \widetilde{X} \leq 12.5\}$, which, by the CLT with the continuity correction, we expect to be fairly close to the answer of part (b).

(d) Compute $f_{\widetilde{X}}(12)$, the pdf of $\widetilde{X}$ evaluated at 12. (We expect the answer to be fairly close to the answer of part (b).)

4.35. **[Gaussian approximation for confidence intervals]**
Recall that if X has the binomial distribution with parameters n and p, the Chebychev inequality implies that

$$P\{|X - np| \geq a\sigma\} \leq \frac{1}{a^2}, \tag{4.42}$$

where σ^2 is the variance of X: $\sigma = \sqrt{np(1-p)} \leq \frac{\sqrt{n}}{2}$. If n is known and p is estimated by $\widehat{p} = \frac{X}{n}$, it follows that the confidence interval with endpoints $\widehat{p} \pm \frac{a}{2\sqrt{n}}$ contains p with probability at least $1 - \frac{1}{a^2}$. (See Section 2.9.) A less conservative, commonly used approach is to note that by the central limit theorem,

$$P\{|X - np| \geq a\sigma\} \approx 2Q(a). \tag{4.43}$$

Example 2.9.2 showed that $n = 625$ is large enough for the random interval with endpoints $\widehat{p} \pm 10\%$ to contain the true value p with probability at least 96%. Calculate the value of n that would be sufficient for the same precision (i.e. within 10% of p) and confidence (i.e. 96%) based on (4.43) rather than (4.42). Explain your reasoning.

4.36. **[Conditional means for a joint Gaussian pdf]**
Suppose X and Y have a bivariate Gaussian joint distribution with $E[X] = E[Y] = 0$ and $\text{Var}(X) = 1$. (The variance of Y and the correlation coefficient are not given.) Finally, suppose X is independent of $X + Y$.

(a) Find $\text{Cov}(X, Y)$.

(b) Find $E[X|X + Y = 2]$.

(c) Find $E[Y|X = 2]$.

4.37. **[Transforming joint Gaussians to independent random variables]**
Suppose X and Y are jointly Gaussian such that X is $N(0, 9)$, Y is $N(0, 4)$, and the correlation coefficient is denoted by ρ. The solutions to the questions below may depend on ρ and may fail to exist for some values of ρ.

(a) For what value(s) of a is X independent of $X + aY$?

(b) For what value(s) of b is $X + Y$ independent of $X - bY$?

(c) For what value(s) of c is $X + cY$ independent of $X - cY$?

(d) For what value(s) of d is $X + dY$ independent of $(X - dY)^3$?

4.38. **[Jointly Gaussian Random Variables I]**
Suppose X and Y are jointly Gaussian with $\mu_X = 1$, $\mu_Y = 2$, $\sigma_X^2 = 9$, $\sigma_Y^2 = 16$, and $\text{Cov}(X, Y) = 6$.

(a) Describe the marginal distribution of X in words and write the explicit formula for its pdf, $f_X(u)$.

(b) Describe the conditional distribution of Y given $X = 5$ in words, and write the explicit formula for the conditional pdf, $f_{Y|X}(v|5)$.

(c) Find the numerical value of $P(Y \geq 2|X = 5)$.

(d) Find the numerical value of $E[Y^2|X = 5]$.

4.39. **[Jointly Gaussian Random Variables II]**
Suppose Y and W are jointly Gaussian random variables with $E[Y] = 2$, $E[W] = 0$, $\text{Var}(Y) = 16$, $\text{Var}(W) = 4$, and $\rho = 0.25$. Let $X = 3Y + W + 3$.

(a) Find $E[X]$ and $\text{Var}(X)$.

(b) Calculate the numerical value of $P\{X \geq 20\}$.

(c) Find $E[Y|X]$. Your answer should be a function of X.

(d) Find the mean square error, $E[(Y - E[Y|X])^2]$.

4.40. **[Estimation of jointly Gaussian random variables]**
Suppose X and Y are jointly Gaussian random variables such that X is $N(4, 16)$, Y is $N(5, 25)$. and $\rho = 0.4$. Let $Z = X + 4Y - 1$.

(a) Find $E[Z]$ and $\text{Var}(Z)$.

(b) Calculate the numerical value of $P\{Z \geq 40\}$.

(c) Find the unconstrained estimator $g^*(Z)$ of Y based on Z with the minimum MSE, and find the resulting MSE.

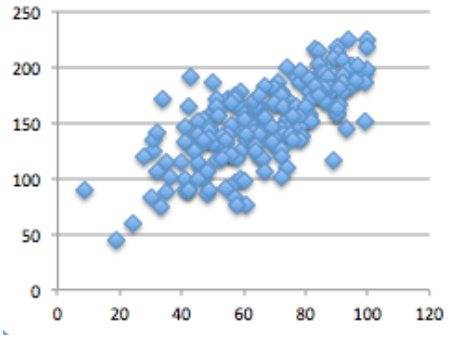
4.41. **[Estimating the third power of a jointly Gaussian random variable]**
See how parts (a) and (b) below combine to yield the answer to part (c).

(a) Suppose $Z = \mu + W$ where W is a $N(0, \sigma^2)$ random variable. Express $E[Z^3]$ in terms of μ and σ^2. (Note that Z is a $N(\mu, \sigma^2)$ random variable, and you are finding an expression for the third moment of such a random variable.)

(b) Suppose X and Y are jointly Gaussian such that X and Y are each standard normal, and the correlation coefficient between X and Y is ρ. Describe the marginal distribution of Y given $X = u$.

(c) Under the same assumptions as (b), express $E[Y^3|X = u]$ in terms of ρ and u.

4.42. **[Joint empirical distribution of ECE 313 scores]**
The scatterplot below shows 205 points, (u_i, v_i), where u_i is the score on exam two, and v_i is the score on the final exam, for the ith student in ECE313 in a recent semester. The empirical mean and standard deviation for exam two are $\mu_X = 67$ and $\sigma_X = 19$, the empirical mean and standard deviation for the final exam are $\mu_Y = 152$ and $\sigma_Y = 35$, and the empirical correlation coefficient (computed using a spreadsheet function) is $\rho = 0.71$. Visual inspection of the data suggests it is reasonable to assume that the joint distribution of the two scores is jointly normal. Let (X, Y) be jointly normal random variables with the above parameter values.



(a) Find $E[Y|X = u]$ as a function of u.

(b) Describe in words the conditional distribution of Y given $X = u$.

Chapter 5

Wrap-up

The topics in these notes are listed in both the table of contents and the index. This chapter briefly summarizes the material, while highlighting some of the connections among the topics.

The probability axioms allow for a mathematical basis for modeling real-world systems with uncertainty, and allow for both discrete-type and continuous-type random variables. Counting problems naturally arise for calculating probabilities when all outcomes are equally likely, and a recurring idea for counting the total number of ways something can be done is to do it sequentially, such as in, "There are n_1 choices for the first ball, and for each choice of the first ball, there are n_2 ways to choose the second ball," and so on. Working with basic probabilities includes working with Karnaugh maps, de Morgan's laws, and definitions of conditional probabilities and mutual independence of two or more events. Our intuition can be strengthened and many calculations made by appealing to the law of total probability and the definition of conditional probability. In particular, we sometimes look back, conditioning on what happened at the end of some scenario, and ask what is the conditional probability that the observation happened in a particular way– using Bayes rule. Binomial coefficients form a link between counting and the binomial probability distribution.

A small number of key discrete-type and continuous-type distributions arise again and again in applications. Knowing the form of the CDFs, pdfs, or pmfs, and formulas for the means and variances, and why each distribution arises frequently in nature and applications, can thus lead to efficient modeling and problem solving. There are relationships among the key distributions. For example, the binomial distribution generalizes the Bernoulli, and the Poisson distribution is the large n, small p limit of the Bernoulli distribution with $np = \lambda$. The exponential distribution is the continuous time version of the geometric distribution; both are memoryless. The exponential distribution is the limit of scaled geometric distributions, and the Gaussian (or normal) distribution, by the central limit theorem, is the limit of standardized sums of large numbers of independent, identically distributed random variables.

The following important concepts apply to both discrete-type random variables and continuous-type random variables:

- Independence of random variables

- Marginals and conditionals
- Functions of one or more random variables, the two or three step procedure to find their distributions, and LOTUS to find their means
- $E[X]$, $\mathrm{Var}(X)$, $E[XY]$, $\mathrm{Cov}(X,Y)$, $\rho_{X,Y}$, σ_X, and relationships among these.
- Binary hypothesis testing (ML rule, MAP rule as likelihood ratio tests)
- Maximum likelihood parameter estimation
- The minimum MSE estimators of Y: $\delta^* = E[Y]$, $L^*(X) = \widehat{E}[Y|X]$, and $g^*(X) = E[Y|X]$.
- Markov, Chebychev, and Schwarz inequalities (The Chebychev inequality can be used for confidence intervals; the Schwarz inequality implies correlation coefficients are between one and minus one.)
- Law of large numbers and central limit theorem

Poisson random processes arise as limits of scaled Bernoulli random processes. Discussion of these processes together entails the Bernoulli, binomial, geometric, negative geometric, exponential, Poisson, and Erlang distributions.

Reliability in these notes is discussed largely in discrete settings–such as the outage probability for an $s-t$ network. Failure rate functions for random variables are discussed for continuous-time positive random variables only, but could be formulated for discrete time.

There are two complementary approaches for dealing with multiple random variables in statistical modeling and analysis, described briefly by the following two lists:

Distribution approach	Moment approach
joint pmf or joint pdf or joint CDF	means, (co)variances, correlation coefficients
marginals, conditionals	
independent	uncorrelated (i.e. $\mathrm{Cov}(X,Y)=0$ or $\rho_{X.Y}=0$)
$E[Y\|X]$	$\widehat{E}[Y\|X]$

That is, on one hand, it sometimes makes sense to postulate or estimate joint distributions. On the other hand, it sometimes makes sense to postulate or estimate joint moments, without explicitly estimating distributions. For jointly Gaussian random variables, the two approaches are equivalent. That is, working with the moments is equivalent to working with the distributions themselves. Independence, which in general is stronger than being uncorrelated, is equivalent to being uncorrelated for the case of jointly Gaussian random variables.

Chapter 6

Appendix

6.1 Some notation

The following notational conventions are used in these notes.

$$
\begin{aligned}
A^c &= \text{ complement of } A \\
AB &= A \cap B \\
A \subset B &\leftrightarrow \text{ any element of } A \text{ is also an element of } B \\
A - B &= AB^c \\
I_A(x) &= \begin{cases} 1 & \text{if } x \in A \\ 0 & \text{else} \end{cases} \\
(a,b) = \{x : a < x < b\} & \qquad (a,b] = \{x : a < x \leq b\} \\
[a,b) = \{x : a \leq x < b\} & \qquad [a,b] = \{x : a \leq x \leq b\} \\
\mathbb{Z} &- \text{ set of integers} \\
\mathbb{Z}_+ &- \text{ set of nonnegative integers} \\
\mathbb{R} &- \text{ set of real numbers} \\
\mathbb{R}_+ &- \text{ set of nonnegative real numbers}
\end{aligned}
$$

$$
\begin{aligned}
\lfloor t \rfloor &= \text{ greatest integer } n \text{ such that } n \leq t \\
\lceil t \rceil &= \text{ least integer } n \text{ such that } n \geq t
\end{aligned}
$$

The *little oh* notation for small numbers is sometimes used: $o(h)$ represents a quantity such that $\lim_{h \to 0} \frac{o(h)}{h} = 0$. For example, the following equivalence holds:

$$
\lim_{h \to 0} \frac{f(t+h) - f(t)}{h} = f'(t) \qquad \Leftrightarrow \qquad f(t+h) = f(t) + hf'(t) + o(h).
$$

6.2 Some sums

The following formulas arise frequently:

$$\begin{aligned}
\sum_{k=0}^{n} r^k &= 1 + r + r^2 + \cdots + r^n = \frac{1 - r^{n+1}}{1 - r} \qquad \text{if } r \neq 1 \\
\sum_{k=0}^{\infty} r^k &= 1 + r + r^2 + \cdots = \frac{1}{1 - r} \qquad \text{if } |r| < 1 \\
\sum_{k=0}^{\infty} \frac{r^k}{k!} &= 1 + r + \frac{r^2}{2} + \frac{r^3}{6} + \cdots = e^r \\
\sum_{k=0}^{n} \binom{n}{k} a^k b^{n-k} &= a^n + n a^{n-1} b + \binom{n}{2} a^{n-2} b^2 + \cdots + n a b^{n-1} + b^n = (a + b)^n \\
\sum_{k=1}^{n} k &= 1 + 2 + \cdots + n = \frac{n(n+1)}{2} \\
\sum_{k=1}^{n} k^2 &= 1 + 4 + 9 + \cdots + n^2 = \frac{n(n+1)(2n+1)}{6}
\end{aligned}$$

6.3 Frequently used distributions

6.3.1 Key discrete-type distributions

Bernoulli(p)**:** $0 \leq p \leq 1$

$$\text{pmf: } p(i) = \begin{cases} p & i = 1 \\ 1 - p & i = 0 \end{cases}$$

$$\text{mean: } p \qquad \text{variance: } p(1 - p)$$

Example: One if heads shows and zero if tails shows for the flip of a coin. The coin is called fair if $p = \frac{1}{2}$ and biased otherwise.

Binomial(n, p)**:** $n \geq 1$, $0 \leq p \leq 1$

$$\text{pmf: } p(i) = \binom{n}{i} p^i (1 - p)^{n-i} \qquad 0 \leq i \leq n$$

$$\text{mean: } np \qquad \text{variance: } np(1 - p)$$

Significance: Sum of n independent Bernoulli random variables with parameter p.

Poisson(λ)**:** $\lambda \geq 0$

$$\text{pmf: } p(i) = \frac{\lambda^i e^{-\lambda}}{i!} \quad i \geq 0$$
$$\text{mean: } \lambda \qquad \text{variance: } \lambda$$

Example: Number of phone calls placed during a ten second interval in a large city.

Significant property: The Poisson pmf is the limit of the binomial pmf as $n \to +\infty$ and $p \to 0$ in such a way that $np \to \lambda$.

Geometric(p)**:** $0 < p \leq 1$

$$\text{pmf: } p(i) = (1-p)^{i-1}p \qquad i \geq 1$$
$$\text{mean: } \frac{1}{p} \qquad \text{variance: } \frac{1-p}{p^2}$$

Example: Number of independent flips of a coin until heads first shows.

Significant property: If L has the geometric distribution with parameter p, $P\{L > i\} = (1-p)^i$ for integers $i \geq 1$. So L has the *memoryless property* in discrete time:

$$P\{L > i + j | L > i\} \quad = \quad P\{L > j\} \text{ for } i, j \geq 0.$$

Any positive integer-valued random variable with this property has the geometric distribution for some p.

6.3.2 Key continuous-type distributions

Uniform(a, b)**:** $-\infty < a < b < \infty$

$$\text{pdf: } f(u) = \begin{cases} \frac{1}{b-a} & a \leq u \leq b \\ 0 & \text{else} \end{cases} \qquad \text{mean: } \frac{a+b}{2} \qquad \text{variance: } \frac{(b-a)^2}{12}$$

Exponential(λ)**:** $\lambda > 0$

$$\text{pdf: } f(t) = \lambda e^{-\lambda t} \quad t \geq 0 \qquad \text{mean: } \frac{1}{\lambda} \qquad \text{variance: } \frac{1}{\lambda^2}$$

Example: Time elapsed between noon sharp and the first time a telephone call is placed after that, in a city, on a given day.

Significant property: If T has the exponential distribution with parameter λ, $P\{T \geq t\} = e^{-\lambda t}$ for $t \geq 0$. So T has the *memoryless property* in continuous time:

$$P\{T \geq s + t | T \geq s\} \quad = \quad P\{T \geq t\} \qquad s, t \geq 0$$

Any nonnegative random variable with the memoryless property in continuous time is exponentially distributed. Failure rate function is constant: $h(t) \equiv \lambda$.

Erlang(r, λ)**:** $r \geq 1$, $\lambda \geq 0$

$$\text{pdf: } f(t) = \frac{\lambda^r t^{r-1} e^{-\lambda t}}{(r-1)!} \qquad t \geq 0 \quad \text{mean: } \frac{r}{\lambda} \qquad \text{variance: } \frac{r}{\lambda^2}$$

Significant property: The distribution of the sum of r independent random variables, each having the exponential distribution with parameter λ.

Gaussian or Normal(μ, σ^2)**:** $\mu \in \mathbb{R}$, $\sigma \geq 0$

$$\text{pdf : } f(u) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(u-\mu)^2}{2\sigma^2}\right) \qquad \text{mean: } \mu \qquad \text{variance: } \sigma^2$$

Notation: $Q(c) = 1 - \Phi(c) = \int_c^\infty \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}} du$
Significant property (CLT): For independent, identically distributed r.v.'s with mean μ, variance σ^2:

$$\lim_{n\to\infty} P\left\{\frac{X_1 + \cdots + X_n - n\mu}{\sqrt{n\sigma^2}} \leq c\right\} \quad = \quad \Phi(c)$$

Rayleigh(σ^2)**:** $\sigma^2 > 0$

$$\text{pdf: } f(r) = \frac{r}{\sigma^2} \exp\left(-\frac{r^2}{2\sigma^2}\right) \qquad r > 0 \quad \text{CDF} : 1 - \exp\left(-\frac{r^2}{2\sigma^2}\right)$$
$$\text{mean: } \sigma\sqrt{\frac{\pi}{2}} \qquad \text{variance: } \sigma^2\left(2 - \frac{\pi}{2}\right)$$

Example: Instantaneous value of the envelope of a mean zero, narrow band noise signal.

