2008春季　統計學（一）講義

張福春

中山大學應用數學系

fuchuen@gmail.com

PROBABILITY AND STATISTICS

Kevin J. Hastings (1997)

2008-02-18 ∼ 2008-06-15

This page intentionally left blank

CONTENTS

Preface

Contents

統計學（一）課程介紹（2008春季）

上課時間	週二10:10∼12:00、週三09:10∼10:00
上課地點	理3001
討論時間	週(四、五) (10:00∼12:00)
授課教師	張福春 教授 理學院 3002-4 Tel: (O) (07) 525-2000 ext 3823 Email: fuchuen@gmail.com
講授方式	課堂講授
教材課本	Kevin J. Hastings (1997). *Probability and Statistics.* ISBN: 0201592789 Publisher: Addison-Wesley 進口商：滄海書局 Tel: 04-2708-8787 中山大學復文書局 Tel: (07) 525-0930
教學目標	熟悉機率空間、隨機變數、常用離散機率分配、連續機率分配、條件分佈、獨立性、隨機變數函數變換
授課方式	教師授課爲主
評分標準	五次作業(每章繳一次): 15%; 五次考試(每章考一次)：75%; 報告: 10%
教學內容	Chap. 1: Sample Spaces and Random Variables Chap. 2: Discrete Probability Chap. 3: Continuous Probability Chap. 4: Conditional Distributions and Independence Chap. 5: Transformations of Random Variables Chap. 6: Asymptotic Theory

教學設備	硬體：電腦、單槍投影機 軟體：Yap, Acrobat Reader

作業

第一章	*1.2.7, 1.2.18, 1.3.6, 1.3.8, 1.4.1, 1.4.9, 1.4.15, 1.5.2, 1.5.6, 1.5.12*
第二章	*2.1.9, 2.1.20, 2.2.15, 2.2.17, 2.3.7, 2.3.12, 2.4.3, 2.4.11, 2.5.10, 2.5.22*
第三章	*3.1.10, 3.1.13, 3.1.16, 3.2.4, 3.2.12, 3.3.13, 3.3.21, 3.4.7, 3.4.12, 3.4.13*
第四章	*4.1.10, 4.1.14, 4.2.6, 4.2.13, 4.2.18, 4.3.6, 4.3.10, 4.3.21, 4.4.3, 4.4.7*
第五章	*5.1.3, 5.1.10, 5.2.2, 5.2.10, 5.3.3, 5.3.12, 5.4.7, 5.4.13, 5.5.4, 5.5.8*
第六章	*6.1.1, 6.1.12, 6.1.17, 6.2.5, 6.2.9*

哥林多前書10:23　凡事都可行，但不都有益處；凡事都可行，但不都建造人。

1 Corinthians 10:23　All things are lawful, but not all things are profitable; all things are lawful, but not all things build up.

有用網頁

1. ***Probability Web***
   The Probability Web is a collection of probability resources on the World Wide Web (WWW). The pages are designed to be especially helpful to researchers, teachers, and people in the probability community.
   `http://www.mathcs.carleton.edu/probweb/probweb.html`

2. ***HyperStat Online***
   This privately maintained site offers an introductory-level hypertext statistics book, suitable for college-level students.
   `http://davidmlane.com/hyperstat/index.html`

3. ***Encyclopedia of Statistical Sciences***
   The online home of the Encyclopedia of Statistical Sciences, the most indispensable reference for statistical content.
   `http://www.mrw.interscience.wiley.com/emrw/9780471667193/home/`

4. ***Statistical Computing (UCLA Academic Technology Services)***
   `http://www.ats.ucla.edu/stat/`

Part I

Lecture Notes

Chapter 1

Sample Spaces and Random Variables

Contents

1.1 Introduction and examples

Theorem 1.1.1 (*Random*) *Pertain to an experiment whose result remains uncertain until the experiment is performed or phenomenon is observed.*

1. Will I or won't I be interrupted by a phone call while I am writing today?
2. How much will IBM common stock go up today, Jan. 13, 1992, on the New York Stock Exchange?
3. What will be the winning three-digit number today, Jan. 13, 1992, on the Michigan's Daily-3 Lottery game?
4. What will be the official high temperature in Atlantic City today?

Question 1.1.1 *Think of at least three other random phenomena. What are their possible outcomes?*

Ans:

1. What is the outcome of tossing a coin? Head or tail.

2. Will I catch a cold in this year? Yes or no.
3. How many traffic tickets will I receive this year? $0, 1, \ldots$ □

Theorem 1.1.2 (*Sample space*) *The set of all possible outcomes of the experiment.*

Question 1.1.2 *Write a formal set-theoretic description of the sample space for the Michigan Daily-3 Lottery game described earlier.*

Ans: $\{(x_1, x_2, x_3) | x_1, x_2, x_3 \in 0, 1, \ldots, 9\}$ □

Theorem 1.1.3 (*Event*) *Subset of the sample space.*

Question 1.1.3 *For each of your phenomena from 1.1.1, give an example of an event.*

Ans:

1. {Head}
2. {I will catch a cold this year}
3. {I will receive more than 1 ticket this year} □

Theorem 1.1.4 (*Probability*) *A measure of the likelihood of events.*

Finite sample space models vs. uncountable infinite models: A runner in the 100-meter dash will not always complete the run in the same amount of time.

1. If the stopwatch that we use to time the run is accurate only to the nearest 1/10 of a second, and our runner cannot run a faster time than 10 seconds or a slower time than 12 seconds, then the outcomes that form the sample space are
$$\{10.0, 10.1, \ldots, 12.0\}.$$
2. If we had a perfect stopwatch, we could measure the runner's time to infinite accuracy, then the sample space $[10, 12]$.

Question 1.1.4 *Think carefully about the next before moving on. How would you sensibly assign probabilities to intervals in the second model?*

Ans:

1. The probability of a region is the ratio of its length to the total length of the interval.
2. For instance, the event $[11, 12]$, the length of the sample space $[10, 12]$, would be given probability $\frac{1}{2}$. □

Three ways to assign probabilities to an event:

1. Equally likely.
2. Base on judgment and past experience.

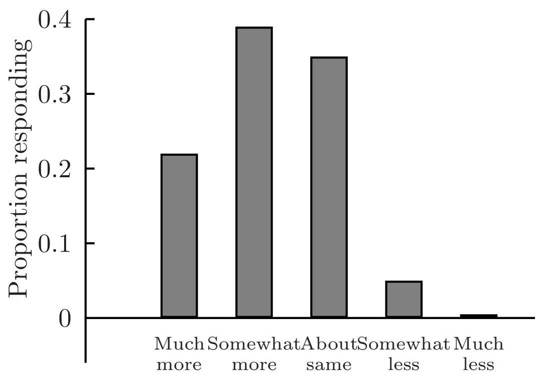


Figure 1.1: ***Histogram of student responses***

3. Empirical probability (histogram, Figure 1.1).

Theorem 1.1.5 (***Random variable***) *A function gives a numerical value to each outcome of a random experiment.*

Question 1.1.5 *Consider the random phenomena that you constructed in 1.1.1. Define at least one random variable relating to each example?*

Ans:

1. Let $X = 1$ and 0 denote the outcome head and tail, respectively.
2. Let $X = 1$ and 0 denote the outcome "catch a cold" and "does not catch a cold", respectively. □

Six terms are introduced above: ***Experiment; Outcome; Sample space; Event; Probability; Random variable.***

1.2 Axioms of probability and their consequences

Definition 1.2.1 (***Sample space***) *The sample space Ω of a random experiment is the collection of all possible outcomes. An event is a subset of the sample space, i.e., a set of outcomes.*

Definition 1.2.2 (***Axioms of probability***) *A probability measure on a sample space Ω of a random experiment is a function $P[\cdot]$ that maps events in Ω to real numbers such that*

1. $P[\Omega] = 1$,
2. $P[F] \geq 0$ *for all events* F,
3. $P[E_1 \cup E_2 \cup \cdots] = P[E_1] + P[E_2] + \cdots$ *where E_i are disjoint events.*

Example 1.2.1 (***Finitely many, equally likely outcomes***) *Two U.S. Senators among four–Bradley, Glenn, Moynihan, and Simon are to appointed as a committee by the Democratic leadership to lobby their republican colleagues for some bill.*

Ans:

- Sample space:

$$\Omega = \{\{B, G\}, \{B, M\}, \{B, S\}, \{G, M\}, \{G, S\}, \{M, S\}\}$$

- An event E: $\{B, G\}, \{B, M\}$, and $\{B, S\}$
- Probability of E is $1/2$. □

Example 1.2.2 (**Countable infinite sample space, unequally likely outcomes**) *Flip a coin indefinitely, until the first tails comes up.*

Ans:

- $\Omega = \{T, HT, HHT, HHHT, \ldots\}$.
- $P[\underbrace{H \cdots H}T] = (1/2)^{n+1}$. □

Example 1.2.3 (**Uncountable sample space**) *A local pizza parlor promises that it will deliver orders in a half hour and, if the delivery is more than 5 min. late, the pizza will be free. Past data indicate that all deliveries are between 10 mins. early and 10 mins. late.*

Ans:

- $\Omega = [-1, 1]$; Time unit is 10 mins.
- Figure 1.2 shows the pizza lateness distribution.
- Fit the density by a parabola with the form

$$f(x) = -ax^2 + b,$$

  then $a = b = 3/4$.
- Three axioms of probability hold:
  - $P[\Omega] = \int_{-1}^{1} f(x)\,\mathrm{d}x = \int_{-1}^{1} \frac{3}{4} - \frac{3x^2}{4}\,\mathrm{d}x = 1$
  - $P[[.5, 1]] = \int_{.5}^{1} \frac{3}{4} - \frac{3x^2}{4}\,\mathrm{d}x = 5/32$
  - $P[\{.2\}] = 0$ □

Question 1.2.2 *What integral represents the probability that the pizza is at least 5 mins. early? no more than 2 mins. late?*

Ans:

- $P[[-1, -0.5]] = \int_{-1}^{-1/2} \frac{3}{4} - \frac{3x^2}{4}\,\mathrm{d}x = 5/32$
- $P[[-1, 0.2]] = \int_{-1}^{1/5} \frac{3}{4} - \frac{3x^2}{4}\,\mathrm{d}x = 56/125$ □

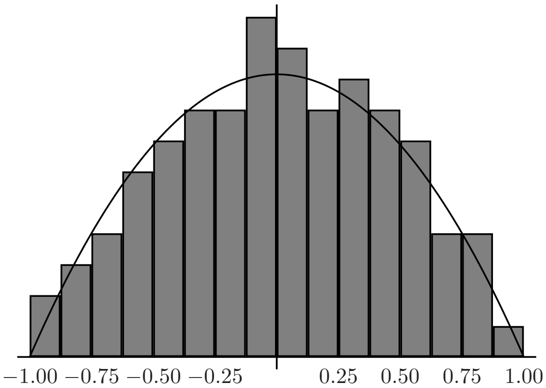


Figure 1.2: ***Pizza lateness distribution***

Proposition 1.2.1 (*Probability of empty set*) $P[\emptyset] = 0$.

Proof:

- Observe first that $\emptyset \cup \Omega = \Omega$, and $\emptyset \cap \Omega = \emptyset$.
- Therefore, by axioms (a) and (c) of probability,

$$1 = P[\Omega] = P[\Omega \cup \emptyset] = P[\Omega] + P[\emptyset] = 1 + P[\emptyset]$$

$$\Rightarrow 1 = 1 + P[\emptyset].$$

- Thus the empty set has zero probability. □

Proposition 1.2.2 *For any event E, denote the complement of E by E^c. Then $P[E] + P[E^c] = 1$.*

Proof:

- The events E and E^c are disjoint, and their union makes up the whole sample space.
- Therefore, by axioms (a) and (c),

$$1 = P[\Omega] = P[E \cup E^c] = P[E] + P[E^c].$$ □

Proposition 1.2.3 *For any event E, $P[E] \leq 1$.*

Proof:

- By axiom (b), $P[E^c] \geq 0$. Thus, by Proposition 1.2.2,

$$P[E] = 1 - P[E^c] \leq 1.$$ □

Proposition 1.2.4 *For any event A and B,*

$$P[A] = P[A \cap B] + P[A \cap B^c].$$

Proof:

- It is easy to verify the decomposition

$$A = (A \cap B) \cup (A \cap B^c).$$

- The sets $A \cap B$ and $A \cap B^c$ are clearly pairwise disjoint, since the B and B^c are. The result follows from the additivity axiom. □

Proposition 1.2.5 *If $E \subseteq F$ are event, then $P[E] \leq P[F]$.*

Proof:

- As Figure 1.3 indicates, we can decompose F into two disjoint subsets as

$$F = E \cup (F \cap E^c).$$

- Then by axioms (c) and (b),

$$P[F] = P[E] + P[F \cap E^c] \geq P[E].$$ □

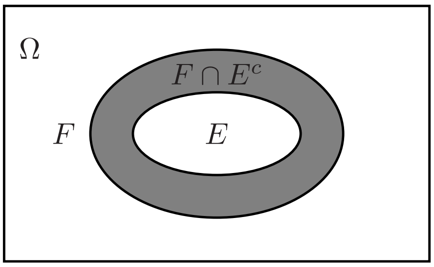


Figure 1.3: ***Subset rule***

Proposition 1.2.6 *If A and B are events, then*

$$P[A \cup B] = P[A] + P[B] - P[A \cap B].$$

Proof:

- The event $A \cup B$ can be decomposed into disjoint subsets as $A \cup B = B \cup (A \cap B^c)$.
- Hence, by axiom (c),

$$P[A \cup B] = P[B] + P[A \cap B^c]. \tag{1.6}$$

- But it is also clear that the two sets $A \cap B$ and $A \cap B^c$ are disjoint and that their union is A; hence by Axiom (c) again,

$$P[A \cap B^c] = P[A] - P[A \cap B]. \tag{1.7}$$

- Combining (1.6) and (1.7) finishes the proof. □

Principle of inclusion-exclusion If $A_i, i = 1, 2, \dots, n$, are events, then

$$P\left[\bigcup_{i=1}^{n} A_i\right] = \sum_{k=1}^{n}(-1)^{k-1} \sum_{1\le i_1<i_2<\cdots<i_k\le n} P[A_{i_1}A_{i_2}\cdots A_{i_k}].$$

1.3 Random variables and distributions

1.3.1 Random variables

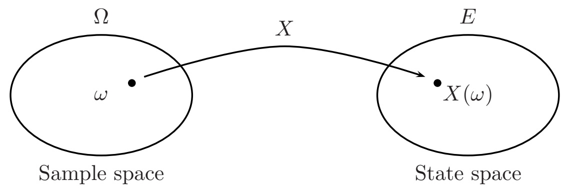


Figure 1.4: ***Action of a random variable***

Definition 1.3.1 (Random variable) *Let Ω be a sample space, and let E be a subset of $\mathbb{R}^n$. A* ***random variable*** *X is a function from Ω into E. We call E the* ***state space*** *of the random variable X. (Figure 1.4)*

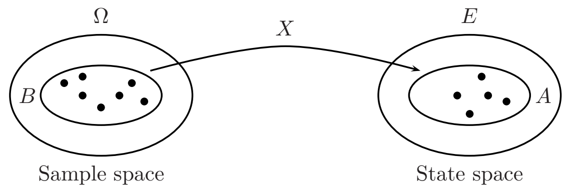


Figure 1.5: ***Random variable taking values in set A***

Example 1.3.1 *In the example of the bank waiting line, the sample space might consist of outcomes of the form*

$$\omega = (s_1, s_2, \dots, s_n),$$

where s_k is the length of the line after the kth arrival or departure.

Ans:

- $(1, 2, 3, 2, 3, 4)$
- The maximum line size random variable $X(\omega) = \max\{s_1, s_2, \dots, s_n\}$.
- $P[X < 3]$ is the probability that the line size was never as large as 3 people. □

Example 1.3.2 *In the bridge example,*

Ans:

- a typical outcome is $\Omega = \{2, 3, 4,$ and 7 of spades; 5, jack, and ace of hearts; $6, 8, 9,$ and king of clubs; 3 and 7 of diamonds$\}$.
- $Y = (Y_1, Y_2, Y_3, Y_4)$ which counts the numbers of each suit in the hand.
- $Y(\omega) = (4, 3, 4, 2)$. $\{Y_1 \geq 5\}$ the event that there at least 5 spades in the hand. □

Question 1.3.1 *Give another example of a random variable, paying attention to the sample space on which it is defined.*

Ans:

- Give a special experiment, there are two outcomes: success and fail.
- Let $\Omega = \{\text{success , fail}\}$, and random variable $X(\text{success}) = 1$, $X(\text{fail}) = 0$. □

1.3.2 Probability distributions

Definition 1.3.2 (*Probability distribution*) *The* ***probability distribution*** *Q of a random variable X is the probability measure on E defined by*

$$Q(A) = P[X \in A], \qquad \text{for } A \subseteq E.$$

Question 1.3.2 *Argue that Q as defined above satisfies axioms (a) and (b) of probability, that is, $Q(E) = 1$, and if $A \subseteq E$, then $Q(A) \geq 0$.*

Ans:

- (a) $Q(E) = P[X \in E] = P[\Omega] = 1$
- (b) $Q(A) = P[X \in A] = P[F] \geq 0$ since F is an event of Ω. □

Example 1.3.3 *Let two six-sided dice be rolled in succession.*

Ans:

- The sample space is

  $$\Omega = \{(1, 1), \ldots, (1, 6), (2, 1), \ldots, (2, 6), \ldots, (6, 6)\}$$

  where each outcome is an ordered pair indicating the results of the first and second rolls, respectively.
- $X(\omega) = X(\omega_1, \omega_2) = \omega_1 + \omega_2$ □

Question 1.3.3 *Fill in the rest of Table 1.2.*

Ans:

Outcomes in Ω	...Mapped by X to ...	Points $x \in E$	$\Omega(\{x\})$
$(1,1)$	$\mapsto$	2	1/36
$(1,2),(2,1)$	$\mapsto$	3	2/36
$(1,3),(2,2),(3,1)$	$\mapsto$	4	3/36
$(1,4),(2,3),(3,2),(4,1)$	$\mapsto$	5	4/36
$(1,5),(2,4),(3,3),(4,2),(5,1)$	$\mapsto$	6	5/36
$(1,6),(2,5),(3,4),(4,3),(5,2),(6,1)$	$\mapsto$	7	6/36
$(2,6),(3,5),(4,4),(5,3),(6,2)$	$\mapsto$	8	5/36
$(3,6),(4,5),(5,4),(6,3)$	$\mapsto$	9	4/36
$(4,6),(5,5),(6,4)$	$\mapsto$	10	3/36
$(5,6),(6,5)$	$\mapsto$	11	2/36
$(6,6)$	$\mapsto$	12	1/36

□

Example 1.3.4 *Consider again Example 1.2.3 on pizza delivery.*

Ans:

- Define X to be the amount of time, in units of 10 minutes, that the pizza is late.
- By the way that we defined the sample space, the value that X gives to an outcome is just the outcome itself – $X(\omega) = \omega$ – which is the lateness of delivery.
- We can then let $E = \Omega = [-1, 1]$.
- The probability distribution of X is

$$\begin{aligned} Q(A) &= P[\{\omega | X(\omega) \in A\}] \\ &= P[\{\omega | \omega \in A\}] \\ &= P[A] \\ &= \int_A \left(\frac{3}{4} - \frac{3}{4}x^2\right) \mathrm{d}x \end{aligned}$$

- For instance,

$$Q([0,1]) = \int_0^1 \left(\frac{3}{4} - \frac{3}{4}x^2\right) \mathrm{d}x = \frac{1}{2}$$

□

1.3.3 Cumulative distribution functions, mass functions, and density functions

Definition 1.3.3 (***Probability mass function***) *Suppose that a random variable X has a discrete (i.e., finite or countable) state space. The function $q : E \to [0, 1]$ is called the* ***probability mass function (p.m.f.)*** *of X if*

$$q(x) = Q(\{x\}) = P[X = x].$$

To define a valid probability distribution, the p.m.f. must satisfy

$$q(x) \geq 0 \quad \forall x \in E \quad \textit{and} \quad \sum_{x \in E} q(x) = 1.$$

Example 1.3.5 (Geometric distribution) *From the coin flipping example (Eg. 1.2.2), if X is the number of flips required to achieve the first tail,*

Ans: The p.m.f. of X is

$$q(x) = P[X = x] = \left(\frac{1}{2}\right)^x, \qquad x = 1, 2, 3, \dots$$

□

Definition 1.3.4 (***Probability density function***) *A random variable X is said to have* ***probability density function (p.d.f.)*** *f if, for all subsets A of the state space,*

$$Q(A) = P[X \in A] = \int_A f(x)\,\mathrm{d}x.$$

To be a valid p.d.f., the function f must satisfy

$$f(x) \geq 0 \quad \forall x \in E \quad \textit{and} \quad \int_E f(x)\,\mathrm{d}x = 1.$$

Definition 1.3.5 (***Cumulative distribution function***) *The* ***cumulative distribution function*** *(c.d.f.) is defined as*

$$F(x) = P[X \leq x] = \int_{-\infty}^{x} f(t)\,\mathrm{d}t$$

where $f(x) = F'(x)$.

Example 1.3.6 (***Uniform distribution***) *A plane has an estimated time of arrival of noon, plus or minus 15 minutes, and within this interval no time should be given preference over any other time. By the latter we mean that the density of probability is the same for each possible time. If we label noon as time 0 and measure time in minutes, then*

Ans:

- State space is $[-15, 15]$.
- $1 = \int_{-15}^{15} c\,\mathrm{d}x = 30c$, implies $c = 1/30$.
- $P[X > 5] = \int_5^{15} c\,\mathrm{d}x = \frac{1}{3}$.

- $F(x) = \int_{-15}^{x} c\,\mathrm{d}t = \frac{1}{30}(x+15), \quad x \in [-15, 15]$ □

Question 1.3.4 *Use the c.d.f. in the last example to find easily the probability that the arrival time is between 11:55 and 12:05.*

Ans: $\int_{-5}^{5} c\,\mathrm{d}x = F(5) - F(-5) = (5-(-5))/30 = 1/3$ □

1.4 Conditional probability

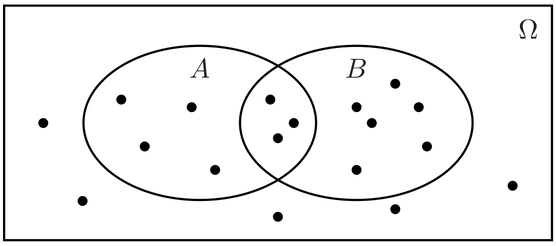


Figure 1.6: ***Reduced sample space in event*** B

Question 1.4.1 *What is the conditional probability of B given A in Figure 1.6?*

Ans:

- # of points in $A = 7$; # of points in $A \cap B = 3$
- $P[B|A] = P[A \cap B]/P[A] = \frac{3}{7}$ □

This discussion suggests the following definition.

Definition 1.4.1 (*Conditional probability*) *If B is an event such that $P[B] > 0$, then the **conditional probability** of an event A given B is defined as*

$$P[A|B] = \frac{P[A \cap B]}{P[B]}.$$

Example 1.4.1 *Grade distribution (Figure 1.7). Suppose that all the institutions listed have 2000 undergraduate.*

		Grade					
		A	B	C	D	F	Other
University	W	26.5	36.6	17.5	3.6	1.6	14.2
	X	23.9	44.9	18.9	4.3	1.2	6.8
	Y	19.9	45.8	23.1	4.1	.9	6.2
	Z	19.4	45.2	25.2	4.8	1.1	4.3

Figure 1.7: ***Grade distribution***

Ans:

- If a student is selected at random from among the 8000 students included in this table, the probability that the student received a C is

$$P[C] = \frac{2000(.175 + .189 + .231 + .252)}{8000} \approx .21$$

-

$$\begin{aligned} P[Y|C] &= \frac{P(Y \cap C)}{P(C)} \\ &= \frac{2000(.231)}{2000(.175 + .189 + .231 + .252)} \approx .27 \end{aligned}$$

- Note that $P[C|Y]$ and $P[Y|C]$ are not the same. □

Example 1.4.2 (*Blackjack*) *Blackjack is a card game in which several players are each dealt one card face down and another card face up from an ordinary 52-card deck. The point values of cards of rank 2-10 are their ranks, face cards have point value 10, and aces can have either value 1 or 11 at the option of the player. Each player is offered the chance of drawing more cards, one at a time, with the object of coming the closest among players to a total of 21 points without going over 21. At a casino blackjack table the object is simply to beat the dealer, who plays for the house.*

Ans:

- Suppose you are player 1 among four players at a blackjack table.
- The first card you receive: $P[\text{king}] = 4/52 = 1/13$.
- $P[\text{king}|\text{heart}] = P[\text{king and heart}]/P[\text{heart}] = \frac{1/52}{13/52} = 1/13$
- This shows that knows that the card is heart does not change the probability that the card is king.
- This is the property that will be calling ***independence*** in the next section.
- Next, suppose that cards have been dealt by player 4 as in Figure 1.8.
- You have 15 and the dealer could have 18.
- Based only on what you know about your own cards, The conditional probability of beating 18 is 11/50 (4s, 5s and 6s).
- Of course, we know three other up cards, none of which is a $4, 5, 6$, so the conditional probability of beating 18 given the full information is 11/47. □

Question 1.4.2 *Given the configuration of cards in Figure 1.8, what is the probability that the dealer has a blackjack (a hand consisting of an ace and a face card)?*

Ans: 0 □

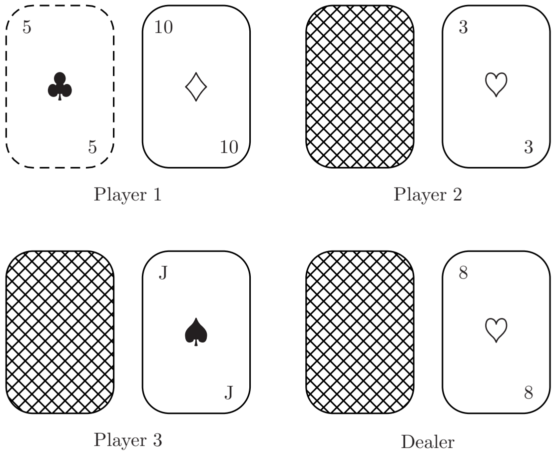


Figure 1.8: ***Blackjack hands***

Example 1.4.3 (***Gamma distribution***) *Suppose that the lifetime X of a lightbulb (in thousands of hours) is a random variable whose distribution is characterized by the probability density function $f(x) = xe^{-x}, x > 0$, then*

Ans:

$$\begin{aligned} P[X > 2|X > 1.5] &= \frac{P[X > 2 \cap X > 1.5]}{P[X > 1.5]} \\ &= \frac{P[X > 2]}{P[X > 1.5]} \\ &= \frac{\int_2^\infty xe^{-x}\,dx}{\int_{1.5}^\infty xe^{-x}\,dx} \\ &= \frac{3e^{-2}}{2.5e^{-1.5}} \approx .728. \end{aligned}$$

□

Alternative definition of conditional probability:

$$P[A \cap B] = P[B] \cdot P[A|B] \tag{1.20}$$

In this form it is usually referred to as the ***multiplication rule*** for conditional probability.

Proposition 1.4.1 (***Generalized multiplication rule***) *If $A_1, \ldots, A_n$ are events such that the following conditional probabilities are defined, then*

$$P[A_1 \cap A_2 \cap \cdots \cap A_n] = P[A_1] \cdot P[A_2|A_1] \cdots P[A_n|A_1 \cap \cdots \cap A_{n-1}].$$

Proof:

- Formula (1.20) is the case $n = 2$.
- You may care to do a general inductive argument using this as the anchoring step.
- Here we will only do the case $n = 3$ as an illustration.
- Applying formula (1.20) to the two events $(A_1 \cap A_2)$ and A_3, we obtain

$$P[A_1 \cap A_2 \cap A_3] = P[(A_1 \cap A_2) \cap A_3] = P[A_1 \cap A_2] \cdot P[A_3|A_1 \cap A_2]. \tag{1.22}$$

- By formula (1.20) again, $P[A_1 \cap A_2] = P[A_1] \cdot P[A_2|A_1]$.
- Substituting this into the right side of formula (1.22) finishes the proof for $n = 3$. □

Example 1.4.4 *Shakespeare wrote 10 tragedies, 17 comedies, and 10 histories. If Professor Brady of the English department puts together a collection of readings from this group of works by randomly picking four of them in sequence without repetition, what is the probability that the first two readings are comedies and the next two are tragedies?*

Ans:

- Let

$$\begin{aligned} A_1 &= \text{"comedy on first"} \\ A_2 &= \text{"comedy on second"} \\ A_3 &= \text{"tragedy on third"} \\ A_4 &= \text{"tragedy on fourth"} \end{aligned}$$

-

$$\begin{aligned} P[A_1 \cap \cdots \cap A_4] &= P[A_1] \cdot P[A_2|A_1] \cdot P[A_3|A_1 \cap A_2] \cdot P[A_4|A_1 \cap A_2 \cap A_3] \\ &= \frac{17}{37} \cdot \frac{16}{36} \cdot \frac{10}{35} \cdot \frac{9}{34} = .015 \end{aligned}$$

□

Powerful computational tool for finding the probability of an event when we know the conditional probability that the event will occur given the occurrence of other events (Figure 1.9).

Proposition 1.4.2 (*Law of total probability*) *Let A be an event, and let $B_1, \ldots, B_n$ be mutually exclusive events of nonzero probability whose union is the sample space Ω. Then*

$$P[A] = P[B_1] \cdot P[A|B_1] + P[B_2] \cdot P[A|B_2] + \cdots + P[B_n] \cdot P[A|B_n].$$

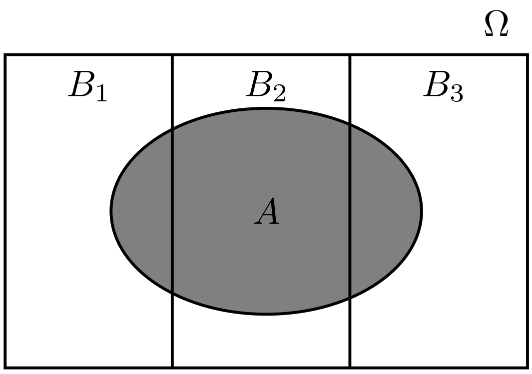
Proof:

- By the assumed conditions on the events B_i, the events $A \cap B_i$ are pairwise disjoint, and their union is the event A.
- Therefore

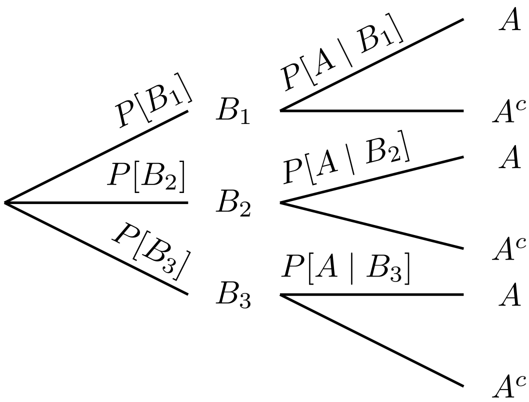
$$\begin{aligned} P[A] &= P[(A \cap B_1) \cup (A \cap B_2) \cup \cdots \cup (A \cap B_n)] \\ &= P[A \cap B_1] + P[A \cap B_2] + \cdots + P[A \cap B_n] \\ &= P[B_1] \cdot P[A|B_1] + P[B_2] \cdot P[A|B_2] + \cdots + P[B_n] \cdot P[A|B_n]. \end{aligned}$$

- The last line comes from the multiplication rule. □

One of the most powerful ways to use the Law of Total Probability is in the phenomena elapsing over time (eg. queueing theory).



(a)



(b)

Figure 1.9: ***Conditioning and unconditioning***

> **Example 1.4.5** *Consider the busy styling salon displayed schematically in Figure 1.10(a). A single stylist is working, and customers arrive at random times seeking service. If the stylist is currently busy, the customer waits in a first-come, first-served line. The stylist requires a random amount of time to serve each customer, and then the customer departs, leaving the stylist free to work on the next customer, if one is waiting. Tree diagram is in Figure 1.10(b).*

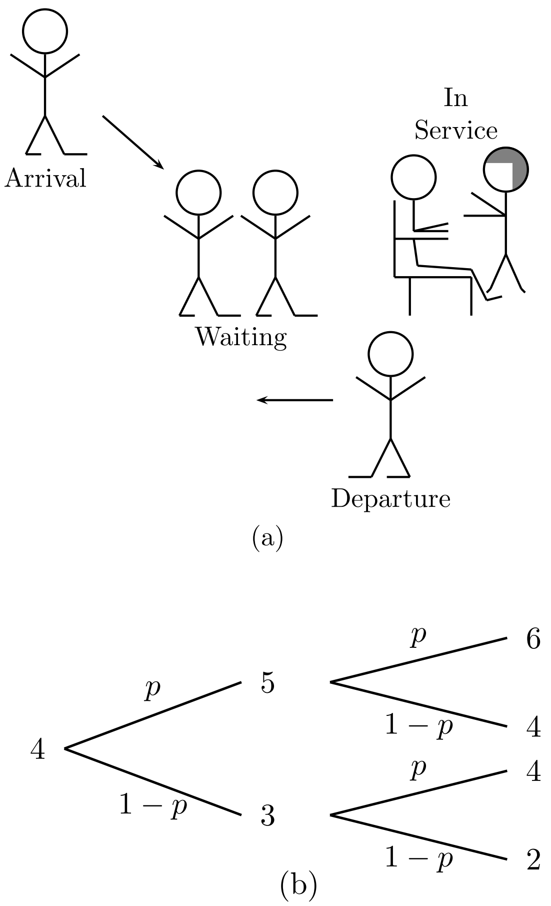


Figure 1.10: ***Queueing system***

Ans:

- Initial system size is 4.
- X_i denotes the system size at time i.
- $X_0 = 4$ and $P[X_1 = 5] = p, P[X_1 = 3] = 1 - p$.
- $P[X_2 = 6] = P[X_2 = 6|X_1 = 5] \cdot P[X_1 = 5] + P[X_2 = 6|X_1 = 3] \cdot P[X_1 = 3] = p \cdot p + 0 \cdot (1 - p) = p^2$
- $P[X_2 = 4] = P[X_2 = 4|X_1 = 5] \cdot P[X_1 = 5] + P[X_2 = 4|X_1 = 3] \cdot P[X_1 = 3] = (1 - p)p + p(1 - p) = 2p(1 - p)$
- $P[X_2 = 2] = P[X_2 = 2|X_1 = 5] \cdot P[X_1 = 5] + P[X_2 = 2|X_1 = 3] \cdot P[X_1 = 3] = 0 \cdot p + (1 - p)(1 - p) = (1 - p)^2$ □

Question 1.4.3 *The computation of the probability distribution of X_2 was done as if the probabilities involved were not dependent on the initial state X_0 of the system. To be more rigorous, we should write conditional probabilities given $X_0 = 4$. Would the computation change?*

Ans: For example, let $P[X_0 = 4] = q$,

$$P[X_2 = 6|X_0 = 4] = \frac{P[X_2 = 6 \cap X_0 = 4]}{P[X_0 = 4]} = \frac{p^2 q}{q} = p^2$$
$$P[X_2 = 6] = P[X_2 = 6|X_0 = 4] = p^2$$

So the computation does not change. □

Proposition 1.4.3 (*Bayes' Theorem*) *Let the sets A and $B_i, i = 1, 2, \ldots, n$ satisfy the hypothesis of Proposition 1.4.2. Then for each $i = 1, 2, \ldots, n$, $P[B_i|A] = \frac{P[A|B_i] \cdot P[B_i]}{P[A|B_1] \cdot P[B_1] + \cdots + P[A|B_n] \cdot P[B_n]}$.*

Proof:

- By the definition of conditional probability,

$$P[B_i|A] = \frac{P[A \cap B_i]}{P[A]}.$$

  Also,

$$P[A \cap B_i] = P[A|B_i] \cdot P[B_i],$$

  by the multiplication rule.

- Substituting this result into the numerator of the first equation gives

$$P[B_i|A] = \frac{P[A|B_i] \cdot P[B_i]}{P[A]}.$$

- The proof is completed by substituting the law of total probability formula (1.22) into the denominator for $P[A]$. □

Example 1.4.6 *Intradermal (真皮內的) skin test for tuberculosis is subject to error. If the injection is too deep, the welt that indicates a positive test may not appear even in a patient with disease. On the other hand, a healed tuberculosis or an infection by a different type of bacteria may cause a positive result on a patient without an active case of disease.*

Ans:

- B_1 = the subject has tuberculosis.
- B_2 = the subject does not have tuberculosis.
- A = the skin test is positive.
- $P[A|B_1] = .90$ and $P[A|B_2] = .10$
- $P[B_1] = .01$ and $P[B_2] = .99$

- Then $P[B_1|A] = \frac{P[A|B_1]\cdot P[B_1]}{P[A|B_1]\cdot P[B_1]+P[A|B_2]\cdot P[B_2]} = .083.$ □

1.5 Independent events

Two events are independent if the conditional probability that the first event occurs given the second is the same as the unconditional probability of the first event.

Definition 1.5.1 (*Independent events*) *Events A and B are said to be **independent** of one another if*

$$P[A|B] = P[A],$$

provided $P[B] > 0$.

Question 1.5.1 *Are disjoint events independent?*

Ans: No! $P[A|B] = \frac{P[A\cap B]}{P[B]} = \frac{0}{P[B]} = 0 \neq P[A]$ □

Example 1.5.1 *For convenience, we reproduce Figure 1.6 below as Figure 1.11. If the outcomes in the sample space shown are equally likely, are the events A and B independent?*

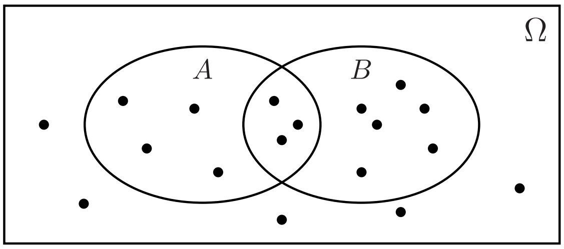


Figure 1.11: ***Non-independent sets A and B***

Ans:

- $P[A] = 7/18,\ P[B] = 9/18$ and $P[A \cap B] = 3/18$.
- $P[A|B] = 3/9 \neq P[A]$ □

Example 1.5.2 *Figure 1.12 displays the state space $E = [0,2]\times[0,2]$ of a random vector $\mathbf{X} = (X_1, X_2)$ that represents the times required for an oil change and a filter change at a quick lubrication station. Let the probability distribution of $\mathbf{X}$ be characterized by $P[\mathbf{X} \in A] =$ (area of A)/(area of E). Show that if A_1 and A_2 are time intervals that are subsets of $[0,2]$, the event that the oil change time belongs to A_1 is independent of the event that the filter change time belongs to A_2.*

Ans:

- $P[X_2 \in A_2|X_1 \in A_1] = \frac{(\text{length } A_1)(\text{length } A_2)/4}{(\text{length } A_1)\cdot 2/4} = \frac{\text{length } A_2}{2} = P[X_2 \in A_2]$ □

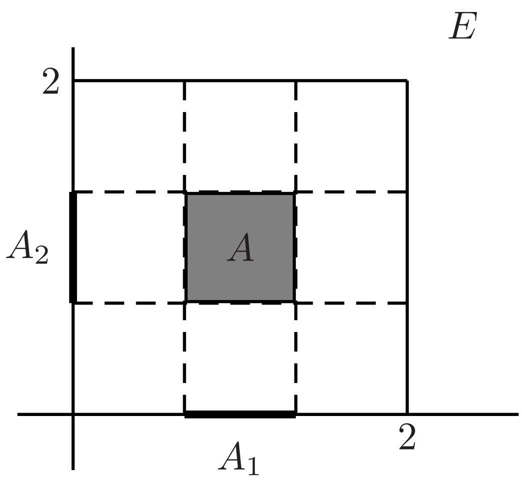


Figure 1.12: ***State space of X***

Question 1.5.2 *In the last example, what is the probability density function of* $\mathbf{X}$*?*

Ans: $f(x_1, x_2) = 1/4$ if $(x_1, x_2) \in E$; 0 otherwise. □

Proposition 1.5.1 *Let A and B be events of positive probability. Then A and B are independent of each other if and only if*

$$P[A \cap B] = P[A] \cdot P[B].$$

Proof:

- Suppose that A and B are independent. Then by the multiplication rule,

$$P[A \cap B] = P[A|B]P[B] = P[A]P[B].$$

- Conversely, suppose that $P[A \cap B] = P[A]P[B]$. Then

$$P[A|B] = \frac{P[A \cap B]}{P[B]} = \frac{P[A]P[B]}{P[B]} = P[A].$$ □

Example 1.5.3 *A carton contains five GI Joe action figures. Mischievous pacifists (和平主義者) have replaced the combat boots on two of the figures with fuzzy pink bunny slippers. One figure is drawn from the carton at random, replaced, and then another figure is drawn at random. Let A be the event that the first figure drawn is wearing fuzzy pink bunny slippers, and let B be the event that the second figure drawn is wearing the slippers. Show that the two events are independent.*

Ans:

- $\Omega = \{(i,j)|i,j \in \{1,2,3,4,5\}\}$
- Joes 1 and 2 are the ones with the fuzzy pink bunny slippers.
- $P[A] = P[\{(i,j)|i \in \{1,2\},\ j \in \{1,2,3,4,5\}\}] = 10/25$

- $P[B] = P[\{(i,j)|i \in \{1,2,3,4,5\},\ j \in \{1,2\}\}] = 10/25$
- $P[A \cap B] = P[\{(i,j)|i,j \in \{1,2\}\}] = 4/25$
- $P[A \cap B] = P[A]P[B]$
- Principle of random sampling: ***If a random sample is drawn with replacement, then the results of successive samples are independent; but if sampling is done without replacement, then successive samples are not independent***. □

Example 1.5.4 *In Figure 1.13 is a 2×2 table of frequencies for 20 subjects classified into one of two categories for each of the two characteristics handedness and quality of language skills. Half the subjects are left-handed and half are right-handed. Twelve of the subjects are determined to have high language skills, and eight to have low skills. What must the table frequencies be in order that the event that a randomly selected subject is left-handed is independent of the event that the subject has high language skills? With those counts, is it also true that the event that a random subject is right-handed is independent of the event that the subject has low language skills?*

		Language	skills	
		High	Low	
Handedness	Left	m	$10-m$	10
	Right	$12-m$	$m-2$	10
		12	8	

Figure 1.13: ***Classification table for independence characteristics***

Ans:

- A_1 (A_2): Left- (right)-handedness
- B_1 (B_2): high (low) skill
- $\frac{m}{20} = P[A_1 \cap B_1] = P[A_1]P[B_1] = \frac{10}{20}\frac{12}{20} = \frac{6}{20}$
- $m = 6$
- $P[A_2 \cap B_2] = \frac{4}{20} = \frac{10}{20}\frac{8}{20} = P[A_2]P[B_2]$ □

Definition 1.5.2 (***Mutually independent events***) *Events $A_1, \ldots, A_n$ are said to be* ***mutually independent*** *if for any subcollection $A_{i_1}, \ldots, A_{i_k}, 1 \le i_1 < \cdots < i_k \le n$, of the events,*

$$P[A_{i_1} \cap A_{i_2} \cap \cdots \cap A_{i_k}] = P[A_{i_1}] \cdot P[A_{i_2}] \cdots P[A_{i_k}].$$

Example 1.5.5 This example is in the area called ***structural reliability***. A machine has two components, the first of which is held stable by three bolts (螺栓), the second by two. The bolts are redundant safety features, in the sense that the component stays stable if at least one of its supporting bolts stays tight. Suppose that each bolt fails (that is, it loosens during a certain fixed time of use) with probability p, and bolts fail independently of one another. We will compute the probability that the machine suffers instability.

Ans:

- Machine will be unstable if all of the component A bolts fails, or if both of the component B bolts fail (Figure 1.14).
- A_i and B_j denote the failure, $i = 1, 2, 3,\ j = 1, 2$.
- $P[\text{machine unstable}] = P[(A_1 \cap A_2 \cap A_3) \cup (B_1 \cap B_2)] = p^3 + p^2 - p^5$ by $P[A \cup B] = P[A] + P[B] - P[A \cap B]$. □

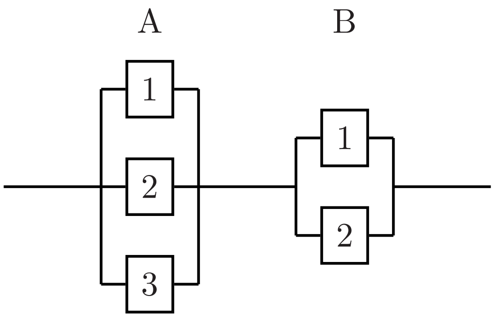


Figure 1.14: ***Reliability system***

Question 1.5.3 *How would the preceding formula for the probability of an unstable machine change if there were only two bolts for component A? What if component A had two bolts and component B had three bolts? What changes result to the original problem if the component A bolt failure probabilities are all p_1, and the component B failure probabilities are both p_2?*

Ans:

$$\begin{aligned} P[\text{machine unstable}] &= P[(A_1 \cap A_2) \cup (B_1 \cap B_2)] \\ &= p^2 + p^2 - p^4 \\ P[\text{machine unstable}] &= P[(A_1 \cap A_2) \cup (B_1 \cap B_2 \cap B_3)] \\ &= p^2 + p^3 - p^5 \\ P[\text{machine unstable}] &= P[(A_1 \cap A_2 \cap A_3) \cup (B_1 \cap B_2)] \\ &= p_1^3 + p_2^2 - p_1^3 \cdot p_2^2 \end{aligned}$$

□

1.6 Summary

1. ***Random***: Pertain to an experiment whose result remains uncertain until the experiment is performed or phenomenon is observed. 4
2. ***Sample space***: The set of all possible outcomes of the experiment. 5
3. ***Event***: Subset of the sample space. 6
4. ***Probability***: A measure of the likelihood of events. 6
5. ***Random variable***: A function gives a numerical value to each outcome of a random experiment. 8

6. ***Sample space***: The sample space Ω of a random experiment is the collection of all possible outcomes. 11

7. ***Axioms of probability***: A probability measure on a sample space Ω of a random experiment is a function $P[\cdot]$ that maps events in Ω to real numbers such that $P[\Omega] = 1$, $P[F] \geq 0$ for all events F, $P[E_1 \cup E_2 \cup \cdots] = P[E_1] + P[E_2] + \cdots$ where E_i are disjoint events. 11

8. ***Probability of empty set***: $P[\emptyset] = 0$. 16

9. For any event E, denote the complement of E by E^c. Then $P[E] + P[E^c] = 1$. . . . 17

10. For any event E, $P[E] \leq 1$. 17

11. For any event A and B, $P[A] = P[A \cap B] + P[A \cap B^c]$. 18

12. If $E \subseteq F$ are event, then $P[E] \leq P[F]$. 18

13. If A and B are events, then $P[A \cup B] = P[A] + P[B] - P[A \cap B]$. 20

14. ***Principle of inclusion-exclusion***: If $A_i, i = 1, 2, \ldots, n$, are events, then $P\left[\bigcup_{i=1}^n A_i\right] = \sum_{k=1}^n (-1)^{k-1} \sum_{1 \leq i_1 < i_2 < \cdots < i_k \leq n} P[A_{i_1} A_{i_2} \cdots A_{i_k}]$. 21

15. ***Random variable***: Let Ω be a sample space, and let E be a subset of $\mathbb{R}^n$. A ***random variable*** X is a function from Ω into E. 22

16. ***State space***: Let Ω be a sample space, and let E be a subset of $\mathbb{R}^n$. We call E the state space of the random variable X. 22

17. ***Probability distribution***: The probability distribution Q of a random variable X is the probability measure on E defined by $Q(A) = P[X \in A]$, for $A \subseteq E$. 25

18. ***Probability mass function (p.m.f.)***: Suppose that a random variable X has a discrete (i.e., finite or countable) state space. The function $q : E \to [0, 1]$ is called the probability mass function (p.m.f.) of X if $q(x) = Q(\{x\}) = P[X = x]$. 30

19. ***Probability density function (p.d.f.)***: A random variable X is said to have probability density function (p.d.f.) f if, for all subsets A of the state space, $Q(A) = P[X \in A] = \int_A f(x)\,dx$. 31

20. ***Cumulative distribution function (c.d.f.)***: The cumulative distribution function (c.d.f.) is defined as $F(x) = P[X \leq x] = \int_{-\infty}^{x} f(t)\,dt$ where $f(x) = F'(x)$. 32

21. ***Conditional probability***: If B is an event such that $P[B] > 0$, then the conditional probability of an event A given B is defined as $P[A|B] = \frac{P[A \cap B]}{P[B]}$. 35

22. ***Alternative definition of conditional probability***: $P[A \cap B] = P[B] \cdot P[A|B]$ 41

23. ***Generalized multiplication rule***: If $A_1, \ldots, A_n$ are events such that the following conditional probabilities are defined, then $P[A_1 \cap A_2 \cap \cdots \cap A_n] = P[A_1] \cdot P[A_2|A_1] \cdots P[A_n|A_1 \cap \cdots \cap A_{n-1}]$. 42

24. ***Law of Total Probability***: Let A be an event, and let $B_1, \ldots, B_n$ be mutually exclusive events of nonzero probability whose union is the sample space Ω. Then $P[A] = P[B_1] \cdot P[A|B_1] + P[B_2] \cdot P[A|B_2] + \cdots + P[B_n] \cdot P[A|B_n]$. 46

25. ***Bayes' Theorem***: Let the sets A and $B_i, i = 1, 2, \ldots, n$ satisfy the hypothesis of Proposition 1.4.2. Then for each $i = 1, 2, \ldots, n$, $P[B_i|A] = \frac{P[A|B_i] \cdot P[B_i]}{P[A|B_1] \cdot P[B_1] + \cdots + P[A|B_n] \cdot P[B_n]}$. 50

26. ***Independent events***: Events A and B are said to be independent of one another if $P[A|B] = P[A]$ provided $P[B] > 0$. 53

27. Let A and B be events of positive probability. Then A and B are independent of each other if and only if $P[A \cap B] = P[A] \cdot P[B]$. 57

28. ***Mutually independent events***: Events $A_1, \ldots, A_n$ are said to be ***mutually independent*** if for any subcollection $A_{i_1}, \ldots, A_{i_k}$, $1 \leq i_1 < \cdots < i_k \leq n$, of the events, $P[A_{i_1} \cap A_{i_2} \cap \cdots \cap A_{i_k}] = P[A_{i_1}] \cdot P[A_{i_2}] \cdots P[A_{i_k}]$. 62

Chapter 2

DISCRETE PROBABILITY

Contents

2.1 Combinatorial probability

2.1.1 Fundamental counting principle

► In many staged phenomena, outcomes are equally likely; hence the probability of an event A would be calculated as

$$P[A] = \frac{n(A)}{n(\Omega)}.$$

- An important class of examples is that in which objects are sampled successively, with or without replacement, from a finite population.

Question 2.1.1 *If an experiment has two stages, the first has seven possible outcomes, and the second has three, how many outcomes does the combination of both stages have?*

Ans: $7 \times 3 = 21$ ☐

Proposition 2.1.1 ***Fundamental counting principle***

(a) ***Rule of product****: Suppose that an experiment has two stages. For the first stage, there are m possible outcomes, and for each of these, the second stage has n possible outcomes. Then the two-stage experiment has $m \cdot n$ outcomes.*

(b) ***Rule of sum****: For a more general two-stage experiment, let the first-stage outcomes be labeled $i = 1, 2, \ldots, m$. Assume that if the first-stage outcome is i, then there are n_i possible outcomes for stage 2. Then the two-stage experiment has*

$$\sum_{i=1}^{m} n_i$$

possible outcomes.

(c) ***Rule of product****: Suppose that an experiment consists of k stages such that the first stage has m_1 possible outcomes, for each outcome of stage 1 there are m_2 possible outcomes of stage 2, for each combined outcome of the first two stages there are m_3 possible outcomes of stages 3, and so on. Then there are $m_1 \cdot m_2 \cdots m_k$ outcomes of the entire experiment.*

Proof:

- Since part (a) follows as a special case of part (b), let us prove the latter first.
- It will be helpful to refer to the tree diagram in Figure 2.1, in which a two-stage procedure is depicted that has two first-stage outcomes, one of which has three possible second-stage outcomes and the other has two.
- An outcome for the two-stage procedure is a path of two edges from the root to a leaf.
- In general, the two-stage experiment has a sample space of pairs (i, j), where i is the first-stage outcome and j is one of its n_i possible second-stage outcomes.
- Let A_i be the event that outcome i occurs on the first stage, for $i = 1, 2, \ldots, m$.
- Then the events A_i are disjoint and their union is all of Ω.
- Therefore
$$n(\Omega) = \sum_{i=1}^{m} n(A_i) = \sum_{i=1}^{m} n_i.$$
- Part (a) follows immediately by setting all $n_i = n$. We leave the proof of part (c) as Exercise 11. ☐

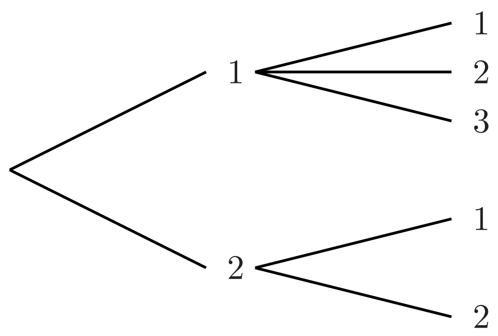


Figure 2.1: ***Fundamental counting principle***

Example 2.1.1 *Recall that in blackjack, a player is dealt one card face down and then a second card face up from an ordinary deck. What is the probability that a player is dealt a blackjack (an ace and a face card)?*

Ans:

$$P[\text{blackjack}] = P[\text{blackjack with ace down}] + P[\text{blackjack with ace up}]$$
$$= \frac{4 \cdot 12 + 12 \cdot 4}{52 \cdot 51} = \frac{8}{221} \qquad \square$$

Example 2.1.2 *A list of names of potential jurors contains 20 names beginning with each of the 6 letters A-F, 25 names beginning with each of the 14 letters G-T, and 10 names beginning with each of the 6 letters U-Z. A prosecuting attorney samples a name by first selecting a letter randomly, then sampling a name within that letter. How many outcomes are in the sample space? Is each person on the list equally likely to be chosen?*

Ans:

- $n(\Omega) = 6 \cdot 20 + 14 \cdot 25 + 6 \cdot 10 = 530$
- $P[A] = 20/530,\ P[G] = 25/530,\ P[U] = 10/530$
- The letters should be equally likely whereas people are not equally likely to be chosen. $\square$

Example 2.1.3 *Writing a* 16*-character string of* 0*s and* 1*s can be viewed as an experiment of* 16 *stages, in which each stage can result in either* 0 *or* 1*, unaffected by other stages. Part (c) of the fundamental counting principle implies that there are* $2^{16} = 65536$ *possible character strings of this type. (This number is equal to* $2^6 \cdot 2^{10} = 64 \cdot 2^{10}$*, often referred to in computer jargon as* $64K$*, where* $K = 2^{10}$ *is around* 1000*.)*

Ans:

- This example is important due to the fact that computer data are encoded as strings of 0s and 1s, often in blocks called *words* of length 16.
- Older computers encoded integers with one word, so that fewer than 33000 positive and 33000 negative integers could be represented.

- Similarly, there are $2^7 = 128$ different 0-1 strings of length 7. Text characters are often ended using 7 positions among an available 8 positions in a half-word (or *byte*).
- Thus 128 characters can be encoded this way, which is enough to include the upper- and lowercase alphabets, digits, punctuation marks, and numerous special symbols. □

Question 2.1.2 *In a 32-bit computer, integers are represented as strings of* 32 *0s and 1s. In view of the fact that one bit will be reserved to characterize the sign of the encoded integer, what range of integers can be so represented?*

Ans: $-2^{31} \sim (2^{31} - 1)$ □

Example 2.1.4 *A confidence man wants to perpetrate (犯罪) a fraudulent (欺騙的) activity in n cities in some order, with no city being visited more than once, for obvious reasons. How many possible routes among cities are there?*

Ans:

- $n!$
- For example, for 10 cities is $10! = 3,628,800$
- Traveling salesman problem in which distances are given between cities and salesman is looking for the shortest tour that hits each city exactly once. □

2.1.2 Permutations and combinations

Definition 2.1.1 *A* ***permutation*** *of n objects $\{y_1, \ldots, y_n\}$, taken r at a time is an ordered list $(x_1, \ldots, x_r)$ selected from the original n objects, such that $x_i \neq x_j, \forall i \neq j$. We denote the number of such permutations by $P_{n,r}$.*

$$P_{n,r} = n(n-1)\cdots(n-r+1) = \frac{n!}{(n-r)!}$$

Example 2.1.5 *If you were at a small party and, at an unfortunate lull (暫時呆滯), you began to wonder whether there were two people at the party with the same birthday, how many partygoers (社交聚會常客) would you expect to be necessary in order to have a good chance to have matching birthdays? (This is a famous problem whose answer will probably surprise you.)*

Ans:

$$\begin{aligned} P[\text{at least 1 pair of matching birthdays}] &= 1 - P[\text{nomatch}] \\ &= 1 - \frac{365 \cdot 364 \cdots (365 - n + 1)}{365^n} \end{aligned}$$

Birthday Matching Probabilities

Number of people	P[at least one match]
20	0.411

Number of people	P[at least one match]
21	0.444
22	0.476
23	0.507
24	0.538
25	0.569
26	0.598
27	0.627
28	0.654
29	0.681
30	0.706

□

Definition 2.1.2 *A* ***combination*** *of n items $\{y_1, \dots, y_n\}, r$ at a time, is a subset $\{x_1, \dots, x_r\}$ selected from the original n items, such that $x_i \neq x_j, \forall i \neq j$. We denote the number of such combinations by $C_{n,r}$, or $\binom{n}{r}$. The latter is read "n choose r."*

$$C_{n,r} = \binom{n}{r} = \frac{P_{n,r}}{r!} = \frac{n!}{(n-r)! \cdot r!} \qquad \text{(Binomial coefficient)}$$

Example 2.1.6 *If a 13-card bridge hand is drawn, what is the probability that there are 7 spades and 6 hearts in the hand?*

Ans:

- $n(\Omega) = \binom{52}{13}$
- $n(A) = \binom{13}{7}\binom{13}{6}$
- $P[A] = \frac{n(A)}{n(\Omega)} = .0000046$ □

Example 2.1.7 *A job discrimination study was performed of a company that had filled four vice-presidential positions by promoting from within. All of those eventually promoted were men. Investigators determined that the pool under consideration consisted of six men and five women, all of comparable qualifications. Would the fact that none of the promotions went to women be unusual if indeed promotions were not gender-biased?*

Ans:

- $P[A] = \frac{C_{6,4}}{C_{11,4}} = .045$ □

Three sampling situations: r samples are taken from a population $\{y_1, y_2, \dots, y_n\}$.

1. If sampling is done in order and ***with replacement*** (i.e., items sampled may be selected again), then the total number of samples is $n \cdot n \cdots n = n^r$.
2. If sampling is done in order and ***without replacement*** (i.e., items sampled may not be selected again), then the total number of samples is

$$P_{n,r} = n(n-1)\cdots(n-r+1) = \frac{n!}{(n-r)!}.$$

3. If sampling is done without regard to order and without replacement, then the total number of samples is
$$C_{n,r} = \binom{n}{r} = \frac{n!}{r! \cdot (n-r)!}.$$

2.1.3 Other combinatorial problems

1. Sampling from indistinguishable objects
2. Partitioning a group of objects into subsets
3. Distributing indistinguishable objects into a group of cells

Example 2.1.8 *Suppose that you are trying to decide whether a sequence of data that is being gathered as time passes is showing a downward trend. The following data, reported in Fisher and Lorie (1977, p.85), are annual rates of return between the years 1966 and 1976 on a common stock portfolio purchased in 1953:*

$$10.7, 12.3, 12.6, 10.4, 9.7, 10.1, 10.0, 8.3, 6.5, 7.7, 8.4$$

The larger values do tend to be earlier in the sequence, but could such a sequence occur randomly with reasonably high probability?

Ans:

- Median value of the sample: 10.0
- A or B according to whether the data point is above or below the median.
- A, A, A, A, B, A, B, B, B, B.
- The probability of such a sequence is $6/252 \approx 0.024$.
- Since this is unlikely to happen under the assumption of randomly ordered sequences, we have strong evidence of a decreasing trend. □

Question 2.1.3 *Explain why we could have obtained the result for the size of the sample space by looking at the problem as one of selecting 5 positions in the list in which to put A symbols.*

Ans: Given 5 positions of A, then all positions will be determined. □

Example 2.1.9 *To analyze the effectiveness of four stain-resistance treatments on carpet, an experiment is to be performed in which* 20 *pieces of carpet are randomly assigned to the four stain resistors, with each resistor to be applied to* 5 *carpet pieces. The levels of staining when the carpet samples are subjected to certain stain-producing actions will then be measured. In how many ways can the carpet samples be divided?*

Ans:

- This is an example of a partitioning problem, in which we have a group of n distinguishable objects, partitioned into some number k of unordered subsets with different sizes $r_1, r_2, \ldots, r_k$.

- $\binom{20}{5}\binom{15}{5}\binom{10}{5}\binom{5}{5} = 8,705,721,024$ □

Question 2.1.4 *Try to simplify the product of binomial coefficients on the left in the last equation. The result that you get might be familiar to you from other contexts as a* ***multinomial coefficient.*** *See also Exercise 23 of this section.*

Ans:

- $\binom{n}{r_1,r_2,\ldots,r_k} = \frac{n!}{r_1!r_2!\cdots r_k!} = \binom{20}{5,5,5,5} = \frac{20!}{5!\cdot 5!\cdot 5!\cdot 5!} = 8,705,721,024$ □

Example 2.1.10 (***Identical objects into distinct cells***) *Our final example illustrates a class of problems known as* ***occupancy problems.*** *In how many ways can r indistinguishable balls be distributed into n cells, with no restrictions on how many balls can be in each cell? Physicists have applied these ideas to distributions of photons among energy levels (the* ***Bose-Einstein statistics****) and other situations.*

Ans:

- Number of nonnegative integer solutions of $x_1 + x_2 + \cdots + x_n = r$.
- $H_n^r = \frac{(n+r-1)!}{(n-1)!r!} = \binom{n+r-1}{n-1} = \binom{n+r-1}{r}$ □

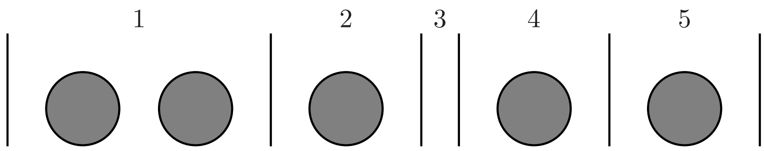


Figure 2.2: ***Partitioning***

2.2 Discrete distributions

2.2.1 Uniform, empirical, and hypergeometric distributions

Probability mass function (p.m.f.):

$$q(x) = P[X = x] \qquad \text{for } x \in E$$

Cumulative distribution function (c.d.f.):

$$F(x) = P[X \le x] = \sum_{t \le x} q(t)$$

A few elementary properties if the distribution function:

- $F(x) - F(x^-) = q(x), \quad \forall x \in E$

- $F(x)$ is a nondecreasing, nonnegative function, step function (discrete)
- $\lim_{x\to\infty} F(x) = 1$
- $P[a < X \leq b] = F(b) - F(a)$

Discrete uniform distribution:

$$q(x) = \frac{1}{n}, \qquad x \in \{x_1, x_2, \ldots, x_n\}$$

Question 2.2.1 *What would the cumulative distribution function of the discrete uniform mass function on* $\{1, 2, \ldots, n\}$ *look like?*

Ans:

- A ladder. □

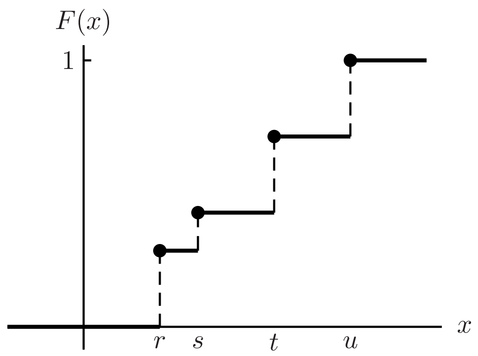


Figure 2.3: ***Discrete cumulative distribution function***

Question 2.2.2 *If you had n independent observations of a discrete random variable, what would be a sensible way to estimate the c.d.f. F? See the following definition for the answer.*

Ans:

- $\hat{F}(x) = \frac{\text{number of } X_i \leq x}{n}$, $x \in \mathbb{R}$. □

Definition 2.2.1 *Let* $X_1, \ldots, X_n$ *be independently sampled random variables all having the same probability distribution. Let* $\{w_1, \ldots, w_m\}$ *be the collection of states taken on by at least one of the* X_i'*s. Then the* ***empirical probability mass function (emf)*** *of the sample is*

$$\hat{q}(w_j) = \frac{\text{number of } X_i = w_j}{n}.$$

The ***empirical cumulative distribution function (edf)*** *of the sample is then the c.d.f. associated with* $\hat{q}$ *through equation (2.7), alternatively,*

$$\hat{F}(w_j) = \frac{\text{number of } X_i \leq w_j}{n}.$$

Example 2.2.1 *A recent survey conducted on the Internet studied grade point averages in calculus courses at 45 responding universities. For calculus I, the GPAs that were reported, rounded to the nearest tenth, were as follows:*

1.9, 1.6, 2.5, 2.3, 2.0, 2.5, 2.5, 1.4, 1.8, 3.1, 3.0, 2.3, 2.2, 1.8, 2.5, 1.9, 2.3, 2.1, 2.2, 2.2, 1.5, 2.7, 2.0, 1.8, 2.2, 2.1, 2.0, 2.5, 2.3, 2.2, 1.7, 1.9, 2.4, 2.2, 2.3, 2.2, 2.1, 2.7, 2.2, 2.1, 2.0, 2.1, 2.7, 2.5, 2.6

Ans:

- {{1, 1.4}, {2, 1.5}, {3, 1.6}, {4, 1.7}, {5, 1.8}, {6, 1.8}, {7, 1.8}, {8, 1.9}, {9, 1.9}, {10, 1.9}, {11, 2.0}, {12, 2.0}, {13, 2.0}, {14, 2.0}, {15, 2.1}, {16, 2.1}, {17, 2.1}, {18, 2.1}, {19, 2.1}, {20, 2.2}, {21, 2.2}, {22, 2.2}, {23, 2.2}, {24, 2.2}, {25, 2.2}, {26, 2.2}, {27, 2.2}, {28, 2.3}, {29, 2.3}, {30, 2.3}, {31, 2.3}, {32, 2.3}, {33, 2.4}, {34, 2.5}, {35, 2.5}, {36, 2.5}, {37, 2.5}, {38, 2.5}, {39, 2.5}, {40, 2.6}, {41, 2.7}, {42, 2.7}, {43, 2.7}, {44, 3.0}, {45, 3.1}}
- $\hat{q}(2.2) = 8/45$
- $\hat{q}(1.9) = 3/45$
- $\hat{F}(1.7) = 4/45$ □

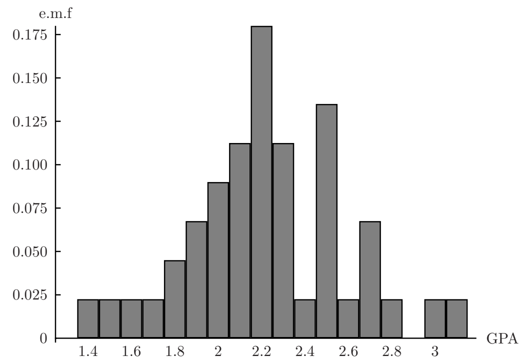


Figure 2.4: ***Empirical mass function***

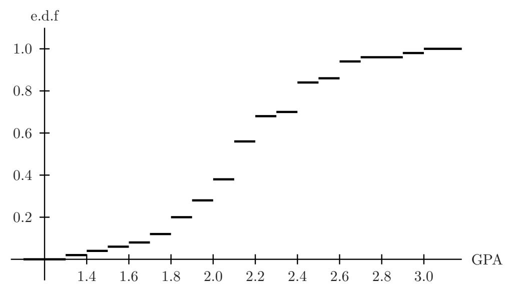


Figure 2.5: ***Empirical distribution function***

The experiments involving sampling without replacement discussed in the last section give rise to a frequently observed distribution called the ***hypergeometric distribution***.

$$q(k) = P[X = k] = \frac{\binom{N}{k}\binom{M-N}{n-k}}{\binom{M}{n}} \quad \text{for } k = \max\{0, n-(M-N)\}, \ldots, \min\{N, n\} \tag{2.17}$$

Question 2.2.3 *Justify (2.17) using the results from Section 2.1.*

Ans:

- Let A be the event of choosing k items from N items and $n-k$ items from $M-N$ items.
- $n(\Omega) = \binom{M}{n}$
- $n(A) = \binom{N}{k}\binom{M-N}{n-k}$
- $P[X = k] = P[A] = \frac{n(A)}{n(\Omega)} = \frac{\binom{N}{k}\binom{M-N}{n-k}}{\binom{M}{n}}$ □

Example 2.2.2 *Exercise 17 of Section 2.1 dealt with a capture-recapture model. Among a group of M mice, N are captured, tagged, and set free to mix with the others.*

Ans:

- If a second sample of n is subsequently caught, then the number of tagged mice in the second sample has the hypergeometric distribution (2.17).
- If the overall population has 100 mice, and the first capture tagged 20 of them, then the probability that there will be 2 or fewer tagged mice in a new sample of 5 is

$$\begin{aligned} P[X \leq 2] &= P[X = 0] + P[X = 1] + P[X = 2] \\ &= \frac{\binom{20}{0}\binom{80}{5} + \binom{20}{1}\binom{80}{4} + \binom{20}{2}\binom{80}{3}}{\binom{100}{5}} \\ &\approx .947 \end{aligned}$$

□

2.2.2 Multivariate discrete distributions

- $\mathbf{X} = (X_1, \ldots, X_n)$
- $q(\mathbf{x}) = P[\mathbf{X} = \mathbf{x}]$: ***joint probability mass function*** of the components of $\mathbf{X}$.
- $F(\mathbf{x}) = F(x_1, \ldots, x_n) = P[X_1 \leq x_1, \ldots, X_n \leq x_n]$

Equal probability weight 1/6 to each point:

$$\{(0,0), (1,0), (2,0), (1,1), (2,1), (2,2)\}$$

- $P[X_1 = 2] = P[(2,0), (2,1), (2,2)] = 3/6$
- $F(1,1) = P[X_1 \leq 1, X_2 \leq 1] = P[(0,0), (1,0), (1,1)] = 3/6$ □

Question 2.2.4 *Draw a picture of the state space of the random vector just described. What is* $P[X_2 = 0]$*?* $P[X_2 = 1]$*?* $P[X_2 = 2]$*? Does anything strike you about this collection of numbers?*

Ans:

- $P[X_2 = 0] = P[(0,0),(1,0),(2,0)] = 3/6$
- $P[X_2 = 1] = P[(1,1),(2,1)] = 2/6$
- $P[X_2 = 2] = P[(2,2)] = 1/6$
- X_2 is a random variable with a probability distribution $q(x)$. □

When a random vector $\mathbf{X} = (X_1, \ldots, X_n)$ has a probability distribution $q(\mathbf{x})$, a probability distribution is also induced on each individual component $\mathbf{X}$; by summing the mass function over all possible states of the other random variables X_j for $j \neq i$.

Definition 2.2.2 *Let* $\mathbf{X} = (X_1, \ldots, X_n)$ *be a random vector with probability mass function* $q(\mathbf{x})$*. The* ***marginal mass function*** *of* X_i *is*

$$q_i(x_i) = P[X_i = x_i] = \sum_{x_1} \cdots \sum_{x_{i-1}} \sum_{x_{i+1}} \cdots \sum_{x_n} q(x_1, \ldots, x_n)$$

Similarly, the ***joint marginal mass function*** *of a subcollection* $X_{i_1}, \ldots, X_{i_k}$ *of the random variables is*

$$\begin{aligned} q(x_{i_1}, \ldots, x_{i_k}) &= P[X_{i_1} = x_{i_1}, \ldots, X_{i_k} = x_{i_k}] \\ &= \sum \cdots \sum q(x_1, \ldots, x_n) \end{aligned}$$

where the sum in (2.20) is taken over all indices i *not in* $\{i_1, \ldots, i_k\}$*.*

Example 2.2.3 *An oil company is drilling at three sites. At each site, holes will be drilled at a sequence of locations until oil is struck. Let* $\mathbf{X} = (X_1, X_2, X_3)$ *represent the numbers of holes drilled at sites* 1, 2, *and 3, and suppose that* $\mathbf{X}$ *has joint probability mass function*

$$q(x_1, x_2, x_3) = \frac{1}{12}\left(\frac{1}{2}\right)^{x_1-1}\left(\frac{3}{4}\right)^{x_2-1}\left(\frac{1}{3}\right)^{x_3-1}, \quad x_1, x_2, x_3 = 1, 2, \ldots$$

Ans:

- Geometric series: $\sum_{n=0}^{\infty} a_0 r^n = \frac{a_0}{1-r}$ if $|r| < 1$.
- $q(x_1, x_2) = \sum_{x_3=1}^{\infty} q(x_1, x_2, x_3) = \frac{1}{8}\left(\frac{1}{2}\right)^{x_1-1}\left(\frac{3}{4}\right)^{x_2-1}, \quad x_1, x_2 \geq 1$
- $q(x_1) = \sum_{x_2=1}^{\infty} q(x_1, x_2) = \frac{1}{2}\left(\frac{1}{2}\right)^{x_1-1}, \quad x_1 = 1, 2, \ldots$ $G(1/2)$
- $q(x_2) = \frac{1}{4}\left(\frac{3}{4}\right)^{x_2-1}, \quad x_2 = 1, 2, \ldots$ $G(1/4)$
- $q(x_3) = \frac{2}{3}\left(\frac{1}{3}\right)^{x_3-1}, \quad x_3 = 1, 2, \ldots$ $G(2/3)$
- X_1, X_2, X_3 are independent since $q(x_1, x_2, x_3) = q(x_1)q(x_2)q(x_3)$. □

Question 2.2.5 *Consider the uniform distribution on the triangular grid of integers* $\{(0,0),(1,0),(2,0),(1,1),(2,1),(2,2)\}$ *with which we started the subsection. Are the marginals of the first coordinate* X_1 *and the second coordinate* X_2 *uniform?*

Ans:

- $q(0) = P[X_1 = 0] = P[(0,0)] = 1/6$
  $q(1) = P[X_1 = 0] = P[(1,0),(1,1)] = 2/6 = 1/3$
  Since $q(0) \neq q(1)$, X_1 is not an uniform distribution.
- $q(0) = P[X_2 = 0] = P[(0,0),(1,0),(2,0)] = 3/6$
  $q(1) = P[X_2 = 1] = P[(1,1),(2,1)] = 2/6$
  Since $q(0) \neq q(1)$, X_2 is not an uniform distribution. □

2.3 Binomial experiments

- A group of discrete probability distributions related to the experiment of flipping a coin sequentially.
- A sequence of small experiments called ***trials*** is performed.
- Each trial results in one of two possible outcomes, labeled arbitrarily as "success" or "failure."
- The events of success and failure on different trials are mutually independent.
- The probability of success, say p, on a single trial does not change from trial to trial; hence the failure probability $q = 1 - p$ is also constant.
- If these conditions hold, we say that the experiment is a sequence of Bernoulli trials, sometimes called a ***binomial experiment***.

Examples:

1. Patients being treated with a new procedure, a success being an improvement in condition.
2. People being polled about their opinions on a controversial issue, a success being a pro as opposed to a con opinion.
3. Airline passengers scheduled for a flight; a success being a passenger who does show up.
4. Shots taken by a basketball player during a game, a success being a made shot.

Question 2.3.1 *Think of at least two other examples of random phenomena that might be modeled as binomial experiments. Review the list of conditions that such an experiment is supposed to satisfy. Are there assumptions that must be made in your examples?*

Ans:

- Flip a coin several times, a success being the coin lands head.
- Launch missiles, a success being a missile hit the target. □

2.3.1 Bernoulli trials

- $q(x) = P[I(\omega) = x] = p^x(1-p)^{1-x}, x = 0, 1$ where (indicator variable)

$$I(\omega) = \begin{cases} 1 & \text{if } \omega = S, \\ 0 & \text{if } \omega = F. \end{cases} \tag{2.21}$$

- Since p is the probability of success, the distribution of I, called the ***Bernoulli distribution***, is clearly:

$$q(x) = P[I = x] = p^x(1-p)^{1-x}, \quad x = 0, 1.$$

- Let I_j be the indicator as in (2.21) for the jth trial; that is, $I_j = 1$ if the jth trial is a success and $I_j = 0$ otherwise.
- The joint p.m.f. of the indicators is

$$\begin{aligned} q(x_1, \ldots, x_n) &= P[I_1 = x_1, I_2 = x_2, \ldots, I_n = x_n] \\ &= P[I_1 = x_1] \cdot P[I_2 = x_2] \cdots P[I_n = x_n] \\ &= p^k q^{n-k}, \quad \text{where } k = \sum_{i=1}^{n} x_i. \end{aligned}$$

- We will mostly be concerned with the random variable X, which counts the total number of successes in binomial experiments.
- Its distribution is called the ***binomial distribution***.

Proposition 2.3.1 *Let X be the total number of successes in n Bernoulli trials. Then X has the* ***binomial probability mass function****:*

$$q(k) = P[X = k] = \binom{n}{k} p^k (1-p)^{n-k}.$$

(A shorthand notation used to refer to this distribution is $b(n, p)$.)

- Figure 2.6 plots the binomial p.m.f. for $n = 5$ in each of the cases $p = .5$, which gives a symmetric distribution, and $p = .2$, which does not.
- Table 1 of Appendix A contains probabilities for the binomial p.m.f. for values of n from 2 to 12, and some common values of p.
- For instance, if X is $b(6, 1/5)$ distributed, then

$$P[X \leq 1] = P[X = 0] + P[X = 1] = .2621 + .3932 = .6553.$$

Example 2.3.1 *Airlines frequently overbook flights, expecting that some of the passengers who have reservations will not show up. Often the airline will offer a free flight or other compensation to passengers with reservations who are bumped from a flight due to overbooking. Suppose that 5% of reserved ticket holders tend to renege. If, on a flight of 200, the airline actually books 210 passengers, what is the probability that the airline will have to give away at least one free ticket?*

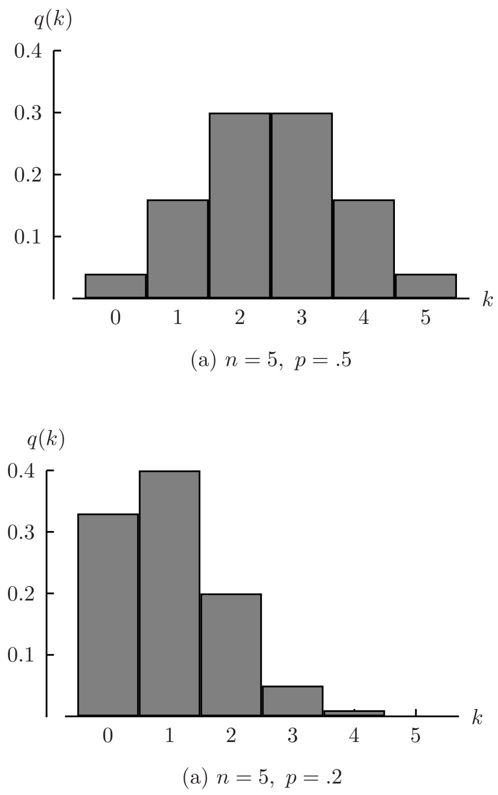


Figure 2.6: ***Binomial mass functions***

Ans:

- Let X denote the number of passengers who renege.
- $X \sim b(210, .05)$
- $P[X \leq 9] = \sum_{k=0}^{9} \binom{210}{k} (.05)^k (.95)^{210-k} = .3926$ □
- Since $F(9) = P[X \leq 9] = .39258$, the probability is about 39% that the airline will have to give away at least one free ticket.

Question 2.3.2 *What is the probability that the airline will give away exactly 3 free tickets? Try to draw some other conclusions about the overbooking problem using Table 2.2.*

Ans:

- $P[X = 7] = \binom{210}{7} (.05)^7 (.95)^{203} = 0.07584$ □

Example 2.3.2 *Suppose that the shots taken by a basketball player form a sequence of Bernoulli trials, with a success defined as a made shot. There is something interesting that we can derive about the pattern of made shots. Let $I_1, \dots, I_n$ be the Bernoulli trial indicator variables as before, and let X be the (binomially distributed) number of successfully made shots.*

Ans:

- Conditional probability

$$P[I_1 = x_1, \dots, I_n = x_n | X = k] = \frac{p^k(1-p)^{n-k}}{\binom{n}{k} p^k (1-p)^{n-k}} = \frac{1}{\binom{n}{k}}$$

- There are $\binom{n}{k}$ arrangements of 0-1 sequences with k 1s, each of which has this probability.
- Given the number of made shots, every possible pattern of made shots is as likely as any other pattern. □

2.3.2 Geometric and negative binomial distributions

We now study an open-ended sequence of Bernoulli trials, performed until such time as the rth success is observed. Consider first the case $r = 1$.