Significant property: If X and Y are independent, $N(0, \sigma^2)$ random variables, then $(X^2 + Y^2)^{\frac{1}{2}}$ has the Rayleigh(σ^2) distribution. Failure rate function is linear: $h(t) = \frac{t}{\sigma^2}$.

6.4 Normal tables

Tables 6.1 and 6.2 below were computed using Abramowitz and Stegun, *Handbook of Mathematical Functions*, Formula 7.1.26, which has maximum error at most 1.5×10^{-7}.

Table 6.1: Φ function, the area under the standard normal pdf to the left of x.

x	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990

Table 6.2: Q function, the area under the standard normal pdf to the right of x.

x	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641
0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
2.9	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
3.0	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010

x	0.0	0.2	0.4	0.6	0.8
0.0	0.5000000	0.4207403	0.3445783	0.2742531	0.2118553
1.0	0.1586553	0.1150697	0.0807567	0.0547993	0.0359303
2.0	0.0227501	0.0139034	0.0081975	0.0046612	0.0025552
3.0	0.0013500	0.0006872	0.0003370	0.0001591	0.0000724
4.0	0.0000317	0.0000134	0.0000054	0.0000021	0.0000008

6.5 Answers to short answer questions

Section 1.2 **1.** 0.32 **2.** 0.30
Section 1.3 **1.** 5040 **2.** 2790 **3.** 6 **4.** 2520
Section 1.4 **1.** 1/5 **2.** 1/20
Section 1.5 **1.** 6/9
Section 2.2 **1.** 16.5 **2.** 1 **3.** 9
Section 2.3 **1.** 1/3 **2.** 2/27 **3.** 3
Section 2.4 **1.** 0.00804 **2.** 9.72 **3.** 0.2276
Section 2.5 **1.** 35 **2.** 6
Section 2.6 **1.** 7,2,5,4 **2.** 2,2,2,2,3 **3.** 7,9,14,18
Section 2.7 **1.** 2.303 **2.** 0.2240
Section 2.8 **1.** $\sqrt{10}$ **2.** 1/3
Section 2.9 **1.** 1/16 **2.** 1,582
Section 2.10 **1.** 0.1705 **2.** 0.2031 **3.** 1/3
Section 2.11 **1.** 0.3, 0.5 **2.** 1/3 **3.** 6
Section 2.12 **1.** 10^{-8} **2.** 1.2×10^{-5}
Section 3.1 **1.** 0.73775 **2.** 0.9345879
Section 3.2 **1.** 1/18 **2.** $\pi/4$
Section 3.3 **1.** 4/45 **2.** 1/3
Section 3.4 **1.** 0.21972 **2.** 34.6573
Section 3.5 **1.** 0.0029166 **2.** 41.589, 6.45 **3.** 0.125
Section 3.6 **1.** 0.175085 **2.** 0.1587 **3.** 0.5281 **4.** 0.1792
Section 3.7 **1.** 5 **2.** 4.5249
Section 3.8 **1.** $\frac{1}{\sqrt{2\pi\sigma^2 c}} \exp\left(-\frac{c}{2\sigma^2}\right) I_{\{c>0\}}$ **2.** $g(u) = \sqrt{u}$
Section 3.9 **1.** $\frac{1}{1+a}$ **2.** $\frac{\alpha}{t} I_{\{t \geq 1\}}$
Section 3.10 **1.** 0, 0.63212 **2.** 0.3191 **3.** 1/3
Section 4.1 **1.** 3/16 **2.** 0.2708333
Section 4.2 **1.** 1/3
Section 4.3 **1.** 12/11 **2.** 8/15 **3.** 0.4135 **4.** 1/3 **5.** 7/27
Section 4.4 **1.** 1/8 **2.** 0.43233 **3.** 1.38629
Section 4.5 **1.** 0.28 **2.** 0.45016
Section 4.6 **1.** 1/6 **2.** 0.12897
Section 4.7 **1.** 0.8
Section 4.8 **1.** 0.75 **2.** 56 **3.** 6 **4.** 4 **5.** 892
Section 4.9 **1.** 2.6666, 2 **2.** $12X$ **3.** 62
Section 4.10 **1.** 0.19 **2.** 0.01105
Section 4.11 **1.** 1 **2.** 0.05466

6.6 Solutions to even numbered problems

1.2. **[Defining a set of outcomes II]**

(a) One choice would be $\Omega = \{(w_1, w_2, w_3) : w_1 \in \{1, 2\}, w_2 \in \{3, 4\}, w_3 \in \{w_1, w_2\}\}$, where w_i denotes which team wins the ith game, for $1 \leq i \leq 3$.
Another choice would be $\Omega = \{(x_1, x_2, x_3) : x_i \in \{L, H\} \text{ for } 1 \leq i \leq 3\}$, where x_i indicates which of the two teams playing the ith game wins; $x_i = H$ indicates that the higher numbered team wins and $x_i = L$ indicates that the lower numbered team wins. For example, (L, L, H) indicates that team one wins the first game, team three wins the second game (so team one plays team three in the third game), and team three wins the third game.

(b) Eight, because there are two possible outcomes for each of the three games, and $2^3 = 8$.

(c) There are three possible tournament brackets , which can be indexed by $j \in \{2, 3, 4\}$, with j indicating which team plays team 1 in the first round. For each possible bracket there are eight possible outcomes for the games, as in part (b). So, accounting for the possible brackets, there are 24 possible outcomes of the tournament. We could represent them by four tuples of the form $\Omega = \{(j; x_1, x_2, x_3) : j \in \{2, 3, 4\},\ x_i \in \{L, H\} \text{ for } 1 \leq i \leq 3\}$. where j represents the team playing team 1 in the first round, and the x_i's have the same meaning as explained in part (a).

(d) There are eight outcomes in which teams 1 and 2 play each other in the first round, namely, the outcomes with $j = 2$. There are two outcomes with $j = 3$ for which teams 1 and 2 play each other in the second round, namely $(3; L, L, L)$ and $(3; L, L, H)$. That is, they must both win in the first round, and then either one could win when they play each other. Similarly, there are two outcomes with $j = 4$ for which teams 1 and 2 play each other in the second round, namely $(4; L, L, L)$ and $(4; L, L, H)$. So overall, teams 1 and 2 play each other in 12 out of the 24 outcomes. (If all outcomes are equally likely, the probability the teams play each other is one half.)

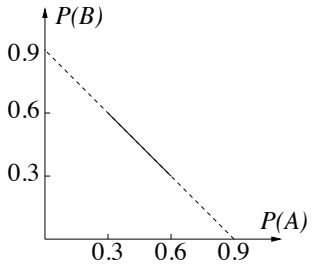
1.4. **[Possible probability assignments]**

(This is one of many ways to get the answer.) Since, by De Morgan's law, the complement of $A \cup B$ is A^cB^c, the fact $P(A \cup B) = 0.6$ is equivalent to $P(A^cB^c) = 0.4$. Thus, we can fill in the Karnaugh diagram for A and B as shown:

B^c	B	
0.4	$0.3-a$	A^c
a	0.3	A

We filled in the variable a for $P(AB^c)$, and then, since the sum of the probabilities is one, it must be that $P(A^cB) = 0.3 - a$. The valid values of a are $0 \leq a \leq 0.3$, and $(P(A), P(B)) = (0.3 + a, 0.6 - a)$. So, in parametric form, the set of possible values of $(P(A), P(B))$ is $\{(0.3 +$

$a, 0.6-a) : 0 \le a \le 0.3\}$. Equivalent ways to write this set are $\{(u,v) : v = 0.9-u$ and $0.3 \le u \le 0.6\}$ or $\{(x, 0.9-x) : 0.3 \le x \le 0.6\}$. The set is also represented by the solid line segment in the following sketch:



1.6. **[Displaying outcomes in a two event Karnaugh map]**

(a)

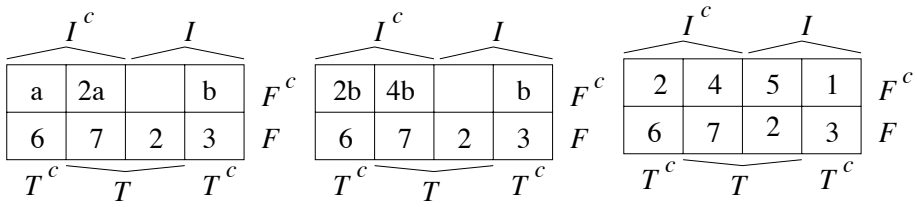
B^c	B	
12,14,16,21, 25,41,45,52, 54,56,61,65	23,32,34, 36,43,63	A^c
11,15,22,24,26, 42,44,46,51,55, 62,64,66	13,31,33, 35,53	A

(b) $P(AB) = \frac{5}{36} \approx 0.13888$.

1.8. **[A classification of students in a class]**

Set up a Karnaugh map using "F" for Facebook, "I" for iPad, and "T" for Twitter, and begin filling in entries by considering the given facts in the following order:

- 3 students on Facebook and having iPads are not on Twitter
- 2 students are on both Facebook and Twitter and have iPads
- 9 are on both Facebook and Twitter
- 12 are not on Facebook (so 18 are on Facebook); all numbers inside F can be filled in, and the sum of the other four numbers is 12)
- 2/3 of the students not on Facebook and without iPads are on Twitter; enter a and 2a in boxes as shown below in leftmost diagram.
- 2/3 of students not on Twitter don't have iPads; Enter b for the number in F^cIT^c as show in the leftmost diagram, and then $a + 6 = 2(b + 3)$ or $a = 2b$.. Eliminate variable a. The situation is shown in the middle diagram, and the sum of the numbers in the top row is 12.



Finally, since at least one student is neither on Twitter nor on Facebook, $2b + b > 0$ or $b > 0$. There are only 12 students total not on Facebook, so $b = 1$. The final diagram is shown on the right. Two students are not on Twitter or Facebook and don't have iPads.

1.10. **[Two more poker hands I]**

(a) There are $\binom{13}{2}$ ways to select the numbers for the two pairs, then 11 ways to choose the number on the fifth card, then $\binom{4}{2} = 6$ ways to choose suits for the cards in one pair, 6 ways to choose suits for the cards of the other pair, and 4 ways to choose the suit of the unpaired card. Thus,

$$\begin{aligned} P(\textit{TWO PAIR}) &= \frac{\binom{13}{2}(11)(6)^2 4}{\binom{52}{5}} \\ &= \frac{3 \cdot 6 \cdot 11}{5 \cdot 17 \cdot 49} \approx 0.0475 \end{aligned}$$

(b) There are 13 ways to choose the number common to three of the cards. Given that choice, there are $\binom{12}{2}$ ways to choose the numbers showing on the remaining two cards, then four ways to choose suits for the three matching cards, and four ways each to choose suits for the other two cards. Thus,

$$\begin{aligned} P(\textit{THREE OF A KIND}) &= \frac{13\binom{12}{2}4^3}{\binom{52}{5}} \\ &= \frac{2 \cdot 4 \cdot 11}{5 \cdot 17 \cdot 49} \approx 0.0211 \end{aligned}$$

(c) There are 13 ways to choose the number common to four of the cards. Given that choice, there are 12 ways to choose the number showing on the remaining card, and four ways to choose the suit of that card. Thus,

$$\begin{aligned} P(\textit{FOUR OF A KIND}) &= \frac{13 \cdot 12 \cdot 4}{\binom{52}{5}} \\ &= \frac{13 \cdot 12 \cdot 4 \cdot 5 \cdot 4 \cdot 3 \cdot 2}{52 \cdot 51 \cdot 50 \cdot 49 \cdot 48} \\ &= \frac{1}{4165} \approx 0.0002401 \end{aligned}$$

1.12. **[Some identities satisfied by binomial coefficients]**

(a) Given a collection of n objects, selecting k of them to be in a set is equivalent to selecting $n - k$ of them to not be in the set.

(b) Suppose the n objects are numbered 1 through n. Every set of k of the n objects either does or does not includes the first object. So the total number of sets of size k is the number of such sets containing the first object, $\binom{n-1}{k-1}$, plus the number of such sets not containing the first object, $\binom{n-1}{k}$.

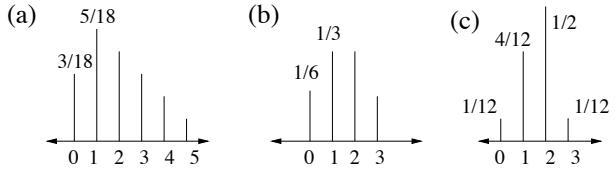
(c) There are $\binom{2n}{n}$ ways to choose n objects from a set of $2n$ objects, half of which are orange and half of which are blue. The number of such ways containing exactly k orange objects and $n-k$ blue objects is $\binom{n}{k}\binom{n}{n-k}$. Summing over k yields the original $\binom{2n}{n}$ possibilities.

(d) Suppose the n objects are numbered 1 through n. Every set of k objects contains a highest numbered object. The number of subsets of the n objects of size k with largest object l is the number of subsets of the first $l-1$ objects of size $k-1$, equal to $\binom{l-1}{k-1}$. Summing over l gives the identity.

Discrete random variables (Sections 2.1 and 2.2)

2.2. **[Distance between two randomly selected vertices]**

(a) There are 36 possible values of (i,j) and $D = |i-j|$. First,
$p_D(0) = P\{D=0\} = \frac{|\{(1,1),\ldots,(6,6)\}|}{36} = \frac{6}{36} = \frac{3}{18} = \frac{1}{6}$.
For $1 \le k \le 5$, $p_D(k) = P\{D=k\} = \frac{|\{(1,k+1),\ldots,(6-k,6)\}\cup\{(k+1,1),\ldots,(6,6-k)\}|}{36} = \frac{2(6-k)}{36} = \frac{6-k}{18}$.
and $p_D(k) = 0$ for $k \notin \{0,\ldots,5\}$.



$E[D] = \frac{3}{18}\cdot 0 + \frac{5}{18}\cdot 1 + \frac{4}{18}\cdot 2 + \frac{3}{18}\cdot 3 + \frac{2}{18}\cdot 4 + \frac{1}{18}\cdot 5 = \frac{35}{18} = 1.944$
$E[D^2] = \frac{6}{18}\cdot 0^2 + \frac{5}{18}\cdot 1^2 + \frac{4}{18}\cdot 2^2 + \frac{3}{18}\cdot 3^2 + \frac{2}{18}\cdot 4^2 + \frac{1}{18}\cdot 5^2 = \frac{105}{18} = 5.8333.$
$\mathrm{Var}(D) = E[D^2] - E[D]^2 = 2.0525$

(b) Let $i=1$ without loss of correctness. Then $D=0$ if $j=1$, $D=1$ if $j \in \{2,6\}$, $D=2$ if $j \in \{3,5\}$, and $D=3$ if $j=4$. So $p_D(0) = p_D(3) = \frac{1}{6}$, $p_D(1) = p_D(2) = \frac{2}{6} = \frac{1}{3}$, and $p_D(k) = 0$ if $k \notin \{0,1,2,3\}$.
$E[D] = \frac{1}{6}\cdot 0 + \frac{1}{3}\cdot 1 + \frac{1}{3}\cdot 2 + \frac{1}{6}\cdot 3 = 1.5$
(The variance of D can be readily calculated using the definition of variance. Here we use the approach of calculating $E[D^2]$ first.) $E[D^2] = \frac{1}{6}\cdot 0 + \frac{1}{3}\cdot 1^2 + \frac{1}{3}\cdot 2^2 + \frac{1}{6}\cdot 3^2 = \frac{19}{6}$
$\mathrm{Var}(D) = E[D^2] - E[D]^2 = \frac{11}{12} = 0.9167.$
(c) Let $i=1$ without loss of correctness. Then $D=0$ if $j=1$, $D=1$ if $j \in \{2,5,9,12\}$, and $D=3$ if $j=7$. For the other six choices of j, $D=2$. Thus, $p_D(0) = p_D(3) = \frac{1}{12}$, $p_D(1) = \frac{4}{12} = \frac{1}{3}$ and $p_D(2) = \frac{6}{12} = \frac{1}{2}$. Proceeding as in parts (a) and (b) yields $E[D] = \frac{19}{12} = 1.5833$, and $\mathrm{Var}(D) = 0.5764$.

2.4. **[A problem on sampling without replacement]**

(a) There are $\binom{2n}{2}$ choices of two shoes from $2n$ of which only n choices yield a pair. Hence, $P(\text{pair}) = \dfrac{n}{\binom{2n}{2}} = \dfrac{n}{2n(2n-1)/(1\times 2)} = \dfrac{1}{2n-1}$. More simply, think of choosing the shoes sequentially. Regardless of what the first shoe drawn is, the chance of getting

its mate on the second draw is $\dfrac{1}{2n-1}$. Note that this has value 1 if $n = 1$, which makes perfect sense.