Proposition 2.3.2 *Let T_1 be the random variable that returns the trial on which the first success occurs in a sequence of Bernoulli trials. Then T_1 has the* ***geometric distribution***

$$P[T_1 = n] = (1-p)^{n-1} p, \qquad n = 1, 2, \dots \tag{2.26}$$

Proof:

- The event that $T_1 = n$ consists of the single outcome $(F, F, F, \dots, S)$, where $n-1$ failures precede a success.
- The formula for the p.m.f. of T_1 follows immediately. □

Question 2.3.3 *Verify that the geometric probability mass function in (2.26) does satisfy the appropriate conditions for a p.m.f.*

Ans:

- $P[T_1 = n] \geq 0, \quad n = 1, 2, \dots$
- $\sum_{n=1}^{\infty} P[T_1 = n] = \sum_{n=1}^{\infty} (1-p)^{n-1} p = \frac{p}{(1-(1-p))} = 1$ □

Example 2.3.3 *Suppose that a court hears civil suit cases and that only 10% of the cases are decided in favor of the plaintiff (原告). What is the probability that the first successful plaintiff in a given week is the seventh case? Conditioned on the next nine cases being ruled against the plaintiffs, what is the probability that the next twelve cases (including the nine) are ruled against?*

Ans:

- In this example the Bernoulli trials, fittingly, are trials.
- A success is a trial that a plaintiff wins, which occurs with probability 1/10. We assume that independent decisions are made by the court.
- Then the probability that the first successful ruling is on the seventh case is $P[T_1 = 7] = \left(\frac{9}{10}\right)^6 \frac{1}{10} \approx .053$

- $P[T_1 > 12 | T_1 > 9] = \frac{(9/10)^{12}}{(9/10)^9} = (9/10)^3$
- This is the same as the unconditional probability that the next three cases (after the ninth) are refused, without regard to the event that nine cases have been refused so far.
- Under our assumptions, the court is memoryless; it is not "due" to rule for a plaintiff just because nine straight plaintiffs have lost. □

Proposition 2.3.3 (***Negative binomial distribution***) *Let T_r be the trial on which the rth success occurs in a sequence of Bernoulli trials. Then T_r has the* ***negative binomial distribution****:*

$$P[T_r = n] = \binom{n-1}{r-1} p^r (1-p)^{n-r}, \qquad n = r, r+1, \ldots \tag{2.28}$$

Proof:

- For values of n at least r, the event that $T_r = n$ is the event that there were $r - 1$ successes on the first $n - 1$ trials, and then a success on the nth trial.
- Applying the binomial distribution and the independence of trial gives us

$$P[T_r = n] = \binom{n-1}{r-1} p^{r-1} (1-p)^{n-r} \cdot p$$

Formula (2.28) results. □

Example 2.3.4 *Rangers (守林人) at a mountain resort set charges at various locations in order to induce controlled avalanches (雪崩) after snowfalls. An avalanche will start after two effective charges detonate, and the chance that a charge will be effective is about .3. Then what is the probability that the avalanche will occur between the detonation of charges 10 and 20 inclusive?*

Ans:

-

$$\begin{aligned} P[10 \le T_2 \le 20] &= \sum_{n=10}^{20} \binom{n-1}{1} \left(\frac{3}{10}\right)^2 \left(\frac{7}{10}\right)^{n-2} \\ &= \left(\frac{3}{10}\right)^2 \sum_{n=10}^{20} (n-1) \left(\frac{7}{10}\right)^{n-2} \end{aligned}$$ □

- The last partial series is of the form $\sum (n-1)x^{n-2}$, which is the derivative of the partial series $\sum x^{n-1}$.
- Since it is well known that the closed form for the partial geometric series $\sum_{n=0}^{m} x^n$

is $(1-x^{m+1})/(1-x)$ we can finish the computation as follows:

$$\begin{aligned}
P[10 \le T_2 \le 20] &= \frac{9}{100}\frac{\mathrm{d}}{\mathrm{d}x}\left[\sum_{n=10}^{20} x^{n-1}\right] \\
&= \frac{9}{100}\frac{\mathrm{d}}{\mathrm{d}x}\left[\sum_{n=0}^{19} x^{n} - \sum_{n=0}^{8} x^{n}\right] \\
&= \frac{9}{100}\frac{\mathrm{d}}{\mathrm{d}x}\left[\frac{1-x^{20}}{1-x} - \frac{1-x^{9}}{1-x}\right] \\
&= \frac{9}{100}\frac{\mathrm{d}}{\mathrm{d}x}\left[\frac{x^{9}-x^{20}}{1-x}\right] \\
&= \frac{9}{100}\left[\frac{9x^8-8x^9-20x^{19}+19x^{20}}{(1-x)^2}\right]_{x=7/10} \approx .188 \qquad \square
\end{aligned}$$

2.3.3 Multinomial distribution

- Extension of the binomial distribution to the multivariate case.
- Let the assumptions of the binomial experiment hold once again, with one exception: Instead of two categories, "success" and "failure," for trial outcomes, let there be k mutually exclusive and exhaustive categories, arbitrarily labeled $1, 2, \ldots, k$.
- The probability that a trial results in category i is denoted by p_i.
- It follows that $\sum_{i=1}^{k} p_i = 1$.
- Let X_i denote the number of trials among the total of n resulting in category i.
- Since there are exactly n trials, it must be that $\sum_{i=1}^{k} X_i = n$.
- The vector random variable $\mathbf{X} = (X_1, X_2, \ldots, X_k)$ has the multinomial distribution

$$q(x_1, \ldots, x_k) = \binom{n}{x_1, \ldots, x_k} p_1^{x_1} \cdots p_k^{x_k}$$

- $\binom{n}{x_1,\ldots,x_k} = \frac{n!}{x_1!\cdots x_k!}$
- $\binom{n}{x_1,\ldots,x_k} = \binom{n}{x_1}\binom{n-x_1}{x_2}\cdots\binom{n-x_1-\cdots-x_{k-1}}{x_k}$

Question 2.3.4 *Look at slices of the figure perpendicular to the coordinate axes. Do you notice anything?*

Ans:

- Given the value of X_i, the graph of p.m.f of X_j is symmetric where $i \neq j$. Thus $X_i|X_j$ is a binomial distribution where $i \neq j$. $\square$

Example 2.3.5 *A politician who is basing his campaign on a certain controversial issue claims that among his constituents, 60% are in favor of the issue, 30% are against, and 10% are undecided. In an attempt to discredit his claim, his rival commissions a survey of a random sample of 20 of the constituents. Among these, just 7 are in favor, 6 are against, and 7 are undecided. Does this survey give strong evidence against the first politician?*

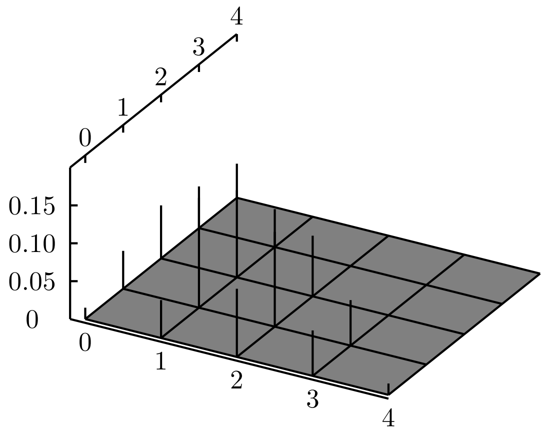


Figure 2.7: ***Multinomial p.m.f.,*** $n = 4, p_1 = p_2 = 1/3$

Ans:

$$
\begin{aligned}
&P[X_1 \le 7, X_2 \le 6] \\
&= \sum_{x_1=0}^{7} \sum_{x_2=0}^{6} \binom{20}{x_1, x_2, 20 - x_1 - x_2} (.6)^{x_1} (.3)^{x_2} (.1)^{20 - x_1 - x_2} \\
&\approx .0004
\end{aligned}
$$

□

- If a trial has possible outcomes $1, 2, \ldots, k$ and we are interested in only the distribution of X_1, we can call category 1 a success and lump together the cases whose trial outcome is one of $2, 3, \ldots, k$ as failure events.

- Then we have reduced the problem to the binomial case, and X_1 should have the $b(n, p_1)$ distribution.

Proposition 2.3.4 *If $X_1, \ldots, X_k$ have the $m(n, p_1, \ldots, p_k)$ distribution, then for each i, X_i has the $b(n, p_i)$ distribution.*

Proof:

- We will prove the theorem only for $k = 3$, for random variable X_1.

- If X_1, X_2, X_3 have the $m(n, p_1, p_2, p_3)$ distribution, we find the marginal distribution of X_1.

- A general proof follows along the same lines.

- Recall first that X_3 is required to take on the value $x_3 = n - x_1 - x_2$, where x_1 and x_2 are the values taken on by X_1 and X_2, respectively.

- Thus we can look at the joint p.m.f. as a function of x_1 and x_2 only:

$$
q(x_1, x_2) = \binom{n}{x_1, x_2, (n - x_1 - x_2)} p_1^{x_1} p_2^{x_2} (1 - p_1 - p_2)^{n - x_1 - x_2} \tag{2.32}
$$

- Summing (2.32) over all possible values of x_2 for a given value of x_1, we obtain

$$
\begin{aligned}
q_1(x_1) &= P[X_1 = x_1] \\
&= \sum_{x_2=0}^{n-x_1} \frac{n!}{x_1!x_2!(n-x_1-x_2)!} p_1^{x_1} p_2^{x_2} (1-p_1-p_2)^{n-x_1-x_2} \\
&= \frac{n!}{x_1!(n-x_1)!} p_1^{x_1} \sum_{x_2=0}^{n-x_1} \frac{(n-x_1)!}{x_2!(n-x_1-x_2)!} p_2^{x_2} (1-p_1-p_2)^{n-x_1-x_2} \\
&= \binom{n}{x_1} p_1^{x_1} (p_2 + (1-p_1-p_2))^{n-x_1} \\
&= \binom{n}{x_1} p_1^{x_1} (1-p_1)^{n-x_1} \qquad (2.33)
\end{aligned}
$$

- We have used the Binomial Theorem to go from the fourth line to the fifth line.
- The last line establishes that X_1 has a marginal $b(n, p_1)$ distribution. □

Question 2.3.5 *What do you think can be said about the joint marginal distribution of X_1 and X_2 if $X_1, \ldots, X_k$ have the $m(n, p_1, \ldots, p_k)$ distribution? Try to generalize.*

Ans:

- $q(x_1, x_2) \sim m(n, p_1, p_2, 1 - (p_1 + p_2))$
- $q(x_{i_1}, x_{i_2}, \ldots, x_{i_m}) \sim m(n, p_{i_1}, p_{i_2}, \ldots, p_{i_m}, 1 - \sum_{j=1}^{m} p_{i_j})$ □

2.4 Poisson random phenomena

2.4.1 Poisson distribution

- A discrete distribution with state space $n = 0, 1, 2 \ldots$.
- Arise as a model of random phenomena in a way that differs from any of the special distributions that we have studied thus far.
- In the cases of the discrete uniform, hypergeometric, Bernoulli, binomial, geometric, and negative binomial probability distributions, the form of the probability mass function could be derived easily from very primitive randomness assumptions using basic combinatorics.
- The Poisson distribution comes from limiting considerations.
- Perhaps more than the other distributions, it is important to use observed data to check how well the Poisson distribution fits reality (that is, to check whether the empirical distribution of a random sample fits the theoretical Poisson distribution reasonably well).
- We will investigate so-called "goodness-of-fit" tests in a later chapter.

- The following example will help to motivate the limiting process from which the Poisson distribution is obtained.

Example 2.4.1 *A student researcher in ecology by the name of Cathy Clover was interested in the geographical distribution of a certain species of jumping mouse in a prairie region. The region was divided into n small, equally sized zones, as in Figure 2.8 where, on inspection, one could either find or not find a trapped mouse. The dots on the diagram are meant to indicate zones where mice were found. (Incidentally, the trapping process was done in such a way as not to harm the animals.)*

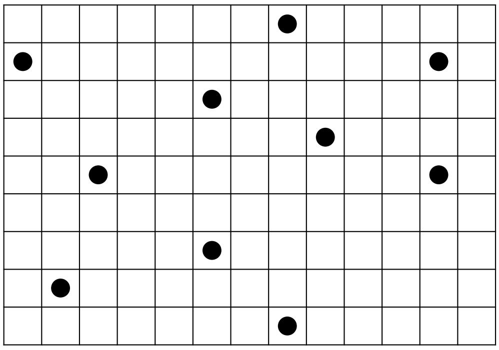


Figure 2.8: ***Trapping zones***

Ans:

- If the mice were randomly distributed around the region, but there were relatively few of them in total, it would be reasonable to model each zone as a Bernoulli trial, with some small success probability p of finding a trapped mouse.

- If we envision increasing the number n of zones while decreasing zone size, it stands to reason that p should decrease.

- Let us suppose that $p = p_n$, is proportional to $1/n$. If we call the proportionality constant λ, then
$$p_n = \frac{\lambda}{n} \Rightarrow \lambda = np_n. \tag{2.34}$$

- Under the assumptions that the success probability is consistent from zone to zone, and that the events of finding trapped mice in different zones are mutually independent, the random variable X = number of trapped mice throughout the region have the $b(n, p_n)$ distribution.

- $b(n, p_n) \rightarrow$ Poisson distribution with parameter λ. □

Proposition 2.4.1 *Under condition (2.34), for each fixed $k \geq 0$,*
$$\binom{n}{k}(p_n)^k(1-p_n)^{n-k} \rightarrow \frac{e^{-\lambda}\lambda^k}{k!} \qquad \text{as } n \rightarrow \infty. \tag{2.35}$$

Proof:

- In the case that $k > 0$, we have that

$$\begin{aligned}&\binom{n}{k}(p_n)^k(1-p_n)^{n-k}\\&=\frac{n!}{(n-k)!k!}\left(\frac{\lambda}{n}\right)^k\left(1-\frac{\lambda}{n}\right)^{n-k}\\&=\frac{\lambda^k}{k!}\cdot\left(1-\frac{\lambda}{n}\right)^n\cdot\frac{n\cdot(n-1)\cdots(n-k+1)}{n\cdot n\cdots n}\left(1-\frac{\lambda}{n}\right)^{-k}\end{aligned}$$

- By the well-known limit result from calculus, $(1-\lambda/n)^n$ approaches $e^{-\lambda}$ as $n \to \infty$.
- Since k is fixed, the third factor is the product of k ratios $n/n, (n-1)/n, \ldots, (n-k+1)/n$, all of which approach 1; hence the product also converges to 1.
- Also, since k and λ are fixed, the fourth factor approaches 1 as $n \to \infty$. This establishes (2.35). □

Question 2.4.1 *What happens to the limit-taking process in the proof in the case $k = 0$?*

Ans:

- $\binom{n}{0}(p_n)^0(1-p_n)^n = \left(1-\frac{\lambda}{n}\right)^n \to e^{-\lambda}$ as $n \to \infty$. □

Definition 2.4.1 *The **Poisson probability mass function** with parameter λ is*

$$q(k) = P[X = k] = \frac{e^{-\lambda}\lambda^k}{k!}, \qquad k = 0, 1, 2, \ldots$$

- Table 2 of Appendix A: probabilities for the Poisson p.m.f.
- For example, if X is Poisson(2.5) distributed, then

$$P[1 \le X \le 3] = .2052 + .2565 + .2138 = .6755.$$

- In the case of the jumping mice and in other cases, the approximability of $b(n, p)$ probabilities by Poisson probabilities with $\lambda = np$ for large n and small p helps to justify using the Poisson distribution as a model for the distribution of the number of occurrences of some event in a fixed region of space or time, since the region can be broken into many small subsets containing at most one occurrence.
- The probability of observing an occurrence in a subset should converge to 0 as the subset size converges to 0.
- Also, the numbers of occurrences in different subsets should be independent in order for the Poisson distribution to be a reasonable model.
- The Poisson mass function (2.36) does satisfy
  (i) $q(k) \ge 0$ for all k.
  (ii) $\sum_{k=0}^{\infty} q(k) = \sum_{k=0}^{\infty} \frac{e^{-\lambda}\lambda^k}{k!} = e^{-\lambda}\sum_{k=0}^{\infty}\frac{\lambda^k}{k!} = e^{-\lambda}\cdot e^{\lambda} = 1$

- Examples of the Poisson probability mass function in Fig. 2.9.

- The distribution is not symmetric but becomes more so as the parameter λ increases.

- The distribution flattens out and the weight moves out to the right as λ increases. There is a reason for this, as we will see when we study mean and variance later.

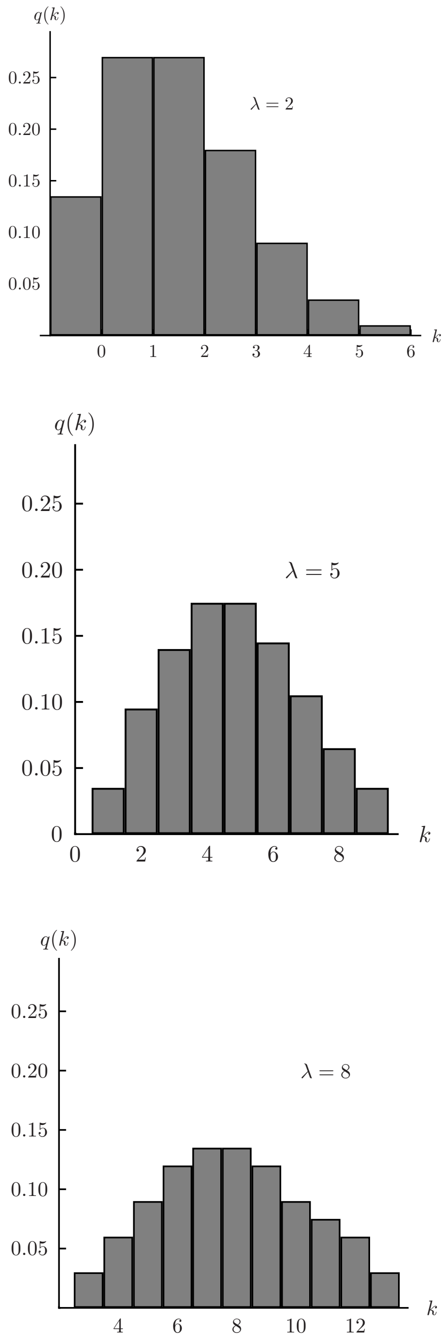


Figure 2.9: ***Poisson p.m.f.'s***

Example 2.4.2 *Suppose that there are a large number of users of a computer lab, but during a particular short time interval each has a very small probability of wanting access to the network server. The number of users X who do want access might be modeled as a Poisson random variable, since it is the sum of many Bernoulli random variables, one for each potential user (with value 1 if the user wants access and 0 otherwise), and the success probability per user is small. Figure 2.10 illustrates the situation. Suppose also that on the average 25% of the time, the server is not busy with any jobs from user machines.*

Ans:

- $q(0) = P[X = 0] = e^{-\lambda} = .25$ implies $\lambda = 1.39$.
- $P[\text{exactly 3 users}] = \frac{e^{-\lambda}\lambda^3}{3!} \approx .11$
- $P[\text{no more than 2 users}] = \frac{e^{-\lambda}\lambda^0}{0!} + \frac{e^{-\lambda}\lambda^1}{1!} + \frac{e^{-\lambda}\lambda^2}{2!} \approx .84$
- $P[\text{at least 3 users}] = 1 - P[\text{no more than two users}] \approx .16$ □

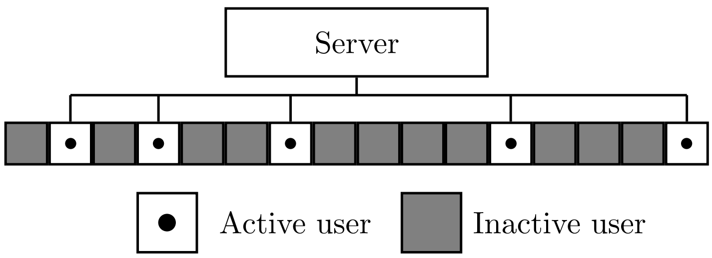


Figure 2.10: ***Computer users***

Question 2.4.2 *Try to think of other example of experimental situations where a Poisson model is appropriate, using the primitive assumption of many Bernoulli trials of small success probability.*

Ans: Radioactive decay. □

2.4.2 Poisson process

- The Poisson distribution finds broad application in random phenomena evolving over time.
- N_t: The number of occurrences of an event up to and including t.
- For example, the arrivals of customers to a coffee shop.
- Outcome ω consists of the arrival instants $(T_1(\omega), T_2(\omega), T_3(\omega), \dots)$.
- $N_t(\omega)$: Total number of customers who have arrived by time t.
- $N_t(\omega)$: Cumulative count as a function of t. (See Fig. 2.11.)

- The graph begins at state 0 at time 0, when no customers have arrived yet. It then jumps by exactly one unit at each arrival time of a customer.
- Breaking up the time interval $[0, t]$ into many small intervals, in each of which there is a very small probability of an arrival, it should be true that under suitable assumptions, N_t has a Poisson distribution.

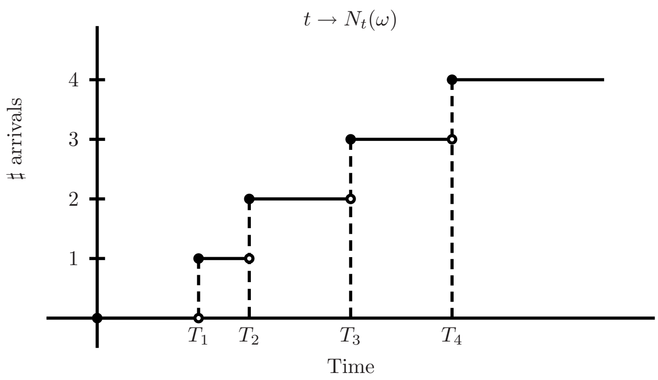


Figure 2.11: ***Poisson process***

Definition 2.4.2 *A* ***Poisson process*** *is a family $(N_t)_{t \geq 0}$ of random variables whose paths are step functions beginning at state 0 at time 0, which jump by 1 at a sequence of random times $T_1, T_2, \ldots$. Additionally, the following conditions are assumed:*

(i) The numbers of jumps within disjoint time intervals are independent.

(ii) The probability distribution of the number of jumps in a time interval depends only on the length of the interval, not on the initial time.

(iii) The probability of exactly one jump on a short time interval of length h is $\lambda h + o(h)$, the probability of two or more jumps is $o(h)$, and hence the probability of no jumps is $1 - \lambda h - o(h)$.

- Condition (i): The Bernoulli trials formed by breaking up the time axis are independent.
- Condition (ii): The success probability will be consistent as long as the subintervals are equally sized.
- Condition (iii): For very short time intervals $[0, h]$ only two events can happen with nonnegligible probability:
  (a) One arrival in $[0, h]$, with approximate probability λh.
  (b) No arrivals in $[0, h]$, with approximate probability $1 - \lambda h$.
- Under condition (iii), trials can have only two possible outcomes, no arrival or arrival, at least approximately.

Proposition 2.4.2 *If $(N_t)_{t\geq 0}$ is a Poisson process with rate parameter λ, then*

$$P[N_t = n] = \frac{e^{-\lambda t}(\lambda t)^n}{n!} \tag{2.37}$$

that is, the number of arrivals up through time t has the Poisson(λt) distribution.

Proof:

- Consider the case $n \geq 1$ first.
- Use the Law of Total Probability to condition and uncondition the event $\{N_{t+h} = n\}$ on the value of N_t, which could be $n, n-1, n-2, \ldots, 0$.
- Since the event of two or more arrivals in $(t, t+h]$ has probability $o(h)$, we can write

$$\begin{aligned} P[N_{t+h} = n] &= P[N_t = n, N_{t+h} - N_t = 0] \\ &\quad + P[N_t = n-1, N_{t+h} - N_t = 1] \\ &\quad + \sum_{k=0}^{n-2} P[N_t = k, N_{t+h} - N_t = 1] \\ &= P[N_t = n]P[N_{t+h} - N_t = 0 | N_t = n] \\ &\quad + P[N_t = n-1]P[N_{t+h} - N_t = 1 | N_t = n-1] \\ &\quad + o(h) \end{aligned}$$

- By Poisson process axiom (i), the number of arrival $N_{t+h} - N_t$ during $(t, t+h]$ is independent of the number of arrivals N_t in $[0, t]$; hence the conditional probabilities on the right side can be changes to ordinary probabilities.
- By axiom (ii), $N_{t+h} - N_t$ has the same distribution as the number of arrivals in $[0, h]$, so that these differences can be replaced by N_h.
- Axiom (iii) lets us express the probability of no arrivals as $1 - \lambda h$ and the probability of one arrival as λh, lumping the $o(h)$ terms into the one already present.
- The last expression above reduces to

$$P[N_{t+h} = n] = P_n(t+h) = P_n(t)(1 - \lambda h) + P_{n-1}(t)\lambda h + o(h). \tag{2.38}$$

- Expanding out the product on the right, moving the $P_n(t)$ term to the left side, and dividing through by h gives

$$\frac{P_n(t+h) - P_n(t)}{h} = -\lambda P_n(t) + \lambda P_{n-1}(t) + \frac{o(h)}{h}$$

- As $h \to 0$, we obtain the system of differential equations:

$$P_n'(t) = -\lambda P_n(t) + \lambda P_{n-1}(t), \qquad n \geq 1. \tag{2.39}$$

- In the case $n = 0$, similar reasoning (see Exercise 18) results in the following differential equation for $P_0(t) = P[N_t = 0]$:

$$P_0'(t) = -\lambda P_0(t), \qquad P_0(0) = 1 \tag{2.40}$$

- The solution of (2.40) is
$$P_0(t) = P[N_t = 0] = e^{-\lambda t} \tag{2.41}$$
which is formula (2.37) for the case $n = 0$.
- Straightforward substitution of formula (2.37) into the differential equation (2.39) verifies that it is a solution. □

Question 2.4.3 *Why must the initial condition $P_0(0) = 1$ in (2.40) be true? What do you think might be the proper initial condition in (2.39) for $P_n(0)$?*

Ans:

- By axiom (iii), $P[N_h = 0] = 1 - \lambda h - o(h)$, $P[N_h = n] = o(h)$.
- $P_0(0) = P[N_0 = 0] = 1$ and $P_n(0) = P[N_0 = n] = 0,\ n \geq 1$. □

Question 2.4.4 *Verify by substitution that $P_1(t) = e^{-\lambda t}(\lambda t)^1/1!$ satisfies (2.39).*

Ans:

- $P_1'(t) = \lambda e^{-\lambda t} - \lambda^2 t e^{-\lambda t} = -\lambda P_1(t) + \lambda P_0(t)$ □

Example 2.4.3 *Let's return to our coffee shop example. Suppose that customers arrive according to a Poisson process with a rate of $\lambda = 10$ per hour. Recall that the meaning of the latter is that for short time intervals of length h hours, the probability of an arrival during the interval is about $10h$. Now the counter attendant, Nigel, is on the verge of a nervous breakdown for reasons that we would not like to get into here, and he will collapse if he has to serve more than four customers per hour.*

Ans:

- By Proposition 2.4.2, the probability of four or fewer customers during the next hour is (using Table 2 of Appendix A)

$$\begin{aligned} P[N_t \leq 4] = \sum_{n=0}^{4} P[N_t = n] &= \sum_{n=0}^{4} \frac{e^{-10\cdot 1}(10\cdot 1)^n}{n!} \\ &= 0 + .0005 + .0023 + .0076 + .0189 \approx .0293 \end{aligned}$$

- Since $P[N_t > 4]$ is therefore about .97, Nigel is living on the edge.
- Find the c.d.f. of the time of arrival of the second customer, and the probability that the second customer arrives between times 10 and 20 minutes, or 1/6 and 1/3 hour.
- To compute the c.d.f. of T_2, note that the event that the second customer arrives by time t is the same as the event that the number of customers who have arrived by time t is at least 2.

$$\begin{aligned} P[T_2 \leq t] &= P[N_t \geq 2] \\ &= 1 - P[N_t \leq 1] \\ &= 1 - \sum_{n=0}^{1} \frac{e^{-10t}(10t)^n}{n!} \\ &= 1 - e^{-10t}[1 + 10t] \end{aligned}$$

- Then

$$P[T_2 \in (1/6, 1/3]] = P[T_2 \le 1/3] - P[T_2 \le 1/6]$$
$$= \frac{8}{3}\, e^{-5/3} - \frac{13}{3}\, e^{-10/3} \approx .35$$

2.5 Expected value and variance

2.5.1 Properties of expectation

- Suppose that a random variable X has two possible states, 1 and 3, as displayed in Fig. 2.12.
- The set of outcomes mapped by X to 1 has some probability p, and consequently the set of outcomes mapped to 3 has probability $1 - p$.
- Weighted average of the possible states:

$$\text{Average value of } X = p \cdot 1 + (1 - p) \cdot 3.$$

- More generally, in the situation displayed in Fig. 2.13 in which there are n possible states, the average value of X should be defined as the weighted average of the possible states e_i of X, where the weight of the ith state is the probability that X assumes that state.

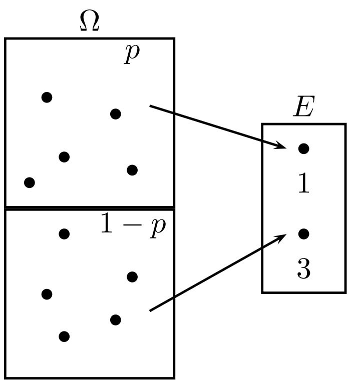


Figure 2.12: ***A random variable with two states***

Definition 2.5.1 *Let X be a discrete, real-value random variable with state space $E = \{e_1, e_2, \ldots\}$ and probability mass function q. Then the* ***expected value or expectation*** *of X is*

$$\mu = E[X] = \sum_i e_i P[X = e_i] = \sum_i e_i q(e_i)$$

provided the series converges.

The expected value is a kind of ***balance point*** for the distribution. This is easiest to see in the case of two states, as shown in Fig. 2.14.
$\mu = E[X]$ is used very frequently in probability. We call μ the ***mean*** of X.

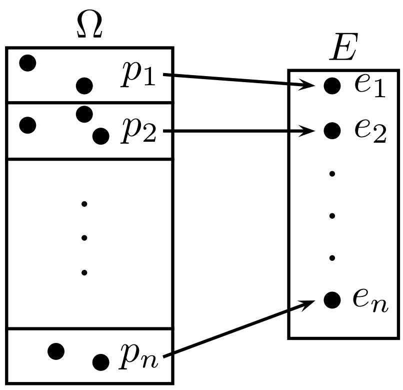


Figure 2.13: ***A random variable with many states***

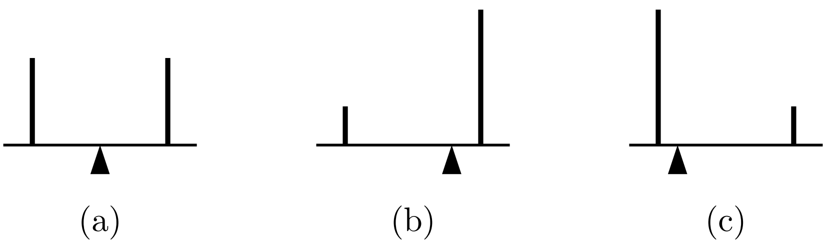


Figure 2.14: ***Expected value as balance point***

Question 2.5.1 *Must a nonnegative valued random variable have a nonnegative mean? Why?*

Ans:

- Yes.
- $e_i \geq 0 \Rightarrow E[X] = \sum_i e_i P[X = e_i] \geq 0$ □

Example 2.5.1 *Fast Eddie likes to play the horses. In an attempt to diversify his investments, he decides to bet \$8 on horse A, which is rated as a 20 : 1 shot to win; \$4 on horse B, which is rated as a 5 : 1 shot; and \$10 on horse C, rated as a 2 : 1 shot to win. If horse A wins, Eddie receives \$18 on the dollar, similarly horse B pays \$4, and horse C pays \$1.50. Find Fast Eddie's expected winnings.*

Ans:

- X: Eddie's net winnings, after deducting the \$22 he has bet.
- Then X takes on these values:
  (a) $18 \cdot 8 - 22 = 122$ if horse A wins;
  (b) $4 \cdot 4 - 22 = -6$ if horse B wins;
  (c) $10 \cdot 1.5 - 22 = -7$ if horse C wins;
  (d) $0 - 22 = -22$ if none of the horses A, B, or C win.
- If the odds against an event are $m : n$, then the probability that the event occurs is $n/(n + m)$.
- The state space of X and the probabilities on the states,

Horse	A	B	C	None
Probability	1/21	1/6	1/3	19/42
Winnings	122	−6	−7	−22

- $E[X] = \frac{1}{21} \cdot 122 + \frac{1}{6} \cdot (-6) + \frac{1}{3} \cdot (-7) + \frac{19}{42} \cdot (-22) = -\frac{157}{21} = -7.47619$ □

Question 2.5.2 *What would Fast Eddie's expected net winnings be if he had placed all of his money on the long shot? What do you think of his diversification system? Supposing (very hypothetically) that you were inclined to bet on one of the three horses described, which would you choose and why?*

Ans: He will lose all his money!

$$\begin{aligned} E[A] &= \frac{1}{21} \cdot 122 + \frac{20}{21} \cdot (-22) = \frac{-318}{21} \qquad \text{(minimum)} \\ E[B] &= \frac{1}{6} \cdot (-6) + \frac{5}{6} \cdot (-22) = \frac{-58}{3} \\ E[C] &= \frac{1}{3} \cdot (-7) + \frac{2}{3} \cdot (-22) = -17 \end{aligned}$$

Choose to bet on A since it has the minimum expectation among betting on three horses.
□

Example 2.5.2 *This example illustrates a case where a random variable does not have finite expectation. Suppose that a fair coin is flipped independently until the first head occurs.*

Ans:

- Then a typical outcome ω is a string of 0 or more tails, followed by a head.
- Let $X(\omega)$ have the value 2^n if ω is such that the first head is on the nth flip.
- Then $P[X = 2^n] = (1/2)^n$; consequently,

$$E[X] = \sum_{n=1}^{\infty} 2^n \left(\frac{1}{2}\right)^n = \sum_{n=1}^{\infty} 1 = \infty.$$ □

Example 2.5.3 *The television game show "The Price Is Right" contains a game called Plinko, which is an amazingly rich illustration of probabilistic ideas. A contestant who has successfully priced some items earns one or more chips and has the opportunity to stand at the top of the Plinko board, sketched in Figure 2.15, and release a chip from any of the topmost slots.*

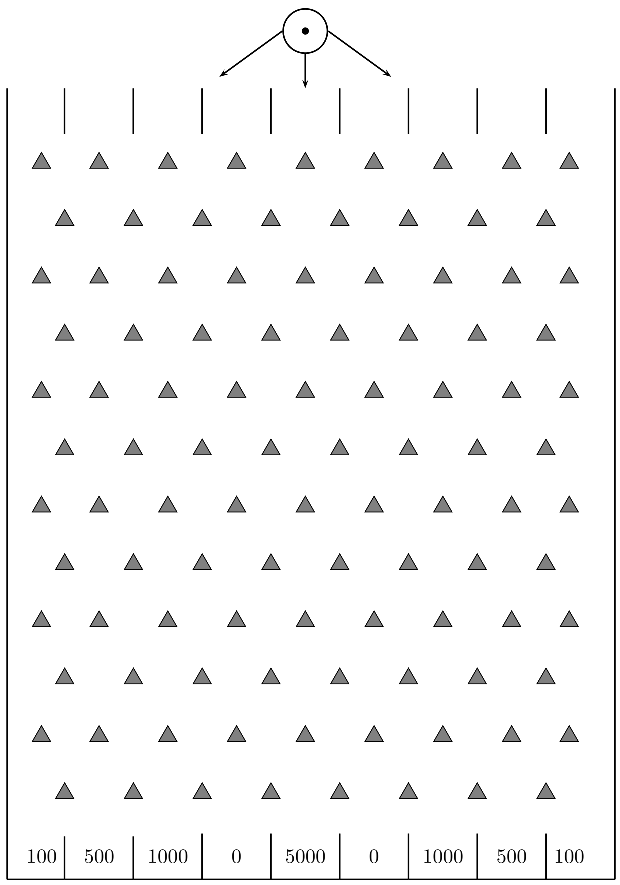


Figure 2.15: ***Plinko board***

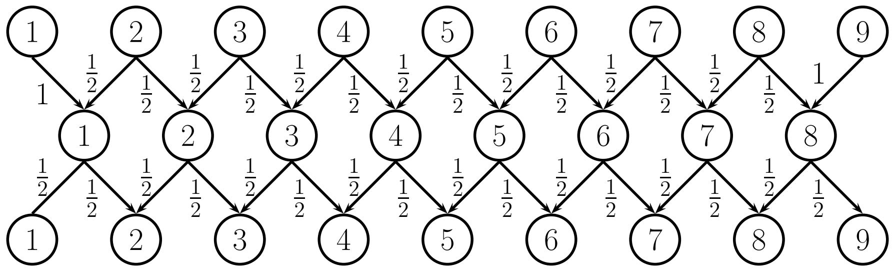


Figure 2.16: ***Successive Plinko rows***

- $p_m(i)$ = probability that chip is in column i when it hits row m.

- You can count 13 rows of the board, including the top slots and the bottom bins. Odd-numbered rows have nine possible column positions from left wall to right (look at the tops of the pegs), and even-numbered rows have eight. Figure 2.16 illustrates consecutive odd, even, and odd rows schematically.

- In row 1, when the contestant chooses slot i from which to drop the chip, it means that $p_1(i) = 1$ and all other $p_1(j) = 0$.

- Recurrence relations for even rows:

$$p_{2k}(1) = 1 \cdot p_{2k-1}(1) + \frac{1}{2} \cdot p_{2k-1}(2)$$
$$p_{2k}(i) = \frac{1}{2} \cdot p_{2k-1}(i) + \frac{1}{2} \cdot p_{2k-1}(i+1), \quad i = 2, \dots, 7$$
$$p_{2k}(8) = \frac{1}{2} \cdot p_{2k-1}(8) + 1 \cdot p_{2k-1}(9)$$

- Recurrence relations for odd rows:

$$p_{2k+1}(1) = \frac{1}{2} \cdot p_{2k}(1)$$
$$p_{2k+1}(i) = \frac{1}{2} \cdot p_{2k}(i-1) + \frac{1}{2} \cdot p_{2k}(i), \quad i = 2, \dots, 8$$
$$p_{2k+1}(9) = \frac{1}{2} \cdot p_{2k}(8)$$

- These difference equations can be solved analytically by using matrix multiplication to express one vector of probabilities p_m, in terns of the preceding vector p_{m-1}.

- Plinko Expected Winnings:

Bin	1	2	3	4	5	6	7	8	9	
Prize	100	500	1000	0	5000	0	1000	500	100	E[win]
Slot 1	.23	.39	.24	.11	.03	.01	0	0	0	608
Slot 2	.19	.35	.25	.14	.06	.02	0	0	0	744
Slot 3	.12	.25	.24	.20	.12	.05	.02	0	0	997
Slot 4	.05	.14	.20	.23	.19	.12	.05	.02	0	1285
Slot 5	.02	.06	.12	.19	.23	.19	.12	.06	.02	1454
Slot 6	0	.02	.05	.12	.19	.23	.20	.14	.05	1285
Slot 7	0	0	.02	.05	.12	.20	.24	.25	.12	997
Slot 8	0	0	0	.02	.06	.14	.25	.35	.19	744
Slot 9	0	0	0	.01	.03	.11	.24	.39	.23	608

Proposition 2.5.1 *If c is a real constant, then*

$$E[c] = c.$$

Proof:

- Two ways to look at the meaning of $E[c]$.
- One of these: Let X be any discrete random variable with p.m.f. q, and to consider the function $f(X) = c$. By (2.46) the expectation is

$$E[c] = \sum_{x \in E} c \cdot q(x) = c \cdot \sum_{x \in E} q(x) = c \cdot 1 = c.$$

- For the second approach, see the next question. □

If X is a discrete random variable with p.m.f. $q(x)$, and f is a real-valued function whose domain includes the state space of X, then the expected value of $f(X)$

$$E[f(X)] = \sum_{x \in E} f(x) \cdot q(x), \tag{2.46}$$

provided the series converges.

- Let X have state space $E = \{-2, -1, 0, 1, 2\}$, and p.m.f.

$$q_X(-2) = 1/10,\ q_X(-1) = 1/5,\ q_X(0) = 2/5,\ q_X(1) = 1/5,\ q_X(2) = 1/10.$$

- Consider the transformed random variable $Y = f(X) = X^2$. Then Y has state space $E_Y = \{0, 1, 4\}$, and $Y = 0$ iff $X = 0; Y = 1$ iff either $X = -1$ or $X = 1$; and $Y = 4$ iff either $X = -2$ or $X = 2$.
- Y has the p.rn.f.

$$q_Y(0) = 2/5, \quad q_Y(1) = 2/5, \quad q_Y(4) = 1/5.$$

- $E[Y] = \frac{2}{5} \cdot 0 + \frac{2}{5} \cdot 1 + \frac{1}{5} \cdot 4 = \frac{6}{5}$

- On the other hand,

$$\begin{aligned} E[Y] &= \frac{2}{5}\cdot 0^2 + \frac{1}{5}\cdot(-1)^2 + \frac{1}{5}\cdot(1)^2 + \frac{1}{10}\cdot(-2)^2 + \frac{1}{10}\cdot(2)^2 \\ &= \sum_{x=-2}^{2} x^2\cdot q_X(x) \\ &= \sum_{x=-2}^{2} f(x)\cdot q_X(x). \end{aligned}$$

Question 2.5.3 *Give an alternative proof of Proposition 2.5.1 using a random variable with only one state c.*

Ans:

- $E = \{c\}$
- $q(c) = P[X = c] = 1$
- $E[c] = c \cdot q(c) = c$ □

Proposition 2.5.2 (*Linearity of expectation*) *If X and Y are random variables with finite expectation, then*

$$E[aX + bY] = aE[X] + bE[Y]. \tag{2.48}$$

Proof:

- Let $q(x, y)$ be the joint p.m.f. of X and Y. Then,

$$\begin{aligned} E[aX + bY] &= \sum_x\sum_y (ax + by)\cdot q(x,y) \\ &= a\sum_x\sum_y x\cdot q(x,y) + b\sum_x\sum_y y\cdot q(x,y) \\ &= a\sum_x x\sum_y q(x,y) + b\sum_y y\sum_x q(x,y) \\ &= a\sum_x xq_X(x) + b\sum_y y\cdot q_Y(y) = aE[X] + bE[Y] \end{aligned}$$

- ***Linearity of expectation*** of two random variables, can be generalized to linear combinations of more than two random variables. □

Example 2.5.4 *To obtain information on the expected time in the recovery room of patients who have undergone a certain surgical procedure, a sample of 14 patients was taken. Their times, recorded to the nearest minute, follow:*

58, 75, 53, 60, 62, 80, 65, 54, 78, 63, 58, ,54, 60, 63

- These times are observed values of random variables $X_1, X_2, \ldots, X_{14} \overset{\text{i.i.d.}}{\sim} X$.
- $\mu = E[X_i] = E[X],\ i = 1, 2, \ldots, 14.$

- The simple arithmetical average of the observed values $\sum_{i=1}^{14} x_i/14 = 63.07$ is a good estimate of the underlying mean μ.
- One reason is that
$$E\left[\frac{1}{14}\sum_{i=1}^{14} X_i\right] = \frac{1}{14}E\left[\sum_{i=1}^{14} X_i\right] = \mu.$$

2.5.2 Variance and moments

- The mean μ is a useful measure of the central tendency of the probability distribution.
- Another important feature of a distribution is its spread, as measured by the expected square difference between the random variable and its mean.