(b) Any one of the n left shoes can be paired with any one of the n right shoes. So, n^2 of the $\binom{2n}{2}$ choices yield one left shoe and one right shoe in the two drawn, giving that

$$P(\text{one L, one R }) = \frac{n^2}{\binom{2n}{2}} = \frac{n^2}{2n(2n-1)/(1\times 2)} = \frac{n}{2n-1}.$$

Again, more simply, regardless of what the first shoe is, the chances of getting a shoe of the opposite footality on the second draw is $\dfrac{n}{2n-1}$.
Note that this has value 1 if $n = 1$, which makes perfect sense.

Suppose now that $n \geq 2$ and that you choose 3 shoes at random from the bag.

(c) Any of the n pairs and any of the other $2n-2$ shoes form a set of 3 shoes. Hence,
$P(\text{pair among three}) = \dfrac{n(2n-2)}{\binom{2n}{3}} = \dfrac{n(2n-2)}{2n(2n-1)(2n-2)/(1\times 2\times 3)} = \dfrac{3}{2n-1}$.
Note that this has value 1 when $n = 2$, which makes perfect sense.

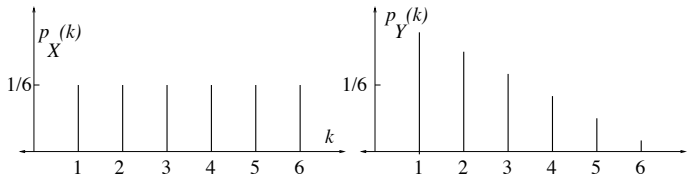
(d) There are $\binom{n}{2}n$ ways to choose two left shoes and one right shoe, and $\binom{n}{2}n$ ways to choose two right shoes and one left shoe. Adding gives $2\binom{n}{2}n = n^2(n-1)$ ways to choose at least one of each. So $P(\text{one L and one R among three}) = \dfrac{n^2(n-1)}{\binom{2n}{3}} = \dfrac{3n}{4n-2}$.
Note that this has value 1 when $n = 2$, which makes perfect sense.

2.6. **[Mean and standard deviation of two simple random variables]**

(a) For $1 \leq k \leq 6$, there are six sample points with $X(i,j) = k$, so $p_X(k) = \frac{6}{36} = \frac{1}{6}$.

(b) $E[X] = \frac{1+2+3+4+5+6}{6} = 3.5$. To compute σ_X we can first find $E[X^2] = \frac{1+2^2+3^2+4^2+5^2+6^2}{6}$ and then $\sigma_X = \sqrt{E[X^2] - E[X]^2} = 1.707825$.

(c) $\{Y = 1\} = \{(1,1),(1,2),(1,3),(1,4),(1,5),(1,6),(2,1),(3,1),(4,1),(5,1),6,1)\}$,
$\{Y = 2\} = \{(2,2),(2,3),(2,4),(2,5),(2,6),(3,2),(4,2),(5,2),(6,2)\}$,
$\{Y = 3\} = \{((3,3),(3,4),(3,5),(3,6),(4,3),(5,3),(6,3)\}$, and so on. In general, for $1 \leq k \leq 6$, there are $13-2k$ sample points for which $Y = k$. Thus, $p_Y(k) = \frac{13-2k}{36}$ for $1 \leq k \leq 6$.



(d) Using part (c) or using a spreadsheet/program, we find $E[Y] = 2.5277\ldots$ and $\sigma_Y^2 = 1.9714062$, so $\sigma_Y = 1.404$.

(e) $\sigma_Y < \sigma_X$. This is consistent with the sketches. Intuitively, the triangular shape of p_Y is more concentrated than the rectangular shape of p_X. To elaborate, this is not a mathematical statement; the mathematical fact is simply that $\sigma(X) \geq \sigma(Y)$. For intuition, look at the sketches of the pmfs in part (c) above. Suppose you start out with the pmf of X, which has mass uniformly spread out over the six points. Then you take your right hand and push some of the mass on the right side over to the left. There is very little mass left at points 5 and 6. The mass gets more bunched up. Maybe it helps to squint a bit when you look at the figure. The standard deviation would be even smaller if you took both hands and pushed mass together from both sides to get a symmetric triangular shape.

(f) Here is one example. Consider a random variable Z with $p_Z(1) = p_Z(6) = 0.5$. Then $E[Z] = \frac{1+6}{2} = 3.5$, $E[Z^2] = \frac{1+36}{2} = 18.5$, and $\sigma_Z = \sqrt{18.5 - (3.5)^2} = 2.5$. In fact, this distribution has the largest variance and standard deviation of any distribution on the set $\{1, 2, 3, 4, 5, 6\}$.

2.8. **[Selecting supply for a random demand]**

(a) Begin by calculating:

$$\begin{aligned} E[\min\{U, L\}] &= \frac{1}{M}\sum_{i=1}^{M} \min\{i, L\} \\ &= \frac{1}{M}\left(\sum_{i=1}^{L} \min\{i, L\} + \sum_{i=L+1}^{M} \min\{i, L\}\right) \\ &= \frac{1}{M}\left(\sum_{i=1}^{L} i + \sum_{i=L+1}^{M} L\right) \\ &= \frac{1}{M}\left(\frac{L(L+1)}{2} + (M-L)L\right) \\ &= L - \frac{L^2 - L}{2M} \end{aligned}$$

Therefore,

$$E[\text{profit}] = (b-a)L - b\frac{L^2 - L}{2M}$$

(b) As seen in part (a), the expected profit is $h(L)$, where $h(L) = (b-a)L - b\frac{L^2-L}{2M}$. The graph of h is a parabola facing downward. If we ignore the requirement that L be an integer we can solve $h'(L) = 0$ and find that h is maximized at L_r, where

$$L_r = \frac{1}{2} + M\left(1 - \frac{a}{b}\right).$$

Since the function h is symmetric about L_r, the maximizing integer value L is obtained by rounding L_r to the nearest integer. That is, the revenue maximizing integer value of L is the integer nearest to L_r. (If $M\left(1 - \frac{a}{b}\right)$ happens to be an integer, then the fractional

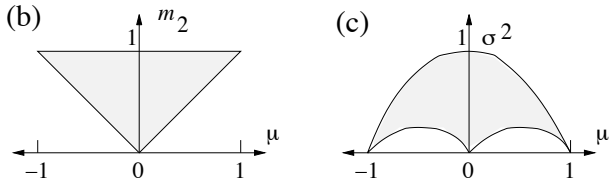
part of L_r is 0.5, and $M\left(1-\frac{a}{b}\right)$ and $M\left(1-\frac{a}{b}\right)+1$ are both choices for L that maximize expected profit.)

Alternative approach To find the integer L that maximizes $h(L)$ we calculate that $h(L+1)-h(L)=(b-a)-\frac{bL}{M}$, and note that it is strictly decreasing in L. A maximizing integer L^* is the minimum integer L such that $h(L+1)-h(L)\leq 0$, or equivalently, the minimum integer L such that $L \geq M\left(1-\frac{a}{b}\right)$, or equivalently, $L^* = \lceil M(1-\frac{a}{b})\rceil$. (If $M\left(1-\frac{a}{b}\right)$ happens to be an integer, then $L^* = M\left(1-\frac{a}{b}\right)$ and $h(L^*+1)-h(L^*)=0$, so $M\left(1-\frac{a}{b}\right)$ and $M\left(1-\frac{a}{b}\right)+1$ are both choices for L that maximize expected profit.) **Comment:** The expression $h(L+1)-h(L)=b(1-\frac{L}{M})-a$ has a nice interpretation. The term $(1-\frac{L}{M})$ equals $P\{U \geq L+1\}$, which is the probability an $(L+1)^{th}$ room would be sold, and a is the price to the reseller of reserving and prepaying for one additional room.

2.10. **[First and second moments of a ternary random variable]**

(a) Since $m_2=\mu^2+\sigma^2$, the given requirement is equivalent to $\mu=\frac{1}{2}$ and $m_2=1$. In general, μ and m_2 can be expressed in terms of a and b as follows: $\mu=b-a$ and $m_2=a+b$. So $b-a=\frac{1}{2}$ and $a+b=1$, or $a=0.25$ and $b=0.75$. (Note that this choice for (a,b) is valid, i.e. $a\geq 0$, $b\geq 0$ and $a+b\leq 1$.)

(b) As noted in the solution to part (a), in general, μ and m_2 can be expressed in terms of a and b as follows: $\mu=b-a$ and $m_2=a+b$. Given μ and m_2, this gives two equations for the two unknowns a and b. Solving yields $a=\frac{m_2-\mu}{2}$ and $b=\frac{m_2+\mu}{2}$. The constraint $a\geq 0$ translates to $\mu \leq m_2$ The constraint $b\geq 0$ translates to $\mu \geq -m_2$. These two constraints are equivalent to the combined constraint $|\mu|\leq m_2$. The constraint $a+b\leq 1$ translates to $m_2\leq 1$. In summary, there is a valid choice for (a,b) if and only if $|\mu|\leq m_2\leq 1$. See the sketch of this region.



(c) As seen in part (b), the mean μ must be in the range $-1\leq\mu\leq 1$. For any such μ, m_2 can be anywhere in the interval $[|\mu|,1]$, and so, by the hint, σ^2 can be anywhere in the interval $[|\mu|-\mu^2, 1-\mu^2]$. See the sketch of this region.

Conditional probability, independence, and the binomial distribution Sections 2.3–2.4

2.12. **[Conditional probability]**

(a) Let D be the event doubles are rolled and S_8 be the event the sum is eight. Then $P(D|S_8)=P(DS_8)/P(S_8)=\frac{1}{36}/\frac{5}{36}=\frac{1}{5}$.

(b) $P(S_8|D)=P(S_8D)/P(D)=\frac{1}{36}/\frac{6}{36}=\frac{1}{6}$.

(c) This event contains the eleven outcomes $\{61, 62, 63, 64, 65, 66, 56, 46, 36, 26, 16\}$ so it has probability $\frac{11}{36}$.

(d) $\frac{P(\text{at least one 6, not doubles})}{P(\text{not doubles})} = \frac{10/36}{30/36} = \frac{10}{30} = \frac{1}{3}$. (This answer is slightly larger than the answer to (c) as expected.)

2.14. **[Independence]**

(a) For brevity, we write ij for the outcome that the orange die shows i and the blue die shows j. Then, $A = \{26, 34, 43, 62\}, B = \{ij : 1 \le j < i \le 6\}, C = \{13, 22, 31, 26, 35, 44, 53, 62, 66\}$, and $D = \{1j : 1 \le j \le 6\} \cup \{3j : 1 \le j \le 6\}$.

$$P(A) = \frac{1}{9}, \quad P(B) = \frac{5}{12}, \quad P(C) = \frac{1}{4}, \quad P(D) = \frac{1}{3}$$

$$P(AB) = P\{43, 62\} = \frac{1}{18}, \ P(AC) = P\{26, 62\} = \frac{1}{18}, \ P(AD) = P\{34\} = \frac{1}{36}$$

$$P(BC) = P\{31, 53, 62\} = \frac{1}{12}, \ P(BD) = P\{31, 32\} = \frac{1}{18}, \ P(CD) = P\{13, 31, 35\} = \frac{1}{12}$$

So C and D are mutually independent; no other pair is mutually independent.

(b) No three events are even pairwise independent by part (a), so no three events are mutually independent.

2.16. **[A team selection problem]**

(a) There are $\binom{7}{3} = \frac{7 \cdot 6 \cdot 5}{3 \cdot 2 \cdot 1} = 35$ ways to select the four people, and the number of ways to select a team with Alice on it is the number of ways to select three of the other six people to go with Alice, or $\binom{6}{3} = \frac{6 \cdot 5 \cdot 4}{3 \cdot 2 \cdot 1} = 20$. So $P(A) = \frac{20}{35} = \frac{4}{7}$. ALTERNATIVELY, a faction 4/7 of the entire debate team is selected so by symmetry, each person has a probability 4/7 of being selected.

(b) Given Bob is selected, there are 20 ways to fill out the rest of the team, but if Alice is also selected, there are $\binom{5}{2} = 10$ ways to fill out the rest of the team, so $P(A|B) = \frac{10}{20} = 0.5$. ALTERNATIVELY by the definition, $P(A|B) = \frac{P(AB)}{P(B)} =$ (number of selections including both Alice and Bob)/(number of selections including Bob) $= \frac{10}{20} = 0.5$. OR ALTERNATIVELY given Bob is selected, half of the other six people are also selected so by symmetry, Alice has chance 0.5 to be among them.

(c) $P(A \cup B) = P(A) + P(B) - P(AB) = \frac{20}{35} + \frac{20}{35} - \frac{10}{35} = \frac{6|}{7}$. ALTERNATIVELY, the number of selections that exclude both Alice and Bob is the number of ways to select four out of the other five people, which is 5. So $P(A \cup B) = 1 - P((A \cup B)^c) = 1 - \frac{5}{35} = \frac{6}{7.}$

2.18. **[Binomial distribution II]**

(a) The number of weeks that your investment doubles in value is a binomial random variable Y with parameters $(5, \frac{1}{2})$. Since the investment halves in value during the remaining $5 - Y$ weeks, and each halving cancels one doubling, we have that $X = 32 \cdot 2^{2Y-5}$. The possible values of X are 1, 4, 16, 64, 256, and 1024, corresponding to $Y = 0$, 1, 2, 3, 4, 5 respectively.

(b) $P\{X=1\}=P\{Y=0\}=\frac{1}{32}$. $P\{X=4\}=P\{Y=1\}=\frac{5}{32}$.
$P\{X=16\}=P\{Y=2\}=\frac{10}{32}$. $P\{X=64\}=P\{Y=3\}=\frac{10}{32}$.
$P\{X=256\}=P\{Y=4\}=\frac{5}{32}$. $P\{X=1024\}=P\{Y=5\}=\frac{1}{32}$.
(To elaborate on the above solution, perhaps it would help to consider some particular outcomes for the five weeks. Since there are two possibilities for each week, there are $2^5=32$ possibilities for the five weeks. For example, if the value is cut in half all five weeks (represent this by 00000) then $X=1$ (start out with 32 and divide it by 2 five times). If the investment doubles the first week and is cut in half the other four weeks (represent this by 10000) then $X=4$ (there'd be 64 after one week and that gets cut in half four times. Similarly, $X=4$ if the investment doubles the second week and is cut in half the other four weeks. Continuing, we see there are 5 ways for $X=4$. The possible values of X are 1,4, 16, 64, 256, 1024. X is not a binomially distributed random variable. But X is determined by the number of good weeks, when the investment doubles. It doesn't depend on the order of good and bad weeks. The number of good weeks out of the five weeks has the binomial distribution–it is the number of successes for a specified number of independent trials with the same success probability.)

(c) $E[X]=1\cdot\frac{1}{32}+4\cdot\frac{5}{32}+16\cdot\frac{10}{32}+64\cdot\frac{10}{32}+256\cdot\frac{5}{32}+1024\cdot\frac{1}{32}=97.65625$. The TV commercial *understates* the performance - undoubtedly a first!

(d) $P\{X<32\}=P\{X=1\}+P\{X=4\}+P\{X=16\}=1/2$.

2.20. **[Binomial Random Variable]**

(a) The probability that the aircraft will fail on a flight is given by

$$\sum_{k=2}^{4}\binom{4}{k}(10^{-3})^k(1-10^{-3})^{4-k}=6\times10^{-6}+4\times10^{-9}+10^{-12}\approx6\times10^{-6}$$

(b) Let n be the number of flights. The aircraft can crash with a probability of $p=\times10^{-9}$ in any given flight. As each flight is independent of the other, and if X represents the number of crashes in n flights, then its pmf is binomially distributed with parameters (n,p). Thus, the probability of having at least one crash in n flights is given by $1-\binom{n}{0}(1-p)^n=1-(1-p)^n$. Searching over values of n starting from 1, we find that $n=100,006$ flights are needed in order for the probability of the aircraft experiencing at least one crash reaches 0.01%.

(c) Sweeping p_c from say 0 upwards in the equation below until it evaluates to 10^{-9}:

$$\sum_{k=2}^{4}\binom{4}{k}(p_c)^k(1-p_c)^{4-k}$$

we get $p_c=1.35\times10^{-5}$. This number dictates how reliable each aircraft subsystem needs to be designed.

Geometric and Poisson distributions, Bernoulli processes, ML parameter estimation and confidence intervals Sections 2.5–2.9

2.22. **[Repeated rolls of four dice]**

(a) This is the same as the probability exactly two even numbers show. Each die shows an even number with probability 0.5, so the number of even numbers showing, X, has the binomial distribution with parameter $p = 0.5$. Therefore, $P\{X = 2\} = \binom{4}{2}(0.5)^2(0.5)^2 = \frac{6}{16} = \frac{3}{8}$.

(b) Each roll does not produce two even and two odd numbers with probability 5/8, and for that to happen on three rolls has probability $(5/8)^3 = \frac{125}{512}$.

2.24. **[A knockout game]**

P$\{X = 0\}$ = P{of players 1,2: 1 has the higher number } = 1/2
P$\{X = 1\}$ = P{of players 1,2,3: 3 has largest, 1 the next largest} = (1/3)(1/2) = 1/6
P$\{X = 2\}$ = P{of 1,2,3,4: 4 has largest, 1 the next largest} = (1/4)(1/3) = 1/12

2.26. **[ML parameter estimation for independent geometrically distributed rvs]**

(a) Since $P\{L_i = k_i\} = p(1-p)^{k_i-1}$ for each i, the likelihood is $p(1-p)^{k_1-1} \cdots p(1-p)^{k_n-1} = p^n(1-p)^{s-n}$, where $s = k_1 + \cdots + k_n$. That is, s is the total number of attempts needed for the completion of the n tasks.