Definition 2.5.2 *The* ***variance*** *of a real random variable X is*

$$\sigma^2 = Var(X) = E[(X-\mu)^2]$$

provided the expectation is finite. The square root of the variance, σ, is referred to as the ***standard deviation*** *of the random variable.*

Example 2.5.5 *To gain an intuitive appreciation for what the variance measures, consider the three probability mass functions sketched in Figure 2.17.*

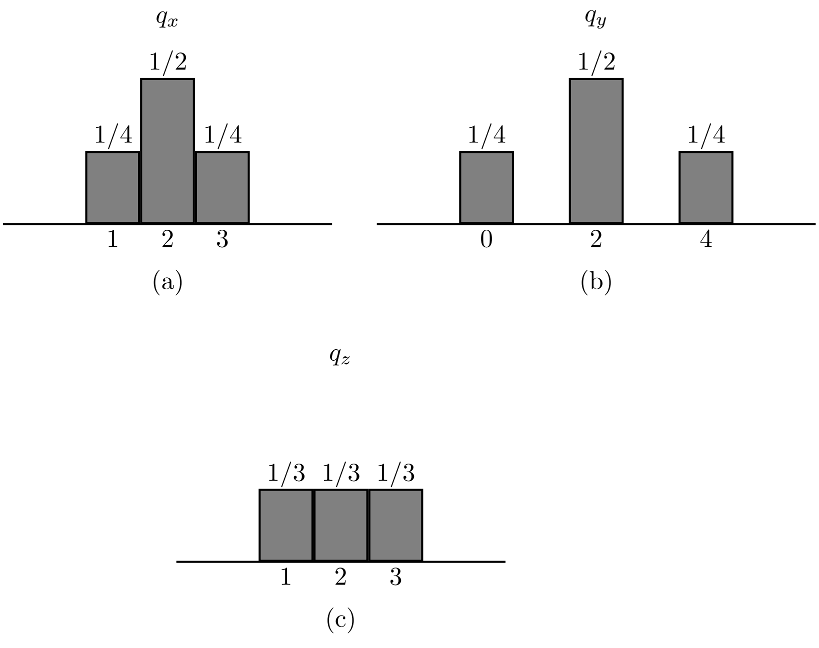


Figure 2.17: ***Three p.m.f.'s with different variances***

Ans:

- $\sigma_X^2 = \frac{1}{4}(1-2)^2 + \frac{1}{2}(2-2)^2 + \frac{1}{4}(3-2)^2 = \frac{1}{2}$
- $\sigma_Y^2 = \frac{1}{4}(0-2)^2 + \frac{1}{2}(2-2)^2 + \frac{1}{4}(4-2)^2 = 2$
- $\sigma_Z^2 = \frac{1}{3}(1-2)^2 + \frac{1}{3}(2-2)^2 + \frac{1}{3}(3-2)^2 = \frac{2}{3}$ □

Question 2.5.4 *Referring to the last example, compute the variance for a distribution that disperses probability in both senses; that is, the states are* $0, 2,$ *and* 4*, and the probabilities are each* $1/3$*. How does it compare to the three preceding variances?*

Ans:

- $\sigma^2 = \frac{1}{3}(0-2)^2 + \frac{1}{3}(2-2)^2 + \frac{1}{3}(4-2)^2 = \frac{8}{3} = 2^2 \cdot \frac{2}{3}$
- Maximum. □

Useful formula: $\sigma^2 = E[(X-\mu)^2] = E[X^2 - 2\mu X + \mu^2] = E[X^2] - \mu^2$

Example 2.5.6 *Let* X *be a random variable with the uniform distribution on* $\{1, 2, 3, 4, 5, 6\}$*.*

Ans: The mean, variance, and standard deviation of X are

$$\mu = E[X] = \frac{1}{6}(1+2+3+4+5+6) = \frac{7}{2}$$
$$\sigma^2 = E[X^2] - \mu^2 = \frac{35}{12}$$
$$\sigma = \sqrt{\sigma^2} = \sqrt{35/12} \approx 1.71$$

□

Proposition 2.5.3 *If* X *is a real random variable with finite variance and* a *and* b *are real constants, then*

$$Var(aX + b) = a^2\, Var(X).$$

Proof:

- The mean μ_y of the random variable $Y = aX + b$ is $\mu_y = E[aX + b] = a\mu_x + b$, where μ_x is the expected value of X.
- Then,

$$\begin{aligned} \mathrm{Var}(aX+b) &= E[(Y-\mu_y)^2] \\ &= E[((aX+b)-(a\mu_x+b))^2] \\ &= E[(a(X-\mu_x))^2] \\ &= a^2E[(X-\mu_x)^2] \\ &= a^2\mathrm{Var}(X) \end{aligned}$$

□

Question 2.5.5 *What is the variance of a constant random variable* $X(\omega) = b, \forall\omega \in \Omega$*?*

Ans:

- $\mathrm{Var}(X) = E[X]^2 - b^2 = b^2 - b^2 = 0$ □

- **Sample mean**: $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$

- ***Variance of $\overline{X}$***:

$$\operatorname{Var}(\overline{X}) = \operatorname{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n}\operatorname{Var}(X_i) = \frac{1}{n^2}\sum_{i=1}^{n}\sigma^2 = \frac{1}{n}\sigma^2$$

- $\operatorname{Var}(\overline{X}) \xrightarrow{n\to\infty} 0$

Definition 2.5.3 *The rth* ***moment*** *of the distribution of a real random variable X is $E[X^r]$, provided the expectation exists. Similarly the rth* ***central moment or moment about the mean*** *is $E[(X-\mu)^r]$.*

- The first moment of the distribution of X is the ***mean***.
- The second moment about the mean is the ***variance*** of the distribution.
- The third moment about the mean $E[(X-\mu)^3]$, called the ***skewness***.

Example 2.5.7 *Consider the three probability distribution displayed in Figure 2.18.*

Ans:

- You can verify easily that the means are

$$E[X] = 3, \quad E[Y] = 37/15, \quad E[Z] = 53/15.$$

- The first distribution is symmetric about its mean, the second is asymmetric in the sense that it has a long right tail, and the third is asymmetric in the sense that it has a long left tail.
- The distribution in part (b) would be called ***skewed to the right or positively skewed***, and the distribution in part (c), ***skewed to the left or negatively skewed***.

$$E\left[(X-\mu_x)^3\right] = 0$$
$$E\left[(Y-\mu_y)^3\right] = 1.08 > 0 \qquad \text{(Skewed to the right)}$$
$$E\left[(Z-\mu_z)^3\right] = -1.08 < 0 \qquad \text{(Skewed to the left)}$$

- Most authors prefer a standardized measure of skewness:

$$\frac{E[(X-\mu_x)^3]}{\sigma^3}$$

2.5.3 Special distributions

List the means and variances of some of the important discrete distributions.

Proposition 2.5.4 *If X has the Bernoulli distribution with success parameter p, then*

$$E[X] = p; \quad Var(X) = p(1-p).$$

Proof:

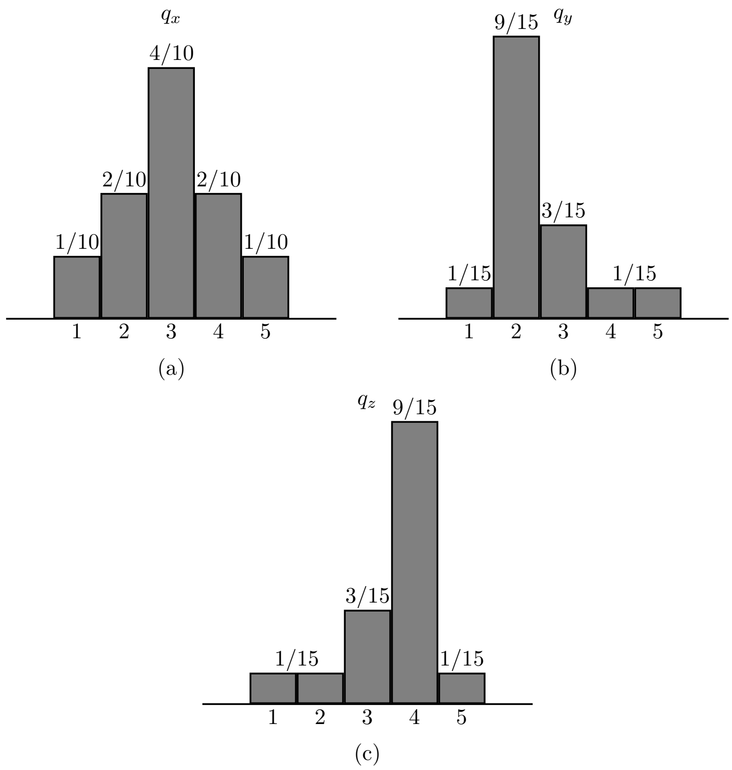


Figure 2.18: ***Three p.m.f.'s with different skewness***

- $E[X] = 0 \cdot (1-p) + 1 \cdot p = p$
- $\text{Var}(X) = E[X^2] - (E[X])^2 = 0^2 \cdot (1-p) + 1^2 \cdot p - p^2 = p - p^2 = p(1-p)$ □

Proposition 2.5.5 *If X has the $b(n,p)$ distribution, then*

$$E[X] = np \quad \textit{and} \quad Var(X) = np(1-p).$$

Proof:

- By the definition of expected value,

$$\begin{aligned} E[X] &= \sum_{k=0}^{n} k \cdot \binom{n}{k} p^k (1-p)^{n-k} \\ &= n \cdot p \cdot \sum_{k=1}^{n} \frac{(n-1)!}{(k-1)!(n-k)!} p^{k-1}(1-p)^{n-k} \\ &= n \cdot p \cdot \sum_{l=0}^{n-1} \frac{(n-1)!}{l!(n-1-l)!} p^l (1-p)^{n-1-l} \\ &= n \cdot p \cdot 1 \end{aligned}$$

- In the second line the $k = 0$ term is dropped, n and p are factored out of the sum, and k is canceled from the $k!$ factor in the denominator.
- The third line results from substituting $l = k - 1$.

- The sum compresses to 1 in the fourth line because it is the sum of all values of the $b(n-1,p)$ probability mass function.
- To find the variance, we take the circuitous, but surprisingly helpful route of finding $E[X(X-1)]$.
- This can be done very similarly to the mean computation:

$$\begin{aligned}
E[X(X-1)] &= \sum_{k=0}^{n} k(k-1)\cdot\binom{n}{k}p^k(1-p)^{n-k} \\
&= n(n-1)p^2\cdot\sum_{k=2}^{n}\frac{(n-2)!}{(k-2)!(n-k)!}p^{k-2}(1-p)^{n-k} \\
&= n(n-1)p^2\cdot\sum_{l=0}^{n-2}\frac{(n-2)!}{l!(n-2-l)!}p^{l}(1-p)^{n-2-l} \\
&= n(n-1)p^2
\end{aligned}$$

- (Make sure that you understand each line of this derivation.) Now we can compute

$$\begin{aligned}
\mathrm{Var}(X) &= E[X^2]-(E[X])^2 \\
&= E[X(X-1)]+E[X]-(E[X])^2 \\
&= n(n-1)p^2+np-(np)^2 \\
&= n^2p^2-np^2+np-n^2p^2 \\
&= np-np^2 \\
&= np(1-p) \qquad \square
\end{aligned}$$

- Note that for a multinomial random variable $\mathbf{X}$ with parameters n, $p_1, p_2, \ldots, p_k$, the marginal distribution of X_i is $b(n, p_i)$; hence the expected number of objects of category i in a multinomial experiment is np_i.
- This fact will become important later when we discuss goodness-of-fit tests.

Question 2.5.6 *Give an alternative proof of the mean and variance formulas for the binomial distribution using the results for the Bernoulli distribution, linearity, and independence.*

Ans:

- $X = X_1 + X_2 + \cdots + X_n$ where $X_i \overset{\text{iid}}{\sim} \text{Bernoulli}(p)$

$$E[X] = E\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} E[X_i] = np$$

$$\mathrm{Var}(X) = \mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n}\mathrm{Var}(X_i) = np(1-p) \qquad \square$$

Example 2.5.8 *A contractor submits bids on ten building projects, on each of which four other contractors will also bid. How many bids would the contractor expect to win? State assumptions carefully.*

Ans:

- There is a sequence of ten trials, one for each project.
- A success corresponds to the event that the contractor in question gets a bid.
- Under the assumptions that awards on different projects are independent of one another, and that each bidder is as likely as any other to win the contract, we have a binomial experiment with ten trials and success probability per trial of 1/5.
- If $P(X_i = 1) = 1/5$, then $\sum_{i=1}^{10} E[X_i] = 10 \cdot 1/5 = 2$.
- If $P(X_i = 1) = 1/4$, then $\sum_{i=1}^{10} E[X_i] = 10 \cdot 1/4 = 5/2$. □

Proposition 2.5.6 *If X has the Poisson(λ) distribution, then*

$$E[X] = \lambda \quad \text{and} \quad Var(X) = \lambda.$$

Proof:

- The Poisson distribution is another distribution amenable to the sort of computational tricks we used on the binomial distribution.
- First, the expected value is

$$\begin{aligned} E[X] &= \sum_{n=0}^{\infty} n \cdot \frac{e^{-\lambda}\lambda^n}{n!} \\ &= \lambda \cdot \sum_{n=1}^{\infty} \frac{e^{-\lambda}\lambda^{n-1}}{(n-1)!} \\ &= \lambda \cdot \sum_{m=0}^{\infty} \frac{e^{-\lambda}\lambda^m}{m!} \\ &= \lambda \end{aligned}$$

- In the second line a common factor of λ is pulled out, the $n = 0$ term, which is 0, is dropped, and an n is canceled from $n!$.
- The third line changes the variable of summation to $m = n - 1$.
- Then the resulting series sums to 1, since it is the sum of all values of the Poisson(λ) p.m.f.
- A similar trick works for the variance; first,

$$\begin{aligned} E[X(X-1)] &= \sum_{n=0}^{\infty} n(n-1) \cdot \frac{e^{-\lambda}\lambda^n}{n!} \\ &= \lambda^2 \cdot \sum_{n=2}^{\infty} \frac{e^{-\lambda}\lambda^{n-2}}{(n-2)!} \\ &= \lambda^2 \cdot \sum_{m=0}^{\infty} \frac{e^{-\lambda}\lambda^m}{m!} \\ &= \lambda^2 \end{aligned}$$

- Then as in the derivation for the binomial distribution,

$$\text{Var}(X) = E[X(X-1)] + E[X] - (E[X])^2 = \lambda^2 + \lambda - \lambda^2 = \lambda.$$

- Referring back to Section 2.4, since λ is the mean of the Poisson distribution, it is reasonable to estimate it by the sample mean of a collection of n observations.
- First, the larger n is, the more precise the estimate is, that is, the smaller is its variance.
- Second, since λ is both the mean and the variance of the distribution, as λ grows the probability weight not only shifts to the right but also spreads out, as we noticed in an earlier figure.
- Third, recall that the number N_t of arrivals in a Poisson process by time t has the Poisson distribution with parameter λt.
- We now know that λt is the expected number of arrivals by time t; hence λ itself is the expected number of arrivals per unit time.
- This is another reason for calling λ the arrival rate of the process. □

Proposition 2.5.7 *If X has the geometric distribution with parameter p, then*

$$E[X] = \frac{1}{p} \quad \textit{and} \quad Var(X) = \frac{1-p}{p^2}.$$

If X has the negative binomial distribution with parameters r and p, then

$$E[X] = \frac{r}{p} \quad \textit{and} \quad Var(X) = r \cdot \frac{1-p}{p^2}.$$

Proof:

- For the geometric distribution

$$\begin{aligned} E[X] &= \textstyle\sum_{k=1}^{\infty} k \cdot p \cdot (1-p)^{k-1} \\ &= p \cdot \frac{\mathrm{d}}{\mathrm{d}x} [\textstyle\sum_{k=1}^{\infty} x^k] \\ &= p \cdot \frac{\mathrm{d}}{\mathrm{d}x} \left[\frac{1}{1-x} - 1\right]_{x=1-p} \\ &= p \cdot \left[\frac{1}{(1-x)^2}\right]_{x=1-p} \\ &= \frac{p}{p^2} \\ &= \frac{1}{p} \end{aligned}$$

- A similar approach yields the following expectation.

$$\begin{aligned} E[X(X-1)] &= \textstyle\sum_{k=1}^{\infty} k(k-1) \cdot p \cdot (1-p)^{k-1} \\ &= p \cdot (1-p) \cdot \textstyle\sum_{k=2}^{\infty} k(k-1)(1-p)^{k-2} \\ &= p \cdot (1-p) \cdot \frac{d^2}{\mathrm{d}x^2} \left[\textstyle\sum_{k=2}^{\infty} x^k\right]_{x=1-p} \\ &= p \cdot (1-p) \cdot \frac{d^2}{\mathrm{d}x^2} \left[\frac{1}{1-x} - 1 - x\right]_{x=1-p} \\ &= p \cdot (1-p) \cdot \left[\frac{2}{(1-x)^3}\right]_{x=1-p} \\ &= \frac{2(1-p)}{p^2} \end{aligned}$$

- Therefore the variance of the geometric distribution is

$$\begin{aligned}\text{Var}(X) &= E[X(X-1)] + E[X] - (E[X])^2 \\ &= \tfrac{2(1-p)}{p^2} + \tfrac{p}{p^2} - \tfrac{1}{p^2} \\ &= \tfrac{(1-p)}{p^2}\end{aligned}$$

- Since a negative binomial random variable with parameters r and p is the sum of r independent geometric random variables with parameter p (because it is the total all r intersuccess times in the Bernoulli trial sequence) its mean and variance are just r times those of the geometric distribution.

- $X = X_1 + X_2 + \cdots + X_r$ where $X_i \overset{\text{iid}}{\sim} \text{Geometric}(p)$

$$E[X] = E[X_1 + X_2 + \cdots + X_r] = r\frac{1}{p}$$

$$\text{Var}(X) = \sum_{i=1}^{r} \text{Var}(X_i) = r \cdot \frac{1-p}{p^2}$$

□

2.5.4 Multivariate expectation

Definition 2.5.4 *If $\mathbf{X} = [X_1\ X_2\ \ldots\ X_n]'$ is a random vector, then the* ***expected value*** *of $\mathbf{X}$ is the vector*

$$E[\mathbf{X}] = \begin{pmatrix} E[X_1] \\ E[X_2] \\ \vdots \\ E[X_n] \end{pmatrix}$$

provided the individual component expectations exist.

$$\begin{aligned}E[X_1] &= \sum_{x_1}\sum_{x_2}\cdots\sum_{x_n} x_1 \cdot q(x_1, x_2, \ldots, x_n) \\ &= \sum_{x_1} x_1 \sum_{x_2}\cdots\sum_{x_n} q(x_1, x_2, \ldots, x_n) \\ &= \sum_{x_1} x_1 \cdot q_1(x_1)\end{aligned}$$

Example 2.5.9 *If X_1 and X_2 have the discrete uniform distribution on the triangular integer grid shown in Figure 2.19, find $E[X_2]$ and $E[X_1 \cdot X_2]$.*

Ans:

- Ten equally likely points (i, j) in the state space of $\mathbf{X} = (X_1, X_2)$, each of which have probability $q(i, j) = 1/10$.

$$\begin{aligned}
E[X_2] &= \sum_i \sum_j j \cdot (1/10) \\
&= (1+1+1+1+2+2+2+3+3+4)/10 = 2 \\
E[X_1 X_2] &= \sum_i \sum_j i \cdot j \cdot (1/10) \\
&= \frac{1\cdot 1 + 2\cdot 1 + 3\cdot 1 + 4\cdot 1 + 1\cdot 2 + 2\cdot 2 + 3\cdot 2 + 1\cdot 3 + 2\cdot 3 + 1\cdot 4}{10} \\
&= 35/10
\end{aligned}$$

□

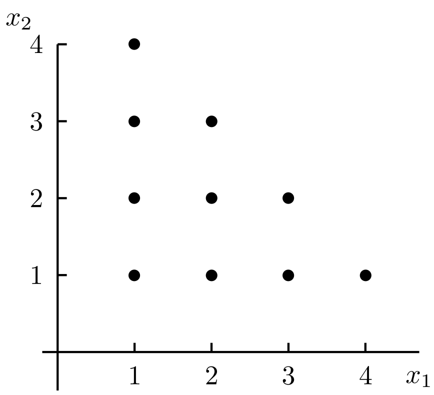


Figure 2.19: ***A two-variable discrete uniform distribution***

Question 2.5.7 *In Example 2.5.9, what is* $E[\mathbf{X}]$*?*

Ans:

- $E[\mathbf{X}] = \left[\frac{1}{10}\ \frac{1}{10}\right]'$ □

Proposition 2.5.8 (*Linearity of expectation*)

1. *Let* $\mathbf{X}$ *and* $\mathbf{Y}$ *be random vectors of the same dimension* n*, and let* a *and* b *be scalar constants. Then,*
$$E[a\mathbf{X} + b\mathbf{Y}] = aE[\mathbf{X}] + bE[\mathbf{Y}]$$
provided the expectations of $\mathbf{X}$ *and* $\mathbf{Y}$ *exist.*

2. *Let* $\mathbf{X}$ *be an* n*-dimensional random vector and let* A *be an* $m \times n$ *matrix of constants. Then,*
$$E[A\mathbf{X}] = AE[\mathbf{X}] \tag{2.64}$$
provided the expectation of $\mathbf{X}$ *exists.*

Proof:

(a) See Exercise 25

(b) It will be helpful to think of matrix multiplication in the following way.

- Let A be a matrix as hypothesized in part (b) and let $\mathbf{b}$ be an n-dimensional column vector.
- Denote the rows of A as $\mathbf{a}_1$, $\mathbf{a}_2$, ..., $\mathbf{a}_m$.
- Then the product $A\mathbf{b}$ has m rows $\mathbf{a}_1 \cdot \mathbf{b}$, $\mathbf{a}_2 \cdot \mathbf{b}$,..., $\mathbf{a}_m \cdot \mathbf{b}$.
- The symbol "$\cdot$" indicates vector dot product here.
- Therefore

$$E[A\mathbf{X}] = E\left[\begin{bmatrix} \mathbf{a_1} \cdot \mathbf{X} \\ \mathbf{a_2} \cdot \mathbf{X} \\ \vdots \\ \mathbf{a_m} \cdot \mathbf{X} \end{bmatrix}\right] = \begin{bmatrix} E[\mathbf{a_1} \cdot \mathbf{X}] \\ E[\mathbf{a_2} \cdot \mathbf{X}] \\ \vdots \\ E[\mathbf{a_m} \cdot \mathbf{X}] \end{bmatrix}$$

- It is easy to prove (see Exercise 24) that

$$E[\mathbf{a_i} \cdot \mathbf{X}] = \mathbf{a_i} \cdot E[\mathbf{X}] \tag{2.65}$$

hence

$$E[A\mathbf{X}] = \begin{bmatrix} \mathbf{a_1} \cdot E[\mathbf{X}] \\ \mathbf{a_2} \cdot E[\mathbf{X}] \\ \vdots \\ \mathbf{a_m} \cdot E[\mathbf{X}] \end{bmatrix} = A \cdot E[\mathbf{X}]$$

- We will find the matrix version of linearity to be particularly useful later in the book when we study regression and experimental design.
- As a brief note on the use of (2.64), we can find in one computation the expected values of several linear combinations of the components of a vector random variable.
- Suppose $\mathbf{X} = [X_1\ X_2\ X_3]'$ has expected value $[0\ 2\ -1]'$.
- Then the expected values of $X_1 - X_2, X_2 - X_3$, and $2X_1 + X_2 + X_3$ can be found by computing as follows:

$$\begin{aligned} E\left[\begin{bmatrix} X_1 - X_2 \\ X_2 - X_3 \\ 2X_1 + X_2 + X_3 \end{bmatrix}\right] &= E\left[\begin{bmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \\ 2 & 1 & 1 \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix}\right] \\ &= \begin{bmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \\ 2 & 1 & 1 \end{bmatrix} \begin{bmatrix} 0 \\ 2 \\ -1 \end{bmatrix} \\ &= \begin{bmatrix} -2 \\ 3 \\ 1 \end{bmatrix} \end{aligned}$$

□

2.6 Summary

1. ***Fundamental counting principle***:

   (a) ***Rule of product***: Suppose that an experiment has two stages. For the first stage, there are m possible outcomes, and for each of these, the second stage has n possible outcomes. Then the two-stage experiment has $m \cdot n$ outcomes.

(b) ***Rule of sum***: For a more general two-stage experiment, let the first-stage outcomes be labeled $i = 1, 2, \ldots, m$. Assume that if the first-stage outcome is i, then there are n_i possible outcomes for stage 2. Then the two-stage experiment has

$$\sum_{i=1}^{m} n_i$$

possible outcomes.

(c) ***Rule of product***: Suppose that an experiment consists of k stages such that the first stage has m_1 possible outcomes, for each outcome of stage 1 there are m_2 possible outcomes of stage 2, for each combined outcome of the first two stages there are m_3 possible outcomes of stages 3, and so on. Then there are $m_1 \cdot m_2 \cdots m_k$ outcomes of the entire experiment.

. . . . 7

2. ***Permutation***: A permutation of n objects $\{y_1, \ldots, y_n\}$, taken r at a time is an ordered list $(x_1, \ldots, x_r)$ selected from the original n objects, such that $x_i \neq x_j, \forall i \neq j$. We denote the number of such permutations by $P_{n,r}$. 14

3. ***Combination***: A combination of n items $\{y_1, \ldots, y_n\}, r$ at a time, is a subset $\{x_1, \ldots, x_r\}$ selected from the original n items, such that $x_i \neq x_j, \forall i \neq j$. We denote the number of such combinations by $C_{n,r}$, or $\binom{n}{r}$. The latter is read "n choose r." 17

4. $P_{n,r} = n(n-1)\cdots(n-r+1) = \frac{n!}{(n-r)!}$ 19

5. $C_{n,r} = \binom{n}{r} = \frac{n!}{r!\cdot(n-r)!}$ 20

6. ***Probability mass function (p.m.f.)***: $q(x) = P[X = x] \quad \text{for } x \in E$. 26

7. ***Cumulative distribution function (c.d.f.)***: (Discrete case) $F(x) = P[X \leq x] = \sum_{t \leq x} q(t)$. 26

8. ***Discrete uniform distribution***: $q(x) = \frac{1}{n}, x \in \{x_1, x_2, \ldots, x_n\}$ 27

9. ***Empirical probability mass function (emf)***: The empirical cumulative distribution function (edf) of the sample is then the c.d.f. associated with $\hat{q}$ through equation (2.7), alternatively, $\hat{F}(w_j) = \frac{\text{number of } X_i = w_j}{n}$. 29

10. ***Empirical cumulative distribution function (edf)***: The empirical probability mass function (emf) of the sample is $\hat{q}(w_j) = \frac{\text{number of } X_i \leq w_j}{n}$. 29

11. ***Hypergeometric distribution***: The experiments involving sampling without replacement discussed in the last section give rise to a frequently observed distribution. 31

12. ***Marginal mass function***: The marginal mass function of X_i is $q_i(x_i) = P[X_i = x_i] = \sum_{x_1} \cdots \sum_{x_{i-1}} \sum_{x_{i+1}} \cdots \sum_{x_n} q(x_1, \ldots, x_n)$. 38

13. ***Joint marginal mass function***: The joint marginal mass function of a subcollection $X_{i_1}, \ldots, X_{i_k}$ of the random variables is $q(x_{i_1}, \ldots, x_{i_k}) = P[X_{i_1} = x_{i_1}, \ldots, X_{i_k} = x_{i_k}] = \sum \cdots \sum q(x_1, \ldots, x_n)$. 38

14. ***Binomial probability mass function***: Let X be the total number of successes in n Bernoulli trials. Then X has the binomial probability mass function: $q(k) = P[X = k] = \binom{n}{k} p^k (1-p)^{n-k}$. 46

15. ***Geometric distribution***: Let T_1 be the random variable that returns the trial on which the first success occurs in a sequence of Bernoulli trials. Then T_1 has the geometric distribution $P[T_1 = n] = (1-p)^{n-1}p, n = 1, 2, \ldots$ 51

16. ***Geometric distribution without memory***:

$$P[X > n + k | X > n] = P[n > k] = (1-p)^k.$$

. . . . 54

17. ***Negative binomial distribution***: Let T_r be the trial on which the rth success occurs in a sequence of Bernoulli trials. Then T_r has the negative binomial distribution: $P[T_r = n] = \binom{n-1}{r-1} p^r (1-p)^{n-r}, n = r, r+1, \ldots$ 54

18. If $X_1, \ldots, X_k$ have the $m(n, p_1, \ldots, p_k)$ distribution, then for each i, X_i has the $b(n, p_i) = m(n, p_1, 1 - p_1)$ distribution. 64

19. If $X_1, \ldots, X_k$ have the $m(n, p_1, \ldots, p_k)$ distribution, then for $X_{i_1}, X_{i_2}, \ldots, X_{i_m}$ has the $m(n, p_{i_1}, p_{i_2}, \ldots, p_{i_m}, 1 - \sum_{j=1}^m p_{i_j})$ distribution. 67

20. ***Poisson probability mass function***: The Poisson probability mass function with parameter λ is $q(k) = P[X = k] = \frac{e^{-\lambda}\lambda^k}{k!}, k = 0, 1, 2, \ldots$ 74

21. ***Poisson process***: A Poisson process is a family $(N_t)_{t \geq 0}$ of random variables whose paths are step functions beginning at state 0 at time 0, which jump by 1 at a sequence of random times $T_1, T_2, \ldots$. Additionally, the some conditions are assumed. If $(N_t)_{t \geq 0}$ is a Poisson process with rate parameter λ, then $P[N_t = n] = \frac{e^{-\lambda t}(\lambda t)^n}{n!}$. 82

22. ***Poisson(λt) distribution***: If $(N_t)_{t \geq 0}$ is a Poisson process with rate parameter λ, then $P[N_t = n] = \frac{e^{-\lambda t}(\lambda t)^n}{n!}$, that is, the number of arrivals up through time t has the Poisson(λt) distribution. 84

23. ***Expected value or expectation***: Let X be a discrete, real-value random variable with state space $E = \{e_1, e_2, \ldots\}$ and probability mass function q. Then the expected value or expectation of X is $\mu = E[X] = \sum_i e_i P[X = e_i] = \sum_i e_i q(e_i)$ provided the series converges. 95

24. ***Mean***: $\mu = E[X]$ 95

25. If c is a real constant, then $E[c] = c$. 107

26. ***Linearity of expectation***: If X and Y are random variables with finite expectation, then $E[aX + bY] = aE[X] + bE[Y]$. 110

27. ***Variance***: The variance of a real random variable X is $\sigma^2 = \text{Var}(X) = E[(X - \mu)^2]$ provided the expectation is finite. 114

28. ***Standard deviation***: The square root of the variance, σ, is referred to as the standard deviation of the random variable. 114

29. $\sigma^2 = E[X^2] - \mu^2$. 116

30. If X is a real random variable with finite variance and a and b are real constants, then $\text{Var}(aX + b) = a^2\text{Var}(X)$. 117

31. ***Sample mean***: $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ 118

32. ***Variance of $\overline{X}$***: $\text{Var}(\overline{X}) = \frac{1}{n}\sigma^2$ 119

33. ***Moment***: The rth moment of the distribution of a real random variable X is $E[X^r]$, provided the expectation exists. 119

34. ***Central moment or moment about the mean***: The rth central moment or moment about the mean is $E[(X - \mu)^r]$. 119

35. ***Skewness***: The third moment about the mean $E[(X - \mu)^3]$. 120

36. ***Skewed to the right***: A distribution has a long right tail. 122

37. ***Skewed to the left***: A distribution has a long left tail. 122

38. If X has the Bernoulli distribution with success parameter p, then $E[X] = p$; $\text{Var}(X) = p(1 - p)$. 123

39. If X has the $b(n, p)$ distribution, then $E[X] = np$ and $\text{Var}(X) = np(1 - p)$. . 123

40. If X has the Poisson(λ) distribution, then $E[X] = \lambda$; $\text{Var}(X) = \lambda$. 130

41. If X has the geometric distribution with parameter p, then $E[X] = \frac{1}{p}$ and $\text{Var}(X) = \frac{1-p}{p^2}$. If X has the negative binomial distribution with parameters r and p, then $E[X] = \frac{r}{p}$ and $\text{Var}(X) = r \cdot \frac{1-p}{p^2}$. 134

42. If $\mathbf{X} = [X_1 \ X_2 \ \dots \ X_n]'$ is a random vector, then the expected value of $\mathbf{X}$ is the vector $E[\mathbf{X}] = \begin{pmatrix} E[X_1] \\ E[X_2] \\ \vdots \\ E[X_n] \end{pmatrix}$ provided the individual component expectations exist. 139

43. ***Linearity of expectation***: Let $\mathbf{X}$ and $\mathbf{Y}$ be random vectors of the same dimension n, and let a and b be scalar constants. Then, $E[a\mathbf{X} + b\mathbf{Y}] = aE[\mathbf{X}] + bE[\mathbf{Y}]$ provided the expectations of $\mathbf{X}$ and $\mathbf{Y}$ exist. 143

44. Let $\mathbf{X}$ be an n-dimensional random vector and let A be an $m \times n$ matrix of constants. Then, $E[A\mathbf{X}] = AE[\mathbf{X}]$ provided the expectation of $\mathbf{X}$ exists. 143

Chapter 3

Continuous Probability

Contents

3.1 Densities and distribution functions

3.1.1 Motivation and review

This chapter will parallel the development of Chap. 2 for the family of ***continuous distributions***, those whose probabilistic law is characterized by a density function. Actually the word ***continuous*** might more properly be replaced by ***differentiable***, since the common feature of members of this class is that their c.d.f.'s have derivatives.

- For continuous distributions, the way in which probabilities are calculated is very analogous to the calculation of the physical mass of objects that are not necessarily uniformly dense.
- Consider a cylindrical bar of cross-sectional area one, which has a continuous density $f(x)$ at the point at a distance of x from its left end. (See Fig. 3.1.)
- Over a short interval $[x, x+\Delta x]$ the bar is approximately uniformly dense with density $f(x)$.

- The mass of the bar in this interval is approximately its density times its volume,

$$f(x) \cdot \Delta x \cdot 1,$$

since the slice is a cylinder with height Δx and cross-sectional area one.

- The mass of the bar over an interval $[a, b]$ can be approximated closely by breaking the interval into many small intervals $[x_i, x_{i+1}]$, $i = 0, 1, \ldots, n-1$, where $x_{i+1} = x_i + \Delta x$.

- The total mass of the bar on $[a, b]$ is the sum of the masses on all the subintervals, which is approximately

$$\sum_{i=0}^{n-1} f(x_i)\Delta x.$$

- It is known from calculus that; because f is continuous, the preceding sum converges to

$$\int_a^b f(x)\,\mathrm{d}x,$$

as the number of subintervals approaches infinity.

- Thus the mass of the bar in $[a, b]$ is the definite integral of its density function over $[a, b]$.

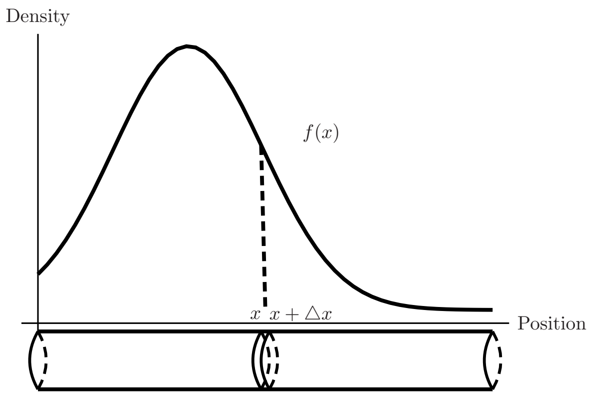


Figure 3.1: ***Cross-sectional area of bar*** $= 1$

- Let $E = \{x_1, x_2, \ldots, x_n\}$ be a state space of equally spaced points on the real line, with spacing Δx.

- Let probability masses $p_1, p_2, \ldots, p_n$ be placed on these points.

- Construct a histogram whose bars have height $p_i/\Delta x$ and whose bases, of length Δx, are centered about the points x_i, as in Fig. 3.2.

- Note that the heights are then probabilities per unit length, which brings to mind the density analogy.

- The area of the ith rectangle is

$$\frac{p_i}{\Delta x} \cdot \Delta x = p_i = P[X = x_i]\,.$$

- For very large n and very small spacing Δx, the state space resembles a continuous interval and, if the assignment of probability is smooth enough, the step function formed by the tops of the rectangles resembles a continuous function.
- The area under that function in an interval is approximately the total area of the thin rectangles within that interval, which is the probability that the random variable associated with the distribution falls into the interval.

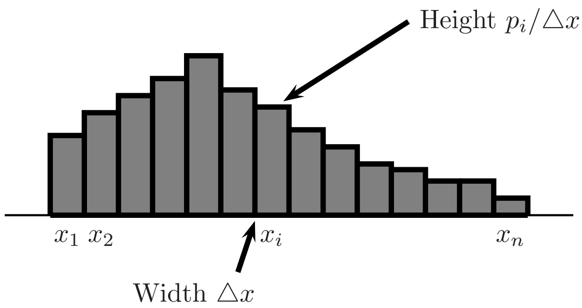


Figure 3.2: ***Approximating a continuous distribution***

- $P[x < X \leq x + \triangle x] = F(x + \triangle x) - F(x)$
- $\lim_{\triangle x \to 0} \frac{F(x+\triangle x)-F(x)}{\triangle x} = f(x)$ (density)

Question 3.1.1 *What would have happened to the foregoing argument if the histogram rectangles were given a height of p_i instead of $p_i/\triangle$? What would happen to the histogram as the number of states increases?*

Ans:

- $P[X = x_i] = p_i \cdot \Delta x$
- If the number of states increases, Δx is very small, then the state space resembles a continuous interval and, if the assignment of probability is smooth enough, the step function formed by the tops of the rectangles resembles a continuous function. □

- A random variable X has probability density function (p.d.f.) f if its probability distribution is

$$Q(A) = P[X \in A] = \int_A f(x)\,\mathrm{d}x.$$

  - $f(x) \geq 0, \forall x \in E$
  - $\int_E f(x)\,\mathrm{d}x = 1$
  - $P[X = c] = \int_c^c f(x)\,\mathrm{d}x = 0$

- ***Cumulative distribution function*** (c.d.f.) $F(x)$:

$$F(x) = P[X \leq x] = \int_{-\infty}^{x} f(t)\,\mathrm{d}t.$$

- ▷ $F'(x) = f(x)$.
- ▷ F is a nondecreasing, nonnegative function.
- ▷ $\lim_{x\to-\infty} F(x) = 0$
- ▷ $\lim_{x\to\infty} F(x) = 1$

$$\begin{aligned} P[a < X \le b] &= P[a \le X \le b] = P[a \le X < b] \\ &= P[a < X < b] = F(b) - F(a) \end{aligned}$$ □

Example 3.1.1 (*Mixture distribution*) *Some probability distributions are hybrids of the two classes that we have considered, discrete and continuous. Consider this simple reliability replacement scenario. The failure time T (in months) of a streetlight bulb has the probability density*

$$f(t) = \frac{1}{5}\, e^{-1/5t}, \qquad t \in (0, \infty).$$

The local city manager has a plan whereby streetlight bulbs are replaced either in 6 months or when they fail, whichever comes first. Find the c.d.f. of the random variable S, the replacement time.

Ans:

- ▶

$$\begin{aligned} F(s) &= P[S \le s] = P[T \le s] \\ &= \begin{cases} 0 & \text{if } s < 0, \\ 1 - e^{-s/5} & \text{if } 0 \le s < 6, \\ 1 & \text{if } s \ge 6. \end{cases} \end{aligned}$$

- ▶ $P[S = 6] = F(6) - F(6^-) = 1 - (1 - e^{-6/5}) = e^{-6/5} = P[T \ge 6]$
- ▶ The graph of the c.d.f. of S is shown in Fig. 3.3. □

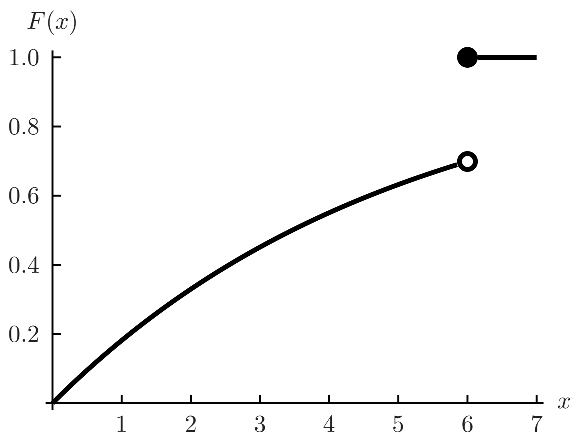


Figure 3.3: ***Hybrid discrete, continuous distribution***

3.1.2 Uniform and empirical distributions

Example 3.1.2 (***Continuous uniform distribution***) *Suppose that a friend tells you that he will meet you for lunch at a restaurant at noon, plus or minus ten minutes. What is the probability that your companion will arrive between 12:05 and 12:08?*

Ans:

- Make some assumptions about the probability distribution of the arrival time.
- Start by labeling noon as time 0 minutes.
- Then the state space E is the time interval $[-10, 10]$. If no time is more likely than any other, then probability ought to be distributed in a uniform fashion across the states.
- $f(t) = c$ for $t \in [-10, 10]$; $f(t) = 0$ otherwise.
- $1 = \int_{-10}^{10} f(t)\,\mathrm{d}t = 20c$ gives $c = 1/20$.
- $P[T \in [5, 8]] = \int_5^8 1/20\,\mathrm{d}t = 3/20$.
- $F(t) = \int_{-10}^{t} 1/20\,\mathrm{d}u = (t + 10)/20, \quad t \in [-10, 10]$.
- $F(t) = 0$ if $t < -10$ and $F(t) = 1$ if $t > 10$.
- $P[T \in [5, 8]] = F(8) - F(5) = 3/20$
- The density and c.d.f. are sketched in Fig. 3.4. □

Continuous uniform density on $[a, b]$:

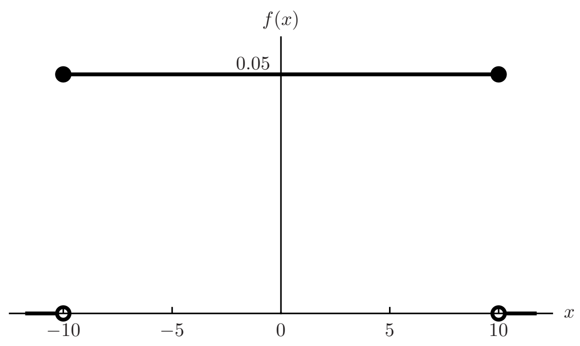
- p.d.f.: $f(x) = \begin{cases} \frac{1}{b-a} & \text{if } x \in [a, b], \\ 0 & \text{otherwise.} \end{cases}$
- c.d.f.: $F(x) = \begin{cases} 0 & \text{if } x < a, \\ \frac{x-a}{b-a} & \text{if } x \in [a, b], \\ 1 & \text{if } x > b. \end{cases}$

Special terms for the center and the spread of the distribution:

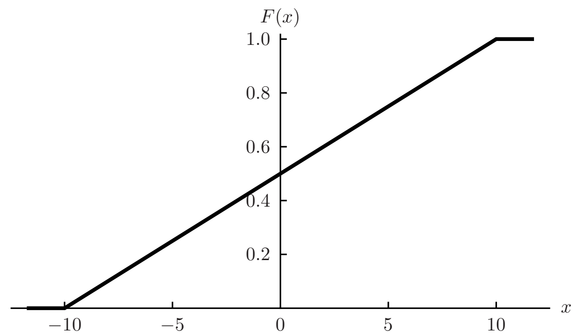
- ***Median*** m of X if $P[X \le m] = F(m) = 1/2$.
- $p \times 100$***th percentile*** x_p if $P[X \le x_p] = F(x_p) = p$.

Uniform distribution on $[0, c]$:

- Median: $1/2 = \int_0^m 1/c\,\mathrm{d}x = m/c$ gives $m = c/2$.
- $p \times 100$th percentile: $p = \int_0^{x_p} 1/c\,\mathrm{d}x = x_p/c$ gives $x_p = cp$. (Fig. 3.5.)
- The median is the 50th percentile of the distribution.



(a) Density



(b) c.d.f.

Figure 3.4: ***A continuous uniform distribution***

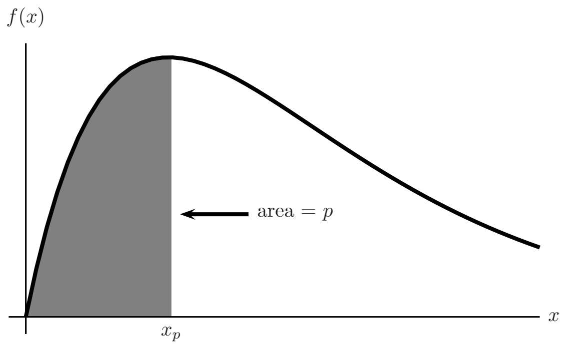


Figure 3.5: ***Percentiles of a distribution***

Empirical distribution function based on a sample of observations $X_1, \ldots, X_n$ from a given distribution is the same in the continuous case as in the discrete case:

$$\widehat{F}_n(x) = \frac{\text{number of } X_i's \leq x}{n}.$$

There are theoretical reasons to believe that for large sample size n, $\widehat{F}_n(x)$ will be close to the true c.d.f. value $F(x)$ for all x.

Example 3.1.3 *Following are 100 simulated observations from a distribution with p.d.f* $f(x) = 3x^2, x \in [0, 1]$. *Such Figures could come, for example, from a simple model of a death age distribution in which there is high probability density at the end and a low density at the start.*

- For brevity, the exact observations have not been recorded, only the left endpoints of the intervals $[0, .1), [.1, .2), [.2, .3)$, and so on, to which the observations belonged.
- A histogram of these values is plotted in Fig. 3.6(a), where the bar heights are the sample proportions of 0s, .1s, .2s, and so on, divided by the subinterval size 0.1. Superimposed on this graph is a graph of the probability density function $f(x)$ listed here.
- Part (b) of the figure plots the empirical distribution function together with the theoretical c.d.f. $F(x) = x^3$ for $x \in [0, 1]$.
- There is fairly close agreement between the theoretical and empirical distributions. □

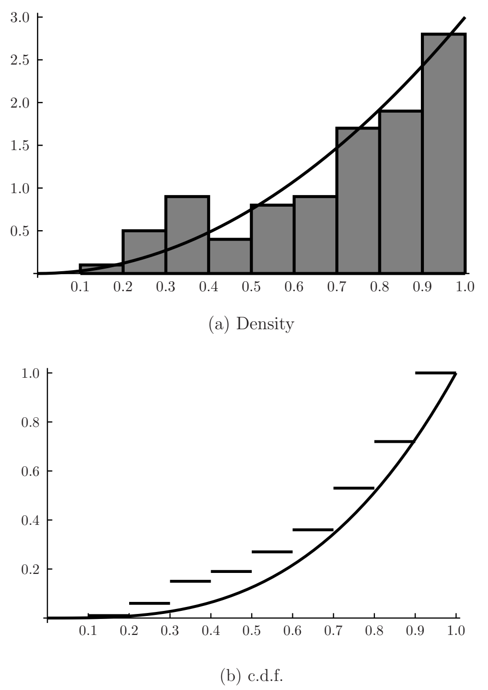


Figure 3.6: ***Empirical distribution***

3.1.3 Multivariate continuous distribution

- When $\mathbf{X}$ is a multivariate random vector, the density $f(\mathbf{x})$ in (3.1) is a real-valued function of a vector variable and the integral in (3.1) is a multiple integral.

- For example, the state space of $\mathbf{X} = (X_1, X_2, X_3)$ is $[0, \infty) \times [0, \infty) \times [0, \infty)$, and the density is $f(x_1, x_2, x_3)$, then the probability that $\mathbf{X}$ belongs to a box-shaped region $[0, 2] \times [1, 3] \times [1, 2]$ could be written as

$$P[\mathbf{X} \in [0, 2] \times [1, 3] \times [1, 2]] = \int_0^2 \int_1^3 \int_1^2 f(x_1, x_2, x_3)\, \mathrm{d}x_3\, \mathrm{d}x_2\, \mathrm{d}x_1.$$

- $F(\mathbf{x}) = P[\mathbf{X} \le \mathbf{x}] = P[X_1 \le x_1, X_2 \le x_2, \ldots, X_n \le x_n]$

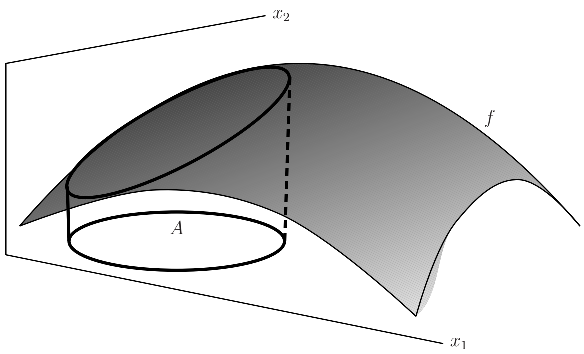


Figure 3.7: ***Joint probability as volume***

Example 3.1.4 (**Compute p.d.f. by c.d.f**) *Suppose that a random vector* $\mathbf{X} = (X_1, X_2)$ *describes, respectively, the longer and shorter utilization times of a mainframe computer by two jobs. The operating system is constructed so as to give jobs a time slice of at most 2 time units. Assume that* $\mathbf{X}$ *has the bivariate p.d.f.*

$$f(x_1, x_2) = \begin{cases} c & \text{if } x_1 - x_2 \ge 0, x_1 \le 2, x_1, x_2 \ge 0, \\ 0 & \text{otherwise.} \end{cases}$$

Find c, the probability that the longer exceeds the shorter by at least 1, and $P[X_1 > x]$ *for all real* x.

Ans:

- $1 = \int_0^2 \int_0^{x_1} c\, \mathrm{d}x_2\, \mathrm{d}x_1 = 2c$ gives $c = 1/2$.

- $P[X_1 \ge X_2 + 1] = \int_1^2 \int_0^{x_1 - 1} 1/2\, \mathrm{d}x_2\, \mathrm{d}x_1 = 1/4$

- $P[X_1 > x] = \int_x^2 \int_0^{x_1} 1/2\, \mathrm{d}x_2\, \mathrm{d}x_1 = 1 - x^2/4$

- $F_1(x) = x^2/4$ if $x \in [0, 2]$.

- $f_1(x) = F_1'(x) = x/2$ if $x \in (0, 2)$. □

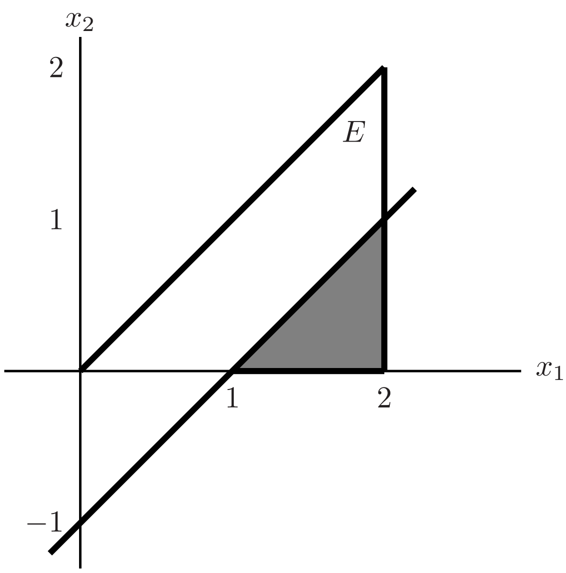


Figure 3.8: ***Event*** $X_1 \geq X_2 + 1$

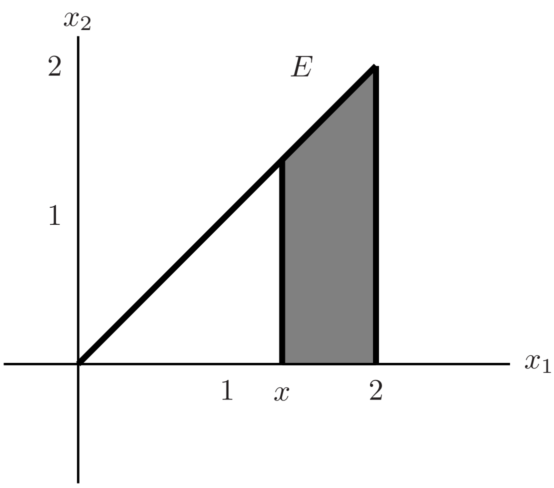


Figure 3.9: ***Event*** $X_1 \geq x$ ***for*** $0 \leq x \leq 2$

Question 3.1.2 *Find the density of X_2 in the last example.*

Ans:

- When $x < 0, P[X_2 > x] = 1$, and when $x > 2, P[X_2 > x] = 0$.
- When $x \in [0, 2]$,

$$P[X_2 > x] = \int_x^2 \int_{x_2}^2 1/2 \, \mathrm{d}x_1 \, \mathrm{d}x_2 = (x-2)^2/4.$$

- $F_2(x) = -\frac{x^2}{4} + x$ if $x \in [0, 2]$, and $f_2(x) = 1 - \frac{x}{2}$, if $x \in (0, 2)$. □

Proposition 3.1.1 (**Marginal density**) *Let $\mathbf{X} = (X_1, \ldots, X_n)$ be a random vector with probability density function $f(\mathbf{x})$ and state space E. The marginal density of X_i is*

$$f_i(x_i) = \int \int_{E(x_i)} \cdots \int f(x_1, \ldots, x_n) \, \mathrm{d}x_n \cdots \mathrm{d}x_{i+1} \, \mathrm{d}x_{i-1} \cdots \mathrm{d}x_1,$$

where $E(x_i) = \{(x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n) | (x_1, \ldots, x_n) \in E\}$. In other words, the marginal density of X_i is found by integrating the joint density over all combinations of possible states for the other component random variables given a fixed state x_i of X_i.

Example 3.1.5 (**Compute marginal density**) *Let $\mathbf{X} = (X, Y, Z)$ represent three delivery times of supplies to a college cafeteria (one unit of time is 10 hours). Suppose that the component random variables have the joint p.d.f*

$$f(x, y, z) = \begin{cases} 8xyz & \text{if } x, y, z \in [0, 1], \\ 0 & \text{otherwise.} \end{cases}$$

Find the joint marginal density and c.d.f of Y and Z, and find the probability that delivers Y and Z are within an hour of each other.

Ans:

- $f_{23}(y, z) = \int_0^1 8xyz \, \mathrm{d}x = 4yz, \quad y, z \in [0, 1]$
- $F_{23}(y, z) = \int_0^y \int_0^z 4uv \, \mathrm{d}u \, \mathrm{d}v = y^2 z^2, \quad y, z \in [0, 1]$
-

$$\begin{aligned} P[|Y - Z| \leq .1] &= \int_0^{.1} \int_0^{y+.1} 4yz \, \mathrm{d}z \, \mathrm{d}y + \int_{.1}^{.9} \int_{y-.1}^{y+.1} 4yz \, \mathrm{d}z \, \mathrm{d}y \\ &\quad + \int_{.9}^1 \int_{y-.1}^1 4yz \, \mathrm{d}z \, \mathrm{d}y \\ &= \int_0^{.1} 2y(y+.1)^2 \, \mathrm{d}y + \int_{.1}^{.9} 2y(.4y) \, \mathrm{d}y + \int_{.9}^1 2y(1-(y-.1)^2) \, \mathrm{d}y \\ &\approx .247. \end{aligned}$$

□

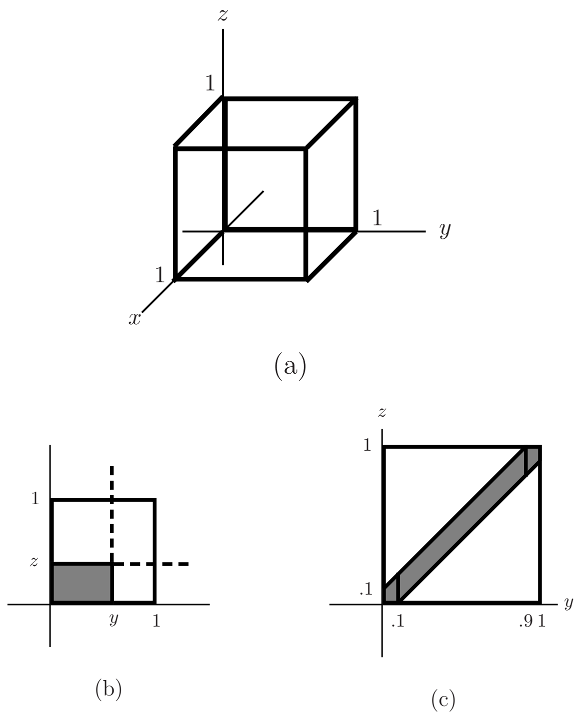


Figure 3.10: ***Joint distributions and probabilities***

3.2 Expectation of continuous random variables

Definition 3.2.1 (*Expected value*) *The* ***expected value*** *of a real-valued, continuous random variable X with p.d.f. f and state space E is*

$$E[X] = \int_E x \cdot f(x)\,\mathrm{d}x,$$

provided the integral exists. It is possible to prove that the expected value of a function $g(X)$ is

$$E[g(X)] = \int_E g(x) \cdot f(x)\,\mathrm{d}x.$$

We adopt terminology and notation as follows, the ***mean*** *of X is*

$$\mu = E[X].$$

Definition 3.2.1 (*Variance, standard deviation, moment*) *The* ***variance*** *of X is*

$$Var(X) = \sigma^2 = E[(X-\mu)^2] = \int_E (x-\mu)^2 f(x)\,dx;$$

the ***standard deviation*** *of X is $\sigma = \sqrt{\sigma^2}$; the* ***rth moment*** *of X is*

$$\mu_r = E[X^r] = \int_E x^r f(x)\,dx;$$

and the ***rth moment about the mean*** *of X is*

$$\mu_r' = E[(X-\mu)^r] = \int_E (x-\mu)^r f(x)\,dx.$$

Note that $\text{Var}(X) = E[X^2] - \mu^2$.

- The mean is a measure of the center of the probability distribution.
- The variance is a measure of the distribution's spread, since σ^2 will be large if large values of $(x-\mu)^2$ have high probability density.
- The skewness is the third moment about μ, and it is a rough measure of asymmetry of the distribution.

Question 3.2.1 *Look back at the results of Section 2.5 on discrete expectation. Which ones seem to carry through to the continuous case? Which of the proofs are independent of the exact definition of expectation, depending only on basic properties such as linearity? (We will summarize the result shortly.)*

Ans:

- Linearity of expectation: $E[aX+bY] = aE[X] + bE[Y]$
- $\text{Var}(aX+b) = a^2\text{Var}(X)$ (affine transformation) □

Example 3.2.1 (*Mean, variance and skewness of continuous uniform distribution*) *Let X have the continuous uniform distribution on $[a,b]$. Find the mean, variance, and skewness of X. We are given that the density of X is*

$$f(x) = \begin{cases} 1/(b-a) & \text{if } x \in [a,b], \\ 0 & \text{otherwise.} \end{cases}$$

- $E[X] = \int_a^b x \cdot \frac{1}{b-a}\,dx = \frac{1}{b-a} \cdot \frac{x^2}{2}\Big|_a^b = \frac{a+b}{2}$
- $\text{Var}(X) = \int_a^b (x-\mu)^2 \cdot \frac{1}{b-a}\,dx = \frac{1}{b-a} \cdot \frac{(x-\mu)^3}{3}\Big|_a^b = \frac{(b-a)^2}{12}$
- The skewness is

$$E[(X-\mu)^3] = \int_a^b (x-\mu)^3 \cdot \frac{1}{b-a}\,dx = \frac{1}{4(b-a)} \cdot (x-\mu)^4\Big|_a^b = 0 \qquad \square$$

Question 3.2.2 *What can you say about all odd-order moments about the mean of the uniform distribution?*

Ans:

- $$E[(X-\mu)^r] = \int_a^b (x-\mu)^r \cdot \frac{1}{b-a}\,\mathrm{d}x = \frac{1}{4(b-a)} \cdot (x-\mu)^{r+1}\Big|_a^b$$
$$= \frac{1}{(r+1)(b-a)}\left[(b-\mu)^{r+1} - (a-\mu)^{r+1}\right]$$
$$= \frac{1}{(r+1)(b-a)}\left[\left(\frac{b-a}{2}\right)^{r+1} - \left(\frac{a-b}{2}\right)^{r+1}\right]$$

- If r is odd, the order moment about the mean of uniform distribution is 0. □

Example 3.2.2 (*rth moment*) *Consider a density $f(x) = c/x^m$ for $x \in [1,\infty)$, where c is chosen in order to make the density integrate to 1 over the state space $[1,\infty)$. Up to what order r does the rth moment of the distribution exist?*

Ans:

- The rth moment, is

$$E[X^r] = \int_1^\infty x^r \cdot \frac{c}{x^m}\,\mathrm{d}x$$
$$= c\cdot \int_1^\infty x^{r-m}\,\mathrm{d}x$$
$$= c\cdot \lim_{B\to\infty}\int_1^B x^{r-m}\,\mathrm{d}x$$
$$= c\cdot \begin{cases} \lim_{B\to\infty} \ln x|_1^B & \text{if } r = m-1, \\ \lim_{B\to\infty} \frac{x^{r-m+1}}{r-m+1}\Big|_1^B & \text{if } r \neq m-1. \end{cases}$$

- The rth moment does not exist when $r \geq m-1$.
- For instance, the density c/x^2 on $[1,\infty)$ has neither a mean $(r = 1 = m-1)$ nor a second moment.
- The density c/x^3 has a mean $(r = 1 = m-2)$ but not a second moment $(r = 2 = m-1)$.
- The density c/x^4 has both a mean $(r = 1 = m-3)$ and a second moment $(r = 2 = m-2)$. □

Question 3.2.3 *What must c equal for $f(x) = c/x^m,\ x\in[1,\infty)$ to be a density?*

Ans:

- If $m > 1, c = m-1$.

- If $m \leq 1$, c does not exist. □

- If c is a real constant, then
$$E[c] = c.$$

- Let X and Y be the random variables of continuous type, and let a and b be constants. Then,
$$E[aX + bY] = aE[X] + bE[Y],$$
provided the expectations of X and Y exist.

- If X is a real random variable of continuous type with finite variance and a and b are real constants, then
$$\text{Var}(aX + b) = a^2\text{Var}(X)$$
In particular the variance of a constant is 0.

- If X is a real random variable of continuous type with finite variance and mean μ, then
$$\text{Var}(X) = E\left[X^2\right] - \mu^2.$$

- If $X_1, X_2, \ldots, X_n$ is a random sample from a continuous distribution with mean μ and variance σ^2, and $\bar{X} = \frac{\sum_{i=1}^n X_i}{n}$, then
$$E[\bar{X}] = \mu \quad \text{and} \quad \text{Var}(\bar{X}) = \frac{\sigma^2}{n}.$$

Example 3.2.3 (*Linearity property of random variables*) *The latest blockbuster super-hero movie, Tick Boy, has come out, and Fast Eddie is looking to make a few dollars marketing T-shirts. He needs to set a price p per shirt. Eddie is relatively sure that the cost per shirt for materials and labor would be \$5. He is uncertain about two areas, however: (1) the number N if shirts he would sell if the price is p, and (2) the setup cost B for licensing and acquisition of equipment to make the shirts. Suppose that N has the Poisson distribution with parameter* $\mu = 600 - 25p$*, where p is the price in dollars, and B has the continuous uniform density on* $[\$100, \$200]$*. What should Fast Eddie do?*

Ans:

- $P = Np - (5N + B)$

- $h(p) = E[P] = (p - 5)E[N] - E[B]$

- $E[N] = 600 - 25p$ and $E[B] = (100 + 200)/2 = 150$

- $h(p) = -25p^2 + 725p - 3125$ has a maximum at $p = 14.5$. □

Expectation in the multivariate continuous case is also analogous to the discrete case.

Example 3.2.4 (***Expectation of bivariate continuous random vectors***) *In the city of Evanston, just north of Chicago, express and nonexpress commuter trains to the Chicago Loop run on weekday mornings. Suppose that at time 0, we are sitting at the Davis Street stop, and the joint density of the times T_e and T_n (in minutes) of arrival of the next express and nonexpress train is*

$$f(t_e, t_n) = \frac{1}{50} e^{-\frac{1}{10}t_e - \frac{1}{5}t_n}, \qquad t_e, t_n > 0.$$

Compute the expected time interval between arrivals of the two trains.

Ans:

$$\begin{aligned} E[|T_n - T_e|] &= \int_0^\infty \int_0^{t_n} (t_n - t_e) f(t_e, t_n)\, \mathrm{d}t_e\, \mathrm{d}t_n + \int_0^\infty \int_{t_n}^\infty (t_n - t_e) f(t_e, t_n)\, \mathrm{d}t_e\, \mathrm{d}t_n \\ &= \frac{1}{5} \int_0^\infty t_n e^{-t_n/5}\, \mathrm{d}t_n - 2 \int_0^\infty e^{-t_n/5}\, \mathrm{d}t_n + 4 \int_0^\infty e^{-3t_n/10}\, \mathrm{d}t_n \\ &= \frac{25}{3} \end{aligned}$$

□

3.3 Examples of continuous distributions

The continuous uniform density $f(x) = 1/(b-a), x \in [a, b]$ was described earlier in this chapter.
To treat a broader range of applications, we must also look at some nonconstant density functions.

3.3.1 Gamma family

These densities are intimately connected with the ***Poisson process***.

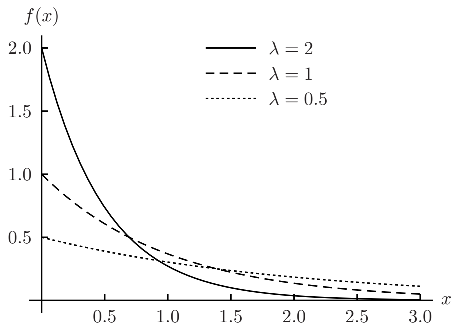
Definition 3.3.1 (***Exponential density***) *The exponential density is*

$$f(t) = \lambda e^{-\lambda t}, \qquad t > 0.$$

The exponential distribution is often used to model the time required for an event to occur, such as the arrival of a plane to an airport or a request for processing to a computer.

Example 3.3.1 (***Application of exponential distribution***) *A group of literary researchers concerned with poetry writing styles analyzed several poems, including Frost's "The Road Not Taken" (Badalamente et al., 1994).*

- They found the frequency distribution of the number of words required to reach the next word not already used before in the poem.
- There is gold in there but it is too deep.
- $0, 0, 0, 0, 1, 0, 1, 0.$
- Histogram for "The Road Not Taken" fitted by an exponential distribution with $\lambda = 1.86$ (Fig. 3.12). □



(a)

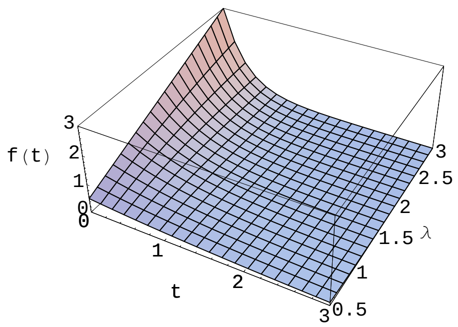


Figure 3.11: ***Exponential densities***

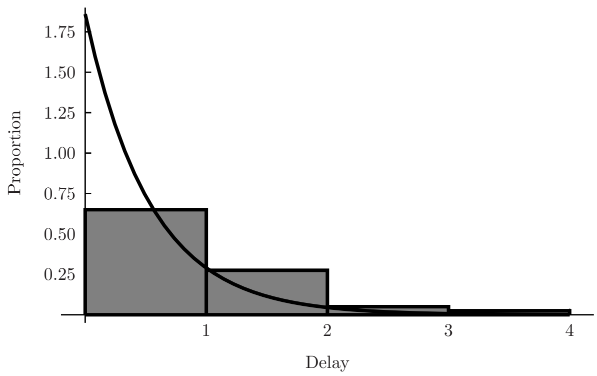


Figure 3.12: ***Distribution of delays until new word in "The Road Not Taken"***

The exponential distribution has an interesting property that can allow you to accept it or reject it on logical grounds as a model for a random time: It is ***memoryless***.

Question 3.3.1 *Show that the c.d.f. of the exponential distribution is*

$$F(t) = 1 - e^{-\lambda t}, \qquad \text{if } t \geq 0.$$

Ans:

$$\begin{aligned} F(t) &= P(T \leq t) \\ &= \int_0^t \lambda e^{-\lambda x}\, \mathrm{d}x \\ &= 1 - e^{-\lambda t}, \qquad \text{if } t \geq 0 \end{aligned}$$

□

Example 3.3.2 (*Memoryless*) *Suppose that the time interval between arrivals of buses to a particular stop is a random variable T (in hours) with the* $\exp(2)$ *distribution. You arrive, panting, to the stop just after a bus leaves. What is the probability that you will have to wait at least 15 minutes for the next bus? Suppose the next bus has not arrived 15 minutes later when a second person comes along. What is the probability that the newcomer will have to wait at least 15 more minutes?*

- $P\left[T \geq \frac{1}{4}\right] = \int_{1/4}^{\infty} 2e^{-2t}\, \mathrm{d}t = -e^{-2t}\big|_{1/4}^{\infty} = e^{-1/2} \approx .607$
- $P[T > 1/2 | T > 1/4] = \frac{P[T>1/2, T>1/4]}{P[T>1/4]} = \frac{e^{-2(1/2)}}{e^{-2(1/4)}} = e^{-1/2}$
- Memoryless □

Proposition 3.3.1 (*Mean and variance of exponential distribution*) *If X has the exponential density (3.32), then*

$$E[X] = \frac{1}{\lambda} \quad \text{and} \quad Var(X) = \frac{1}{\lambda^2}.$$

Proof: The proof is left as Exercise 2 at the end of the section. □

The exponential distribution is related to the Poisson process.

Proposition 3.3.2 (*Distribution of interarrival times*) *Let $T_1, T_2, T_3, \ldots$ be the arrival times of a Poisson process with rate λ. Then the interarrival time random variables $T_1, T_2 - T_1, T_3 - T_2, \ldots$ each have the* $\exp(\lambda)$ *distribution. Furthermore, the interarrival times are independent.*

A generalization of the exponential density is the following.

Definition 3.3.2 (*Gamma density*) *The **gamma density** is*

$$f(t) = \frac{\lambda^\alpha}{\Gamma(\alpha)}\, t^{\alpha-1} e^{-\lambda t}, \qquad t > 0$$

*where the **gamma function** is defined by*

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x}\, \mathrm{d}x.$$

`http://demonstrations.wolfram.com/TheGammaDistribution/`

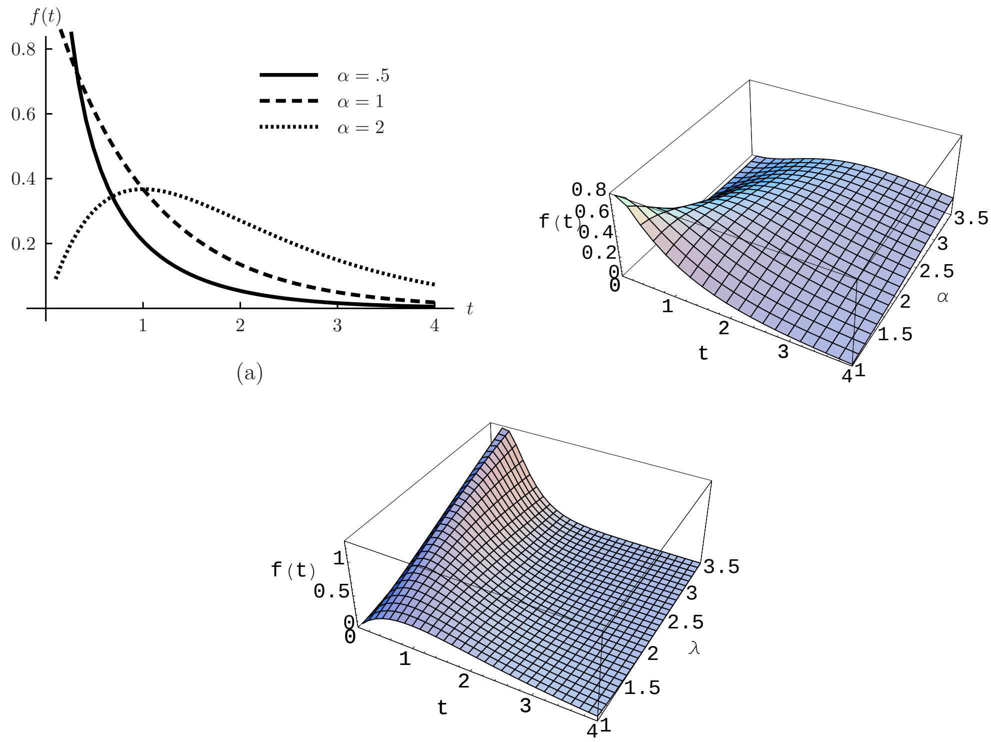


Figure 3.13: ***Gamma densities***

- $\Gamma(1) = 1$
- $\Gamma(\alpha + 1) = \alpha \cdot \Gamma(\alpha) = \alpha!, \quad \alpha \neq -1, -2, \ldots$ (Exercise 11)
- $\Gamma(1/2) = \sqrt{\pi}$ (Exercise 11)

Question 3.3.2 *What is $\Gamma(2)$? $\Gamma(3)$? $\Gamma(4)$? Show that $\Gamma(n) = (n-1)!$ when n is positive integer. Now find $\Gamma(3/2)$ and $\Gamma(5/2)$, and discover a similar result of $\Gamma(m/2)$, where m is an odd positive integer.*

Ans:

- $\Gamma(2) = 1!, \Gamma(3) = 2!, \Gamma(4) = 3!$
- $$\Gamma(n) = \int_0^\infty x^{n-1} e^{-x}\, \mathrm{d}x = (n-1) \int_0^\infty x^{n-2} e^{-x}\, \mathrm{d}x$$
  $$= \cdots = (n-1)! \int_0^\infty e^{-x}\, \mathrm{d}x = (n-1)!$$
- $\Gamma(3/2) = \sqrt{\pi}/2, \Gamma(5/2) = 3\sqrt{\pi}/4$
- $$\Gamma(m/2) = \int_0^\infty x^{m/2-1} e^{-x}\, \mathrm{d}x = (m/2-1) \int_0^\infty x^{m/2-2} e^{-x}\, \mathrm{d}x$$
  $$= \cdots = (m/2-1)! \int_0^\infty e^{-x}\, \mathrm{d}x = (m/2-1)!$$

□

When the parameter α is positive integer n, the gamma density is often called the ***Erlang density*** in honor of an applied mathematician who did early work on queueing theory in which this distribution played an important role. Like the exponential distribution, the gamma distribution has a connection to Poisson processes.

Proposition 3.3.3 (***Distribution of nth arrival time***) *Let $T_1, T_2, T_3, \ldots$ be the arrival times of a Poisson process with rate λ. Then T_n has $\Gamma(n, \lambda)$ distribution. Furthermore, the time between the mth and $(m+n)$th arrivals $T_{m+n} - T_m$ has the $\Gamma(n, \lambda)$ distribution for $m, n > 0$.*

Example 3.3.3 (***Probability of gamma distribution***) *Suppose that cars pass expressway exit according to a Poisson process with a rate of 10 per minute. Find the probability that the second car passes the next between times 5 and 8 seconds.*

- T_2 has the $\Gamma(2, 10)$ distribution.

- $P[1/12 \leq T_2 \leq 2/15] = \int_{1/12}^{2/15} 100te^{-10t}\,\mathrm{d}t = \frac{11}{6}e^{-5/6} - \frac{7}{3}e^{-4/3} \approx .18$ □

Proposition 3.3.4 (***Mean and variance of Gamma distribution***) *If T has the $\Gamma(\alpha, \lambda)$ distribution, then*

$$E[T] = \frac{\alpha}{\lambda} \quad \textit{and} \quad Var(T) = \frac{\alpha}{\lambda^2}.$$

Proof:

- We can calculate the mean by a simple substitution and an appeal to properties of the gamma function as follows:

$$\begin{aligned}
E[T] &= \int_0^\infty t \cdot \frac{\lambda^\alpha}{\Gamma(\alpha)} t^{\alpha-1} e^{-\lambda t}\,\mathrm{d}t \\
&= \frac{1}{\Gamma(\alpha)} \int_0^\infty (\lambda t)^\alpha e^{-\lambda t}\,\mathrm{d}t \qquad (x = \lambda t,\ \ \mathrm{d}x = \lambda\,\mathrm{d}t) \\
&= \frac{1}{\lambda\Gamma(\alpha)} \int_0^\infty x^\alpha e^{-x}\,\mathrm{d}x \\
&= \frac{1}{\lambda\Gamma(\alpha)} \cdot \Gamma(\alpha+1) \\
&= \frac{1}{\lambda\Gamma(\alpha)} \cdot \alpha\Gamma(\alpha) \\
&= \frac{\alpha}{\lambda}
\end{aligned}$$

- Recall that $\mathrm{Var}(T) = E[T^2] - (E[T])^2$.

- The second moment of T is

$$
\begin{aligned}
E[T^2] &= \int_0^\infty t^2 \cdot \frac{\lambda^\alpha}{\Gamma(\alpha)} t^{\alpha-1} e^{-\lambda t}\,\mathrm{d}t \\
&= \frac{1}{\lambda\Gamma(\alpha)} \int_0^\infty (\lambda t)^{\alpha+1} e^{-\lambda t}\,\mathrm{d}t \qquad (x = \lambda t,\ \ \mathrm{d}x = \lambda\,\mathrm{d}t) \\
&= \frac{1}{\lambda^2\Gamma(\alpha)} \int_0^\infty x^{\alpha+1} e^{-x}\,\mathrm{d}x \\
&= \frac{1}{\lambda^2\Gamma(\alpha)} \cdot \Gamma(\alpha+2) \\
&= \frac{1}{\lambda^2\Gamma(\alpha)} \cdot (\alpha+1)\Gamma(\alpha+1) \\
&= \frac{1}{\lambda^2\Gamma(\alpha)} \cdot (\alpha+1)\alpha\Gamma(\alpha) \\
&= \frac{\alpha(\alpha+1)}{\lambda^2}
\end{aligned}
$$

- $\mathrm{Var}(T) = E[T^2] - E[T]^2 = \frac{\alpha(\alpha+1)}{\lambda^2} - \left(\frac{\alpha}{\lambda}\right)^2 = \frac{\alpha}{\lambda^2}$ □

Another important subcase of the gamma family is chi-square density.

Definition 3.3.3 (*Chi-square distribution*) *The* ***chi-square density*** *is*

$$f(x) = \frac{1}{2^{n/2}\Gamma(n/2)} x^{n/2-1} e^{-x/2}, \qquad x > 0$$

where n is a positive integer parameter that is called the ***degrees of freedom*** *of the distribution.*

http://demonstrations.wolfram.com/TheExponentialDistribution/

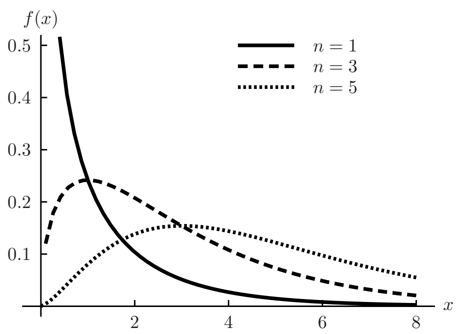


Figure 3.14: ***Chi-square densities***

- Shorthand $\chi^2(n)$
- $\chi^2(n) = \Gamma(n/2, 1/2)$
- $E[X] = n$ and $\mathrm{Var}(X) = 2n$
- Some instances of the density are sketched in Fig. 3.14.

- It plays a role nearly as important as that of the normal distribution in the statistical analysis of variability of data.
- Table 4 in Appendix A gives points x such that $P[X \leq x] = \alpha$, for a number of specified values of the degrees of freedom r, and specified probability values α, where X is a $\chi^2(r)$ random variable.
- For example, when $n = 14$, the appropriate line of the table indicates that

$$P[X \leq 4.66] \approx .01, \quad P[X \leq 6.57] \approx .05, \quad P[X \leq 29.14] \approx .95$$

- $$P[X > 4.66] \approx .99, \quad P[X \in (6.57, 29.14]] \approx .99 - .05 = .94$$

Question 3.3.3 *If X is $\chi^2(10)$, for what x is $P[X \leq x] = 0.05$? What is $P[X > 4.87]$? If X is $\chi^2(18)$, what is $P[X \in [9.39, 28.87]]$?*

Ans:

- By Table 4 in appendix A, if X is $\chi^2(10)$, $P[X \leq 3.94] = .05$
- $P[X > 3.94] = .95$
- If X is $\chi^2(18)$, $P[X \in [9.39, 28.87]] = .95 - .05 = .9$. □

3.3.2 Normal distribution

- Fig. 3.15 displays a histogram of some actual data on natural birthrates in 33 countries that have been subject to forced mass migration recently (Wood, 1994).
- The data are grouped into intervals of length .3, starting at birthrate 0.7.
- The histogram of frequencies takes on a rough bell shape that is characteristic of histograms of many data samples, including physical quantities like heights and weights, measurement errors in scientific experiments, and test scores and grades (see, for example, Fig. 2.4 on the calculus GPAs).
- ***Central Limit Theorem***: Chap. 6 will show in general that the distribution of any sum of n independent and identically distributed random variables converges as $n \to \infty$ to the normal distribution.
- The normal distribution is the most important distribution in probability and statistics.

Definition 3.3.4 (*Normal density*) *The **normal density** is*

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \qquad -\infty < x < \infty. \tag{3.44}$$

`http://demonstrations.wolfram.com/AreaOfANormalDistribution/`
$N(\mu, \sigma^2)$, location parameter $\mu \in \mathbb{R}$, spread (scale) parameter $\sigma^2 > 0$

- Sketches for fixed $\sigma^2 = 1$ and several values of μ are given in Fig. 3.16(a).

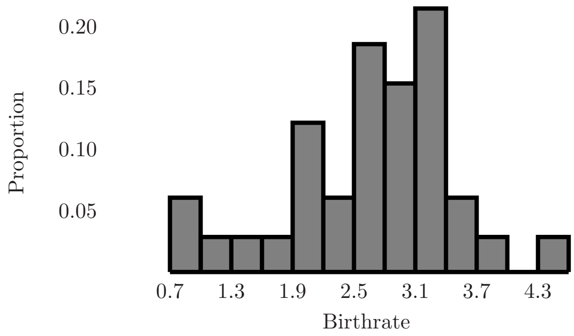
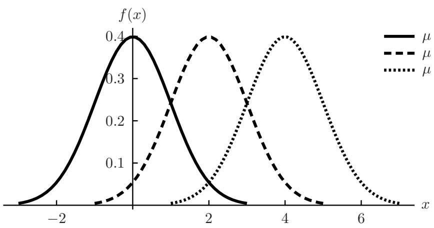


Figure 3.15: ***Histogram of birthrates***



(a) $\sigma^2 = 1$

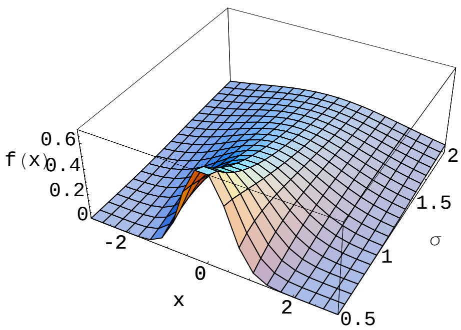


Figure 3.16: ***Normal densities***

- As μ increases, the density curve translates to the right.
- A three-dimensional look at the dependence of the density on σ^2 for fixed $\mu = 0$ is in part (b) of the figure. As the variance increases, the curve spreads out.