(b) By definition, $\widehat{p}_{ML}$ is the value of p that maximizes the likelihood found in (a). A special case is $s = n$, i.e., each completion requires only one attempt. The likelihood of that is p^n which is maximized when $p = 1$. Suppose now that $s > n$. Differentiating the likelihood with respect to p yields:

$$\begin{aligned} \frac{d(p^n(1-p)^{s-n})}{dp} &= np^{n-1}(1-p)^{s-n} - (s-n)p^n(1-p)^{s-n-1} \\ &= (n(1-p) - (s-n)p)p^{n-1}(1-p)^{s-n-1} \\ &= (n-sp)p^{n-1}(1-p)^{s-n-1} \end{aligned}$$

The derivative is zero at $p = \frac{n}{s}$ (and is positive for p smaller than $\frac{n}{s}$ and negative for p larger than $\frac{n}{s}$). So $\widehat{p}_{ML} = \frac{n}{s}$. This formula also works for the case $s = n$, already discussed, so it works in general. This formula makes intuitive sense; during the observation period, n successes are observed out of a total of s attempts, so $\widehat{p}_{ML}$ is the fraction of observed attempts that are successful.)

2.28. **[Scaling of a confidence interval]**

(a) The width of the original window is $\frac{a}{\sqrt{n}}$, where a determines the confidence level (which we don't need to find yet). To reduce this width by a factor of two, n should be increased by a factor of four. So 1200 samples would be needed.

(b) The confidence interval has width 0.1 if $\frac{a}{\sqrt{n}} = 0.1$. Solving for a with $n = 300$ yields $a = (0.1)\sqrt{300}$, so the confidence level is given by $1 - \frac{1}{a^2} = 1 - \frac{1}{(0.01)(300)} = \frac{2}{3}$.

(c) In order for $1 - \frac{1}{a^2} = 0.96$ we need $a^2 = \frac{1}{0.04} = 25$, or $a = 5$. Then, in order to have $\frac{a}{\sqrt{n}} = 0.1$, we need $\sqrt{n} = \frac{5}{0.1} = 50$, or $n = 2500$.

2.30. **[Parameter estimation for the binomial distribution]**

(a) The distribution of X is approximately Poisson with parameter $\lambda = np = 3$. So $P\{X \geq 2\} = 1 - p_X(0) - p_X(1) \approx 1 - \frac{e^{-\lambda}\lambda^0}{0!} - \frac{e^{-\lambda}\lambda^1}{1!} = 1 - e^{-3} - 3e^{-3} = 1 - 4e^{-3}$.

(b) The half-width of the confidence interval is $0.025 = \frac{a}{2\sqrt{n}} = \frac{a}{2\sqrt{10{,}000}}$, so $a = 5$. Therefore, we can claim a confidence level of $1 - \frac{1}{a^2} = 1 - \frac{1}{5^2} = 0.96$.

(c) The likelihood (i.e. probability) of observing $X = 7$ is zero if $n \leq 6$. If $n \geq 7$, the likelihood is given by $p_X(7) = \binom{n}{7}(0.03)^7(0.97)^{n-7}$. The desired estimate $\hat{n}_{ML}$ is the value of n that maximizes this. That is, $\hat{n}_{ML}$ is the value of n that maximizes $L(n) = \binom{n}{7}(0.03)^7(0.97)^{n-7}$ over the range $n \geq 7$. Consider the ratio:

$$\frac{L(n)}{L(n-1)} = \frac{\binom{n}{7}(0.97)}{\binom{n-1}{7}} = \frac{n(0.97)}{n-7}$$

Note that $\frac{L(n)}{L(n-1)} > 1$ if $n(0.97) > n - 7$, or $7 \geq (0.03)n$, or $n < 233.33$. Similarly, $\frac{L(n)}{L(n-1)} < 1$ if $n > 233.33$. So $L(n)$ is strictly increasing in the range $7 \leq n \leq 233$ and strictly decreasing in the range $233 \leq n < \infty$. So $\hat{n}_{ML} = 233$. (For general p and observed value k of X, an ML estimate of n is given by $\left\lfloor \frac{k}{p} \right\rfloor$.)

Bayes' Formula and binary hypothesis testing Sections 2.10 & 2.11

2.32. **[The weight of a positive]**

(a)

$$\begin{aligned} P(\text{positive}) &= P(\text{positive}|\text{cancer})P(\text{cancer}) + P(\text{positive}|\text{no cancer})P(\text{no cancer}) \\ &= (0.9)(0.008) + (0.07)(0.992) = 0.07664 \approx 7.7\% \end{aligned}$$

(b)

$$\begin{aligned} P(\text{cancer}|\text{positive}) &= \frac{P(\text{cancer and positive})}{P(\text{positive})} \\ &= \frac{P(\text{positive}|\text{cancer})P(\text{cancer})}{P(\text{positive})} \\ &= \frac{(0.9)(0.008)}{0.07664} = 0.0939 \approx 9.4\% \end{aligned}$$

(c) Out of 1000 woman getting a mammogram, we expect 1000*0.008=8 women to have breast cancer, and 8*(0.9)=7.2 of those to get a positive mammogram. Expect 1000*(0.992)*(0.07)=69.4 women to get a false positive. There is a debate within the health industry as to whether women in this age range should get mammograms.

2.34. **[Conditional distribution of half-way point]**

(a) In this problem we use repeatedly the fact that the probability that k out of n consecutive steps are right steps is $\binom{n}{k}2^{-n}$. That is because any particular sequence of n steps has probability 2^{-n} and there are $\binom{n}{k}$ sequences of n steps with exactly k right steps. Similarly, the probability that k out of n consecutive steps are left steps is $\binom{n}{k}2^{-n}$. The event F is that four of the first eight steps are right steps, so $P(F) = \binom{8}{4}2^{-8} = 70/256 \approx 0.2734$.

(b) For $0 \le k \le 4$, the event $X = 2k-4$ is the same as the event that k out of the first four steps are right steps. Thus, $p_X(2k-4) = \binom{4}{k}2^{-4}$, for $0 \le k \le 4$. That is, the support of the distribution of X is $\{-4,-2,0,2,4\}$ and

$$p_X(0) = \binom{4}{2}2^{-4} = \frac{6}{16} \qquad p_X(2) = p_X(-2) = \binom{4}{1}2^{-4} = \frac{4}{16}, \qquad p_X(-4) = p_X(4) = \frac{1}{16}.$$

(c) The event $\{X = i\}F$ is the event that the robot is at i at time four and at zero at time eight. The only possible values of i have the form $2k-4$ with $0 \le k \le 4$, as in part (a). And $\{X = 2k-4\}F$ is true if and only if k of the first four steps are right steps, and k of the next four steps are left steps. Each of these two parts has probability equal to $\binom{4}{k}2^{-4}$, In addition, the first four steps are independent of the next four steps, so we multiply the probabilities for the first four steps and second four steps to get: $P(\{X = 2k-4\}F) = \left(\binom{4}{k}2^{-4}\right)^2 = \binom{4}{k}^2 2^{-8}$. That is, $P(\{X = 2k-4\}F)$ is nonzero only for $i \in \{-4,-2,0,2,4\}$ and

$$P(\{X=0\}F) = \binom{4}{2}^2 2^{-8} = \frac{36}{256} = \frac{9}{64} \quad P(\{X=2\}F) = P(\{X=-2\}F) = \binom{4}{1}^2 2^{-8} = \frac{16}{256} = \frac{1}{16}$$

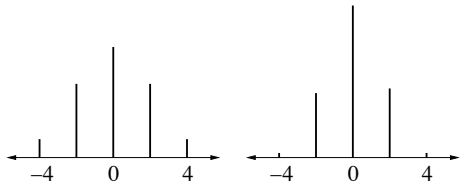
$$P(\{X=4\}F) = P(\{X=-4\}F) = \frac{1}{256}.$$

(d) By the definition of conditional probability, $p_X(i|F) = \frac{P(\{X=i\}F)}{P(F)}$. So by the answers to parts (a) and (c), we find that the support of the conditional pmf is $\{-4,-2,0,2,4\}$ and

$$p_X(0|A) = \frac{\frac{36}{256}}{\frac{70}{256}} = \frac{36}{70} \quad p_X(2) = p_X(-2) = \frac{\frac{16}{256}}{\frac{70}{256}} = \frac{16}{70}$$

$$p_X(4) = p_X(-4) = \frac{\frac{1}{256}}{\frac{70}{256}} = \frac{1}{70}$$

Comparing sketches of the unconditional and conditional pmfs:



shows that the conditional pmf of X given F is more concentrated around zero than the unconditional pmf of X.

2.36. **[A simple hypothesis testing problem with discrete observations]**

(a) The likelihood ratio function is $\Lambda(i) = \frac{p_1(i)}{p_0(i)} = \frac{9i^2}{60}$. The ML rule is as follows: if $X = i$, decide H_1 if $\Lambda(i) \geq 1$, or equivalently, if $|i| \geq 2.58$, or equivalently, because i is integer valued, if $|i| \geq 3$. Equivalently, $\Gamma_1 = \{-4, -3, 3, 4\}$ and $\Gamma_0 = \{-2, -1, 0, 1, 2\}$. (The meaning of Γ_1 and Γ_0 is that if $X \in \Gamma_1$ we declare that H_1 is true, and if $X \in \Gamma_0$ we declare that H_0 is true.)

(b) For the ML rule, $p_{false\ alarm} = P(|X| \geq 3|H_0) = \frac{4}{9}$, and $p_{miss} = P(|X| \leq 2|H_1) = \frac{2^2+1^2+0^2+1^2+2^2}{60} = \frac{1}{6}$.

(c) The MAP rule for $X = i$ is to decide H_1 if $\Lambda(i) \geq \frac{\pi_0}{\pi_1}$, or equivalently if $\frac{9i^2}{60} \geq 2$, or equivalently, if $|i| \geq 3.65$, or equivalently, because i is integer valued, if $|i| \geq 4$. Equivalently, $\Gamma_1 = \{-4, 4\}$ and $\Gamma_0 = \{-3, -2, -1, 0, 1, 2, 3\}$.

(d) For the MAP rule, $p_{false\ alarm} = P(|X| \geq 4|H_0) = \frac{2}{9}$ and $p_{miss} = P(|X| \leq 3|H_1) = \frac{3^2+2^2+1^2+0^2+1^2+2^2+3^2}{60} = \frac{28}{60} = \frac{7}{15}$. The average probability of error is $p_e = \pi_0 p_{false\ alarm} + \pi_1 p_{miss} = \frac{2}{3}\frac{2}{9} + \frac{1}{3}\frac{7}{15} = \frac{41}{135} \approx .0.3037$.

(e) Since $\max_i \Lambda(i) = \frac{9(4)^2}{60} = 2.4$, the MAP rule always decides H_0 if and only if $\frac{\pi_0}{\pi_1} > 2.4$.

2.38. **[Field goal percentages – home vs. away]**

(a) Using $a = \sqrt{20}$ so that $1 - \frac{1}{a^2} = 0.95$, the symmetric, two-sided confidence interval for the parameter p of a binomial distribution with known n and observed p has endpoints $\frac{k}{n} \pm \frac{\sqrt{20}}{2\sqrt{n}}$. The given data yields the confidence interval $[0.290, 0.557]$ for p_h and the interval $[0.309, 0.505]$ for p_a. These intervals intersect; we say the data is not conclusive.

(b) It can be said that if H_0 were true, and if p denotes the common value of p_h and p_a, from a perspective before the experiment is conducted, the probability the confidence intervals will not intersect is less than or equal to the probability that at least one of the confidence intervals will not contain p. The probability the confidence interval for p_h will not contain p is less than or equal to 5%, and the probability the confidence interval for p_a will not contain p is less than or equal to 5%. So the probability that at least one of the two confidence intervals will not contain p is less than or equal to 10%. So therefore, if H_0 is true, before we know about the data, we would say the probability the intervals do not intersect is less than or equal to 10%. Thus, if and when we observe nonintersecting confidence intervals, we can say either H_0 is false, or we just observed data with an unlikely extreme. Note that we cannot conclude that there is a 90% chance that H_1 is true.

2.40. **[Hypothesis testing for independent geometrically distributed observations]**

(a) In Problem 2.26 we've seen that if p is the true parameter, then $P\{(L_1, \ldots, L_n) = (k_1, \ldots, k_n)\} = p^n(1-p)^{s-n}$, where $s = k_1 + \cdots + k_n$. The likelihood ratio for a given observed vector $(k_1, \ldots, k_n)$ is the ratio of this probability for $p = 0.25$ to the probability for $p = 0.5$:

$$\Lambda(k_1, \ldots, k_n) = \frac{(0.25)^n(075)^{s-n}}{(0.5)^n(0.5)^{s-n}} = \frac{(1.5)^s}{3^n}.$$

The maximum likelihood decision rule declares in favor of H_1 if $\Lambda(k_1, \ldots, k_n) > 1$, or equivalently if $(1.5)^s \geq 3^n$, or equivalently if $s \geq \frac{n \ln 3}{\ln(1.5)}$ or if $s \geq (2.71)n$, and in favor of H_0 otherwise.

Does this rule make sense? Equivalently, in terms of $S = L_1 + \cdots + L_n$, the rule is to declare H_1 is true, given S, if $\frac{S}{n} \geq 2.71$. This seems reasonable because if H_1 is true, $\frac{S}{n}$ has mean 4 and if H_0 is true, $\frac{S}{n}$ has mean 2, so the threshold 2.71 is close to halfway between the two possible means.

(b) The MAP rule can also be expressed as a likelihood ratio test, but for it, the threshold to compare Λ to is $\frac{\pi_0}{\pi_1} = 8$. So the MAP rule decides H_1 is true if $(1.5)^s \geq 8(3^n)$, or equivalently, if $s \geq \frac{\ln 8 + n \ln 3}{\ln(1.5)}$, or $s \geq 5.13 + (2.71)n$. And declare H_0 otherwise.

Reliability Section 2.12

2.42. **[The reliability of a hierarchical backup system]**

(a) The number of server failures X in a given subsystem has the binomial distribution with parameters $n = 9$ and p. So $p_0 = P\{X \geq 2\} = 1 - P\{X = 0\} - P\{X = 1\} = 1 - (1-p)^9 - 9p(1-p)^8$. For $p = 0.001$, $p_0 = 0.000035832$.

(b) By the same reasoning as in part (a), $p_1 = 1 - (1-p_0)^9 - 9p_0(1-p_0)^8$. For $p = 0.001$, $p_1 = 0.46215 \times 10^{-7}$.

(c) Consider a particular subsystem, and let $E_{i,j}$ be the event that servers i and j both fail for $1 \leq i < j \leq 9$. Then $P\{E_{i,j}\} = p^2$ for $1 \leq i < j \leq 9$, and the event F that the subsystem fails satisfies $F = \cup_{\{(i,j):1 \leq i < j \leq 9\}} E_{i,j}$. There are $\binom{9}{2} = 36$ possibilities for (i, j). Hence, $p_0 \leq 36p^2 = 0.000036$ in case $p = 0.001$.
By the same reasoning, $p_1 \leq 36p_0^2$. Combining these two bounds yields $p_1 \leq 36(36p^2)^2 = (36)^3 p^4 = 0.46656 \times 10^{-7}$ in case $p = 0.001$. Note that the bounds are quite close to the exact values for $p = 0.001$.

2.44. **[Fault detection in a Boolean circuit]**

(a) The correct output is one only for the three input sequences 1101, 1110, 1111. With the stuck at one fault present, the output is one only for the four input sequences 1100, 1101, 1110, 1111. Thus, the output is incorrect only for the input sequence 1100. So the probability of incorrect output due to the fault is $\frac{1}{16}$.

(b) Let F be the event that the circuit is faulty, and A be the event that the output is correct for all three test patterns. If F is true, then A is true if and only if the pattern 1100 is not among the three randomly selected test patterns, so $P(A|F) = \frac{13}{16}$. (One way to see that $P(A|F) = \frac{13}{16}$ is to suppose the three test patterns are selected by generating a random permutation of all 16 patterns, and then using the first three patterns in the permutation. The probability 1100 is in one of the first three positions of the permutation, and hence the probability it is included in the set of three test patterns, is $\frac{3}{16}$. Therefore, $P(A|F) = \frac{13}{16}$. A more tedious but straightforward way to prove that $P(A|F) = \frac{13}{16}$ is to consider the test patterns to be selected one at a time. The first test pattern is not equal to 1100 with probability $\frac{15}{16}$. Given that the first test pattern is not

equal to 1100, the second test pattern is not equal to 1100 with probability $\frac{14}{15}$. Given that the first two test patterns are not equal to 1100, the third test pattern is not equal to 1100 with probability $\frac{13}{14}$. Multiplying these three probabilities gives $P(A|F) = \frac{13}{16}$.) Thus, by Bayes rule, $P(F|A) = \frac{P(F,A)}{P(A)} = \frac{P(A|F)P(F)}{P(A|F)P(F)+P(A|F^c)P(F^c)} = \frac{\frac{13}{16}\cdot\frac{1}{2}}{\frac{13}{16}\cdot\frac{1}{2}+1\cdot\frac{1}{2}} = \frac{13}{29} = 0.4483$.

2.46. **[Distribution of capacity of an $s-t$ flow network]**

The possible values of Y are $0, 8, 10$. The event $\{Y = 10\}$ is true if and only if links 1, 3, and 4 all do not fail, so $p_Y(10) = (1-p)^3$. The event $\{Y > 0\}$ is true if and only if at least one link does not fail in each of the three stages, so $p_Y(0) = 1 - P\{Y > 0\} = 1-(1-p^2)(1-p)(1-p^2)$. Finally, $p_Y(8) = P\{Y > 0\} - p_Y(10) = (1-p^2)(1-p)(1-p^2)-(1-p)^3$. Writing these answers as polynomials in p yields:

$$\begin{aligned} p_Y(0) &= p + 2p^2 - 2p^3 - p^4 + p^5 \\ p_Y(8) &= 2p - 5p^2 + 3p^3 + p^4 - p^5 \\ p_Y(10) &= 1 - 3p + 3p^2 - p^3 \end{aligned}$$

For $p = 0.5$ we have $(p_Y(0), p_Y(8), p_Y(10)) = (23/32, 5/32, 4/32)$.

Cumulative distribution functions (CDFs) Section 3.1

3.2. **[Using a CDF II]**

(a) $P\{X = 5\}$ is equal to the size of the jump of F_X at 5, which is 0.25.

(b) There is no jump of F_X at zero, so $P\{X = 0\} = 0$.

(c) $P\{|X| \le 5\} = P\{-5 \le X \le 5\} = P\{X \le 5\} - P\{X < -5\} = F_X(5) - F_X((-5)-) = 1 - 0.25 = 0.75$. (Here, $F_X((-5)-)$ represents the left-hand limit of F_X at the point -5.) Another way to think about it is that $P\{|X| \le 5\} = P\{X = 5\}+P\{X = -5\}+P\{-5 < X < 5\} = 0.25 + 0.25 + 0.25 = 0.75$.

(d) $P\{X^2 \le 4\} = P\{-2 \le X \le 2\} = F_X(2) - F_X((-2)-) = 0.6 - 0.5 = 0.1$.

Continuous-type random variables, uniform and exponential distributions, and Poisson processes Sections 3.2-3.5

3.4. **[Continuous-type random variables I]**

(a) $A = C = 0, B = 1$.

(b)

$$F(c) = \begin{cases} 0, & c \le 0 \\ c^2/2, & 0 \le c < 1 \\ -c^2/2 + 2c - 1, & 1 \le c < 2 \\ 1, & c \ge 2 \end{cases}$$

There are several ways to derive the above solution. One is to perform the integration of the pdf following the definition. Other ways are to argue geometrically, using the fact

that the area of a triangle is one half the base times height. For example, if $c \in [1,2]$, then $F(c)$ is one minus the area of the triangle formed by the pdf over the interval $[c,2]$. That triangle has base and height $2-c$, so its area is $(1/2)(2-c)^2$, and the CDF is $1-(1/2)(2-c)^2$. A different way for $c \in [1,2]$ is to add the area of the left half of the pdf (which is $1/2$) to the area of the trapezoid formed by the pdf over the interval $[1,c]$. This gives $F(c) = (1/2) + (c-1)(1+2-c)/2$, because the area of a trapezoid is the base times the average height.

3.6. **[Selecting pdf parameters to match a mean and CDF]**

The density is symmetric about a so the mean is a. Therefore $a = \mu_X = 2$.

In order for f_X to be a valid pdf, it must integrate to one. The region under f_X consists of a rectangle of area $2b(1)$ and two triangles of area $c(1)/2$ each, so its total area is $2b+c$; therefore, $2b+c=1$.

By inspection of the figure, $F_X(a-b)$ is the area of the first triangle, so $F_X(a-b) = \frac{c}{2}$. By the formula given for F_X, $F_X(a-b) = \frac{5c^2}{6}$. So $\frac{c}{2} = \frac{5c^2}{6}$, or $c = 0.6$. (ALTERNATIVELY, taking the derivative of the expression given form F_X shows that $f_X(u) = \frac{5}{3}(u-(a-b-c))$ in the interval $[a-b-c, a-b]$. Setting $f_X(a-b) = 1$ givens $\frac{5c}{3} = 1$ or $c = 0.6$. ALTERNATIVELY, taking the derivative of F_X twice yields $F_X'' = f_X' = \frac{5}{3}$ over the interval $[a-b-c, a-b]$. By the picture, the slope of f_X over the interval is $\frac{1}{c}$. So $\frac{5}{3} = \frac{1}{c}$ or $c = 0.6$. There are other ways to derive $c = 0.6$; they all involve comparing the picture of f_X over the interval $[a-b-c, a-b]$ to the formula given for F_X.) Since $2b+c=1$, $b = 0.2$.

3.8. **[A continuous approximation of the Zipf distribution]**

(a) In order for f_Y to integrate to one,

$$C = \int_{0.5}^{M+0.5} u^{-\alpha} du = \frac{u^{1-\alpha}}{1-\alpha}\Big|_{0.5}^{M+0.5} = \frac{(M+0.5)^{1-\alpha} - (0.5)^{1-\alpha}}{1-\alpha}$$

(b) Using the formula for the integral from part (a), we find

$$P\{Y \le 500.5\} = \frac{\int_{0.5}^{500+0.5} u^{-\alpha} du}{\int_{0.5}^{2000+0.5} u^{-\alpha} du} = \frac{(500.5)^{0.2} - (0.5)^{0.2}}{(2000.5)^{0.2} - (0.5)^{0.2}} = 0.7011$$

(This is close to $P\{X \le 500\}$, which is 0.6997.)

3.10. **[Uniform and exponential distribution II]**

(a) Let B be the event the radio comes from the first batch and let L be the lifetime of the radio. Let's calculate $P\{L \ge c\}$ for any constant $c \ge 0$, because it will be needed for several values of c. By the law of total probability,

$$P\{L \ge c\} = P(L \ge c|B)P(B) + P(L \ge c|B^c)P(B^c).$$

Since L is uniformly distributed over the interval $[0,2]$ if B is true,

$$P(L \ge c|B) = 1 - P(L \le c|B) = \begin{cases} 1 - \frac{c}{2} & 0 \le c \le 2 \\ 0 & c \ge 2 \end{cases}.$$

Since L has the Exp(0.1) distribution if B^c is true, $P(L \geq c|B^c) = e^{-c/10}$ for all $c \geq 0$. Combining these facts and using $P(B) = P(B^c) = \frac{1}{2}$, yields

$$P(L \geq c) = \begin{cases} \frac{1}{2}(1 - \frac{c}{2} + e^{-c/10}) & 0 \leq c \leq 2 \\ \frac{1}{2}(e^{-c/10}) & c \geq 2 \end{cases} \tag{6.1}$$

$$P(L \geq 5+3|L \geq 5) = \frac{\frac{1}{2}e^{-0.8}}{\frac{1}{2}e^{-0.5}} = e^{-0.3} = 0.7408$$

A shorter answer is the following. Given $L \geq 5$, the battery must be from the second batch, and thus have an exponentially distributed lifetime with parameter λ. By the memoryless property of the exponential distribution, the battery after five years is as good a new. So the probability it lasts at least three more years is $e^{-3\lambda} = e^{-0.3}$.

(b) By the definition of conditional probabilities and (6.1),

$$P(L \geq 1+3|L \geq 1) = \frac{\frac{1}{2}e^{-0.4}}{\frac{1}{2}(\frac{1}{2} + e^{-0.1})} = 0.4771$$

3.12. [**Disk crashes modeled by a Poisson process**]

(a) The (duration of time) times (the rate per unit time) is $(24)\lambda$.

(b) The number of crashes in a five hour period has the Poisson distribution with mean 5λ, so the probability of exactly three crashes in a five hour period is $\frac{(5\lambda)^3 e^{-5\lambda}}{3!}$.

(c) The mean time until a disk crash is $\frac{1}{\lambda}$, so the mean time until three disk crashes is $\frac{3}{\lambda}$.

(d) The time of the third disk crash has the Erlang distribution with parameters λ and $r = 3$, so the pdf is $f_T(t) = \begin{cases} \frac{\lambda^3 t^2 e^{-\lambda t}}{2} & t \geq 0 \\ 0 & \text{else} \end{cases}$.

(a) $\frac{P\{N_1=1, N_3-N_1=0\}}{P\{N_3=1\}} = \frac{P\{N_1=1\}P\{N_3-N_1=0\}}{P\{N_3=1\}} = \frac{\lambda e^{-\lambda} e^{-2\lambda}}{3\lambda e^{-3\lambda}} = \frac{1}{3}$.
ALTERNATIVELY, though it is not emphasized in the course, it turns out that given the number of arrivals of a Poisson process within an interval of time, the distribution of arrival times within the interval are as if they are independent and uniform. Thus, given one arrival during [0,3], it falls within the first third of the interval with probability one third.

(b) Erlang distribution with parameters 2 and λ. The pdf is $f_{T_2}(t) = \lambda^2 t e^{-\lambda t}, t \geq 0$. In general, the rth arrival has the Erlang distribution with parameters λ and r.

(c) Zero. The number of arrivals in the length zero interval $[1, 1]$ has the Poisson distribution with parameter $\lambda(1-1) = 0$, and hence there are no arrivals in that interval, with probability one. Another way to see this is to let T_k be the arrival time of the kth bus, which is a continuous-type random variables for each $k \geq 1$. Hence the probability any of them takes a given value is zero. Formally,

$$P\{\text{there is a bus arriving at } t = 1\} = P\left(\bigcup_{k=1}^{\infty} \{T_k = 1\}\right) \leq \sum_{k=1}^{\infty} P\{T_k = 1\} = 0.$$

3.14. **[Poisson tweets]**

(a) The number of tweets in one week is a Poisson random variable with parameter $(1/7)7 = 1$. Let W_i be the event you receive exactly one tweet on week i. Then $P(W_i) = e^{-1}1^1/1! = e^{-1}$, and by the independence of non-overlapping intervals, $P(W_1, W_2, W_3, W_4) = P(W_1)P(W_2)P(W_3)P(W_4) = \left(e^{-1}\right)^4 = e^{-4}$.

(b) Let T be the time (in weeks) until the third tweet arrives, and let N_2 be the number of tweets received by the end of the second week, which is a Poisson random variable with parameter $(1/7)14 = 2$. Then, $P\{T > 2\} = P\{N_2 \leq 2\} = \sum_{k=0}^{2} e^{-2}2^k/k! = e^{-2}2^0/0! + e^{-2}2^1/1! + e^{-2}2^2/2! = 5e^{-2}$.

(c) The time between consecutive tweets has the exponential distribution with parameter $1/7$, and thus mean 7 days. By the memoryless property of the exponential distribution, the mean amount of additional time until the arrival of the fifth tweet is $7 * 5 = 35$ days, or five weeks.

Linear Scaling, Gaussian Distribution, ML Parameter Estimation
Sections 3.6 & 3.7

3.16. **[Gaussian distribution]**

(a) Since X is a continuous-type random variable, $P\{X = c\} = 0$ for any c, including $c = 0$.

(b)

$$\begin{aligned} P\{|X+4| \geq 2\} &= P\{X \leq -6 \text{ or } X \geq -2\} \\ &= P\{X \leq -6\} + P\{X \geq -2\} \\ &= 2P\{X \geq -2\} \quad \text{(distribution is symmetric about -4)} \\ &= 2P\left\{\frac{X+4}{3} \geq \frac{-2+4}{3}\right\} = 2Q(2/3). \end{aligned}$$

Note that we could get the same answer without using the symmetry property:

$$P\{X \leq -6\} = P\left\{\frac{X+4}{3} \leq \frac{-6+4}{3}\right\} = \Phi\left(-\frac{2}{3}\right) = Q\left(\frac{2}{3}\right).$$

(c)

$$\begin{aligned} P\{0 < X < 2\} &= P\{X > 0\} - P\{X \geq 2\} \\ &= P\left\{\frac{X+4}{3} > 4/3\right\} - P\left\{\frac{X+4}{3} \geq 2\right\} \\ &= Q(4/3) - Q(2). \end{aligned}$$

We remark that in the above solution we could have written "$\geq$" instead of "$>$" and/or "$>$" instead of '$\geq$," because X is a continuous type random variable.

(d)

$$\begin{aligned} P\{X^2 < 9\} &= P\{-3 < X < 3\} \\ &= P\left\{\frac{1}{3} < \frac{X+4}{3} < 7/3\right\} \\ &= Q(1/3) - Q(7/3). \end{aligned}$$

3.18. **[Blind guessing answers on an exam]**

(a) By the problem description we take X to have the binomial distribution with parameters $n = 10$ and $p = 0.5$. Since $S = 3X - 3(10 - X) = 6X - 30$,

$$P\{S \geq 12\} = P\{6X - 30 \geq 12\} = P\{X \geq 7\}.$$

Since $E[X] = 5$, the Markov inequality yields:

$$P\{S \geq 12\} = P\{X \geq 7\} \leq \frac{E[X]}{7} = \frac{5}{7}.$$

(b) By the observed symmetry and the connection to X,

$$P\{S \geq 12\} = (0.5)P\{|S| \geq 12\} = (0.5)P\{|X - 5| \geq 2\}.$$

Since $E[X] = 5$ and $\mathrm{Var}(X) = 10(0.5)(0.5) = \frac{5}{2}$, Chebyshev inequality yields

$$P\{S \geq 12)\} = (0.5)P\{|X - 5| \geq 2\} \leq \frac{\mathrm{Var}(X)}{2 \cdot 4} = \frac{5}{16}.$$

(c)

$$P\{S \geq 12\} = P\{X \geq 7\} = P\{X \geq 6.5\} = P\left\{\frac{X-5}{\sqrt{2.5}} \geq \frac{1.5}{\sqrt{2.5}}\right\} \approx Q\left(\frac{1.5}{\sqrt{2.5}}\right) = 0.1714.$$

3.20. **[Betting with doubling or halving]**

(a) If he wins a game, his money increases by a factor of $\frac{3}{2}$. If he loses a game, his money decreases by a factor of 2. By definition, he wins X games and loses $196 - X$ games in the first 196 games. Thus, his money after 196 games is

$$S = 2^{20}\left(\frac{3}{2}\right)^X \left(\frac{1}{2}\right)^{196-X} = 2^{-176}3^X.$$

(b) From part (a),

$$P\{S \le 1\} = P\{\log_2 S \le 0\} = P\{-176 + X\log_2(3) \le 0\} = P\left\{X \le \frac{176}{\log_2(3)}\right\} = P\{X \le 110\}.$$

Since $E[X] = 98$ and $\mathrm{Var}(X) = 49$,

$$P\{S \le 1\} = P\left(\frac{X-98}{7} \le \frac{110-98}{7}\right) \approx \Phi\left(\frac{12}{7}\right) = 1 - Q\left(\frac{12}{7}\right) = 0.9568.$$

Using the continuity correction as follows gives a more accurate estimate. To derive it, note that $P\{X \le 110\} = P\{X \le 110.5\}$ because X is integer valued, and approximate $P\{X \le 110.5\}$:

$$P\{S \le 1\} = P\left(\frac{X-98}{7} \le \frac{110.5-98}{7}\right) \approx \Phi\left(\frac{12.5}{7}\right) = 1 - Q\left(\frac{12.5}{7}\right) = 0.9629.$$

(Using the binomial distribution directly gives $P\{S \le 1\} = P\{X \le 110\} = 0.9631$.)

3.22. **[ML parameter estimation for Rayleigh and uniform distributions]**

(a) By definition, $\widehat{\theta}_{ML}(10)$ is the value of θ that maximizes the likelihood of $X = 10$. The likelihood of $X = 10$ is $f_\theta(10) = \left(\frac{10}{\theta}\right) e^{-\frac{100}{2\theta}}$, and the log likelihood is $\ln(10) - \ln(\theta) - \frac{50}{\theta}$. Differentiation with respect to θ yields $\frac{d\ln f_\theta(10)}{d\theta} = -\frac{1}{\theta} + \frac{50}{\theta^2}$. This derivative is zero for $\theta = 50$, and it is positive for $\theta < 50$ and negative for $\theta > 50$. Hence, $\widehat{\theta}_{ML}(10) = 50$.
(Note: If θ is replaced by σ^2, then f is the Rayleigh pdf with parameter σ^2. The same reasoning as above shows that, in general, for observation $X = u$, $\widehat{\theta}_{ML} = \widehat{(\sigma^2)}_{ML} = \frac{u^2}{2}$.)

(b) The pdf of Y is given by

$$f_a(u) = \begin{cases} \frac{1}{\frac{2}{a} - \frac{1}{a}} = a & \text{if } \frac{1}{a} \le u \le \frac{2}{a} \\ 0 & \text{else} \end{cases}$$

In order for $f_a(3) > 0$, the support of f_a must include the observed value $u = 3$, or $\frac{1}{a} \le 3 \le \frac{2}{a}$, or equivalently $\frac{1}{3} \le a \le \frac{2}{3}$. So $f_a(3)$ is an increasing function of a over the interval $\frac{1}{3} \le a \le \frac{2}{3}$, and it is zero elsewhere. Therefore, the ML estimate is the largest value of a in this interval. Thus, $\hat{a}_{ML} = \frac{2}{3}$.