Question 3.3.4 *Show that the normal density is symmetric about μ (For an interesting geometric interpretation of σ^2, read Exercise 15.)*

Ans: $f(\mu + x) = f(\mu - x)$, so the normal density is symmetric about μ. □

Check that the function in (3.44) does define a valid p.d.f. Clearly $f(x) \geq 0,\ \forall x \in \mathbb{R}$. Substituting $u = (x-\mu)/\sigma,\ \ \mathrm{d}u = \mathrm{d}x/\sigma$, we obtain

$$\begin{aligned}
&\left(\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \mathrm{d}u\right)^2 \\
&= \left(\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-u^2/2}\,\mathrm{d}u\right)^2 \\
&= \left(\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-u^2/2}\,\mathrm{d}u\right)\left(\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-v^2/2}\,\mathrm{d}v\right) \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \frac{1}{2\pi} e^{-u^2/2-v^2/2}\,\mathrm{d}v\,\mathrm{d}u
\end{aligned}$$

In the double integral, change to ***polar coordinates*** with the substitution $u = r\cos\theta,\ v = r\sin\theta,\ \ \mathrm{d}v\,\mathrm{d}u = r\,\mathrm{d}r\,\mathrm{d}\theta$.

$$\begin{aligned}
\left(\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \mathrm{d}u\right)^2 &= \int_0^{\infty}\int_0^{2\pi} \frac{1}{2\pi} e^{-r^2/2} r\,\mathrm{d}r\,\mathrm{d}\theta \\
&= \left(\int_0^{\infty} r e^{-r^2/2}\,\mathrm{d}r\right)\left(\int_0^{2\pi} \frac{1}{2\pi}\,\mathrm{d}\theta\right) \\
&= \left(-e^{-r^2/2}\Big|_0^{\infty}\right)\left(\frac{1}{2\pi}\theta\Big|_0^{2\pi}\right) \\
&= 1 \cdot 1 = 1.
\end{aligned}$$

Proposition 3.3.5 (*Mean and variance of normal distribution*) *If X has the $N(\mu, \sigma^2)$ distribution, then*

$$E[X] = \mu \quad \textit{and} \quad Var(X) = \sigma^2.$$

Proof:

- To prove that $E[X] = \mu$, we write

$$\begin{aligned}
E[X-\mu] &= \int_{-\infty}^{+\infty} (x-\mu) \cdot \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \mathrm{d}x \qquad \left(u = \frac{x-\mu}{\sigma},\ \mathrm{d}u = \mathrm{d}x/\sigma\right) \\
&= \sigma \cdot \int_{-\infty}^{+\infty} u \cdot \frac{1}{\sqrt{2\pi}} e^{-u^2/2}\,\mathrm{d}u \\
&= \frac{\sigma}{\sqrt{2\pi}} (-e^{-u^2/2})\Big|_{-\infty}^{+\infty} \\
&= 0.
\end{aligned}$$

- By linearity of expectation, $E[X] = \mu$
- By using the general properties of variance, we can recast the desired result into something computationally easier as follows:

$$\begin{aligned} \text{Var}(X) = \sigma^2 &\iff \text{Var}(X-\mu) = \sigma^2 \\ &\iff \frac{\text{Var}(X-\mu)}{\sigma^2} = 1 \\ &\iff \text{Var}\left(\frac{X-\mu}{\sigma}\right) = 1. \end{aligned}$$

- $E\left[\frac{X-\mu}{\sigma}\right] = 0$, moreover, so it suffices to show that $E[((X-\mu)/\sigma^2)] = 1$. (Justify this statement)

$$\begin{aligned} E\left[\left(\frac{X-\mu}{\sigma}\right)^2\right] &= \int_{-\infty}^{+\infty} \left(\frac{x-\mu}{\sigma}\right)^2 \cdot \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \mathrm{d}x \\ &\quad (y = \frac{x-\mu}{\sigma}, \mathrm{d}y = \mathrm{d}x/\sigma) \\ &= \frac{1}{\sqrt{2\pi}} \cdot \int_{-\infty}^{+\infty} y^2 \cdot e^{-y^2/2} \mathrm{d}y \\ &\quad (\text{parts}, u = y, dv = ye^{-y^2/2} \mathrm{d}y) \\ &= \frac{1}{\sqrt{2\pi}} \left(-ye^{-y^2/2}\Big|_{-\infty}^{+\infty} + \int_{-\infty}^{+\infty} e^{-y^2/2} \mathrm{d}y \right) \\ &= 0 + 1 \\ &= 1 \end{aligned}$$

- The last integral on the right is 1, since, after bringing the $1/\sqrt{2\pi}$ factor back into the integral, the integrand is the normal density with $\mu = 0$ and $\sigma^2 = 1$, which integrates to 1 over the state space $(-\infty, +\infty)$.
- This completes the proof. □

Proposition 3.3.6 (*Standardizing*) *If a random variable X has the $N(\mu, \sigma^2)$ distribution, then the random variable Z defined by*

$$Z = \frac{X-\mu}{\sigma}$$

has the standard normal distribution. Therefore

$$P[X \le b] = P\left[Z = \frac{X-\mu}{\sigma} \le \frac{b-\mu}{\sigma}\right] = \int_{-\infty}^{(b-\mu)/\sigma} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} \,\mathrm{d}z.$$

(The algebraic operation of subtracting μ and dividing by σ is known as ***standardizing.****)*

Proof:

- We can compute the c.d.f. of Z as follows:

$$F(z) = P[Z \le z] = P\left[\frac{X-\mu}{\sigma} \le z\right] = P[X \le \mu + \sigma z]$$

- Since X was assumed to be normally distributed,

$$F(z) = \int_{-\infty}^{\mu+\sigma z} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \, dx.$$

- By the Fundamental Theorem of Calculus and the chain rule, the density of Z is

$$f(z) = F'(z) = \sigma \cdot \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(\mu+\sigma z-\mu)^2}{2\sigma^2}\right) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2},$$

- That is, Z is standard normal. □

Example 3.3.4 *Suppose the load X required to break a 1×10 board is normally distributed with mean 2.50 and standard deviation 0.24. Then the probability that the piece breaks at a load of 2.61 or less is*

$P[X \le 2.61] = P\left[Z = \frac{X-\mu}{\sigma} \le \frac{2.61-2.50}{0.24}\right] = P[Z \le 0.46] \approx 0.6772.$

- $P[X > 2.39] = P[Z = (X-\mu)/\sigma > -.46]$
- $P[X \ge 2.86] = P[Z = (X-\mu)/\sigma > 1.50] \approx .0668$
- $P[2.61 \le X \le 2.86] = P[.46 \le Z \le 1.50] = .256$ □

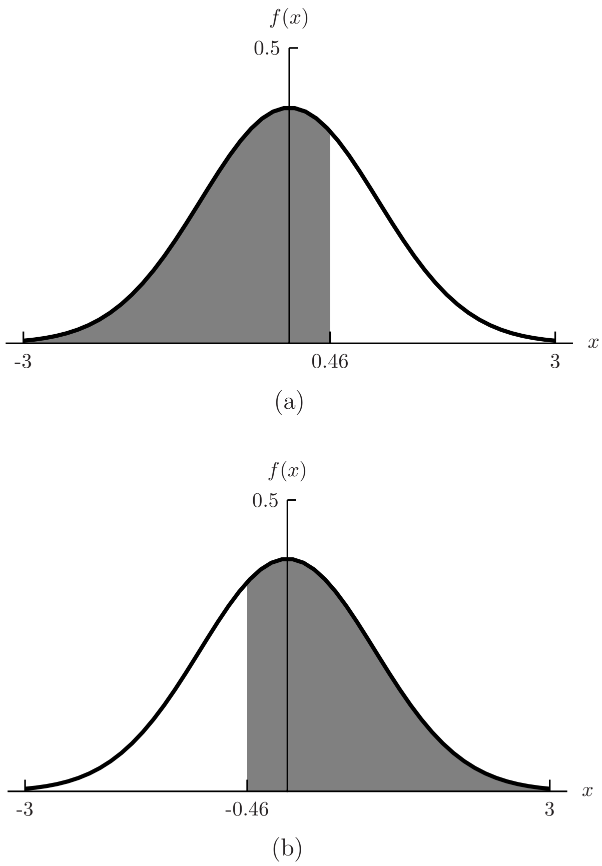


Figure 3.17: ***Normal probabilities***

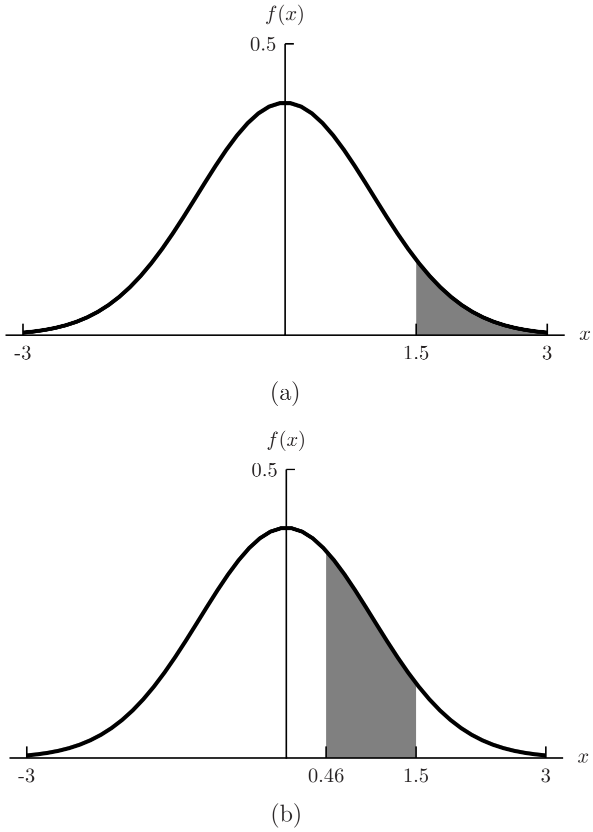


Figure 3.18: ***More normal probabilities***

Question 3.3.5 *From the last example, what is* $P[Z \leq -1.50]$? $P[-1.50 \leq Z \leq -0.46]$?

Ans:

- $P[Z \leq -1.5] = 1 - P[Z \leq 1.5] = 1 - .9332 = .0668$
- $P[-1.50 \leq Z \leq -0.46] = 1 - P[Z \leq .46] - (1 - P[Z \leq 1.5]) = 1 - .6772 + 1 + .9332 = .256$ □

Example 3.3.5 *Recall that the* $p \times 100$*th percentile* x_p *of a distribution satisfies* $P[X \leq x_p] = p$, *in other words, the area under the p.d.f of* X *to the left of* x_p *is exactly* p. *Suppose that for the birthrate data with which we begin the section, observations are approximately normally distributed with mean 2.58 and standard deviation 0.836. What are the 90th and 80th percentiles?*

- $.9 = P[X \leq x_{.9}] = P\left[Z \leq \frac{x_{.9} - 2.58}{.836}\right]$ gives $\frac{x_{.9} - 2.58}{.836} = z_{.9} \approx 1.28$, then $x_{.9} \approx 3.65$.
- $x_{.8} \approx 3.28$ □

Question 3.3.6 *Justify that* $x_p = \mu + \sigma z_p$ *is a general formula relating the percentiles of the* $N(\mu, \sigma^2)$ *distribution to the standard normal percentiles.*

Ans:

-

$$\begin{aligned} p &= P[X \leq x_p] \\ &= P\left[\frac{X-\mu}{\sigma} \leq \frac{x_p-\mu}{\sigma}\right] \\ &= P\left[Z \leq \frac{x_p-\mu}{\sigma}\right] \\ &= P[Z \leq z_p] \end{aligned}$$

- $z_p = \frac{x_p-\mu}{\sigma}$, $x_p = \mu + \sigma z_p$ □

3.3.3 Other distributions

- The Weibull distribution, which is widely used to model time to failure of components or systems in reliability problems.

- ***Weibull distribution***: Weibull(λ, β), $\lambda, \beta > 0$

$$f(t) = \beta\lambda^\beta t^{\beta-1} e^{-(\lambda t)^\beta}, \qquad t > 0$$

`http://demonstrations.wolfram.com/ReliabilityDistributions/`

  - ▷ Scale (λ) and shape (β) parameters.
  - ▷ If $\beta = 1$ becomes $\exp(\lambda)$.
  - ▷ Let T be a failure time random variable with some density $f(t)$ and c.d.f. $F(t)$.
  - ▷ Then the probability of failure in the time interval $(t, t+\Delta t]$, given survival until time t, is

$$P[T \in (t, t+\Delta t] | T > t] = \frac{P[T \in (t, t+\Delta t]]}{P[T > t]} \approx \frac{f(t)\Delta t}{1-F(t)}$$

    for small Δt.
  - ▷ Failure rate function: $h(t) = \lim_{\Delta t \to 0} \frac{P[T \in (t,t+\Delta t]|T>t]}{\Delta t} = \frac{f(t)}{1-F(t)} = \beta\lambda^\beta t^{\beta-1}$
  - ▷ $F(t) = 1 - e^{-(\lambda t)^\beta}$.
  - ▷ Failure rate is decreasing for $0 < \beta < 1$, constant when $\beta = 1$, and increasing for $\beta > 1$.

- ***Lognormal density***: $\mu \in \mathbb{R}$, $\sigma^2 > 0$

$$f(y) = \frac{1}{y}\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(\ln(y)-\mu)^2}{2\sigma^2}\right], \qquad y > 0$$

  - ▷ Model price behavior of financial instruments.
  - ▷ Relationship to the normal density: $X = \log Y \sim N(\mu, \sigma^2)$

- ***Pareto density***: $\theta > 0$

$$f(x) = \frac{\theta}{(1+x)^{\theta+1}}, \qquad x \geq 0, \theta > 0$$

  - A reasonable empirical fit to incomes above a threshold value and other economic random phenomena.

- ***Cauchy density***:

$$f(x) = \frac{1}{\pi[1+(x-\mu)^2]}, \qquad -\infty < x < \infty$$

  - Its shape is similar to that of the normal density except for its heavier tails.
  - Mean and variance are undefined.

- ***Beta density***: $\alpha, \beta > 0$

$$f(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\, x^{\alpha-1}(1-x)^{\beta-1}, \qquad 0 < x < 1$$

  - In Chapter 5, it can be shown that the beta distribution is the distribution of the proportion that one gamma random variable has within the total of two independent gamma random variables.
  - T_n: nth arrival times in a Poisson process

$$\frac{T_m}{T_n} = \frac{T_m}{T_m + (T_n - T_m)} \sim \text{Beta}(m, n-m), \quad n > m$$

- t- and F-distributions are deferred to Chap. 5 since the derivations of them need the ***theory of transformations***.

3.4 Multivariate normal distribution

- A multivariate analog of the normal distribution.
- More detailed in Chapters 4 and 5.
- The multivariate normal distribution is used as a model in multidimensional measurements of the same kinds of random numerical phenomena for which the single-variable normal distribution is appropriate (e.g., heights and weights, SAT and ACT scores, temperature, humidity, and barometric pressure, etc.)

Example 3.4.1 *Van Buren, interested in the economic Exploitation of workers in Mexico, analyzed average prices and worker salaries in the United States and Mexico and used the results to come up with estimates of the numbers of minutes that U.S. and Mexican workers must work to earn enough to buy certain standard quantities of grocery staples.*

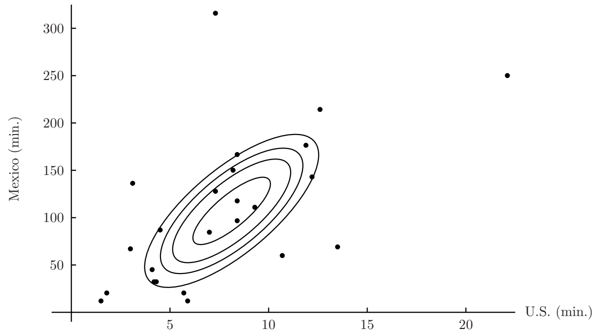


Figure 3.19: ***Time to work for food staples, U.S. and Mexico***

- (U.S., Mexican)

- Scattergram: Figure 3.19

  (4.3, 32.3), (22.1, 250.0), (10.7, 60.0), (4.1 45.1) (7.3, 315.8), (1.5, 11.9), (12.6, 214.3), (4.5, 87.0), (8.4, 117.6), (9.3, 111.1), (5.9, 11.9) (11.9, 176.5), (7.0, 84.5), (3.1, 136.4), (7.3, 127.7), (4.2, 32.3), (12.2, 142.9), (3.0, 51.7), (8.4, 166.7), (8.2, 150.0), (13.5, 69.0), (1.8, 20.4), (8.4, 96.8), (5.7, 20.4)

- There does seem to be increasing relationship between the two variables, roughly linear, but with high variability.

- Data seem to spread out in a sort of elliptical cloud, denser toward the center, with major axis having a slope about 10.

- It follows that for every extra minute that the U.S. worker puts in for added consumer benefit, the Mexican worker must put in around ten, depending on some random wage, price, availability, and general market factors. □

Bivariate normal density with $\rho = 0$:

$$f(x_1, x_2) = \frac{1}{2\pi\sigma_1\sigma_2} \exp\left(-\frac{1}{2}\left[\frac{x_1^2}{\sigma_1^2} + \frac{x_2^2}{\sigma_2^2}\right]\right), \quad x_1, x_2 \in \mathbb{R} \tag{3.59}$$

Figure 3.20(a) shows the graph of this function for $\sigma_1 = \sigma_2 = 1$, and part (b) shows the case $\sigma_1 = 2,\ \sigma_2 = 1$.

Question 3.4.1 *What does the graph of the density look like in the case $\sigma_1 = 1, \sigma_2 = 2$? What shape is taken on by slices of the surface parallel to the $x_1 - x_2$ plane? How would the graph be affected if, in the exponent of (3.59), the ${x_i}^2$ terms were replaced by $(x_i - \mu_i)^2$, where μ_1 and μ_2 are constants?*

$$f(x_1, x_2) = \frac{1}{\sqrt{2\pi\sigma_1^2}}\, e^{-x_1^2/2\sigma_1^2} \cdot \frac{1}{\sqrt{2\pi\sigma_2^2}}\, e^{-x_2^2/2\sigma_2^2}, \quad x_1, x_2 \in \mathbb{R}$$

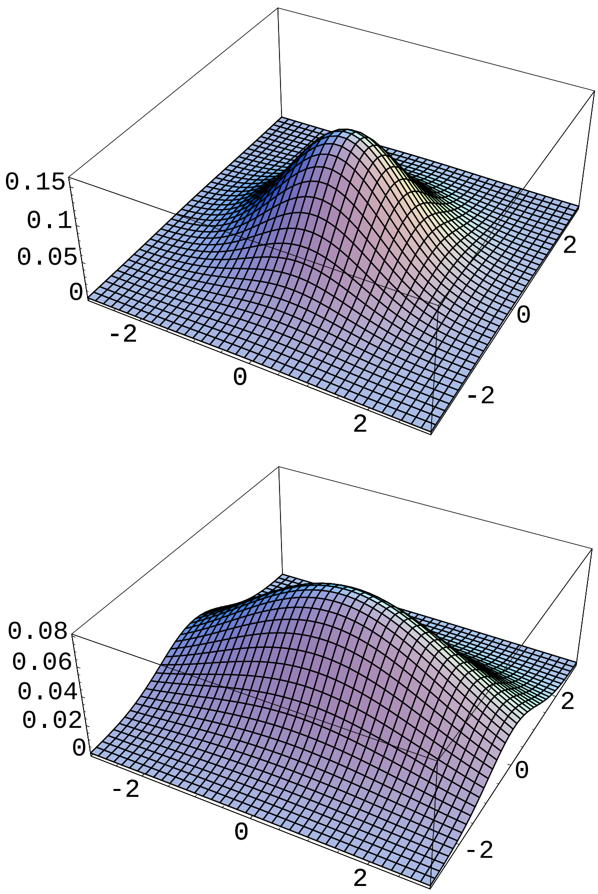


Figure 3.20: ***Bivariate normal density, case*** $\rho = 0$

The marginal density of X_1 is

$$\begin{aligned} f_1(x_1) &= \int_{-\infty}^{\infty} f(x_1, x_2)\,\mathrm{d}x_2 \\ &= \frac{1}{\sqrt{2\pi\sigma_1^2}}\, e^{-x_1^2/2\sigma_1^2} \cdot \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma_2^2}}\, e^{-x_2^2/2\sigma_2^2}\,\mathrm{d}x_2 \\ &= \frac{1}{\sqrt{2\pi\sigma_1^2}}\, e^{-x_1^2/2\sigma_1^2} \end{aligned}$$

Similarly the marginal density of X_2 is

$$f_2(x_2) = \frac{1}{\sqrt{2\pi\sigma_2^2}}\, e^{-x_2^2/2\sigma_2^2}$$

The joint density (3.59) can be written in matrix form as follows:

$$f(\mathbf{x}) = \frac{1}{2\pi\sqrt{\det(\Sigma)}}\, e^{(-1/2)\mathbf{x}'\Sigma^{-1}\mathbf{x}}, \qquad \mathbf{x} \in \mathbb{R}^2 \tag{3.62}$$

where $\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$, $\Sigma = \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}$, $\det(\Sigma) = \sigma_1^2\sigma_2^2$, $\Sigma^{-1} = \begin{bmatrix} 1/\sigma_1^2 & 0 \\ 0 & 1/\sigma_2^2 \end{bmatrix}$.

Question 3.4.2 *Check the representation in (3.62).*

A slight generalization of formula (3.62) to nonzero means leads to the following definition.

Definition 3.4.1 *A random vector $\mathbf{X} = (X_1, X_2, \ldots, X_n)'$ is said to have the* ***multivariate normal distribution*** *with parameters $\boldsymbol{\mu}$ and Σ if its density is*

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{n/2}\sqrt{\det(\Sigma)}}\, e^{-(1/2)(\mathbf{x}-\boldsymbol{\mu})'\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})}, \qquad \mathbf{x} \in \mathbb{R}^n$$

where Σ is an $n \times n$ symmetric matrix, assumed to be positive definite (i.e., $\mathbf{y}'\Sigma\mathbf{y} > 0\ \forall \mathbf{y} \neq 0$), and $\boldsymbol{\mu}$ is an n-vector. The parameter $\boldsymbol{\mu}$ is called the mean vector, and Σ is called the ***covariance matrix*** *of the distribution.*

Example 3.4.2 *Suppose that $\mathbf{X} = (X_1, X_2, X_3)'$ has a trivariate normal distribution with mean $\boldsymbol{\mu} = [1, 2, 0]'$ and diagonal covariance matrix*

$$\Sigma = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 9 \end{bmatrix}.$$

Ans: Then,

$$\Sigma^{-1} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1/4 & 0 \\ 0 & 0 & 1/9 \end{bmatrix}, \qquad \det(\Sigma) = 1 \cdot 4 \cdot 9.$$

$$\begin{aligned} (\mathbf{x}-\boldsymbol{\mu})'\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu}) &= \begin{bmatrix} x_1 - 1 & x_2 - 2 & x_3 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1/4 & 0 \\ 0 & 0 & 1/9 \end{bmatrix} \begin{bmatrix} x_1 - 1 \\ x_2 - 2 \\ x_3 \end{bmatrix} \\ &= (x_1 - 1)^2 + \frac{1}{4}(x_2 - 2)^2 + \frac{x_3^2}{9} \end{aligned}$$

□

Proposition 3.4.1 *Let* $\mathbf{X} = (X_1, X_2, \ldots, X_n)'$ *have the multivariate normal distribution with mean* $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_n)'$ *and covariance* Σ. *If* Σ *is a diagonal matrix with entries* $\sigma_1^2, \ldots, \sigma_n^2$ *on its diagonal, then the joint density is*

$$f(\mathbf{x}) = f_1(x_1) \cdot f_2(x_2) \cdots f_n(x_n)$$

where f_i *is the* $N(\mu_i, {\sigma_i}^2)$ *density. Consequently, under these assumptions the marginal distribution of each* X_i *is* $N(\mu_i, {\sigma_i}^2)$.

Proof: If the covariance matrix is of the form

$$\Sigma = \begin{bmatrix} \sigma_1^2 & 0 & \ldots & 0 \\ 0 & \sigma_2^2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \sigma_n^2 \end{bmatrix},$$

then it is easy to see that $\det(\Sigma) = \sigma_1^2\sigma_2^2 \ldots \sigma_n^2$ and

$$\Sigma^{-1} = \begin{bmatrix} 1/\sigma_1^2 & 0 & \ldots & 0 \\ 0 & 1/\sigma_2^2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & 1/\sigma_n^2 \end{bmatrix}$$

Thus,

$$\begin{aligned} f(\mathbf{x}) &= \frac{1}{(2\pi)^{n/2}\sqrt{\sigma_1^2\sigma_2^2 \ldots \sigma_n^2}} \cdot \exp\left[-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})' \begin{bmatrix} 1/\sigma_1^2 & 0 & \ldots & 0 \\ 0 & 1/\sigma_2^2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & 1/\sigma_n^2 \end{bmatrix} (\mathbf{x} - \boldsymbol{\mu}) \right] \\ &= \frac{1}{(2\pi)^{n/2}\sqrt{\sigma_1^2\sigma_2^2 \ldots \sigma_n^2}} \cdot \exp\left[-\frac{1}{2}\sum_{i=1}^{n}(x_i - \mu_i)^2/\sigma_i^2 \right] \\ &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma_i^2}} \ \exp\left[-\frac{1}{2}(x_i - \mu_i)^2/\sigma_i^2 \right]. \end{aligned}$$

The last expression is the product of the desired normal marginal densities. □

- We concentrate on the $n = 2$ case, called the ***bivariate normal distribution***.
- The covariance matrix has the form

  $$\Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{bmatrix}$$

  where $\sigma_{12} = \sigma_{21}$.
- We call the two random variables X and Y rather than X_1 and X_2, and change the notation for the means and variances.
- $$\Sigma = \begin{bmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{bmatrix}$$

- The condition that Σ is positive definite, hance

$$\sigma_x^2\sigma_y^2 - \rho^2\sigma_x^2\sigma_y^2 = \sigma_x^2\sigma_y^2(1-\rho^2) > 0 \Rightarrow -1 < \rho < 1$$

- The parameter ρ will be called the ***correlation coefficient*** of X and Y.

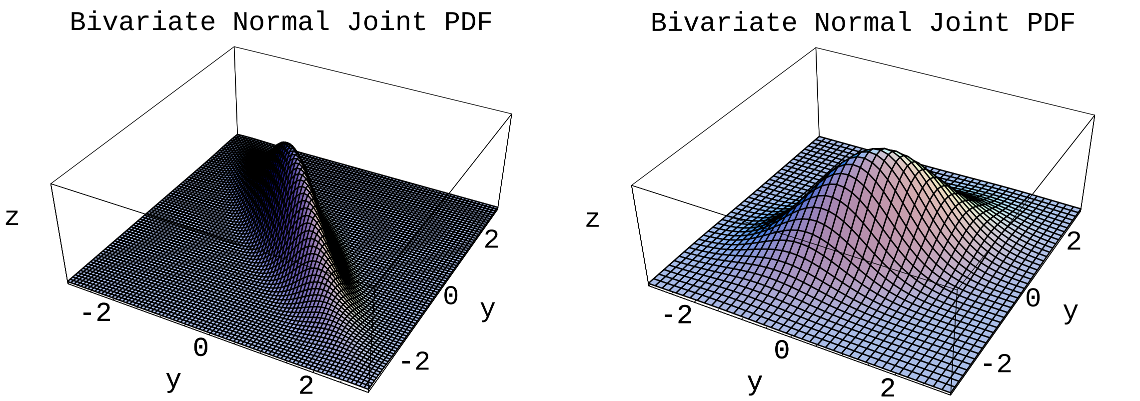
- $\det(\Sigma) = \sigma_x^2\sigma_y^2(1-\rho^2)$, $\quad \Sigma^{-1} = \frac{1}{\sigma_x^2\sigma_y^2(1-\rho^2)}\begin{bmatrix} \sigma_y^2 & -\rho\sigma_x\sigma_y \\ -\rho\sigma_x\sigma_y & \sigma_x^2 \end{bmatrix}$

$$f(x,y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}}\exp[-Q],$$

where

$$Q = \frac{1}{2(1-\rho^2)}\left[\frac{(x-\mu_x)^2}{\sigma_x^2} - 2\rho\frac{(x-\mu_x)(y-\mu_y)}{\sigma_x\sigma_y} + \frac{(y-\mu_y)^2}{\sigma_y^2}\right]$$

Bivariate normal distribution: `http://lstat.kuleuven.be/java/version2.0/Applet030.html`



(a) $\rho = -.9$ (b) $\rho = .5$

Figure 3.21: ***Bivariate normal densities,*** $\sigma_x^2 = \sigma_y^2 = 1$

Question 3.4.3 *Use a computer grapher to see what happens to the graph when the variances are unequal. Check that the equal variance assumption is directly responsible for the fact that the lines around which probability weight is concentrated have slopes* -1 *and* 1.

Proposition 3.4.2 (*Marginal distribution of bivariate normal distribution*) *Let* $\mathbf{X} = (X, Y)$ *have the bivariate normal density (3.69). Then* X *has the* $N(\mu_x, {\sigma_x}^2)$ *distribution, and* Y *has the* $N(\mu_y, {\sigma_y}^2)$ *distribution.*

Proof:

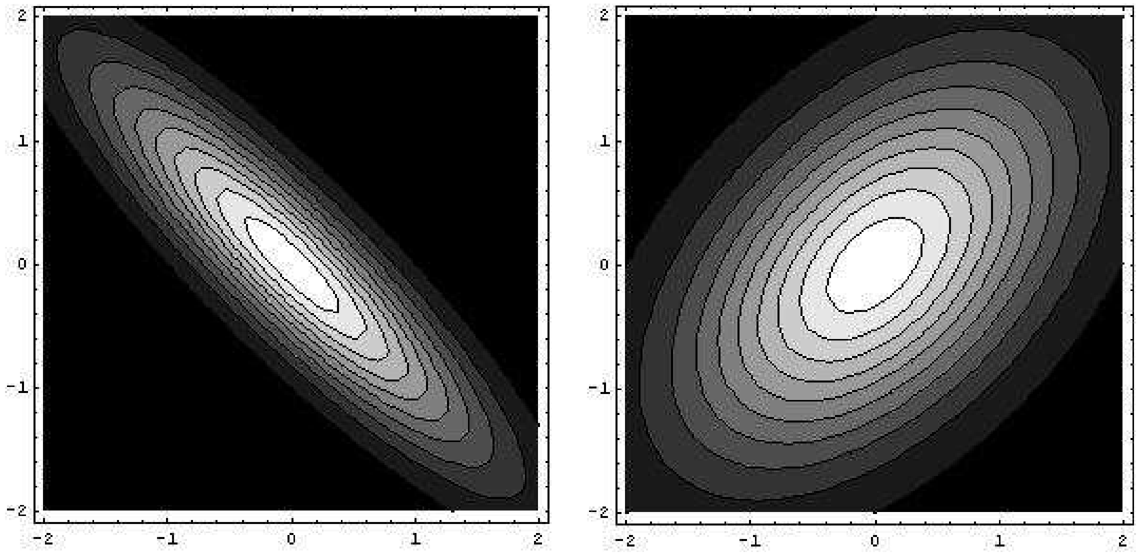
- In Exercise 9 you are to show that completing the square on y in the exponent Q in (3.70) yields

$$Q = \frac{(y-\mu_{y|x})^2}{2\sigma_y^2(1-\rho^2)} + \frac{(x-\mu_x)^2}{2\sigma_x^2}, \tag{3.77}$$

where

$$\mu_{y|x} \equiv \mu_y + \rho\frac{\sigma_y}{\sigma_x}(x-\mu_x). \tag{3.78}$$

(The reason for the odd notation $\mu_{y|x}$ will be explained in the next chapter.)



(a) $\rho = -.9$ (b) $\rho = .5$

Figure 3.22: ***Contours for surfaces in Figure 3.21***

- Then f can be rewritten as

$$f(x,y) = \frac{1}{\sqrt{2\pi\sigma_x^2}} \exp\left[-\frac{(x-\mu_x)^2}{2\pi\sigma_x^2}\right] \cdot \frac{1}{\sqrt{2\pi\sigma_y^2(1-\rho^2)}} \exp\left[-\frac{(y-\mu_{y|x})^2}{2\sigma_y^2(1-\rho^2)}\right]. \quad (3.79)$$

- The marginal density of X is the integral between $-\infty$ and $+\infty$ of (3.79) with respect to y.

- But the leading terms in x, which make up the $N(\mu_x, \sigma_x^2)$ density, can be drawn outside the integral, and the remaining integral is the integral of a normal density, which equals 1.

- This shows that X has the desired density.

- Working similarly for Y, we can also write the exponent Qas

$$Q = \frac{(x-\mu_{x|y})^2}{2\sigma_x^2(1-\rho^2)} + \frac{(y-\mu_y)^2}{2\sigma_y^2}, \quad (3.80)$$

where

$$\mu_{x|y} \equiv \mu_x + \rho\frac{\sigma_x}{\sigma_y}(x-\mu_y). \quad (3.81)$$

- Then the bivariate density f has the form

$$f(x,y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left[-\frac{(y-\mu_y)^2}{2\pi\sigma_y^2}\right] \cdot \frac{1}{\sqrt{2\pi\sigma_x^2(1-\rho^2)}} \exp\left[-\frac{(x-\mu_{x|y})^2}{2\sigma_x^2(1-\rho^2)}\right]. \quad (3.82)$$

- The marginal density of Y is the integral of (3.82) with respect to x, which, by the foregoing reasoning, reduces to the $N(\mu_y, \sigma_y^2)$ density.

- This completes the proof. □

3.5 Summary

1. ***Probability density function (p.d.f.)***: $Q(A) = P[X \in A] = \int_A f(x)\,\mathrm{d}x$. (a) $f(x) \geq 0, \forall x \in E$; (b) $\int_E f(x)\,\mathrm{d}x = 1$ 11

2. ***Cumulative distribution function (c.d.f.)***: $F(x) = P[X \leq x] = \int_{-\infty}^{x} f(t)\,\mathrm{d}t$. (a) $F'(x) = f(x)$; (b) F is a nondecreasing, nonnegative function; (c) $\lim_{x\to-\infty} F(x) = 0$; (d) $\lim_{x\to\infty} F(x) = 1$; (e) $P[a < X \leq b] = F(b) - F(a)$ 11

3. ***Continuous uniform distribution*** on $[a, b]$:

   p.d.f.: $f(x) = \begin{cases} \frac{1}{b-a} & \text{if } x \in [a,b], \\ 0 & \text{otherwise.} \end{cases}$, c.d.f.: $F(x) = \begin{cases} 0 & \text{if } x < a, \\ \frac{x-a}{b-a} & \text{if } x \in [a,b], \\ 1 & \text{if } x > b. \end{cases}$ 19

4. ***Median***: m of X if $P[X \leq m] = F(m) = 1/2$. 19

5. ***$p \times 100$th percentile***: x_p if $P[X \leq x_p] = F(x_p) = p$. 19

6. ***Uniform distribution on $[0, c]$***: Median: $1/2 = \int_0^m 1/c\,\mathrm{d}x = m/c$ gives $m = c/2$. $p \times 100$th percentile: $p = \int_0^{x_p} 1/c\,\mathrm{d}x = x_p/c$ gives $x_p = cp$. 19

7. ***Empirical distribution function***: $\widehat{F}_n(x) = \frac{\text{number of } X_i's \leq x}{n}$ 21

8. ***Multivariate c.d.f.***: $F(\mathbf{x}) = P[\mathbf{X} \leq \mathbf{x}] = P[X_1 \leq x_1, X_2 \leq x_2, \ldots, X_n \leq x_n]$ 24

9. ***Marginal density***: Let $\mathbf{X} = (X_1, \ldots, X_n)$ be a random vector with probability density function $f(\mathbf{x})$ and state space E. The marginal density of X_i is $f_i(x_i) = \int \int_{E(x_i)} \cdots \int f(x_1, \ldots, x_n)\,\mathrm{d}x_n$ $\cdots \mathrm{d}x_{i+1}\,\mathrm{d}x_{i-1} \cdots \mathrm{d}x_1$ where $E(x_i) = \{(x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n) | (x_1, \ldots, x_n) \in E\}$. 30

10. ***Expected value***: The expected value of a real-valued, continuous random variable X with p.d.f. f and state space E is $E[X] = \int_E x \cdot f(x)\,\mathrm{d}x$ provided the integral exists. the expected value of a function $g(X)$ is $E[g(X)] = \int_E g(x) \cdot f(x)\,\mathrm{d}x$. 34

11. ***Mean***: The mean of X is $\mu = E[X]$. 34

12. ***Variance***: The variance of X is $\mathrm{Var}(X) = \sigma^2 = E[(X-\mu)^2] = \int_E (x-\mu)^2 f(x)\,\mathrm{d}x$. 35

13. ***Standard deviation***: The standard deviation of X is $\sigma = \sqrt{\mathrm{Var}(X)}$. 35

14. ***Moment***: The rth moment of X is $\mu_r = E[X^r] = \int_E x^r f(x)\,\mathrm{d}x$. 35

15. ***Moment about mean***: The rth moment about the mean of X is $\mu_r' = E[(X-\mu)^r] = \int_E (x-\mu)^r f(x)\,\mathrm{d}x$. 35

16. ***Properties of mean and variance of linear combination of random variables***: (i) $E[aX + bY] = aE[X] + bE[Y]$; (ii) $\mathrm{Var}(aX + b) = a^2\mathrm{Var}(X)$ 42

17. If X is a real random variable of continuous type with finite variance and mean μ, then
$$\mathrm{Var}(X) = E\left[X^2\right] - \mu^2.$$
. 43

34. ***Cauchy density***: $f(x) = \frac{1}{\pi[1+(x-\mu)^2]}, -\infty < x < \infty$. 86

35. ***Beta density***: $f(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\, x^{\alpha-1}(1-x)^{\beta-1}, \alpha, \beta > 0, 0 < x < 1$. 86

36. ***Multivariate normal distribution***: A random vector $\mathbf{X} = (X_1, X_2, \ldots, X_n)'$ is said to have the multivariate normal distribution with parameters $\boldsymbol{\mu}$ and Σ if its density is $f(\mathbf{x}) = \frac{1}{(2\pi)^{n/2}\sqrt{\det(\Sigma)}}\, e^{-(1/2)(\mathbf{x}-\boldsymbol{\mu})'\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})}, \mathbf{x} \in \mathbb{R}^n$. 95

37. Let $\mathbf{X} = (X_1, X_2, \ldots, X_n)'$ have the multivariate normal distribution with mean $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_n)'$ and covariance Σ. If Σ is a diagonal matrix with entries $\sigma_1^2, \ldots, \sigma_n^2$ on its diagonal, then the joint density is $f(\mathbf{x}) = f_1(x_1) \cdot f_2(x_2) \cdots f_n(x_n)$ where f_i is the $N(\mu_i, {\sigma_i}^2)$ density. Consequently, under these assumptions the marginal distribution of each X_i is $N(\mu_i, {\sigma_i}^2)$. 97

38. ***Bivariate normal distribution***: $N(\boldsymbol{\mu}, \Sigma)$ where $\Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{bmatrix}$

$$f(\mathbf{x}) = \frac{1}{2\pi\sqrt{\det \Sigma}} \cdot \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})'\Sigma(\mathbf{x}-\boldsymbol{\mu})\right]$$

. 100

39. ***Marginal distribution of bivariate normal distribution***: Let $\mathbf{X} = (X, Y)$ have the bivariate normal distribution $(\mu_x, \mu_y, \sigma_x^2, \sigma_y^2, \rho)$. Then X has the $N(\mu_x, {\sigma_x}^2)$ distribution, and Y has the $N(\mu_y, {\sigma_y}^2)$ distribution. 103

Chapter 4

Conditional Distribution and Independence

Contents

4.1 Independence of random variables

- Events E and F independent if

$$P[E|F] = \frac{P[E \cap F]}{P[F]} = P[E] \,.$$

- X and Y to be independent if

$$P[X \in A|Y \in B] = \frac{P[X \in A,\ Y \in B]}{P[Y \in B]} = P[X \in A]$$

for all subsets A and B of the state spaces of X and Y, respectively, such that $P[Y \in B] \neq 0$.

- X and Y to be independent if $P[X \in A,\ Y \in B] = P[X \in A]\,P[Y \in B]$.

Definition 4.1.1 *Random variables X and Y are said to be* ***independent*** *of one another if*

$$P[X \in A, Y \in B] = P[X \in A] \cdot P[Y \in B] \tag{4.3}$$

for all sets A and B. Random variables $X_1, X_2, \ldots, X_n$ are called ***mutually independent*** *if for any subcollection of them $X_{i1}, X_{i2}, \ldots, X_{ik}, k \leq n$, and corresponding subsets $B_{i1}, B_{i2}, \ldots, B_{ik}$ of their state spaces,*

$$\begin{aligned} &P[X_{i1} \in B_{i1}, X_{i2} \in B_{i2}, \ldots, X_{ik} \in B_{ik}] \\ =&P[X_{i1} \in B_{i1}] \cdot P[X_{i2} \in B_{i2}] \cdots P[X_{ik} \in B_{ik}] \end{aligned} \tag{4.4}$$

Random variables that are not independent are called ***dependent****.*

Exercise 8 at the end of this section shows that in order to prove independence of many random variables, it is enough to prove the factorization (4.4) for the entire collection of random variables $X_1, X_2, \ldots, X_n$ rather than for every possible subcollection.

Question 4.1.1 *Show that Definition 4.1.1 implies that for any pair X_i and X_j of the collection of independent random variables, $P[X_i \in A | X_j \in B] = P[X_i \in A]$.*

Ans:

$$\begin{aligned} P[X_i \in A | X_j \in B] &= \frac{P[X_i \in A, X_j \in B]}{P[X_j \in B]} \\ &= \frac{P[X_i \in A] \cdot P[X_j \in B]}{P[X_j \in B]} = P[X_i \in A] \end{aligned}$$

□

Example 4.1.1 *A state lottery game called Pick 4 operates by putting Ping-Pong balls numbered 0-9 in each of four mixing machines, then drawing out a ball from each machine, in succession, to choose the digits of the winning number. If the mixing is good and the machines operate independently, it makes sense to assume that the random variables X_1, X_2, X_3, X_4, defined as the observed digits on machines 1-4, respectively, are independent and have discrete uniform distributions on $\{0, 1, \ldots, 9\}$.*

Ans:

- $P[X_1 > 5, X_2 < 4] = P[X_1 > 5]P[X_2 < 4] = (4/10)^2$
- $P[X_1 = 0, X_2 < 2, X_3 = 8, X_4 > 7] = (1/10)(2/10)(1/10)(2/10)$
- $P[X_1 = i, X_2 = j, X_3 = k, X_4 = l] = (1/10)^4 = 1/10000$
- The joint distribution of the X's, or the distribution of the random vector $\mathbf{X} = (X_1, X_2, X_3, X_4)$ equivalent to the winning number, is therefore uniform on the set of $10,000$ possible Pick 4 winners 0-9999. □

The next proposition characterizes independence in terms of the joint probability distribution of the random variables.