3.24. **[ML parameter estimation for independent exponentially distributed rvs]**

(a) Since $f_\lambda(u_i) = \lambda\exp(-\lambda u_i)$ for each i, the likelihood is $\lambda\exp(-\lambda u_1)\cdots\lambda\exp(-\lambda u_n) = \lambda^n\exp(-\lambda t)$ where $t = u_1 + \cdots + u_n$. That is, t is the sum of the lifetimes of all n lasers.

(b) By definition, $\widehat{\lambda}_{ML}$ is the value of λ that maximizes the likelihood found in (a). Differentiating the likelihood with respect to λ yields:

$$\begin{aligned} \frac{d(\lambda^n\exp(-\lambda t))}{d\lambda} &= n\lambda^{n-1}\exp(-\lambda t) - s\lambda^n\exp(-\lambda t) \\ &= (n - s\lambda)\lambda^{n-1}\exp(-\lambda t) \end{aligned}$$

The derivative is zero at $\lambda = \frac{n}{t}$ (and is positive for λ smaller than $\frac{n}{t}$ and negative for λ larger than $\frac{n}{t}$). So $\widehat{\lambda}_{ML} = \frac{n}{t}$. This formula makes intuitive sense; n laser failures happened during operation of total duration t, so $\widehat{\lambda}_{ML}$ is the observed rate of failures per unit time.

Functions of a random variable, failure rate functions, and binary hypothesis testing for continuous-type observations Sections 3.8-3.10

3.26. **[Some simple questions about a uniformly distributed random variable]**

(a) $E[X^2] = \int_0^3 \frac{u^2}{3} du = \left.\frac{u^3}{9}\right|_0^3 = 3.$

(b) $P\{\lfloor X^2 \rfloor = 3\} = P\{3 \le X^2 < 4\} = P\{\sqrt{3} \le X < 2\} = \frac{2-\sqrt{3}}{3}$.

(c) The range of Y is $(-\infty, \ln 3]$. Thus, using the fact the pdf of X is $\frac{1}{3}$ over $[0,3]$,

$$F_Y(c) = \begin{cases} P\{\ln X \le c\} = P\{X \le e^c\} = \frac{e^c}{3} & -\infty < c \le \ln 3 \\ 1 & c \ge \ln 3 \end{cases}.$$

3.28. **[A binary quantizer with Laplacian input]**

(a) Y takes values in $\{-\alpha, \alpha\}$; its pmf is

$$p_Y(u) = \begin{cases} P\{X \ge 0\} = 0.5 & \text{if } u = \alpha \\ P\{X < 0\} = 0.5 & \text{if } u = -\alpha \\ 0 & \text{else} \end{cases}$$

(b) Note that $(X-Y)^2$ is a function of X. If $u < 0$ and $X = u$ then the quantizer output is -α, so the squared error is $(u - (-\alpha))^2 = (u+\alpha)^2$. For example, if $\alpha = -1$, and if the input is -1.2, the output will be -1, and the squared error is $(-0.2)^2$. If $u > 0$ then the quantizer output is α so the squared error is $(u-\alpha)^2$. Thus, in general $(X-Y)^2 = h(X)$ where h is the function defined by

$$h(u) = \begin{cases} (u+\alpha)^2 & \text{if } u \ge 0 \\ (u-\alpha)^2 & \text{if } u < 0 \end{cases}$$

So using LOTUS and symmetry yields

$$\begin{aligned} E[(X-Y)^2] = E[h(X)] &= \int_{-\infty}^{0} (u+\alpha)^2 \frac{e^u}{2} du + \int_0^{\infty} (u-\alpha)^2 \frac{e^{-u}}{2} du \\ &= \int_0^{\infty} (u-\alpha)^2 e^{-u} du \\ &= \int_0^{\infty} (u^2 - 2\alpha u + \alpha^2) e^{-u} du \\ &= 2 - 2\alpha + \alpha^2 \end{aligned}$$

(c) The derivative of the mean square error with respect to α is $2(-1+\alpha)$, so the mean square error is decreasing in α for $\alpha < 1$ and increasing for $\alpha > 1$. So $\alpha = 1$ minimizes the mean square error. ALTERNATIVELY, the mean square error can be expressed as $1 + (1-\alpha)^2$, which makes it clear that $\alpha = 1$ minimizes it.

3.30. **[Log uniform and log normal random variables]**

(a) Z takes values in the set $[e^a, e^b]$; for $e^a \le c \le e^b$:

$$F_Z(c) = P\{e^U \le c\} = P\{U \le \ln c\} = \frac{\ln c - a}{b-a}.$$

Differentiating yields

$$f_Z(c) = \begin{cases} \frac{1}{c(b-a)} & e^a \le c \le e^b \\ 0 & \text{else.} \end{cases}$$

(b) By LOTUS, $E[Z] = \int_a^b \frac{e^u}{b-a} du = \frac{e^b - e^a}{b-a}$.

(c) Y is positive valued; for $c \ge 0$, $F_Y(c) = P\{e^X \le c\} = P\{X \le \ln c\} = F_X(\ln(c))$. Differentiating and using the chain rule yields

$$f_Y(c) = \begin{cases} \frac{f_X(\ln c)}{c} = \frac{1}{c\sqrt{2\pi}} \exp\left(-\frac{(\ln(c))^2}{2}\right) & c \ge 0 \\ 0 & \text{else.} \end{cases}$$

(d) By LOTUS,

$$\begin{aligned} E[Y] &= \int_{-\infty}^{\infty} e^u \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u^2}{2}\right) du \\ &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u^2 - 2u}{2}\right) du \\ &= \exp\left(\frac{1}{2}\right) \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(u-1)^2}{2}\right) du \\ &= \exp\left(\frac{1}{2}\right) = \sqrt{e} = 1.64872. \end{aligned}$$

3.32. **[Function of a random variable]**

(a) By LOTUS and symmetry,

$$\begin{aligned} E[Y] &= \int_{-\infty}^{\infty} \sqrt{|u|} f_X(u) du \\ &= \int_{-\infty}^{-1} \sqrt{-u} \frac{1}{2u^2} du + \int_1^{\infty} \sqrt{u} \frac{1}{2u^2} du \\ &= \int_1^{\infty} \sqrt{u} \frac{1}{u^2} du \\ &= \int_1^{\infty} u^{-3/2} du \\ &= (-2u^{-1/2})\Big|_1^{\infty} = 2 \end{aligned}$$

(b) Note that Y ranges over $[1, \infty)$. For $1 \leq c < \infty$,

$$F_Y(c) = P\{Y \leq c\} = 2P\{1 \leq X \leq c^2\} = \int_1^{c^2} \frac{1}{u^2} du = -\frac{1}{u}\Big|_1^{c^2} = 1 - \frac{1}{c^2}.$$

Differentiating with respect to c yields

$$f_Y(c) = \begin{cases} 2/c^3 & \text{if } c \geq 1 \\ 0 & \text{else.} \end{cases}$$

(c) By Example 3.8.11 of the notes, the function $h(u)$ is just the CDF of X : $h(c) = F_X(c)$. For $c \leq -1$,

$$F_X(c) = \int_{-\infty}^{c} \frac{1}{2u^2} du = -\frac{1}{2u}\Big|_{-\infty}^{c} = -\frac{1}{2c}$$

In particular, $F_X(-1) = 0.5$. Since $f_X(u) = 0$ for $u \in (-1, 1)$, it follows that F_X is flat over the interval $[-1, 1]$, with $F_X(1) = 0.5$. For $c \geq 1$

$$F_X(c) = 0.5 + \int_1^{c} \frac{1}{2u^2} du = 0.5 + \left(-\frac{1}{2u}\Big|_1^{c}\right) = 1 - \frac{1}{2c}$$

Combining the above,

$$h(c) = F_X(c) = \begin{cases} -\frac{1}{2c} & \text{if } c \leq -1 \\ 0.5 & \text{if } -1 < c \leq -1 \\ 1 - \frac{1}{2c} & \text{if } c > 1. \end{cases}$$

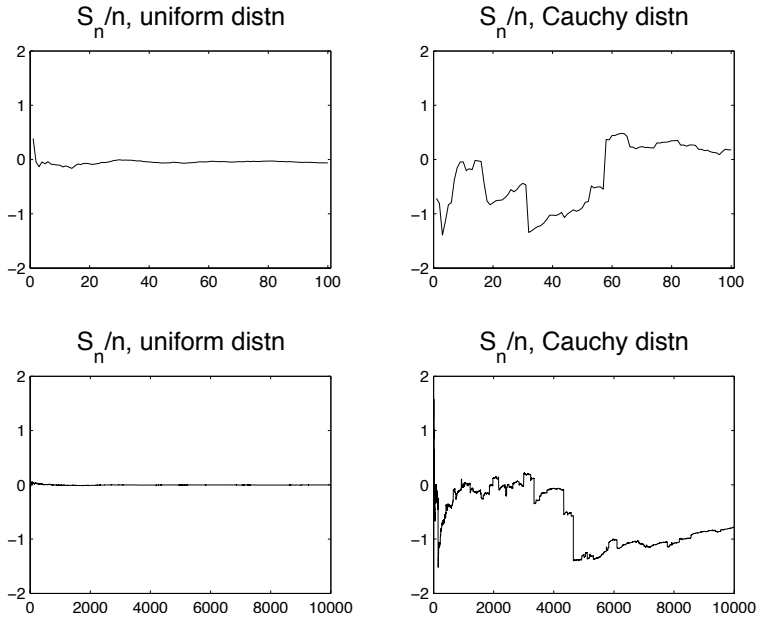
3.34. **[Generation of a random variable with a given failure rate function]**

(a) Clearly T is a nonnegative random variable. By the hint and the fact $F_X(c) = 1 - e^{-c}$ for $c \geq 0$, for any $t \geq 0$,

$$F_T(t) = P\{T \leq t\} = P\left\{X \leq \int_0^t r(s)ds\right\} = F_X\left(\int_0^t r(s)ds\right) = 1 - e^{-\int_0^t r(s)ds}$$

(b) By the fundamental theorem of calculus, the derivative of $\int_0^t r(s)ds$ with respect to t is $r(t)$. Using that fact and the chain rule yields $f_T(t) = F_T'(t) = r(t)e^{-\int_0^t r(s)ds}$. Thus, the failure rate function of T is given by $h(t) = \frac{f_T(t)}{1-F_T(t)} = r(t)$. That is, r is the failure rate function of T.

3.36. **[(COMPUTER EXERCISE) Running averages of independent, identically distributed random variables]**



3.38. **[Failure rate of a network with two parallel links]**

(a) $\{T > t\} = \{T_1 > t\} \cup \{T_2 > t\}$ and the events $\{T_1 > t\}$ and $\{T_2 > t\}$ are independent (because T_1 and T_2 are independent) so
$P\{T > t\} = P\{T_1 > t\} + P\{T_2 > t\} - P\{T_1 > t\}P\{T_2 > t\} = 2e^{-\lambda t} - e^{-2\lambda t}$.

(b) $f_T(t) = -(P\{T > t\})' = 2\lambda(e^{-\lambda t} - e^{-2\lambda t})$ for $t \geq 0$. Also, $f_T(t) = 0$ for $t < 0$.

(c) Using the answers to parts (a) and (b),

$$h(t) = \frac{f_T(t)}{P\{T > t\}} = \frac{2\lambda(e^{-\lambda t} - e^{-2\lambda t})}{2e^{-t\lambda} - e^{-2\lambda t}} = \frac{2\lambda(1 - e^{-\lambda t})}{2 - e^{-\lambda t}} = \lambda\left(1 - \frac{1}{2e^{\lambda t} - 1}\right).$$

(d) By the definition of conditional probabilities,

$$P\{\min\{T_1, T_2\} < t | T > t\} = \frac{P\{\min\{T_1, T_2\} < t, T > t\}}{P\{T > t\}}.$$

Observe that $\{\min\{T_1, T_2\} < t, T > t\} = \{T_1 < t, T_2 > t\} \cup \{T_2 < t, T_1 > t\}$ and the two events $\{T_1 < t, T_2 > t\}$ and $\{T_1 < t, T_2 > t\}$ are mutually exclusive, and by symmetry, they have equal probability. So $P\{\min\{T_1, T_2\} < t, T > t\} = 2P\{T_1 < t, T_2 > t\} = 2P\{T_1 < t\}P\{T_2 > t\} = 2(1 - e^{-\lambda t})e^{-\lambda t}$. Combining this with part (a) yields an expression for $P\{\min\{T_1, T_2\} < t | T > t\}$, which when multiplied by λ is the same as the expression for h found in part (c).

Jointly distributed random variables including independent random variables Sections 4.1-4.4

4.2. **[A joint distribution]**

(a) Clearly $f_X(u) = 0$ for $u < 0$. For $u \geq 0$,

$$f_X(u) = \int_0^\infty ve^{-(1+u)v}dv = \frac{1}{(1+u)^2}\int_0^\infty (1+u)^2 ve^{-(1+u)v}dv = \frac{1}{(1+u)^2},$$

where we used the fact that the gamma density with parameters $r = 2$ and $\lambda = 1 + u$ integrates to one.
Clearly $f_Y(v) = 0$ for $v < 0$. For $v > 0$,

$$f_V(v) = \int_0^\infty ve^{-(1+u)v}du = ve^{-v}\int_0^\infty e^{-uv}du = ve^{-v}\frac{1}{v} = e^{-v}.$$

That is, V has the exponential distribution with parameter one.

(b) Since the support of f_X is $\mathbb{R}_+$, the conditional pdf $f_{Y|X}(v|u)$ is well-defined only for $u \geq 0$. For such u,

$$f_{Y|X}(v|u) = \begin{cases} \frac{ve^{-(1+u)v}}{\frac{1}{(1+u)^2}} = (1+u)^2 ve^{-(1+u)v} & v \geq 0 \\ 0 & v < 0. \end{cases}$$

That is, the conditional distribution of Y given $X = u$ is the gamma distribution with parameters $r = 2$ and $\lambda = 1 + u$.
Since the support of f_Y is $\mathbb{R}_+$, the conditional pdf $f_{X|Y}(u|v)$ is well-defined only for $v \geq 0$. For such v,

$$f_{X|Y}(u|v) = \begin{cases} \frac{ve^{-(1+u)v}}{e^{-v}} = ve^{-uv} & u \geq 0 \\ 0 & u < 0. \end{cases}$$

That is, the conditional distribution of X given $Y = v$ is the exponential distribution with parameter v.

(c) The joint CDF $F_{X,Y}(u_o, v_o)$ is zero if either $u_o < 0$ or $v_o < 0$. For $u_o \geq 0$ and $v_o \geq 0$,

$$\begin{aligned} F_{X,Y}(u_0, v_0) &= \int_0^{v_o}\int_0^{u_o} ve^{-(1+u)v}dudv \\ &= \int_0^{v_o} e^{-v}\left\{\int_0^{u_o} ve^{-uv}du\right\}dv \\ &= \int_0^{v_o} e^{-v}(1 - e^{-u_o v})dv \\ &= \int_0^{v_o} (e^{-v} - e^{-v(1+u_o)})dv \\ &= \frac{1}{1+u_o}\left\{u_o + e^{-v_o(1+u_o)}\right\} - e^{-v_o} \end{aligned}$$

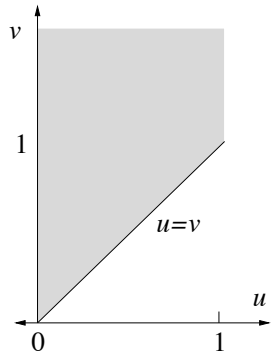
(d) No. For example, the answer to part (b) shows that $f_{Y|X}(v|u)$ is well defined for $u > 0$ and it is not a function of v alone.

4.4. **[Working with the joint pdf of two independent variables]**

(a) By the assumptions on X and Y

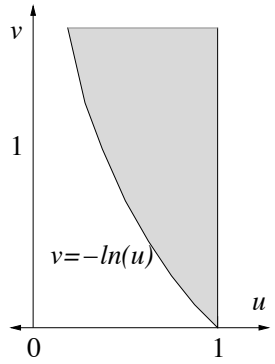
$$f_{X,Y}(u,v) = f_X(u)f_Y(v) = \begin{cases} \lambda e^{-\lambda v} & 0 \le u \le 1, v \ge 0 \\ 0 & \text{else.} \end{cases}$$

(b) The probability to be found, $P\{Y \ge X\}$, is the integral of the joint density over the set $\{(u,v) : v \ge u\}$ intersected with the support of $f_{X,Y}$, which is the shaded region shown:



Thus, $P\{Y \ge X\} = \int_0^1 \int_u^\infty \lambda e^{-\lambda v} dv du = \int_0^1 e^{-\lambda u} du = \frac{1-e^{-\lambda}}{\lambda}$. So $\lim_{\lambda \to 0} P\{Y \ge X\} = 1$ (by l'Hopital's rule) and $\lim_{\lambda \to \infty} P\{Y \ge X\} = 0$. (These limits make sense intuitively. Since the mean of Y is $1/\lambda$, taking $\lambda \to 0$ means that Y tends to be very large, and taking $\lambda \to \infty$ means Y tends to be close to zero.)

(c) The probability to be found, $P\{Xe^Y \ge 1\}$, is the integral of the joint density over the set $\{(u,v) : ue^v \ge 1\}$ intersected with the support of $f_{X,Y}$. Note that for u and v positive, the condition $ue^v \ge 1$ is equivalent to $v \ge -\ln(u)$, so the region to be integrated over is shown:



Thus, $P\{Xe^Y \ge 1\} = \int_0^1 \int_{-\ln(u)}^\infty \lambda e^{-\lambda v} dv du = \int_0^1 u^\lambda du = \frac{1}{1+\lambda}$. So $\lim_{\lambda \to 0} P\{Xe^Y \ge 1\} = 1$ and $\lim_{\lambda \to \infty} P\{Xe^Y \ge 1\} = 0$.

4.6. **[Working with a simple joint pdf]**

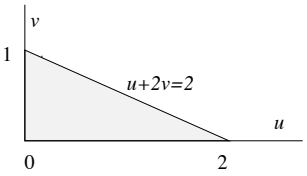
(a) The joint pdf is the product of the pdfs of an exponential with parameter $\lambda = 2$ and an Erlang with parameters $r = 2$ and $\lambda = 1$. Therefore, $f_X(u) = ue^{-u}$ for $u \ge 0$ and zero else.
Alternatively, $f_X(u) = \int_0^\infty 2ue^{-u-2v} dv = ue^{-u}$ for $u \ge 0$.

(b) X and Y are independent. Therefore, for $u > 0$,

$$f_{Y|X}(v|u) = \begin{cases} 2e^{-2v} & \text{for } v \geq 0 \\ 0 & \text{for } v < 0 \end{cases}$$

It's undefined for $u \leq 0$.
An alternative approach is to use that for $u > 0$, and $v \geq 0$, $f_{Y|X}(v|u) = \frac{f_{X,Y}(u,v)}{f_X(u)} = \frac{2ue^{-u-2v}}{ue^{-u}} = 2e^{-2v}$.

(c) The probability in question is the integral of the joint pdf over the region $\{u \geq 0, v \geq 0, u + 2v \leq 2\}$.



So $P\{X+2Y \leq 2\}$ can be expressed as $\int_0^2 \int_0^{1-u/2} 2ue^{-u-2v}dvdu$ or $\int_0^1 \int_0^{1-2v} 2ue^{-u-2v}dudv$
Integrating yields:

$$\begin{aligned} P\{X + 2Y \leq 2\} &= \int_0^2 \int_0^{1-u/2} 2ue^{-u-2v}dvdu \\ &= \int_0^2 ue^{-u} \int_0^{1-u/2} 2e^{-2v}dvdu \\ &= \int_0^2 (ue^{-u})(1 - e^{-2+u})du \\ &= \int_0^2 ue^{-u} - ue^{-2}du \\ &= -ue^{-u}\Big|_0^2 + \int_0^2 e^{-u}du - e^{-2}\int_0^2 udu \\ &= -2e^{-2} + 1 - e^{-2} - 2e^{-2} = 1 - 5e^{-2} \end{aligned}$$

(d) Using the fact that the exponential density and the Erlang density with parameter $r = 2$ each integrate to one:

$$\begin{aligned} E\left[\frac{Y}{X}\right] &= \int_0^\infty \int_0^\infty (v/u)2ue^{-u-2v}dvdu \\ &= \frac{1}{2}\left(\int_0^\infty e^{-u}du\right)\left(\int_0^\infty 2^2ve^{-2v}dvdu\right) \\ &= \frac{1}{2} \end{aligned}$$

4.8. **[Independent or not?]**

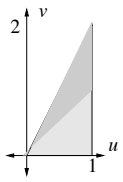
(a) No. The sum $u^2 + v^2$ does not factor into the product of a function of u and a function of v. (A test for a function $F(u, v)$ to factor is $F(u, v)F(u', v') = F(u, v')F(u', v)$ for all u, v, u', v'.)

(b) Yes, the support set is a product set and the value of the pdf over the product set factors into a function of u times a function of v: $Cuv(1+\cos(\pi u)) = (C_1 u(1+\cos(\pi u)))\,(C_2 v)$, where the constants C_1 and C_2 can be chosen to make the first factor integrate to one over $[0,1]$ and the second factor integrate to one over $[2,3]$.

(c) No. Even though $2\exp(-u+2v)$ factors, the support of the pdf is not a product set.

4.10. **[Working with a joint pdf II]**

(a) The support of the pdf is shown.



They are not independent because, for example, the support of $f_{X,Y}$ is not a product set (see Propositions 4.4.3 and 4.4.4).

(b) For $0 \le u \le 1$, $f_{X,Y}(u,v)$ is zero if $v > 2u$. Thus,

$$f_X(u) = \int_0^{2u} \alpha(u+v^2)dv = \alpha\left(2u^2 + \frac{(2u)^3}{3}\right).$$

for $0 \le u \le 1$, and $f_X(u) = 0$ elsewhere.

(c) By integrating $f_X(u)$ from 0 to 1, we have

$$1 = \int_0^1 2\alpha\left(u^2 + \frac{4u^3}{3}\right)du = 2\alpha\left(\frac{1}{3}+\frac{1}{3}\right) = \frac{4}{3}\alpha.$$

Since the right-hand side is equal to one, $\alpha = \frac{3}{4}$.

(d) The support of $f_X(u)$ is $u \in (0,1]$. Thus, the conditional pdf $f_{Y|X}(v|u)$ is defined for $u \in (0,1]$. (Note: In such a question one might wonder whether $u = 0$ should be included. That is a fine point and graders are hereby asked not to pay attention to it.)

(e) This requires integration over the subset of the support where $u < v$, which is shaded more darkly in the sketch of the support of the pdf.

$$\begin{aligned} P\{X<Y\} &= \int_0^1 \int_u^{2u} \frac{3}{4}(u+v^2)dvdu \\ &= \frac{3}{4}\int_0^1 \left(u^2 + \frac{7u^3}{3}\right)du \\ &= \frac{3}{4}\left(\frac{u^3}{3}\Big|_0^1 + \frac{7u^4}{12}\Big|_0^1\right) = \frac{1}{4}+\frac{7}{16} = \frac{11}{16} \end{aligned}$$

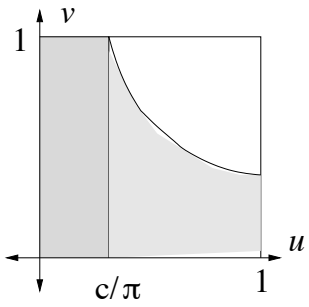
4.12. **[The volume of a random cylinder]**

(a) The joint pdf is one on the unit square and zero elsewhere. So, by LOTUS,

$$\begin{aligned} E[V] &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{H,R}(u,v)\pi uv^2 dudv \\ &= \int_0^1\int_0^1 \pi uv^2 dudv \\ &= \pi\left(\int_0^1 udu\right)\left(\int_0^1 v^2 dv\right) = \pi\left(\frac{1}{2}\right)\left(\frac{1}{3}\right) = \frac{\pi}{6} \end{aligned}$$

(b) Always $V \geq 0$ because $H \geq 0$ and $R \geq 0$. V can be close to zero because the same is true for H and R. The largest V can be is π, which happens if $H = R = 1$. Thus, the distribution of V has support $[0, \pi]$.

(c) $F_V(c) = 0$ for $c \leq 0$ and $F_V(c) = 1$ for $c \geq \pi$. So fix c with $0 \leq c \leq \pi$ and focus on calculating $F_V(c) = P\{\pi HR^2 \leq c\}$. It is the same as the area of the region $\{(u,v) \in [0,1]^2 : \pi uv^2 \leq c\}$, which is pictured below as a shaded region.



The curved part of the boundary is given by the equation $\pi uv^2 = c$, or equivalently, $v = \sqrt{\frac{c}{\pi u}}$. The curve intersects the horizontal line $v = 1$ at $u = \frac{c}{\pi}$. Thus, the shaded region is the union of the rectangle with $u \in \left[0, \frac{c}{\pi}\right]$ and the region under the curve with $u \in \left[\frac{c}{\pi}, 1\right]$. Therefore,

$$\begin{aligned} F_V(c) &= P\left\{HR^2 \leq \frac{c}{\pi}\right\} \\ &= \int_0^{\frac{c}{\pi}}\int_0^1 dvdu + \int_{\frac{c}{\pi}}^1\int_0^{\sqrt{\frac{c}{\pi u}}} dvdu \\ &= \frac{c}{\pi} + \sqrt{\frac{c}{\pi}}\int_{\frac{c}{\pi}}^1 u^{-\frac{1}{2}} du \\ &= \frac{c}{\pi} + \sqrt{\frac{c}{\pi}}2\left(1 - \sqrt{\frac{c}{\pi}}\right) \\ &= 2\sqrt{\frac{c}{\pi}} - \frac{c}{\pi} \end{aligned}$$

Summarizing,

$$F_V(c) = \begin{cases} 0 & c \leq 0 \\ 2\sqrt{\frac{c}{\pi}} - \frac{c}{\pi} & 0 \leq c \leq \pi \\ 1 & c \geq \pi \end{cases}$$

(d) Differentiating F_V yields

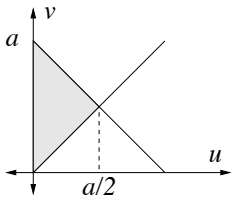
$$f_V(c) = \begin{cases} \frac{1}{\sqrt{\pi c}} - \frac{1}{\pi} & 0 < c \leq \pi \\ 0 & \text{else} \end{cases}$$

(Note: Alternatively we could have let the pdf equal infinity at $c = 0$. The value of a pdf at a single point doesn't matter, because the value of the pdf at a single point doesn't change the integrals of the pdf.)

Joint pdfs and functions of two random variables Sections 4.5&4.6

4.14. **[Functions of random variables]**

The variable Z takes values in the positive reals, and for $a \geq 0$, $F_Z(a) = P\{Z \leq a\}$ is equal to the integral of the joint pdf over the shaded region:



The integral can be computed using either integration with respect to u on the outside or integration with respect to v on the outside. Using u on the outside permits us to do the computation using only one double integral:

$$F_Z(a) = \int_0^{a/2} \int_u^{a-u} 2e^{-u-v} dv \; du = 1 - (1+a)e^{-a}.$$

Therefore,

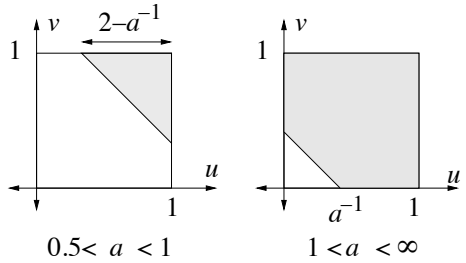
$$F_Z(a) = \begin{cases} 1 - (1+a)e^{-a} & a \geq 0 \\ 0 & \text{else.} \end{cases}$$

Differentiating the CDF yields $f_Z(a) = ae^{-a}$ for $a \geq 0$. That is, Z has the gamma distribution with parameters $r = 2$ and $\lambda = 1$. (This could have been deduced without calculation. The joint CDF is the same as the joint CDF of two independent exponential random variables, conditioned on the first to be smaller than the second. Since the two random variables are independent and identically distributed, such conditioning does not change the distribution of the sum, and the sum of two independent exponentials has the gamma distribution with $r = 2$.)

4.16. **[A function of two random variables]**

The current I is given by $I = \frac{1}{R_1 + R_2}$. One approach to this problem is to use the fact that the distribution of $R_1 + R_2$ has the triangular pdf as found in Example 4.5.4. Here we take a direct approach. Note that I takes values in the set $[\frac{1}{2}, \infty)$. For a in that range, $F_I(a) = P\{I \leq a\} = P\{1/(R_1 + R_2) \leq a\} = P\{R_1 + R_2 \geq a^{-1}\}$. Since (R_1, R_2) is uniformly distributed over

the unit square region $[0,1]^2$, the probability $P\{R_1 + R_2 \geq a^{-1}\}$ is the area of the shaded region, consisting of the intersection of $[0,1]^2$ and the half-plane $\{(u,v) : u + v \geq a^{-1}\}$. The shape of the intersection is qualitatively different for $a < 1$ and $a > 1$; an example of each of these two cases is shown in the figure.



We thus consider these two cases separately. In case $0.5 < a < 1$ the shaded region is triangular with area one half base times height. The edge of the half plane, given by $\{u+v = a^{-1}\}$, intersects the line $v = 1$ at $u = a^{-1} - 1$. Therefore, the width of the triangular region is $1 - (a^{-1} - 1)$ or $2 - a^{-1}$, and the height is the same. In case $1 < a < \infty$ the shaded region is $[0,1]^2$ with a triangular region removed from the lower left corner, so its area is one minus the area of the triangle. The width and height of the triangle are a^{-1}. Therefore,

$$F_I(a) = \begin{cases} 0 & a < 1/2 \\ (2 - a^{-1})^2/2 & 1/2 \leq a \leq 1 \\ 1 - a^{-2}/2, & 1 < a < \infty \end{cases}$$

Differentiation yields:

$$f_I(a) = \begin{cases} 2a^{-2} - a^{-3}, & 1/2 \leq a \leq 1 \\ a^{-3}, & 1 < a < \infty \end{cases}$$

Moments of jointly distributed random variables, minimum mean square error estimation Sections 4.8-4.9.

4.18. **[Variance and correlation]**

(a) Yes. Using the substitutions $\mathrm{Var}(X + 3Y) = \mathrm{Var}(X) + 6\mathrm{Cov}(X,Y) + 9\mathrm{Var}(Y)$ and $\mathrm{Var}(X - 3Y) = \mathrm{Var}(X) - 6\mathrm{Cov}(X,Y) + 9\mathrm{Var}(Y)$ in the given equation and canceling the $\mathrm{Var}(X)$ and $\mathrm{Var}(Y)$ terms yields $6\mathrm{Cov}(X,Y) = -6\mathrm{Cov}(X,Y)$, or $\mathrm{Cov}(X,Y) = 0$. Hence, X and Y are uncorrelated.

(b) No. For example, suppose X is any random variable with positive variance and $P\{X = Y\} = 1$. Then $\mathrm{Var}(X) = \mathrm{Var}(Y)$ but $\mathrm{Cov}(X,Y) = \mathrm{Var}(X) \neq 0$.

(c) Yes, because $\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\mathrm{Cov}(X,Y)$, so the given statement is equivalent to $\mathrm{Var}(X) + \mathrm{Var}(Y) + 2\mathrm{Cov}(X,Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$, or $\mathrm{Cov}(X,Y) = 0$.

(d) No. For example, suppose $P\{X = Y = 1\} = P\{X = Y = -1\} = 0.5$. Then X^2 and Y^2 are both equal to one with probability one, so $\mathrm{Cov}(X^2, Y^2) = \mathrm{Cov}(1,1) = 0$, but $\mathrm{Cov}(X,Y) = E[XY] - E[X]E[Y] = 1 - 0 \cdot 0 = 1 \neq 0$.

4.20. **[Variances and covariances of sums]**

(a) $\text{Cov}(3X+2, 5Y-1) = \text{Cov}(3X, 5Y) = 15\text{Cov}(X,Y)$.

(b)

$$\begin{aligned}\text{Cov}(2X+1, X+5Y-1) &= \text{Cov}(2X, X+5Y) = \text{Cov}(2X,X) + \text{Cov}(2X,5Y) \\ &= 2\text{Cov}(X,X) + 10\text{Cov}(X,Y) = 2\text{Var}(X) + 10\text{Cov}(X,Y)\end{aligned}$$

(c)

$$\begin{aligned}\text{Cov}(2X+3Z, Y+2Z) &= \text{Cov}(2X,Y) + \text{Cov}(2X+2Z) + \text{Cov}(3Z,Y) + \text{Cov}(3Z,2Z) \\ &= 2\text{Cov}(X,Y) + 4\text{Cov}(X,Z) + 3\text{Cov}(Z,Y) + 6\text{Cov}(Z,Z) \\ &= 2\text{Cov}(X,Y) + 6\text{Var}(Z)\end{aligned}$$

4.22. **[Signal to noise ratio with correlated observations]**

(a) In general, for $S = \frac{X_1+X_2}{2}$.

$$\begin{aligned}E[S] &= \mu_S = E\left[\frac{X_1+X_2}{2}\right] = \mu \\ \sigma_S^2 &= \frac{\text{Var}(X_1+X_2)}{4} = \frac{2\sigma^2 + 2\text{Cov}(X_1,X_2)}{4} = \frac{\sigma^2 + \text{Cov}(X_1,X_2)}{2} \\ SNR_S &= \frac{2\mu^2}{\sigma^2 + \text{Cov}(X_1,X_2)}\end{aligned}$$

Thus, if X_1 and X_2 are uncorrelated, $SNR_S = \frac{2\mu^2}{\sigma^2} = 2SNR_X$. Thus, averaging improves the SNR by a factor equal to the number of observations being averaged, if the observations are uncorrelated.

(b) Since $\text{Cov}(X_1,X_2) = \sigma^2\rho_{X_1,X_2}$, the formula above for SNR_S is equivalent to

$$SNR_S = \frac{2\mu^2}{\sigma^2(1+\rho_{X_1X_2})}.$$

Setting SNR_S equal to $1.5\frac{\mu^2}{\sigma^2}$ yields $\rho_{XY} = \frac{1}{3}$.