Proposition 4.1.1 *The following are equivalent:*

1. $X_1, X_2, \ldots, X_n$ *are independent random variables.*
2. *If* $F(x_1, x_2, \ldots, x_n)$ *is the joint c.d.f of* $X_1, X_2, \ldots, X_n$ *and* $F_1(x_1), F_2(x_2), \ldots, F_n(x_n)$ *are the marginal c.d.f.'s, then*
$$F(x_1, x_2, \ldots, x_n) = F_1(x_1) \cdot F_2(x_2) \cdots F_n(x_n)$$
3. *If* $f(x_1, x_2, \ldots, x_n)$ *is the joint probability density function (mass function in the discrete) of* $X_1, X_2, \ldots, X_n$ *and* $f_1(x_1), f_2(x_2), \ldots, f_n(x_n)$ *are the marginal densities(or mass functions), then*
$$f(x_1, x_2, \ldots, x_n) = f_1(x_1) \cdot f_2(x_2) \cdots f_n(x_n)$$

Proof:

- The strategy for showing this triple equivalence will be to show that $(a) \Rightarrow (b) \Rightarrow (c) \Rightarrow (a)$.
- $(a) \Rightarrow (b)$: If the random variables are independent, then we can apply Equation (4.4) to the sets $B_1 = (-\infty, x_1]$, $B_2 = (-\infty, x_2]$, …, $B_n = (-\infty, x_n]$, to obtain
$$\begin{aligned} & F(x_1, x_2, \ldots, x_n) \\ =& P[X_1 \leq x_1, X_2 \leq x_2, \ldots, X_n \leq x_n] \\ =& P[X_1 \in (-\infty, x_1], X_2 \in (-\infty, x_2], \ldots, X_n \in (-\infty, x_n]] \\ =& P[X_1 \in (-\infty, x_1]] \cdot P[X_2 \in (-\infty, x_2]] \cdots P[X_n \in (-\infty, x_n]] \\ =& F_1(x_1) \cdot F_2(x_2) \cdots F_n(x_n) \end{aligned}$$
- $(b) \Rightarrow (c)$: We will prove this implication only for $n = 2$; the argument in the continuous case extends in a straightforward way, but the discrete argument is tedious. (You will be asked in Exercise 9 for a proof when $n = 3$, which should convince you that the methods can be adapted for any desired n.)
- So, let X_1 and X_2 be random variables whose joint distribution function factors into the product of the marginal distribution functions.
- If X_1 and X_2 are of the continuous class, then the Fundamental Theorem of Calculus applied twice yields that
$$\frac{\partial^2 F}{\partial x_1 \partial x_2} = \frac{\partial^2}{\partial x_1 \partial x_2} \int_{-\infty}^{x_1} \int_{-\infty}^{x_2} f(t_1, t_2)\, \mathrm{d}t_1\, \mathrm{d}t_2 = f(x_1, x_2) \tag{4.7}$$
where f is the joint density of X_1 and X_2.
- But also,
$$\frac{\partial^2 F}{\partial x_1 \partial x_2} = \frac{\partial^2}{\partial x_1 \partial x_2}(F_1(x_1) \cdot F_2(x_2)) = f_1(x_1) \cdot f_2(x_2)$$
where f_1 and f_2 are the marginals.
- Equating the two representations of the second partial of F proves the result in the continuous case.

- If X_1 and X_2 are of the discrete class, let $B_1 = (a_1, b_1]$ and $B_2 = (a_2, b_2]$.

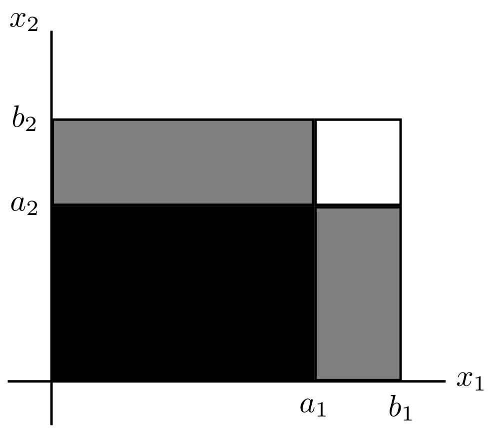


Figure 4.1: ***Light = Total − Two Grays + Dark***

- From Figure 4.1, it is easy to see that the probability that $X_1 \in B_1$ and $X_2 \in B_2$ can be broken apart as follows:

$$\begin{aligned} P[x_1 \in B_1, X_2 \in B_2] &= P[a_1 < X_1 \le b_1, a_2 < X_2 \le b_2] \\ &= P[X_1 \le b_1, X_2 \le b_2] - P[X_1 \le a_1, X_2 \le b_2] \\ &\quad - P[X_1 \le b_1, X_2 \le a_2] + P[X_1 \le a_1, X_2 \le a_2] \\ &= F(b_1, b_2) - F(a_1, b_2) - F(b_1, a_2) + F(a_1, a_2) \\ &= F_1(b_1)F_2(b_2) - F_1(a_1)F_2(b_2) \\ &\quad - F_1(b_1)F_2(a_2) + F_1(a_1)F_2(a_2) \\ &= (F_1(b_1) - F_1(a_1))(F_2(b_2) - F_2(a_2)) \\ &= P[a_1 < X_1 \le b_1]P[a_2 < X_2 \le b_2] \\ &= P[X_1 \in B_1] \cdot P[X_2 \in B_2] \end{aligned} \tag{4.8}$$

- If b_1 and b_2 are arbitrary points of positive probability for the discrete distribution, then in the limit as $a_1 \to b_1$ and $a_2 \to b_2$ from below, the foregoing computation shows that

$$\begin{aligned} P[X_1 = b_1, X_2 = b_2] &= P[X_1 = b_1] \cdot P[X_2 = b_2] \\ &\Rightarrow f(b_1, b_2) = f_1(b_1) \cdot f_2(b_2) \end{aligned}$$

where f is the joint mass function of X_1 and X_2 and f_1 and f_2 are the marginals. This finishes the proof of $(b) \Rightarrow (c)$.

- $(c) \Rightarrow (a)$: Suppose that the joint density factors as in (c).

- Then we can write, in the continuous case,

$$\begin{aligned} P[X_1 \in B_1, \dots, X_n \in B_n] &= \int_{B_1} \cdots \int_{B_n} f(x_1, \dots, x_n)\,\mathrm{d}x_n \dots \mathrm{d}x_1 \\ &= \int_{B_1} \cdots \int_{B_n} f_1(x_1) \cdots f_n(x_n)\,\mathrm{d}x_n \dots \mathrm{d}x_1 \\ &= \int_{B_1} f_1(x_1)\,\mathrm{d}x_1 \cdot \dots \cdot \int_{B_n} f_n(x_n)\,\mathrm{d}x_n \\ &= P[X_1 \in B_1] \cdot \dots \cdot P[X_n \in B_n] \end{aligned}$$

- Thus $(c) \Rightarrow (a)$ for continuous random variables.
- The discrete case is similar. □

Question 4.1.2 *Show that if the joint density factors into a product of any functions of the individual variables, not just the marginals, then the associated random variables are independent.*

Ans:

- Here, we consider continuous case.
- $f(x_1, x_2, \dots, x_n) = \prod_{i=1}^n g_i(x_i)$ for some g_i's.
- Define $c_i = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \prod_{j=1, j\neq i}^n g_j(x_j)\,\mathrm{d}x_j$, $\quad i = 1, 2, \dots, n.$
- $$\begin{aligned} \prod_{i=1}^n c_i &= \left(\prod_{i=1}^n \int_{-\infty}^{\infty} g_i(x_i)\,\mathrm{d}x_i\right)^{n-1} \\ &= \left(\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(x_1, x_2, \dots, x_n)\,\mathrm{d}x_n \dots \mathrm{d}x_1\right)^{n-1} = 1 \end{aligned}$$
- $$\begin{aligned} f_i(x_i) &= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \prod_{j=1}^n g_j(x_j)\,\mathrm{d}x_1 \dots \mathrm{d}x_{i-1}\,\mathrm{d}x_{i+1} \dots \mathrm{d}x_n \\ &= c_i \cdot g_i(x_i), \quad i = 1, 2, \cdots, n \end{aligned}$$
- Thus,
$$f(x_1, x_2, \dots, x_n) = \prod_{i=1}^n g_i(x_i) = \prod_{i=1}^n g_i(x_i) \cdot \prod_{i=1}^n c_i = \prod_{i=1}^n f_i(x_i)$$
shows that X_i's are independent.
- The discrete case is similar. □

Example 4.1.2 *Suppose that an ecologist is interested in whether a species of prairie mouse and a species of vole (野鼠) tend to avoid each other. One way of checking that they do not is to see whether the mouse and vole populations in small regions are independent. Numerous traps are set in a field, and for each of* 100 *days the catch is recorded as a pair* (M, V)*, where* M *is the (random) number of mice caught and* V *is the (random) number of voles caught. Suppose that the frequency of days on which each catch combination was observed is as appears in Figure 4.2(a). Using an empirical estimate of the distribution of* (M, V)*, does it seem as if the number of mice caught is independent of the number of voles caught?*

Ans:

- Goodness-of-fit for $f = f_M f_V$
- The marginal probability mass functions are found by totaling the joint probabilities across rows for f_M, and columns for f_V. (Fig. 4.2(b))

Mice \ Voles	0	1	2	3
0	16	8	7	6
1	20	10	4	2
2	5	5	2	1
3	9	2	2	1

(a)

Mice \ Voles	0	1	2	3	f_m
0	.16 (.185)	.08 (.093)	.07 (.056)	.06 (.037)	.37
1	.20 (.180)	.10 (.090)	.04 (.054)	.02 (.036)	.36
2	.05 (.065)	.05 (.033)	.02 (.019)	.01 (.013)	.13
3	.09 (.070)	.02 (.035)	.02 (.021)	.01 (.014)	.14
f_v	.5	.25	.15	.10	

(b)

Figure 4.2: ***Mouse and vole trap probabilities***

- In parentheses in this figure, we have computed for each (m, v) pair the product $f_M(m)f_V(v)$.
- The most serious discrepancies are at the extremes when either 3 mice or 3 voles are caught, but otherwise the agreement is close.
- We deal with goodness-of-fit tests in Chapter 12. □

Example 4.1.3 *Three components are connected in series to make up a device as in Figure 4.3. The components fail at random times* T_1, T_2, T_3*, which are independent of each other and have the Weibull distribution with parameters* $\lambda = 1, \beta = 2$*. Therefore the components have an increasing failure rate. Find the c.d.f. and density function of the system failure time. Does the system have an increasing failure rate?*

Ans:

- Let T be the system failure time. Because of the series structure, the system still survives at time t if and only if all of the components survive at time t.

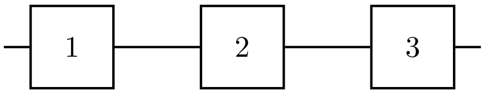


Figure 4.3: ***Series reliability system***

- $P[T > t] = P[T_1 > t,\ T_2 > t,\ T_3 > t] = P[T_1 > t]\, P[T_2 > t]\, P[T_3 > t]$
- $f(t) = \beta\lambda^\beta t^{\beta-1} e^{-(\lambda t)^\beta}$
- $F(t) = 1 - e^{-(\lambda t)^\beta}$
- $P[T > t] = P[T_1 > t, T_2 > t, T_3 > t] = (P[T_1 > t])^3 = e^{-3t^2}$
- $F(t) = 1 - P[T > t] = 1 - e^{-(\sqrt{3}t)^2}$ is Weibull($\lambda = \sqrt{3}, \beta = 2$).
- $f(t) = 6te^{-3t^2}$
- Failure rate: $h(t) = \dfrac{f(t)}{1 - F(t)} = 6t$ is an increasing function of t. □

Example 4.1.4 *Sugar Blasters cereal comes in boxes that are listed as 32 oz (by weight, before settling, of course). Past experience suggests that the actual fill weight is normally distributed with a standard deviation of 1 oz. If four such boxes were to be inspected, and all were found to have a weight of under 31 oz, would there be cause for suspicion of underfilling?*

Ans:

- $$\begin{aligned} P[X_1 < 31, \ldots, X_4 < 31] &= P\left[\frac{X_1 - 32}{1} < -1\right] \cdots P\left[\frac{X_4 - 32}{1} < -1\right] \\ &= (P[Z < -1])^4 \approx (.16)^4 = .00065 \end{aligned}$$
- Highly unlikely to occur if the assumptions are true.
- We really need tests for means, variances, and independence, which will occupy a good deal of our attention in Chapter 9. □

Question 4.1.3 *In the computation of the last example, we used independence to factor in the first line, and then we standardized each inequality in the second line. Do you think that the order of these two operations could have reversed? Give a preliminary answer now, then reconsider the question later after you have read Proposition 4.1.2.*

Ans:

- Yes, you can reverse the order of these two operations.
- The answer follows from Proposition 4.1.2. □

Proposition 4.1.2 *Suppose that $X_1, X_2, \ldots, X_n$ are mutually independent random variables, and suppose that $f_1, f_2, \ldots, f_n$ are functions whose domains include the state spaces of the corresponding $X_1, X_2, \ldots, X_n$. Then $f_1(X_1), f_2(X_2), \ldots, f_n(X_n)$ are mutually independent random variables.*

Proof:

- If A is a set in the range of a function f, define the ***inverse image*** of A as the following subset of the domain of f:

$$f^{-1}(A) = \{x | f(x) \in A\}$$

- Then note that $f_i(X_i) \in A_i$ if and only if $X_i \in f^{-1}(A_i)$. Consequently,

$$\begin{aligned} & P[f_1(X_1) \in A_1, \ldots, f_n(X_n) \in A_n] \\ &= P[X_1 \in f_1^{-1}(A_1), \ldots, X_n \in f_n^{-1}(A_n)] \\ &= P[X_1 \in f_1^{-1}(A_1)] \cdots P[X_n \in f_n^{-1}(A_n)] \\ &= P[f_1(X_1) \in A_1] \cdots P[f_n(X_n) \in A_n] \end{aligned}$$

- By Exercise 8, this sufficient to show the independence of the random variables $f_i(X_i)$. □

Example 4.1.5 *One of many useful consequences of Proposition 4.1.2 is that independence of random variables is a fundamental structural property that is not lost if the random variables are measured in a different system of units. If X and Y are independent, then $X' = aX + b$ and $Y' = cY + d$ are also independent.* □

Proposition 4.1.3 *Suppose that $X_1, X_2, \ldots, X_n$ are independent random variables, and suppose that $h_1, h_2, \ldots, h_n$ are functions whose domains include the state spaces of the corresponding $X_1, X_2, \ldots, X_n$. Then*

$$\begin{aligned} & E[h_1(X_1) \cdot h_2(X_2) \cdots h_n(X_n)] \\ &= E[h_1(X_1)] \cdot E[h_2(X_2)] \cdots E[h_n(X_n)] \end{aligned}$$

provided the expectations exist.

Proof:

- Since $Y_1 = h_1(X_1),\ Y_2 = h_2(X_2),\ \ldots,\ Y_n = h_n(X_n)$ are independent, it is enough to show that for any family of independent random variables $Y_1, Y_2, \ldots, Y_n$

$$E[Y_1 \cdot Y_2 \cdots Y_n] = E[Y_1] \cdot E[Y_2] \cdots E[Y_n] \tag{4.10}$$

- In the discrete case, by (4.6),

$$\begin{aligned} E[Y_1 \cdot Y_2 \cdots Y_n] &= \sum \sum \cdots \sum y_1 \cdot y_2 \cdots y_n f(y_1, y_2, \ldots, y_n) \\ &= \sum \sum \cdots \sum y_1 \cdot y_2 \cdots y_n f_1(y_1) \cdot f_2(y_2) \cdots f_n(y_n) \\ &= \left(\sum y_1 f_1(y_1)\right) \cdot \left(\sum y_2 f_2(y_2)\right) \cdots \left(\sum y_n f_n(y_n)\right) \\ &= E[Y_1] \cdot E[Y_2] \cdots E[Y_n] \end{aligned}$$

- The continuous case is similar. □

An important corollary of Proposition 4.1.3 is the next result on variances.

Proposition 4.1.4 (***Variance of sum of independent R.V.***) *If $X_1, X_2, \ldots, X_n$ are independent random variables, then*

$$Var\left(\sum_{i=1}^{n} c_i X_i\right) = \sum_{i=1}^{n} c_i^2\, Var(X_i)$$

provided the variances exist.

Proof:

- By Proposition 4.1.2, the random variables $Y_i = c_i X_i$ are independent, and by Proposition 2.5.3, $\text{Var}(Y_i) = \text{Var}(c_i X_i) = c_i^2 \text{Var}(X_i)$.
- Therefore it suffices to show that if $Y_1, Y_2, \ldots, Y_n$ is any collection of independent random variables, then

$$\text{Var}\left(\sum_{i=1}^{n} Y_i\right) = \sum_{i=1}^{n} \text{Var}(Y_i) \tag{4.12}$$

- To do this, note that $E[\sum_{i=1}^{n} Y_i] = \sum_{i=1}^{n} \mu_i$, where μ_i is the mean of Y_i. Thus,

$$\begin{aligned}
\text{Var}\left(\sum_{i=1}^{n} Y_i\right) &= E\left[\left(\sum_{i=1}^{n} Y_i - \sum_{i=1}^{n} \mu_i\right)^2\right] \\
&= E\left[\left(\sum_{i=1}^{n} (Y_i - \mu_i)\right)^2\right] \\
&= E\left[\sum_{i=1}^{n} (Y_i - \mu_i)^2 + \sum_{1\le j,k\le n; j\ne k} (Y_j - \mu_j)(Y_k - \mu_k)\right] \\
&= \sum_{i=1}^{n} E[(Y_i - \mu_i)^2] + \sum_{1\le j,k\le n; j\ne k} E[(Y_j - \mu_j)(Y_k - \mu_k)] \qquad (4.13)
\end{aligned}$$

  The first summation is the sum of the variances of the Y_i's, which is the desired quantity.

- The second summation is zero, since, by the independence of the Y_i's and Proposition 4.1.3,

$$E[(Y_j - \mu_j)(Y_k - \mu_k)] = E[Y_j - \mu_j] \cdot E[Y_k - \mu_k] = 0 \cdot 0 = 0$$

□

Question 4.1.4 *Write down formula (4.13) for just two random variables Y_1 and Y_2. Try to interpret the meaning of the mixed product term that is added to the sum of the variances.*

Ans:

- $\text{Var}(Y_1 + Y_2) = E[(Y_1 - \mu_1)^2] + E[(Y_2 - \mu_2)^2] + 2E[(Y_1 - \mu_1)(Y_2 - \mu_2)]$

- The intensity of the linear relation between Y_1 and Y_2.
- $\text{Cov}(Y_1, Y_2) = E[(Y_1 - \mu_1)(Y_2 - \mu_2)]$ □

Example 4.1.6 *When investors decide how to spread their wealth among common stocks and other risky assets, they must consider the probabilistic behavior of the* ***rate of return*** *on the stocks, that is, the dollars that will be profited per dollars invested.*

Ans:

- R_i: random rates of return on ith stock, $i = 1, \dots, n$
- w_i: fraction of his wealth to stock i.
- The rate of return on the combination or portfolio of assets is the random variable

$$R = w_1 R_1 + \cdots + w_n R_n.$$

- $E[R] = w_1 E[R_1] + \cdots + w_n E[R_n]$
- $\text{Var}(R) = w_1^2 \text{Var}(R_1) + \cdots + w_n^2 \text{Var}(R_n)$
- Maximize expected return while minimize risk: $\max E[R] - c\text{Var}(R)$ subject to $w_1 + \cdots + w_n = 1$. □

- A ***random sample*** $X_1, \dots, X_n$ is a collection of n independent and identically distributed (i.i.d.) random variables.
- In an actual experiment of sampling from a finite population, sampling must be done with replacement, and consistently from observation to observation in order for the independence and identical distribution assumptions to hold.
- The joint density (or mass function) of the sample variables is

$$f(x_1, x_2, \dots, x_n) = g(x_1) \cdot g(x_2) \cdots g(x_n)$$

where g is the common density of each X_i.

4.2 Conditional distributions of random variables

- Study the probability distribution of a random variable given the observed value of another.
- Discrete and continuous conditional distribution, conditional expectation.

4.2.1 Discrete conditional distributions

Example 4.2.1 *As I write this example, my doctor is conducting a survey of patients who have recently had an appointment. The survey form contains many items or statements about service, to which respondents are to answer either (1) very satisfied, (2) somewhat satisfied, (3) somewhat dissatisfied, or (4) very dissatisfied. One item involves the amount of time spent waiting in the reception area, and another asks the amount of time the doctor took to answer questions. Suppose that the number of patients responding in each of the possible combination comes out as Figure 4.4. A look at the numbers seems to suggest that the waiting time increases, patients are more likely to be dissatisfied with the doctor's performance.*

Ans:

		Waiting time				
		1	2	3	4	Total
	1	15	22	16	8	61
Doctor time	2	10	25	18	12	65
	3	5	18	20	17	60
	4	3	9	10	15	37
	Total	33	74	64	52	223

Figure 4.4: ***Doctor survey***

- X: waiting time
- Y: doctor time

$$P[Y=1|X=2] = 22/74 \approx .297$$
$$P[Y=2|X=2] = 25/74 \approx .338$$
$$P[Y=3|X=2] = 18/74 \approx .243$$
$$P[Y=4|X=2] = 9/74 \approx .122$$

but

$$P[Y=1|X=4] = 8/52 \approx .154$$
$$P[Y=2|X=4] = 12/52 \approx .231$$
$$P[Y=3|X=4] = 17/52 \approx .327$$
$$P[Y=4|X=4] = 15/52 \approx .288$$

- So the probability distribution of the patient's evaluation of the doctor changes according to the patient's experience in the waiting room.
- In each offset group, the total probability adds to 1, so each list of numbers defines a valid probability mass function.

- Also notice that, since the total number of patients surveyed was 223, these probabilities could have been computed in the following way:

$$P[Y=3|X=2]=\frac{18}{74}=\frac{18/223}{74/223}=\frac{P[Y=3\cap X=2]}{P[X=2]}. \qquad \square$$

Definition 4.2.1 *If X and Y are discrete random variables with joint probability mass function $f(x,y)$, and f_X and f_Y are the marginal mass functions, then the* ***conditional probability mass function*** *of Y given $X=x$ is*

$$f(y|x)=\frac{f(x,y)}{f_X(x)},$$

provided $f_X(x)>0$. Similarly,

$$f(x|y)=\frac{f(x,y)}{f_Y(y)},$$

provided $f_Y(y)>0$.

Question 4.2.1 *Use the table of Figure 4.4 to find the conditional p.m.f. of the waiting time evaluation given that the doctor's evaluation is 1, and the conditional p.m.f. of the waiting time evaluation given that the doctor's evaluation is 3.*

Ans:

- $f(x=1|y=1)=P[X=1,Y=1]/P[Y=1]=15/61$
  $f(x=2|y=1)=P[X=2,Y=1]/P[Y=1]=22/61$
  $f(x=3|y=1)=P[X=3,Y=1]/P[Y=1]=16/61$
  $f(x=4|y=1)=P[X=4,Y=1]/P[Y=1]=8/61$

- $f(x=1|y=3)=P[X=1,Y=1]/P[Y=3]=5/60$
  $f(x=2|y=3)=P[X=2,Y=1]/P[Y=3]=18/60$
  $f(x=3|y=3)=P[X=3,Y=1]/P[Y=3]=20/60$
  $f(x=4|y=3)=P[X=4,Y=1]/P[Y=3]=17/60$ $\square$

Question 4.2.2 *Verify that if $f(y|x)=f_Y(y)$, then X and Y are independent. Similarly, argue that if $f(x|y)=f_X(x)$ then X and Y are independent.*

Ans:

- $f_Y(y)=f(y|x)=f(x,y)/f_X(x)\Rightarrow f(x,y)=f_X(x)f_Y(y)$

- $f_X(x)=f(x|y)=f(x,y)/f_Y(y)\Rightarrow f(x,y)=f_X(x)f_Y(y)$ $\square$

Consider a random vector $\mathbf{X}=(X_1,X_2,\dots,X_n)$ with joint p.m.f. $f(x_1,x_2,\dots,x_n)$. The ***joint conditional p.m.f.*** of $X_{m+1},\dots,X_n$ given $X_1,\dots,X_m$ is

$$f(x_{m+1},\dots,x_n|x_1,\dots,x_m)=\frac{f(x_1,\dots,x_n)}{f_{1,\dots,m}(x_1,\dots,x_m)}$$

where $f_{1,\dots,m}$ is the joint marginal p.m.f. of $X_1,X_2,\dots,X_m$.

Example 4.2.2 *Random variables X, Y, and Z representing, respectively, the observed face on the first of three rolled dice, the sum of the faces on the first two dice, and the sum of the faces on all three dice have joint p.m.f*

$$f(x,y,z) = \left(\frac{1}{6}\right)^3, \quad 1 \le x \le 6,\ x+1 \le y \le x+6,\ y+1 \le z \le y+6$$

where x, y, and z are positive integers. (Justify this.) Find the conditional distribution of Z given X and Y.

Ans:

- $f(z|x,y) = \frac{f(x,y,z)}{f_{XY}(x,y)}$
- $f_{XY}(x,y) = \sum_{z=y+1}^{y+6} \left(\frac{1}{6}\right)^3 = \frac{1}{36}, \quad 1 \le x \le 6,\ x+1 \le y \le x+6$
- $f(z|x,y) = 1/6, \quad y+1 \le z \le y+6$
- The conditional mass function $f(z|x,y)$ is the same as $f(z|y)$; that is, conditioned on Y, the additional knowledge of the value of X does not change the probability distribution of Z.
- In this situation we call Z and X ***conditionally independent*** of each other, given Y. □

4.2.2 Continuous conditional distributions

$$\int_{-\infty}^{\infty} f(x,y)\,\mathrm{d}y = f_X(x)$$
$$\int_{-\infty}^{\infty} \frac{f(x,y)}{f_X(x)}\,\mathrm{d}y = 1$$

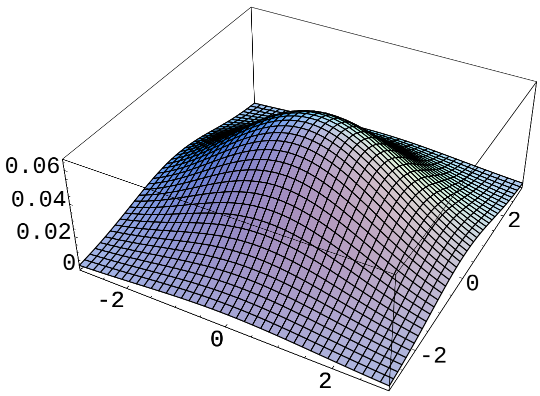


Figure 4.5: ***Bivariate normal density,*** $\sigma_x^2 = 3, \sigma_y^2 = 2, \rho = 0$

Question 4.2.3 *In the example that we have just considered, find a simplified form for the conditional density of Y given $X = x$. Geometrically, does the curve obtained after stretching the cross section depend on x?*

Ans:

- X and Y are bivariate normal distribution with both mean 0, correlation 0, and variance 3 and 2, respectively.
- $f_{XY}(x,y) = (1/2\pi\sqrt{6}) \cdot e^{-(x^2/6+y^2/4)}, \quad -\infty \le x,y \le \infty$
- $f_X(x) = (1/\sqrt{6\pi}) \cdot e^{-x^2/6}, \quad -\infty \le x \le \infty$
- $f(y|x) = f_{XY}(x,y)/f_X(x) = (1/2\sqrt{\pi}) \cdot e^{-y^2/4}, \quad -\infty \le y \le \infty$
- No. □

Definition 4.2.2 *If X and Y are continuous random variables with joint probability density function $f(x,y)$, and f_X and f_Y are the marginal density functions, then the conditional probability density function of Y given $X = x$ is*

$$f(y|x) = \frac{f(x,y)}{f_X(x)}$$

provided $f_X(x) > 0$. Similarly the conditional probability density function of X given $Y = y$ is

$$f(x|y) = \frac{f(x,y)}{f_Y(y)}$$

provided $f_Y(y) > 0$.

Example 4.2.3 *Consider the joint density*

$$f(x,y) = \begin{cases} 1/2 & \text{if } x-y \ge 0, x \le 2, x,y \ge 0, \\ 0 & \text{otherwise.} \end{cases}$$

The state space E is the triangle and its interior displayed in Fig. 4.6.

Ans:

- $f(y|x) = \frac{f(x,y)}{f_X(x)} = \frac{1/2}{x/2} = \frac{1}{x}, \quad 0 \le y < x$
- $P[Y \in [0,1/2]|X = 1] = \int_0^{1/2} \mathrm{d}y = 1/2$ □

Example 4.2.4 *If an event has nonzero probability, the conditional distribution of a random variable given that the event has occurred is easy to characterize using cumulative distribution and our original definition of conditional probability.*
Given that $N_t = 1$, the first arrival must occur somewhere in the time interval $[0,t]$.

Ans:

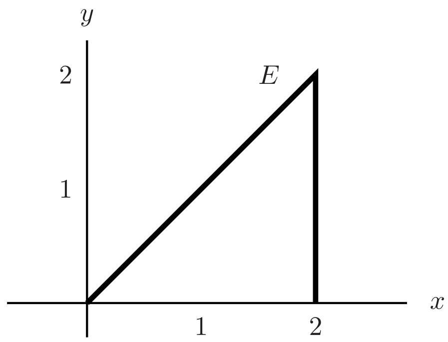


Figure 4.6: ***State space of*** (X, Y)

- The conditional c.d.f. of T_1:

$$\begin{aligned} F(s|N_t = 1) &= P[T_1 \le s|N_t = 1] \\ &= \frac{P[T_1 \le s, N_t = 1]}{P[N_t = 1]} \\ &= \frac{P[N_s = 1, N_t - N_s = 0]}{P[N_t = 1]} \\ &= \frac{(e^{-\lambda s}(\lambda s)^1/1!)(e^{-\lambda(t-s)}(\lambda(t-s))^0/0!)}{e^{-\lambda t}(\lambda t)^1/1!} \\ &= \frac{s}{t}, \quad s \in [0, t] \end{aligned}$$

- $f(s|N_t = 1) = 1/t, \quad s \in [0, t]$

- If we know that exactly one arrival occurs during the first t time units, the time of that arrival is uniformly distributed on $[0, t]$. □

4.2.3 Conditional expectation

Definition 4.2.3 *The* ***conditional expectation*** $E[g(Y)|X = x]$ *of a function of a continuous random variable given the observed value of another continuous random variable is*

$$E[g(Y)|X = x] = \int_{-\infty}^{\infty} g(y) f(y|x)\, \mathrm{d}y.$$

The integral is replaced by a sum in the discrete case.

Two special conditional expectations are the ***conditional mean*** of Y given $X = x$

$$\mu_{Y|x} = E[Y|X = x] = \int_{-\infty}^{\infty} y \cdot f(y|x)\, \mathrm{d}y$$

and the ***conditional variance*** of Y given $X = x$

$$\sigma^2_{Y|x} = E[(Y - \mu_{Y|x})^2 | X = x] = \int_{-\infty}^{\infty} (y - \mu_{Y|x})^2 f(y|x)\, \mathrm{d}y.$$

You can show (Exercise 16) that the following computational formula holds:

$$\sigma^2_{Y|x} = E[Y^2|X = x] - (\mu_{Y|x})^2$$

Question 4.2.4 *What does $E[g(Y)|X = x]$ reduce to if X and Y are independent? (The answer will appear later.)*

Ans:

- Here, we consider continuous case.
- If X and Y are independent, then $f(y|x) = f_Y(y)$.
- $E[g(Y)|X = x] = \int_{-\infty}^{\infty} g(y)f(y|x)\,\mathrm{d}y = \int_{-\infty}^{\infty} g(y)f_Y(y)\,\mathrm{d}y = E[g(Y)]$
- The discrete case is similar. □

Example 4.2.5 *Let random variables X and Y have joint density:*

$$f(x, y) = \frac{1}{64}(x + y), \quad x, y \in [0, 4].$$

Ans:

- $f_X(x) = \int_0^4 \frac{1}{64}(x + y)\,\mathrm{d}y = \frac{1}{16}x + \frac{1}{8}, \quad x \in [0, 4]$
- $f(y|x) = \dfrac{f(x, y)}{f_X(x)} = \dfrac{1/64(x + y)}{1/16(x + 2)} = \dfrac{x + y}{4(x + 2)}, \quad y \in [0, 4]$
- $\mu_{Y|x} = E[Y|X = x] = \int_0^4 y \cdot \frac{x+y}{4(x+2)}\,\mathrm{d}y = \frac{8x+64/3}{4x+8}$
- $E[Y^2|X = x] = \int_0^4 y^2 \cdot \frac{x+y}{4(x+2)}\,\mathrm{d}y = \frac{\frac{64}{3}x+64}{4x+8}$
- $\sigma^2_{Y|x} = E[Y^2|X = x] - \mu^2_{Y|x} = \dfrac{64}{9} \cdot \dfrac{3x^2 + 12x + 8}{(4x + 8)^2}$ □

Except in the special case where X and Y are independent, in which $E[g(Y)|X = x]$ is just $E[g(Y)]$, it is typical that $E[g(Y)|X = x] = h(x)$ is a function of x.

Proposition 4.2.1 (Conditional expectation formula) *If the expectations of $g(Y)$ and $h(X)$ exist, then*

$$E[g(Y)] = E[E[g(Y)|X]].$$

Proof:

- We will do the proof in the discrete case this time; the continuous case is very similar.
- As usual, let $f(x, y)$ be the joint p.m.f. of X and Y, let $f_X(x)$ be the marginal p.m.f. of X, and let $f(y|x)$ be the conditional p.m.f. of Y given $X = x$.
- Then, by (4.34)

$$h(x) = E[g(Y)|X = x] = \sum_y g(y)f(y|x)$$

The expected value of the function $h(X)$ is therefore

$$E[h(X)] = \sum_x h(x) f_X(x) = \sum_x \sum_y g(y) f(y|x) f_X(x) \tag{4.37}$$

- From the definition of conditional mass functions, however, $f(y|x)f_X(x) = f(x,y)$; hence
$$E[E[g(Y)|X]] = E[h(X)] = \sum_x \sum_y g(y) f(x,y) = E[g(Y)]$$ □

Example 4.2.6 *A duplicating machine malfunctions at the times T_i of a Poisson process with rate $1/8$ per week. When a breakdown happens, it requires one of four levels of repair, with probabilities $1/3, 1/3, 1/6,$ and $1/6$; respectively. The costs incurred in having the repair performed are \$50, \$100, \$200 and, \$400 for the four levels. Find the expected total cost of repairs through the ith week. Would a service contract that cost \$15 per week be a good deal?*

Ans:

- N_t (Poisson$(t/8)$): Number of repairs by time t
- C_i: Cost of the ith repairs
- $E[C_i] = 50/3 + 100/3 + 200/6 + 400/6 = 150$
- $C = C_1 + \cdots + C_{N_t} = \sum_{i=1}^{N_t} C_i$
- $h(n) = E[\sum_{i=1}^{N_t} C_i | N_t = n] = 150n$
- $E[\sum_{i=1}^{N_t} C_i] = E[E[\sum_{i=1}^{N_t} C_i | N_t]] = E[h(N_t)] = E[150 N_t] = 18.75t$
- We can expect to pay \$18.75 per week for repairs.
- Therefore a service contract costing \$15 a week would be cost effective. □

4.3 Covariance and correlation

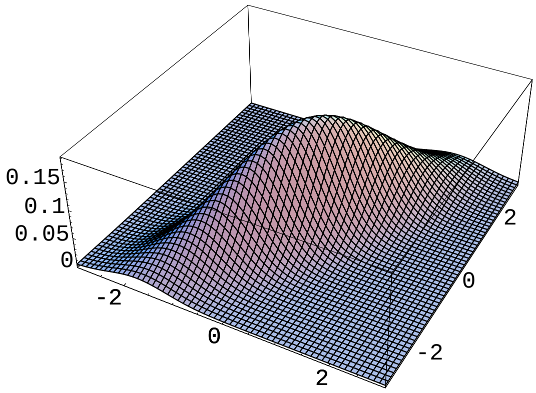
4.3.1 Main ideas

- The covariance and correlation, are relatively good measures of the degree to which one random variable is a linear function of another.
- Consider the bivariate normal joint density with parameters $\mu_X = 0, \mu_Y = 0, \sigma_X^2 = 1, \sigma_Y^2 = 2$, and $\rho = .8$ sketched in Fig. 4.7(a).
- The contour plot of curves of equal probability density is given in part (b) of the figure.
- It uses levels of shading to indicate the height of the density surface in part (a); the lighter the shading, the higher the density.
- Much of the probability weight is located around a line through the origin whose slope is about 2.

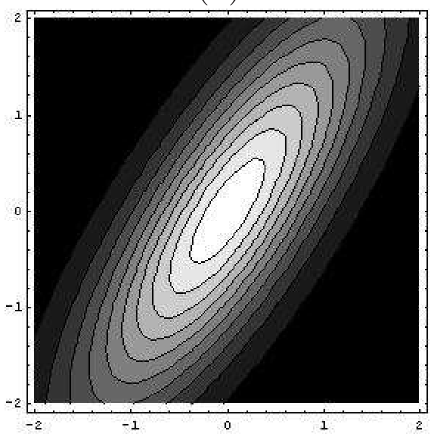
- In terms of the random variables X and Y associated with this density, given an X value, the Y value tends to be around $2X$ with high probability.
- Y tends to be large when X is large, and small when X is small.
- For this density the product

$$(X - \mu_X)(Y - \mu_Y)$$

would be large in magnitude and have positive sign with high probability.



(a)



(b)

Figure 4.7: ***Bivariate normal density:*** $\mu_X = 0,\ \mu_Y = 0,\ \sigma_X^2 = 1,\ \sigma_Y^2 = 2,\ \rho = 0.8$

Definition 4.3.1 *The **covariance** of two real random variables X and Y is*

$$\sigma_{XY} = Cov(X, Y) = E[(X - \mu_X)(Y - \mu_Y)] \tag{4.39}$$

*provided the expectation exists. If the covariance and the marginal variances exist, then the **correlation** between X and Y is*

$$\rho = \rho_{XY} = Corr(X, Y) = E\left[\left(\frac{X - \mu_X}{\sigma_X}\right)\left(\frac{Y - \mu_Y}{\sigma_Y}\right)\right] = \frac{Cov(X, Y)}{\sigma_X \sigma_Y}$$

Notice that the covariance generalizes the variance, because

$$Cov(X, X) = E[(X - \mu_X)(X - \mu_X)] = E[(X - \mu_X)^2] = Var(X).$$

Question 4.3.1 *Expand formula (4.39) to derive the computational formula:* $Cov(X, Y) = E[XY] - E[X]E[Y] = E[XY] - \mu_X \mu_Y$.

Ans:

$$\begin{aligned}\mathrm{Cov}(X, Y) &= E[(X - \mu_X)(Y - \mu_Y)] \\ &= E[XY - \mu_X Y - \mu_Y X + \mu_X \mu_Y] \\ &= E[XY] - \mu_X \mu_Y - \mu_X \mu_Y + \mu_X \mu_Y \\ &= E[XY] - \mu_X \mu_Y\end{aligned}$$

□

Example 4.3.1 *To illustrate the computation, let us find the covariance and correlation of the random variables* X *and* Y *with the joint density in Example 4.2.3:*

$$f(x, y) = 1/2, \quad \text{if } x - y \geq 0,\ x \leq 2,\ x, y \geq 0.$$

Ans:

- $f_X(x) = \frac{1}{2}x, \quad x \in [0, 2]$
- $f_Y(y) = \int_y^2 \frac{1}{2}\,\mathrm{d}x = 1 - \frac{1}{2}y, \quad y \in [0, 2]$
- $\mu_X = \int_0^2 x \cdot \frac{1}{2}x\,\mathrm{d}x = \frac{4}{3}$
- $\sigma_X^2 = E[X^2] - \mu_X^2 = \int_0^2 x^2 \cdot \frac{1}{2}x\,\mathrm{d}x - \mu_X^2 = 2 - \frac{16}{9} = \frac{2}{9}$
- $\mu_Y = 2/3, \quad \sigma_Y^2 = 2/9$
- $E[XY] = \int_0^2 \int_0^x \frac{1}{2}xy\,\mathrm{d}y\,\mathrm{d}x = 1$
- $\mathrm{Cov}(X, Y) = E[XY] - \mu_X \mu_Y = 1 - \frac{4}{3} \cdot \frac{2}{3} = 1/9$
- $\rho = \frac{1/9}{\sqrt{2/9}\sqrt{2/9}} = \frac{\mathrm{Cov}(X,Y)}{\sigma_X \sigma_Y} = 1/2$ □

Example 4.3.2 *Let* X *and* Y *have the joint discrete mass function* $f(x, y)$ *that puts probability weight* $1/4$ *on the point* $(1, 1)$, $1/2$ *on the point* $(2, 3/2)$, *and* $1/4$ *on the point* $(3, 2)$. *Find the covariance and correlation of* X *and* Y.