(c) $SNR_S \to \infty$ as $\rho_{X_1X_2} \to -1$.

4.24. **[Linear minimum MSE estimation from uncorrelated observations]**

(a) The MSE can be written as $E[((Y - bX_1 - cX_2) - a)^2]$, which is the same as the MSE for estimation of $Y - bX_1 - cX_2$ by the constant a. The optimal choice of a is $E[Y - bX_1 - cX_2] = E[Y]$. Substituting $a = E[Y]$, the MSE satisfies

$$\begin{aligned}\text{MSE} &= \text{Var}(Y - bX_1 - cX_2) \\ &= \text{Cov}(Y - bX_1 - cX_2, Y - bX_1 - cX_2) \\ &= \text{Cov}(Y,Y) + b^2\text{Cov}(X_1,X_1) - 2b\text{Cov}(Y,X_1) + c^2\text{Cov}(X_2,X_2) - 2c\text{Cov}(Y,X_2) \\ &= \text{Var}(Y) + \left(b^2\text{Var}(X_1) - 2b\text{Cov}(Y,X_1)\right) + \left(c^2\text{Var}(X_2) - 2c\text{Cov}(Y,X_2)\right). \quad (6.2)\end{aligned}$$

The MSE is quadratic in b and c and the minimizers are easily found to be $b = \frac{\text{Cov}(Y,X_1)}{\text{Var}(X_1)}$ and $c = \frac{\text{Cov}(Y,X_2)}{\text{Var}(X_2)}$. Thus, $L(X_1, X_2) = E[Y] + \frac{\text{Cov}(Y,X_1)}{\text{Var}(X_1)}X_1 + \frac{\text{Cov}(Y,X_2)}{\text{Var}(X_2)}X_2$.

(b) Substituting the values of b and c found into (6.2) yields

$$\text{MSE} = \text{Var}(Y) - \frac{\text{Cov}(Y, X_1)^2}{\text{Var}(X_1)} - \frac{\text{Cov}(Y, X_2)^2}{\text{Var}(X_2)}.$$

4.26. **[What's new? or the innovations method]**

(a) $\text{Cov}(X_1, X_2 - hX_1) = \text{Cov}(X_1, X_2) - h\text{Var}(X_1) = 0.5 - h$, so $h = 0.5$. Thus, $\widetilde{X_2} = X_2 - (0.5)X_1$.

(b)

$$\begin{aligned}
\text{Var}(\widetilde{X}_2) &= \text{Cov}(X_2 - (0.5)X_1, X_2 - (0.5)X_1) \\
&= \text{Var}(X_2) - 2(0.5)\text{Cov}(X_1, X_2) + (0.5)^2\text{Var}(X_1) \\
&= 1 - 0.5 + 0.25 = 0.75. \\
\text{Cov}(Y, \widetilde{X}_2) &= \text{Cov}(Y, X_2 - (0.5)X_1) \\
&= \text{Cov}(Y, X_2) - (0.5)\text{Cov}(Y, X_1) \\
&= 0.8 - (0.5)(0.8) = 0.4
\end{aligned}$$

(c) $a = E[Y] = 0$, $b = \frac{\text{Cov}(Y,X_1)}{\text{Var}(X_1)} = 0.8$, and $c = \frac{\text{Cov}(Y,\widetilde{X}_2)}{\text{Var}(\widetilde{X}_2)} = \frac{0.4}{0.75}$, and

$$\text{MSE} = \text{Var}(Y) - \frac{\text{Cov}(Y, X_1)^2}{\text{Var}(X_1)} - \frac{\text{Cov}(Y, \widetilde{X}_2)^2}{\text{Var}(\widetilde{X}_2)} = 1 - 0.64 - \frac{(0.4)^2}{0.75} = 0.1466\ldots.$$

(d) The estimator is $L^* = (0.8)X_1 + \frac{0.4}{0.75}(X_2 - (0.5)X_1) = \frac{0.4}{0.75}(X_1 + X_2)$.

4.28. **[Simple estimation problems]**

(a) The range of possible values of X is the interval $[0, 1]$. By inspection (or using the definition of conditional density), if $0 \le u \le 1$, we see that given $X = u$, the conditional distribution of Y is the uniform distribution over the interval $[0, u]$. Therefore, $E[Y^2|X = u] = \int_0^u \frac{1}{u}v^2 dv = \frac{u^2}{3}$.

(b) By the same observation used in part (a), $E[Y|X = u] = \frac{u}{2}$. Since this is a linear function of u, it is also equal to $\widehat{E}[Y|X = u]$. That is, $\widehat{E}[Y|X = u] = \frac{u}{2}$. (This result can be worked out using the formula for $\widehat{E}[Y|X = u]$ as well.)

LLN, CLT, and joint Gaussian distribution Sections 4.10 & 4.11

4.30. **[Law of Large Numbers and Central Limit Theorem]**

(a) From LLN, we have

$$\begin{aligned} P\left\{\left|\frac{S_n}{n}-\mu_X\right| \geq 0.01\mu_X\right\} &\leq \frac{\sigma_X^2}{n10^{-4}\mu_X^2} \\ P\left\{\left|\frac{S_n}{n}-\mu_X\right| < 0.01\mu_X\right\} &\geq 1-\frac{\sigma_X^2}{n10^{-4}\mu_X^2} \end{aligned}$$

where $\mu_X = \frac{1+2+3+4+5+6}{6} = 3.5$ and $\sigma_X^2 = \frac{1+4+9+16+25+36}{6} - (3.5)^2 = 2.92$. Thus, substituting for μ_X and σ_X^2 in the equation below,

$$1-\frac{\sigma_X^2}{n10^{-4}\mu_X^2} \geq 0.95$$

gives $n \geq 47674$.

(b) $E[S_n] = 3.5n$, $\text{Var}(S_n) = 2.92n$, and the standard deviation of S_n is $1.7078\sqrt{n}$.

$$\begin{aligned} P\{|S_n - 3.5n| \leq (0.035)n\} &= P\left\{\frac{|S_n-3.5n|}{1.7078\sqrt{n}} \leq \frac{(0.035)n}{1.7078\sqrt{n}}\right\} \\ &\approx 1-2Q\left(\frac{(0.035)n}{1.7078\sqrt{n}}\right) \end{aligned}$$

From the normal tables, $Q(x) = 0.025$ for $x = 1.96$. Thus, $n \geq \lceil((1.96)(1.7078)/(0.035))^2\rceil = 9147$ is sufficient.

4.32. **[Marathon blackjack]**

(a) The expected net gain is $(1000)(\$100)(-0.0029) = -\$290$

(b) The net gain, in dollars, is the sum of the gains for the 1000 games, $S = X_1 + \cdots + X_n$, where each X_i has mean $100(0.0029) = 0.29$ and standard deviation $(100)(1.141) = 114$. As already mentioned, S has mean $1000(-0.29) = -290$, and its standard deviation is $114\sqrt{1000} = 3605$. By the Gaussian approximation backed by the central limit theorem,

$$P\{S > 0\} = P\left\{\frac{S+290}{3605} > \frac{290}{3605}\right\} \approx Q\left(\frac{290}{3605}\right) = 0.4679$$

(c)

$$P\{S > 1000\} = P\left\{\frac{S+290}{3605} > \frac{1290}{3605}\right\} \approx Q\left(\frac{1290}{3605}\right) = 0.3602$$

(d) Replacing 1000 by a general integer n, we have

$$P\{S > 0\} = P\left\{\frac{S+(0.29)n}{114\sqrt{n}} > \frac{(0.29)n}{114\sqrt{n}}\right\} \approx Q\left(\frac{(0.29)n}{114\sqrt{n}}\right)$$

Since $Q(0.253) = 0.4$, we solve $\frac{(0.29)n}{114\sqrt{n}} = 0.253$ or

$$n = \left(\frac{(0.253)(114)}{0.29}\right)^2 = 9891$$

(That is a lot of blackjack!)

4.34. **[The CLT and the Poisson distribution]**

(a) Consider a Poisson random process $(N_t : t \geq 0)$ with rate one. Then X has the same distribution as N_{10}, which in turn is the sum of 10 independent Poisson random variables with mean one, because $N_{10} = N_1 + (N_2 - N_1) + \cdots + (N_{10} - N_9)$. Therefore, the CLT suggests that the distribution of X should be approximately Gaussian.

(b) $p_X(12) = \frac{10^{12}e^{-10}}{12!} = 0.09478$.

(c) Since X has mean and variance equal to λ, the random variable $\widetilde{X}$ has the $N(10, 10)$ distribution. Therefore, $P\{11.5 \leq \widetilde{X} \leq 12.5\} = P\left\{\frac{11.5-10}{\sqrt{10}} \leq \frac{\widetilde{X}-10}{\sqrt{10}} \leq \frac{12.5-10}{\sqrt{10}}\right\} = \Phi\left(\frac{2.5}{\sqrt{10}}\right) - \Phi\left(\frac{1.5}{\sqrt{10}}\right) = 0.10303$

(d) $f_{\widetilde{X}}(12) = \frac{1}{\sqrt{2\pi(10)}} \exp\left(-\frac{(12-10)^2}{2(10)}\right) = 0.10329$. Note: The approximations are off by around 10%. This inaccuracy is not too surprising, since ten is a small number of random variables for the CLT based Gaussian approximation.)

4.36. **[Conditional means for a joint Gaussian pdf]**

(a) Since X and $X + Y$ are independent, they are uncorrelated. So $0 = \mathrm{Cov}(X, X+Y) = \mathrm{Cov}(X, X) + \mathrm{Cov}(X, Y) = \mathrm{Var}(X) + \mathrm{Cov}(X, Y) = 1 + \mathrm{Cov}(X, Y)$. Therefore, $\mathrm{Cov}(X, Y) = -1$.

(b) Since X and $X + Y$ are independent, $E[X|X + Y = 2] = E[X] = 0$.

(c) Since X and Y are jointly Gaussian, $E[Y|X = 2] = \widehat{E}[Y|X = 2] = \mu_Y + \frac{\mathrm{Cov}(Y,X)}{\mathrm{Var}(X)}(2 - \mu_X) = -2$.

4.38. **[Jointly Gaussian Random Variables I]**

(a) X has the $N(1, 9)$ distribution; $f_X(u) = \frac{1}{\sqrt{18\pi}} e^{-(u-1)^2/18}$.

(b) By the formula for wide sense conditional expectation, $\widehat{E}[Y|X = 5] = 2 + \frac{6}{9}(5-1) = 4.667$, and by the formula for the corresponding MSE, $\sigma_e^2 = \sigma_Y^2 - \frac{\mathrm{Cov}(X,Y)^2}{\sigma_X^2} = 16 - \frac{6^2}{9} = 12$. So, the conditional distribution of Y given $X = 5$ is the $N(4.667, 12)$ distribution; $f_{Y|X}(v|5) = \frac{1}{\sqrt{24\pi}} e^{-(v-4.667)^2/24}$.

(c) By part (b), this is the probability that a $N(4.667, 12)$ random variable is greater than or equal to 2, which is $Q\left(\frac{2-4.667}{\sqrt{12}}\right) = Q(-0.9698) = 0.7793$.

(d) The second moment of a random variable Z is $E[Z^2]$, and it is equal to $E[Z]^2 + \text{Var}(Z)$. The idea is to apply that observation to the conditional distribution of Y given $X = 5$. By part (b), that conditional distribution is the $N(4.667, 12)$ distribution, so the second moment for it is given by $E[Y^2|X = 5] = (4.667)^2 + 12 = 33.7778$.

4.40. **[Estimation of jointly Gaussian random variables]**

(a) $E[Z] = E[X + 4Y - 1] = E[X] + 4E[Y] - 1 = 23$.
We use the fact $\text{Cov}(X, Y) = \rho\sigma_X\sigma_Y = (0.4)(4)(5) = 8$ to get

$$\begin{aligned}\text{Var}(Z) &= \text{Var}(X + 4Y - 1)\\ &= \text{Var}(X) + 16 \cdot \text{Var}(Y) + 2 \cdot 4 \cdot \text{Cov}(X, Y)\\ &= 16 + 400 + 64 = 480\end{aligned}$$

(b) Since Z is a linear combination of jointly Gaussian random variables, Z is Gaussian. So

$$P\{Z \geq 40\} = P\left\{\frac{Z - 23}{\sqrt{480}} \geq \frac{40 - 23}{\sqrt{480}}\right\} = Q\left(\frac{40 - 23}{\sqrt{480}}\right) = Q(0.7759) = 0.2189$$

(c) Since Z and Y are linear combinations of the jointly Gaussian random variables X and Y, the variables Z and Y are jointly Gaussian. Therefore, the best unconstrained estimator of Y given Z is the best linear estimator of Y given Z.
So

$$g^*(Z) = L^*(Z) = \widehat{E}[Y|Z] = E[Y] + \frac{\text{Cov}(Y, Z)}{\text{Var}(Z)}(Z - E[Z]).$$

Using

$$\text{Cov}(Y, Z) = \text{Cov}(Y, X + 4Y - 1) = \text{Cov}(Y, X) + 4\text{Var}(Y) = 8 + 100 = 108,$$

we find

$$g^*(Z) = L^*(Z) = 5 + \frac{108}{480}(Z - 23) = 5 + \frac{27}{120}(Z - 23)$$

and

$$\text{MSE} = \text{Var}(Y) - \frac{(\text{Cov}(Y, Z))^2}{\text{Var}(Z)} = 25 - \frac{(108)^2}{480} = 0.7.$$

4.42. **[Joint empirical distribution of ECE 313 scores]**

(a) By the joint Gaussian assumption,
$E[Y|X = u] = \widehat{E}[Y|X = u] = 152 + \frac{(35)(0.71)}{19}(u - 67) = 152 + 1.31(u - 67)$

(b) By the formula for minimum MSE for linear estimation,
$\sigma_e = \sigma_X\sqrt{1 - \rho^2} = 35\sqrt{1 - 0.71^2} = 24.65$. Thus, given $X = u$, the conditional distribution of Y is normal with mean $152 + 1.31(u - 67)$ and standard deviation 24.65 (i.e. variance 607.5).

Index

area rule for expectation, 135
average error probability, 62

Bayes' formula, 52
Bernoulli distribution, 37, 234
Bernoulli process, 43
binomial coefficient, 13
binomial distribution, 37, 234

cardinality, 10
Cauchy distribution, 128
CDF, *see* cumulative distribution function
central limit theorem, 208
Chebychev inequality, 50
complementary CDF, 102
conditional
    expectation, 163
    mean, *see* conditional expectation
    pdf, 163
    probability, 32
confidence interval, 51
confidence level, 51
correlation
    coefficient, 191
count times, 106
coupon collector problem, 42
cumulative distribution function, 93
    inverse of, 133
    joint, 157

decision rule, 59
DeMorgan's laws, 7
determinant, 185
distribution of a function of a random variable, 122

Erlang distribution, 110
event, 6
events, 8
expectation of a random variable, *see* mean of a random variable

failure rate function, 136
false alarm probability, 60
flow network, 69

gamma distribution, 110
Gaussian
    bivariate distribution, 211
Gaussian approximation, 117
    with continuity correction, 118
generating a random variable with given distribution, 133
geometric distribution, 41
geometric series, 18

increment of a counting process, 44
independence
    events, 34
    pairwise, 35
    random variables, 37, 171
intercount times, 107
interval estimator, 51

joint probability matrix, 62
jointly continuous-type, 161

Karnaugh map, 7, 14

Laplace distribution, 142
law of large numbers, 206
law of the unconscious statistician, 28, 99, 162
law of total probability, 52
    for a pdf, 163
likelihood matrix, 59
likelihood ratio test, 61
limit
    left, 94

right, 94
LOTUS, *see* law of the unconscious statistician
LRT, *see* likelihood ratio test

Maclaurin series, 38
MAP decision rule, *see* maximum a posteriori decision rule
Markov inequality, 49
maximum a posteriori decision rule, 62
maximum likelihood
decision rule, 60
parameter estimation, 121
maximum likelihood parameter estimate, 47
mean of a random variable, 27
mean square error, 199
mean square error., 200
memoryless property
in continuous time, 103
in discrete time, 42, 235
miss probability, 60
mode of a pmf, 40
moment, 31

n choose k, 13
negative binomial distribution, 44
network outage, 66

one-to-one function, 186

partition, 7, 52
pdf, *see* probability density function
pmf, *see* probability mass function
Poisson distribution, 235
Poisson process, 107
principle of counting, 10
principle of over counting, 12
prior probabilities, 61
probability density function, 98
conditional, 163
joint, 161
marginal, 162
probability mass function, 25
conditional, 160
joint, 159
marginal, 160
probability measure, 8
probability space, 8
product set, 172

random variable, 25
continuous-type, 97
discrete-type, 25, 96
Rayleigh distribution, 137, 188

sample mean, 196
sample path, 43
sample space, 6, 8
sample variance, 196
scaling rule for pdfs, 111
Schwarz's inequality, 194
sensor fusion, 64
standard deviation, 30
standard normal distribution, 113
standardized version of a random variable, 31
support
of a joint pdf, 162
of a pdf, 98

Taylor series, 38

unbiased estimator, 196
uncorrelated random variables, 191
uniform prior, 62
union bound, 66

variance, 30

wide-sense conditional expectation, 202