Ans:

- From Fig. 4.8 you can see that the states lie along a line, which should cause you to anticipate a high correlation.
- $E[X] = \frac{1}{4} \cdot 1 + \frac{1}{2} \cdot 2 + \frac{1}{4} \cdot 3 = 2$
- $\mathrm{Var}(X) = \frac{1}{4}(1 - 2)^2 + \frac{1}{2}(2 - 2)^2 + \frac{1}{4}(3 - 2)^2 = 1/2$
- $E[Y] = 3/2, \quad \mathrm{Var}(Y) = 1/8$
- $E[XY] = 13/4$
- $\mathrm{Cov}(X, Y) = 1/4, \quad \rho = 1$ □

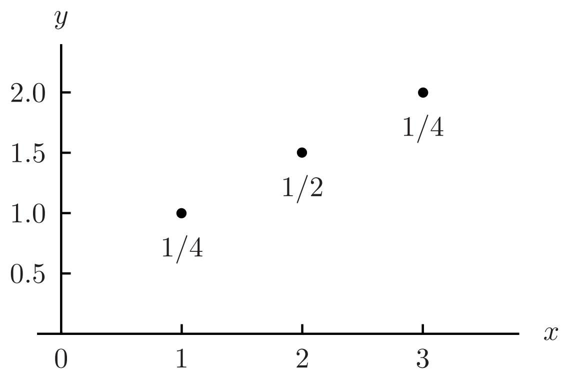


Figure 4.8: ***Perfectly correlated discrete random variables***

Question 4.3.2 *Does the fact that $\rho = 1$ in this example depend in any way on the assignment of probabilities $1/4, 1/2, 1/4$ to the three states? Try a few alternatives to see. Is the covariance affected by a change in these probabilities?*

Ans:

- Let $f(1,1) = 1/2,\ f(2,3/2) = 1/4,\ f(3,2) = 1/4.$
- $E[X] = 7/4,\ \text{Var}(X) = 11/16$
- $E[Y] = 11/8,\ \text{Var}(Y) = 11/64$
- $\text{Cov}(X,Y) = 11/32,\ \rho = 1$
- The covariance is not affected by a change in these probabilities. □

Proposition 4.3.1

1. *If X and Y are real random variables, and a, b, c, d are real constants with $b, d \neq 0$, then (bilinear)*

$$Cov(a + bX, c + dY) = b \cdot d \cdot Cov(X,Y) \tag{4.44}$$

2. *Under the hypotheses of part (1),*

$$Corr(a + bX, c + dY) = \begin{cases} Corr(X,Y) & \text{if } b, d \text{ have the same sign} \\ -Corr(X,Y) & \text{if } b, d \text{ have opposite signs} \end{cases} \tag{4.45}$$

Proof:

(1) By linearity of expectation,

$$E[a + bX] = a + b\mu_X, \quad E[c + dY] = c + d\mu_Y$$

- Thus, by the definition of covariance,

$$\begin{aligned} \text{Cov}(a + bX, c + dY) &= E[((a + bX) - (a + b\mu_X))((c + dY) - (c + d\mu_Y))] \\ &= E[b \cdot d \cdot (X - \mu_X)(Y - \mu_Y)] = b \cdot d \cdot \text{Cov}(X,Y) \end{aligned}$$

(2) By Proposition 2.5.3,

$$\mathrm{Var}(a + bX) = b^2\mathrm{Var}(X), \quad \mathrm{Var}(c + dY) = d^2\mathrm{Var}(Y)$$

- Thus

$$\begin{aligned}\mathrm{Corr}(a + bX, c + dY) &= \frac{\mathrm{Cov}(a + bX, c + dY)}{\sqrt{\mathrm{Var}(a + bX)}\sqrt{\mathrm{Var}(c + dY)}} \\ &= \frac{b \cdot d \cdot \mathrm{Cov}(X, Y)}{|b| \cdot |d|\sigma_X\sigma_Y} = \frac{b \cdot d}{|b \cdot d|} \cdot \mathrm{Corr}(X, Y)\end{aligned}$$

- The quantity $bd/|bd|$ is equal to $+1$ or -1 according to whether to whether $bd > 0$ or $bd < 0$, or respectively, whether or not b and d have the same sign.

- This result establishes (4.45). □

- Although covariance has a slightly simpler definition, correlation is a better measure of dependence than covariance from the standpoint that it is intrinsic to the random variables involved and is not dependent on their units.

- It is also better because it is bounded and its extreme values have a clear meaning.

Proposition 4.3.2

1. *If X and Y are real random variables with correlation ρ, then*

$$|\rho| \leq 1$$

2. *Moreover, $|\rho| = 1$ if and only if there are constants a, b and $b \neq 0$ such that $Y = a + bX$ with probability 1.*

3. *If X and Y are independent, then both $Cov(X, Y) = 0$ and $\rho = 0$.*

Proof:

(1) The following expectation is nonnegative for all real t:

$$E[((X - \mu_X) + t(Y - \mu_Y))^2] \geq 0$$

- Expanding out the square, it follows that

$$E[(X - \mu_X)^2] + 2tE[(X - \mu_X)(Y - \mu_Y)] + t^2E[(Y - \mu_Y)^2] \geq 0$$

that is,

$$\sigma_X^2 + 2t\sigma_{XY} + t^2\sigma_Y^2 \geq 0 \tag{4.46}$$

- The quadratic function of t defined by the left side of (4.46) therefore has either one or no real roots.

- This implies that the discriminant of the quadratic is less than or equal to 0, which means that

$$\begin{aligned}(2\sigma_{XY})^2 - 4\sigma_X^2\sigma_Y^2 \leq 0 &\Leftrightarrow \sigma_{XY}^2 \leq \sigma_X^2\sigma_Y^2 \\ &\Leftrightarrow \left(\frac{\sigma_{XY}}{\sigma_X\sigma_Y}\right)^2 \leq 1 \\ &\Leftrightarrow \rho^2 \leq 1 \end{aligned} \tag{4.47}$$

(2) If constants a, b, with $b \neq 0$ exist such that $Y(\omega) = a + bX(\omega)$ except for some outcomes $\omega \in \Omega$ of no total probability, then also the expected value of Y must be $a + b\mu_X$ and the expected value of the product $(X - \mu_X)(Y - \mu_Y)$ must be the same as the expected value of $(X - \mu_X) \cdot ((a + bX) - (a + b\mu_X))$.

- Thus, by (4.44),

$$\text{Cov}(X, Y) = \text{Cov}(X, a + bX) = b \cdot \text{Cov}(X, X) = b \cdot \sigma_X^2$$

which implies

$$|\text{Corr}(X, Y)| = \left|\frac{\text{Cov}(X, Y)}{\sigma_X\sigma_Y}\right| = \left|\frac{b\sigma_X^2}{\sigma_X\sqrt{b^2\sigma_X^2}}\right| = \left|\frac{b}{|b|}\right| = 1$$

- Conversely, suppose that $|\rho| = 1$.
- Then the chain of inequalities in (4.47) are equalities, and the quadratic function of t in (4.46) has exactly one real root t_0.
- For this t_0, retracing the argument in part (1) in reverse gives

$$\sigma_X^2 + 2t_0\sigma_{XY} + t_0^2\sigma_Y^2 = 0 \Rightarrow E[((X - \mu_X) + t_0(Y - \mu_Y))^2] = 0$$

- The nonnegative random variable $((X - \mu_X) + t_0(Y - \mu_Y))^2$ has expected value equal to 0; hence it is impossible that it can take positive values on a set of positive probability.
- Therefore

$$((X - \mu_X) + t_0(Y - \mu_Y))^2 = 0 \Rightarrow (X - \mu_X) + t_0(Y - \mu_Y) = 0$$

with probability 1.

- For all but perhaps some exceptional outcomes of total probability 0, Y is thus a linear function of X.

(3) Since X and Y are independent,

$$\sigma_{XY} = E[(X - \mu_X)(Y - \mu_Y)] = E[X - \mu_X]E[Y - \mu_Y] = 0$$

hence $\rho = 0$. □

Proposition 4.3.3 *If $X_1, X_2, \ldots, X_n$ are real random variables and $a_1, a_2, \ldots, a_n$ are real constants, then*

$$Var\left(\sum_{i=1}^{n} a_i X_i\right) = \sum_{i=1}^{n} a_i^2 \cdot Var(X_i) + \sum_{j,k=1, j\neq k}^{n} a_j \cdot a_k \cdot Cov(X_j, X_k)$$

provided the variances and covariances exist. Consequently, if each pair X_j, X_k is uncorrelated, then

$$Var\left(\sum_{i=1}^{n} a_i X_i\right) = \sum_{i=1}^{n} a_i^2 \cdot Var(X_i). \tag{4.49}$$

Proof:

- We will do the proof in the case that all $a_i = 1$. Answer Question 4.3.3 following this proof for the general case.

- The proof comes directly from the computation (4.13) in Section 4.1, repeated here for your convenience:

$$\begin{aligned}
\text{Var}\left(\sum_{i=1}^{n} X_i\right) &= E\left[\left(\left(\sum_{i=1}^{n} X_i\right) - \left(\sum_{i=1}^{n} \mu_i\right)\right)^2\right] \\
&= E\left[\left(\sum_{i=1}^{n} (X_i - \mu_i)\right)^2\right] \\
&= E\left[\sum_{i=1}^{n} (X_i - \mu_i)^2 + \sum_{j,k=1; j\neq k}^{n} (X_j - \mu_j)(X_k - \mu_k)\right] \\
&= \sum_{i=1}^{n} E[(X_i - \mu_i)^2] + \sum_{j,k=1; j\neq k}^{n} E[(X_j - \mu_j)(X_k - \mu_k)] \qquad (4.50)
\end{aligned}$$

- The first sum on the right side of the last line is the sum of the individual variances, and the second sum is of the covariances, as desired. □

Question 4.3.3 *Extend the proof of Proposition 4.3.3 to the case where the coefficients a_i are not necessarily equal to 1.*

Ans:

▶

$$\begin{aligned}
&\operatorname{Var}\left(\sum_{i=1}^{n} a_i X_i\right) \\
&= E\left[\left(\left(\sum_{i=1}^{n} a_i X_i\right) - \left(\sum_{i=1}^{n} a_i \cdot \mu_i\right)\right)^2\right] \\
&= E\left[\left(\sum_{i=1}^{n} a_i \cdot (X_i - \mu_i)\right)^2\right] \\
&= E\left[\sum_{i=1}^{n} a_i^2 \cdot (X_i - \mu_i)^2 + \sum_{j,k=1;j\neq k}^{n} a_j \cdot a_k \cdot (X_j - \mu_j)(X_k - \mu_k)\right] \\
&= \sum_{i=1}^{n} a_i^2 \cdot E\left[(X_i - \mu_i)^2\right] + \sum_{j,k=1;j\neq k}^{n} a_j \cdot a_k \cdot E\left[(X_j - \mu_j)(X_k - \mu_k)\right] \qquad \square
\end{aligned}$$

Example 4.3.3 *From a list of M prospects, an insurance salesman selects n names at random in order and without replacement. He plans to call them (during their dinner hour, of course) to sell them accidental dismemberment insurance. Suppose that among the M people, $N < M$ will be interested. Compute the mean and variance of the number of interested people in the sample of n.*

Ans:

- Number of interested people: $X = \sum_{i=1}^{n} X_i$ where

$$X_i = \begin{cases} 1 & \text{if sample person } i \text{ is interested} \\ 0 & \text{otherwise} \end{cases}$$

- $X_i \sim \text{Bernoulli}(N/M)$, but the X_i's are dependent.
- $E[X_i] = 1 \cdot P[X_i = 1] + 0 \cdot P[X_i = 0] = 1 \cdot \frac{N}{M} = \frac{N}{M}$
- $E[X] = \sum_{i=1}^{n} E[X_i] = \frac{nN}{M}$
- $\operatorname{Var}(X_i) = \frac{N}{M}\left(1 - \frac{N}{M}\right) = \frac{N(M-N)}{M^2}$
- $E[X_j X_k] = 1 \cdot P[X_j = 1, X_k = 1] + 0 = \frac{N(N-1)}{M(M-1)}$
- $\operatorname{Cov}(X_j, X_k) = E[X_j X_k] - E[X_j]E[X_k] = \frac{N(N-M)}{M^2(M-1)}$
- ▶

$$\begin{aligned}
\operatorname{Var}(X) &= \sum_{i=1}^{n} \operatorname{Var}(X_i) + \sum_{j\neq k} \operatorname{Cov}(X_j, X_k) \\
&= n \cdot \frac{N(M-N)}{M^2} + n(n-1) \cdot \frac{N(N-M)}{M^2(M-1)} \\
&= \frac{nN(M-N)(M-n)}{M^2(M-1)} \qquad \square
\end{aligned}$$

Question 4.3.4 *In the last example, for what sample size is the variance the smallest, and what is the expected number of successful calls for that sample size?*

Ans:

- Let $h(n) = n(M-n)$.
- $h(n+1) - h(n) \geq 0 \Rightarrow M - 2n - 1 \geq 0 \Rightarrow n \leq (M+1)/2$
- As $n = [(M+1)/2]$, the variance is largest.
- $E[X] = [(M+1)/2] \cdot N/M$
- As $n = M$, the variance is smallest (0).
- $E[X] = M \cdot N/M = N$ □

Example 4.3.4 *Recall Example 4.1.6 on the rate of return $R = \sum_{i=1}^n w_i R_i$ of a portfolio of n assets with individual rates of return R_i and portfolio weights w_i. The expected return us $E[R] = \sum_{i=1}^n w_i E[R_i]$. The objective is to maximize $E[R] - c\,Var(R)$ subject to $w_1 + w_2 + \cdots + w_n = 1$. Previously, we were able to find the portfolio variance only for independent rates of return. We know from (4.49) that the weaker assumption of uncorrelated rates of return would have sufficed. But even better, a direct application of formula (4.48) yields that in general the portfolio variance is*

$$Var(R) = Var\left(\sum_{i=1}^n w_i R_i\right) = \sum_{i=1}^n w_i^2 \cdot Var(R_i) + \sum_{j,k=1;j\neq k}^n w_j \cdot w_k \cdot Cov(R_j, R_k)$$

Ans:

- $E[R_1] = .051$ and $E[R_2] = .05$.
- $\mathrm{Var}(R_1) = .025^2$, $\mathrm{Var}(R_2) = .03^2$ and $\rho = -.8$.
- $w_1 = w$ and $w_2 = 1 - w$.
- $f(w) = E[R] - \mathrm{Var}(R) = -.0491 + .0034w - .002125w^2$ has maximum at $w = .8$. □

Proposition 4.3.4 *Suppose that X and Y are random variables such that the conditional mean of Y given X is a linear function of x. Then*

$$\mu_{Y|x} = \mu_Y + \frac{\rho\sigma_Y}{\sigma_X}(x - \mu_X)$$

Proof:

- By assumption, $h(x) = \mu_{Y|x} = E[Y|X = x] = mx + b$ for some slope coefficient m and intercept b.
- Proposition 4.2.1 implies that

$$\mu_Y = E[Y] = E[E[Y|X]] = E[h(X)] = E[mX + b] = m\mu_X + b$$

- Thus $b = \mu_Y - m\mu_X$, and the linear function h has the form

$$h(x) = \mu_{Y|x} = mx + b = mx + \mu_Y - m\mu_X = \mu_Y + m(x - \mu_X) \tag{4.55}$$

- It remains to show that the slope $m = \rho\sigma_Y/\sigma_X$. We can compute that

$$\begin{aligned}
\rho\sigma_X\sigma_Y = \sigma_{XY} &= E[(X - \mu_X)(Y - \mu_Y)] \\
&= E[E[(X - \mu_X)(Y - \mu_Y)|X]] \\
&= E[(X - \mu_X)E[(Y - \mu_Y)|X]] \\
&= E[(X - \mu_X)(h(X) - \mu_Y)] \\
&= E[(X - \mu_X) \cdot m(X - \mu_X)] \\
&= m \cdot \sigma_X^2 \\
\Longrightarrow m &= \frac{\rho\sigma_Y}{\sigma_X}
\end{aligned} \tag{4.56}$$ □

Example 4.3.5 *In Section 3.4 we saw an example data set in which X was the number of minutes that a U.S worker had to work for a unit of food staple and Y was the number of minutes that a Mexican worker had to work for the same food. In Chap. 8 we will have ways of using data to produce good estimates of means and variances, and correlations.*

Ans:

- Estimates for this economic data:

$$\mu_X = 7.73, \quad \mu_Y = 105.51, \quad \sigma_X = 4.60, \quad \sigma_Y = 79.20, \quad \rho = 0.60$$

- $\mu_{Y|x} = \mu_Y + \frac{\rho\sigma_Y}{\sigma_X}(x - \mu_X) = 128.96$ when $x = 10$ min.

- $\mu_{X|y} = \mu_X + \frac{\rho\sigma_X}{\sigma_Y}(y - \mu_Y) = 6.14$ when $y = 60$ min. □

4.3.2 Multivariate results

Matrix theory simplifies the characterization and study of dependence among many random variables. The key parameters are the matrices in the following definition.

Definition 4.3.2 *Let $\mathbf{X} = [X_1\ X_2\ \ldots\ X_n]'$ be a random vector. The* ***covariance matrix*** *of $\mathbf{X}$ is the $n \times n$ symmetric matrix*

$$\Sigma = Cov(\mathbf{X}) = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \ldots & \sigma_{1n} \\ \sigma_{21} & \sigma_2^2 & \ldots & \sigma_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \ldots & \sigma_n^2 \end{bmatrix}$$

where $\sigma_i^2 = Var(X_i)$ and $\sigma_{ij} = \sigma_{ji} = Cov(X_i, X_j)$.

Definition 4.3.2 *The* **correlation matrix** *of* $\mathbf{X}$ *is the* $n \times n$ *symmetric matrix*

$$\Upsilon = Corr(\mathbf{X}) = \begin{bmatrix} 1 & \rho_{12} & \dots & \rho_{1n} \\ \rho_{21} & 1 & \dots & \rho_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{n1} & \rho_{n2} & \dots & 1 \end{bmatrix}$$

where $\rho_{ij} = \rho_{ji} = Corr(X_i, X_j)$.

Question 4.3.5 *Why are the covariance and correlation matrices symmetric?*

Ans:

- $\sigma_{ij} = \sigma_{ji} = E[(X_i - \mu_i)(X_j - \mu_j)], \quad$ for all i, j.
- $\rho_{ij} = \rho_{ji} = E[(X_i - \mu_i)(X_j - \mu_j)] / \sigma_i \sigma_j, \quad$ for all i, j. □

The covariance matrix contains the marginal variances on its main diagonal, and the pair-wise covariance between the ith and jth random variable in its i, j off-diagonal component. Let $\boldsymbol{\mu} = [\mu_1\ \mu_2 \dots \mu_n]'$ be the mean vector of $\mathbf{X} = [X_1\ X_2 \dots X_n]'$. Then,

$$\begin{aligned}
&(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})' \\
= &\begin{bmatrix} X_1 - \mu_1 \\ X_2 - \mu_2 \\ \vdots \\ X_n - \mu_n \end{bmatrix} [X_1 - \mu_1\ X_2 - \mu_2 \dots\ X_n - \mu_n] \\
= &\begin{bmatrix} (X_1 - \mu_1)^2 & (X_1 - \mu_1)(X_2 - \mu_2) & \dots & (X_1 - \mu_1)(X_n - \mu_n) \\ (X_2 - \mu_2)(X_1 - \mu_1) & (X_2 - \mu_2)^2 & \dots & (X_2 - \mu_2)(X_n - \mu_n) \\ \vdots & \vdots & \ddots & \vdots \\ (X_n - \mu_n)(X_1 - \mu_1) & (X_n - \mu_n)(X_2 - \mu_2) & \dots & (X_n - \mu_n)^2 \end{bmatrix}
\end{aligned}$$

Taking the expected value entry by entry, we obtain

$$E[(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})'] = \text{Cov}(\mathbf{X})$$

Using this result, you can show (see Exercise 20, this section) that the covariance matrix is nonnegative definite. □

Proposition 4.3.5 *Let* $\mathbf{X}$ *be a random vector of* n *components whose covariance matrix* Σ *exists, and let* A *be a constant* $m \times n$ *matrix. Then,*

$$Cov(A \cdot \mathbf{X}) = A \cdot Cov(\mathbf{X}) \cdot A' = A \Sigma A'$$

Proof:

- By (4.60) we can write:

$$\begin{aligned}
\text{Cov}(A\mathbf{X}) &= E[(A\mathbf{X} - A\boldsymbol{\mu})(A\mathbf{X} - A\boldsymbol{\mu})'] \\
&= AE[(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})']A' \\
&= A\Sigma A'
\end{aligned}$$

[The third line follows from the result in Exercise 20(a).] □

One random vector $\mathbf{X} = [X_1\ X_2\ \ldots\ X_m]'$ can be stacked on another $\mathbf{Y} = [Y_1\ Y_2\ \ldots\ Y_n]'$ to produce a vector $\mathbf{Z}$ with $m+n$ components; the covariance matrix of $\mathbf{Z}$ would then have the following block structure:

$$\mathbf{Z} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_m \\ Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} \Rightarrow \Sigma_Z = \text{Cov}(\mathbf{Z}) = \begin{bmatrix} \Sigma_X & \Sigma_{XY} \\ \Sigma_{YX} & \Sigma_Y \end{bmatrix} \tag{4.62}$$

- Σ_X: $m \times m$ covariance matrix of $\mathbf{X}$.
- Σ_Y: $n \times n$ covariance matrix of $\mathbf{Y}$.
- Σ_{XY}: $m \times n$ matrix whose i, j component is $\text{Cov}(X_i, Y_j)$ and $\Sigma_{YX} = \Sigma'_{XY}$.

Proposition 4.3.6 (*Bilinear property of covariance*) *Let $\mathbf{X} = [X_1, X_2, \ldots, X_m]'$ and $\mathbf{Y} = [Y_1, Y_2, \ldots, Y_n]'$ be random vectors, and let $\mathbf{a} = [a_1, a_2, \ldots, a_m]$ and $\mathbf{b} = [b_1, b_2, \ldots, b_n]$ be constant row vectors. Then*

$$Cov\left(\sum_{i=1}^{m} a_i X_i, \sum_{j=1}^{n} b_j Y_j\right) = \sum_{i=1}^{m}\sum_{j=1}^{n} a_i b_j\, Cov(X_i, Y_j) \tag{4.63}$$

Proof:

- From the random vector $\mathbf{Z}$ displayed in (4.62), and let

$$A = \begin{bmatrix} \mathbf{a} & \mathbf{0}_n \\ \mathbf{0}_m & \mathbf{b} \end{bmatrix}$$

  where $\mathbf{0}_n$ and $\mathbf{0}_m$ are row vectors consisting entirely of zeros, of lengths n and m, respectively.

- Then

$$A \cdot \mathbf{Z} = \begin{bmatrix} \sum_{i=1}^{m} a_i X_i \\ \sum_{j=1}^{m} b_j Y_j \end{bmatrix}$$

  and hence the covariance that we are looking for is the $1, 2$ component of the covariance matrix of $A \cdot \mathbf{Z}$.

- By Proposition 4.3.5,

$$\begin{aligned} \text{Cov}(A \cdot \mathbf{Z}) &= A \cdot \text{Cov}(\mathbf{Z}) \cdot A' \\ &= \begin{bmatrix} \mathbf{a} & \mathbf{0}_n \\ \mathbf{0}_m & \mathbf{b} \end{bmatrix} \begin{bmatrix} \Sigma_X & \Sigma_{XY} \\ \Sigma_{YX} & \Sigma_Y \end{bmatrix} \begin{bmatrix} \mathbf{a}' & \mathbf{0}'_m \\ \mathbf{0}'_n & \mathbf{b}' \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{a}\Sigma_X\mathbf{a}' & \mathbf{a}\Sigma_{XY}\mathbf{b}' \\ \mathbf{b}\Sigma_{YX}\mathbf{a}' & \mathbf{b}\Sigma_Y\mathbf{b}' \end{bmatrix} \end{aligned}$$

- The desired covariance is therefore

$$\mathbf{a}\Sigma_{XY}\mathbf{b}' = [a_1\ a_2 \cdots a_m] \begin{bmatrix} \mathrm{Cov}(X_1, Y_1) & \dots & \mathrm{Cov}(X_1, Y_n) \\ \vdots & \ddots & \vdots \\ \mathrm{Cov}(X_m, Y_1) & \dots & \mathrm{Cov}(X_m, Y_n) \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix}$$

- You can verify that when the matrix product is computed, formula (4.63) results. □

Example 4.3.6 *If R_1, R_2, R_3, and R_4 are random rates of return on four risky assets, then the portfolio rates of return on two portfolios that use only the first two and the second two assets, respectively, are $w_1R_1 + w_2R_2$ and $w_3R_3 + w_4R_4$.*

Ans:

- The covariance between these two portfolio return:

$$\begin{aligned}\mathrm{Cov}(w_1R_1 + w_2R_2, w_3R_3 + w_4R_4) &= w_1w_3\mathrm{Cov}(R_1, R_3) + w_1w_4\mathrm{Cov}(R_1, R_4) \\ &\quad + w_2w_3\mathrm{Cov}(R_2, R_3) + w_2w_4\mathrm{Cov}(R_2, R_4)\end{aligned}$$

- Expression above: bilinear form
- Appropriate choosing w_i's may have the two portfolios are uncorrelated. □

4.4 More on the multivariate normal distribution

Multivariate normal density:

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{n/2}\sqrt{\det(\Sigma)}}\, e^{-(1/2)(\mathbf{X}-\boldsymbol{\mu})'\Sigma^{-1}(\mathbf{X}-\boldsymbol{\mu})}, \quad \mathbf{x} \in \mathbb{R}^n \tag{4.65}$$

where $\boldsymbol{\mu}$ is the constant $n \times 1$ vector parameter called the ***mean vector*** and Σ is the constant, symmetric $n \times n$ matrix called the ***covariance matrix***.

Bivariate normal density: $\boldsymbol{\mu} = (\mu_x, \mu_y)$, $\Sigma = \begin{bmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{bmatrix}$

$$\begin{aligned} f(x, y) &= \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \\ &\quad \times \exp\left[-\frac{1}{2(1-\rho^2)}\left[\frac{(x-\mu_x)^2}{\sigma_x^2} - 2\rho\frac{(x-\mu_x)(y-\mu_y)}{\sigma_x\sigma_y} + \frac{(y-\mu_y)^2}{\sigma_y^2}\right]\right] \end{aligned} \tag{4.66}$$

- The magnitude of ρ determines the degree of concentration of probability density around a line in the $x - y$ plane.
- The sign of ρ determines whether that line is positively or negatively sloped.
- Cross sections of the density surface perpendicular to the coordinate axes have a normal bell shape, and contour curves are elliptical, with centers at (μ_x, μ_y) and tilt angles and axis radii dependent on σ_x^2, σ_y^2, and ρ.

Proposition 4.1.1, the factorization theorem for independent random variables, immediately yields the following result.

Proposition 4.4.1 *Let the random vector* $\mathbf{X} = [X_1 \ X_2 \ \dots \ X_n]'$ *have the multivariate normal distribution with mean* $\boldsymbol{\mu} = [\mu_1 \ \mu_2 \ \dots \ \mu_n]'$ *and covariance matrix* Σ. *Then* Σ *is a diagonal matrix with diagonal entries* $\sigma_1^2, \sigma_2^2, \dots, \sigma_n^2$ *if and only if* $X_1, X_2, \dots, X_n$ *are mutually independent and* X_i *has the* $N(\mu_i, \sigma_i^2)$ *distribution.*

Question 4.4.1 *If a bivariate normal density has a diagonal covariance matrix, what will the cross sections of the density surface by planes parallel to the x-y coordinate plane look like?*

Ans:

- An ellipse without slanting.

For the remainder of this section we will concentrate on the dependent case, beginning with the bivariate normal distribution.

Proposition 4.4.2 *If* $\mathbf{X} = [X \ Y]'$ *has the bivariate normal density as described earlier, then the conditional density of* Y *given* $X = x$ *is* $N(\mu_{y|x}, \sigma_{y|x}^2)$, *where*

$$\mu_{y|x} = \mu_y + \rho\frac{\sigma_y}{\sigma_x}(x - \mu_x); \quad \sigma_{y|x}^2 = \sigma_y^2(1 - \rho^2)$$

Similarly, the conditional density of X *given* $Y = y$ *is* $N(\mu_{x|y}, \sigma_{x|y}^2)$, *where*

$$\mu_{x|y} = \mu_x + \rho\frac{\sigma_x}{\sigma_y}(y - \mu_y); \quad \sigma_{x|y}^2 = \sigma_x^2(1 - \rho^2)$$

Proof:

- Recall the factorization (3.79) from the proof of Proposition 3.4.2:

$$f(x, y) = \underbrace{\frac{1}{\sqrt{2\pi\sigma_x^2}}\exp\left(-\frac{(x-\mu_x)^2}{2\sigma_x^2}\right)}_{f_X(x)} \cdot \underbrace{\frac{1}{\sqrt{2\pi\sigma_y^2(1-\rho^2)}}\exp\left(-\frac{(y-\mu_{y|x})^2}{2\sigma_y^2(1-\rho^2)}\right)}_{f(y|x)}$$

- The first two factors makes up the marginal density f_X of the random variable X.
- Here the conditional density of Y given $X = x$ is

$$f(y|x) = \frac{f(x, y)}{f_X(x)} = \frac{1}{\sqrt{2\pi\sigma_y^2(1-\rho^2)}}\exp\left(-\frac{(y-\mu_{y|x})^2}{2\sigma_y^2(1-\rho^2)}\right) \tag{4.69}$$

  which is the desired normal density.
- You can easily do a similar argument to show that the conditional distribution of X given $Y = y$ is $N(\mu_{x|y}, \sigma_{x|y}^2)$ using factorization (3.82). □

- Proposition 4.3.4, in the case where the conditional mean of Y is linear in x, the slope coefficient is the correlation between X and Y times the ratio of standard deviations.

- Equating this to the slope coefficient in the bivariate normal conditional mean, it follows that ρ is indeed the correlation between X and Y.

Example 4.4.1 (Bivariate model) *The following information comes from a medical text (Fischbach, 1980). Fibrinogen (纖維蛋白素原) is a protein in the blood that aids in coagulation (凝結). Its normal range of concentrations is* 200-400 *milligrams per* 100 *milliliters. A simple test for fibrinogen deficiency, the whole blood clotting (凝固) test, measures the clotting time of a small volume of blood placed into a glass tube. The normal range of clotting times is* 5-10 *minutes. Increasing the expected clotting time to* 12 *minutes requires a decrease of fibrinogen concentration to about* 50*. If an individual has a clotting time of* 15 *minutes, what is the chance that his fibrinogen concentration has decreased to a level below* 100*?*

Example 4.4.1 *Let us make some reasonable interpretations and assumptions about what the given information means. We will assume that among the general population of tested individuals, the fibrinogen concentration X and the clotting time Y have the bivariate normal distribution. The phrase normal range will mean an interval of radius two standard deviations centered about the mean. Therefore, for fibrinogen, the interval* $[200, 400]$ *gives us a mean* $\mu_x = 300$ *and a standard deviation* $\sigma_x = 50$*. For clotting time, the interval* $[5, 10]$ *yields a mean* $\mu_y = 7.5$ *and a standard deviation* $\sigma_y = 1.25$*. We will assume that the other information provided to us means that the conditional expectation of clotting time for a value of fibrinogen concentration of* 50 *is* $\mu_{y|50} = 12$*. Our problem is to compute* $P[X < 100|Y = 15]$*.*

Ans:

- $\mu_{y|50} = 12 = \mu_y + \rho\frac{\sigma_y}{\sigma_x}(x - \mu_x) = 7.5 + \rho\frac{1.25}{50}(50 - 300) \Rightarrow \rho = -.72$

- Given $Y = 15$, X is normal with mean and variance

$$\mu_{x|y} = \mu_x + \rho\frac{\sigma_x}{\sigma_y}(y - \mu_y) = 84, \quad \sigma^2_{y|x} = \sigma^2_x(1 - \rho^2) = 1204$$

- $P[X < 100|Y = 15] = P\left[Z < \frac{100 - 84}{\sqrt{1204}}\right] = P[Z < .46] = .6772$ □

- Clever use of some results from matrix algebra will allow us to generalize what we have done from the bivariate to the multivariate case.

- $\mathbf{X} = [X_1, X_2, \ldots, X_n]'$ have the multivariate normal distribution with mean vector $\boldsymbol{\mu} = [\mu_1, \mu_2, \ldots, \mu_n]'$ and positive definite covariance matrix Σ.

- Partition the components of $\mathbf{X}$ into two groups, the first m and the remaining $(n-m)$ random variables.

$$\mathbf{X} = \begin{bmatrix}\mathbf{X}_1 \\ \mathbf{X}_2\end{bmatrix}, \quad \boldsymbol{\mu} = \begin{bmatrix}\boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2\end{bmatrix}, \quad \Sigma = \begin{bmatrix}\Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22}\end{bmatrix}$$

Proposition 4.4.3 *Let $\mathbf{X}, \boldsymbol{\mu}$, and Σ be as already described. Then,*

1. *$\mathbf{X}_1$ has the multivariate normal distribution with mean vector $\boldsymbol{\mu}_1$ and covariance matrix Σ_{11}.*

2. *$\mathbf{X}_2$ has the multivariate normal distribution with mean vector $\boldsymbol{\mu}_2$ and covariance matrix Σ_{22}.*

3. *Conditioned on $\mathbf{X}_1 = \mathbf{x}_1$, $\mathbf{X}_2$ has the multivariate normal distribution with mean vector and covariance matrix*

$$\boldsymbol{\mu}_{2|1} = \boldsymbol{\mu}_2 + \Sigma_{21}\Sigma_{11}^{-1}(\mathbf{x}_1 - \boldsymbol{\mu}_1); \quad \Sigma_{2|1} = \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}.$$

4. *Conditioned on $\mathbf{X}_2 = \mathbf{x}_2$, $\mathbf{X}_1$ has the multivariate normal distribution with mean vector and covariance matrix*

$$\boldsymbol{\mu}_{1|2} = \boldsymbol{\mu}_1 + \Sigma_{12}\Sigma_{22}^{-1}(\mathbf{x}_2 - \boldsymbol{\mu}_2); \quad \Sigma_{1|2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}.$$

Proof:

- We will only prove parts (2) and (4) [parts (1) and (3) are analogous].

- Proofs of both will fall out simultaneously once we obtain a factorization of the joint density $f(\mathbf{x})$ in (4.65).

- To that end, it can be verified (check the matrix products yourself in forthcoming Question 4.4.2) that

$$\begin{bmatrix} I & -\Sigma_{12}\Sigma_{22}^{-1} \\ 0 & I \end{bmatrix} \cdot \Sigma \cdot \begin{bmatrix} I & 0 \\ -\Sigma_{22}^{-1}\Sigma_{21} & I \end{bmatrix} = \begin{bmatrix} \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} & 0 \\ 0 & \Sigma_{22} \end{bmatrix} \tag{4.73}$$

  where the I's are identity matrices of the appropriate size.

- Since both matrices surrounding Σ have determinant equal to 1, equating the determinants of both sides yields

$$\det(\Sigma) = \det(\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}) \cdot \det(\Sigma_{22}) = \det(\Sigma_{1|2}) \cdot \det(\Sigma_{22}) \tag{4.74}$$

- The relation (4.73) also allows us to factor Σ^{-1}. Multiply on the left and right by the inverses of the two matrices that surround Σ to obtain

$$\Rightarrow \Sigma = \begin{bmatrix} I & -\Sigma_{12}\Sigma_{22}^{-1} \\ 0 & I \end{bmatrix}^{-1} \begin{bmatrix} \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} & 0 \\ 0 & \Sigma_{22} \end{bmatrix} \begin{bmatrix} I & 0 \\ -\Sigma_{22}^{-1}\Sigma_{21} & I \end{bmatrix}^{-1}$$

$$\Rightarrow \Sigma^{-1} = \begin{bmatrix} I & 0 \\ -\Sigma_{22}^{-1}\Sigma_{21} & I \end{bmatrix} \begin{bmatrix} (\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})^{-1} & 0 \\ 0 & \Sigma_{22}^{-1} \end{bmatrix} \begin{bmatrix} I & -\Sigma_{12}\Sigma_{22}^{-1} \\ 0 & I \end{bmatrix}. \tag{4.75}$$

- The factorization enables us to write

$$(\mathbf{x} - \boldsymbol{\mu})'\Sigma^{-1}(\mathbf{x} - \boldsymbol{\mu}) = (\mathbf{x}_1 - \boldsymbol{\mu}_{1|2})' \cdot \Sigma_{1|2}^{-1} \cdot (\mathbf{x}_1 - \boldsymbol{\mu}_{1|2}) + (\mathbf{x}_2 - \boldsymbol{\mu}_2)'\Sigma_{22}^{-1}(\mathbf{x}_2 - \boldsymbol{\mu}_2). \tag{4.76}$$

(Prove this result in Exercise 11 at the end of this section.)

- By Eqs. (4.74)-(4.76), the multivariate normal density (4.65) can therefore be factored as

$$f(\mathbf{x}) = \underbrace{\frac{1}{(2\pi)^{m/2}\sqrt{\det(\Sigma_{1|2})}} \cdot e^{-(1/2)(\mathbf{x}_1-\boldsymbol{\mu}_{1|2})'\Sigma_{1|2}^{-1}(\mathbf{x}_1-\boldsymbol{\mu}_{1|2})}}_{f(\mathbf{x}_1|\mathbf{x}_2)} \cdot \underbrace{\frac{1}{(2\pi)^{(n-m)/2}\sqrt{\det(\Sigma_{22})}} \cdot e^{-(1/2)(\mathbf{x}_2-\boldsymbol{\mu}_2)'\Sigma_{22}^{-1}(\mathbf{x}_2-\boldsymbol{\mu}_2)}}_{f_2(\mathbf{x}_2)}. \tag{4.77}$$

- Since the first pair of factors in (4.77) forms a multivariate normal density in $\mathbf{x}_1$, the integral of $f(\mathbf{x})$ over $\mathbf{x}_1$ gives the second pair of factors.

- In other words, the marginal density of $\mathbf{X}_2$ is multivariate normal with mean vector $\boldsymbol{\mu}_2$ and covariance matrix Σ_{22}.

- The first pair of factors is therefore the quotient of the joint density $f(\mathbf{x})$ and the marginal density $f_2(\mathbf{x}_2)$, that is, the conditional density of $\mathbf{X}_1$ given $\mathbf{X}_2 = \mathbf{x}_2$.

- The parameters of this conditional density match those in the statement of part (4); hence both (2) and (4) are proved. □

Question 4.4.2 *Check Equation (4.73).*

Ans:

- $\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$

-

$$\begin{aligned}
&\begin{bmatrix} I & -\Sigma_{12}\Sigma_{22}^{-1} \\ 0 & I \end{bmatrix} \cdot \Sigma \cdot \begin{bmatrix} I & 0 \\ -\Sigma_{22}^{-1}\Sigma_{21} & I \end{bmatrix} \\
&= \begin{bmatrix} I & -\Sigma_{12}\Sigma_{22}^{-1} \\ 0 & I \end{bmatrix} \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \begin{bmatrix} I & 0 \\ -\Sigma_{22}^{-1}\Sigma_{21} & I \end{bmatrix} \\
&= \begin{bmatrix} I & -\Sigma_{12}\Sigma_{22}^{-1} \\ 0 & I \end{bmatrix} \begin{bmatrix} \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} & \Sigma_{12} \\ 0 & \Sigma_{22} \end{bmatrix} \\
&= \begin{bmatrix} \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} & 0 \\ 0 & \Sigma_{22} \end{bmatrix}
\end{aligned}$$

□

Example 4.4.2 (Trivariate normal model) *In Exercise 14 of Section 3.3 we mentioned a data set on ozone levels in Pennsylvania forests. It is reasonable to think that from one day to the next, and possibly to the third day, there is a correlation between ozone levels, but as more time elapses, levels are nearly independent. On this basis, I looked at some of the data and produced a sample of triples* (X_1, X_2, X_3) *on successive days, where one triple was separated by a few days from another. From these I estimated means, variances, and the covariance matrix using methods we will discuss in Chapters 7 and 8. Using a trivariate normal model with the following mean and covariance, let us compute the expected value of the ozone level on the third day and find an interval centered about that mean in which* X_3 *should lie with probability* 95%, *if on the first two days levels of* 70 *and* 75 *were observed:*

- $$\mu = \begin{bmatrix} 52.8 \\ 52.4 \\ \hline 55.3 \end{bmatrix}, \quad \Sigma = \left[\begin{array}{cc|c} 472.75 & 177.83 & 156.77 \\ 177.83 & 311.74 & 157.94 \\ \hline 156.77 & 157.94 & 247.21 \end{array}\right]$$
- The conditional distribution of X_3 is normal.
- To compute its conditional mean and variance, use part (c) of the proposition, with $\mathbf{X}_1 = [X_1 \ X_2]'$ and $\mathbf{X}_2 = [X_3]'$.
- Then
$$\boldsymbol{\mu}_1 = \begin{bmatrix} 52.8 \\ 52.4 \end{bmatrix}, \quad \boldsymbol{\mu}_2 = [55.3], \quad \Sigma_{11} = \begin{bmatrix} 472.75 & 177.83 \\ 177.83 & 311.74 \end{bmatrix},$$
$$\Sigma_{12} = \begin{bmatrix} 156.77 \\ 157.94 \end{bmatrix}, \quad \Sigma_{22} = [247.21].$$
- Condition on $X_1 = 70$ and $X_2 = 75$.
- $\mu_{2|1} = \mu_2 + \Sigma_{21}\Sigma_{11}^{-1}(\mathbf{x}_1 - \mu_1) = 67.5$
- $\Sigma_{2|1} = \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12} = 155.2$
- $$\begin{aligned} 0.95 &= P\left[-1.96 \le \frac{X_3 - 67.5}{\sqrt{155.2}} \le 1.96 \,\middle|\, X_1 = 70, X_2 = 75\right] \\ &= P[43.1 \le X_3 \le 91.9 | X_1 = 70, X_2 = 75] \end{aligned}$$
- The interval $[43.1, 91.91]$ is very wide; hence the information about first two days does not give a very precise prediction of what the ozone level on the third day will be. □

Question 4.4.3 *In the previous example, find the* 80% *prediction interval centered about the mean. Is it substantially shorter, and if so, why would you have expected it to be?*

Ans:

- $$\begin{aligned} 0.8 &= P\left[-1.28 \le \frac{X_3 - 67.5}{\sqrt{155.2}} \le 1.28 \,\middle|\, X_1 = 70, X_2 = 75\right] \\ &= P[51.6 \le X_3 \le 83.4 | X_1 = 70, X_2 = 75] \end{aligned}$$

- ▶ The percent of prediction interval centered about mean is lower, the length of that interval is shorter. □

4.5 Summary

1. ***Mutually independent***: Random variables $X_1, X_2, \ldots, X_n$ are called mutually independent if for any subcollection of them $X_{i1}, X_{i2}, \ldots, X_{ik}, k \leq n$, and corresponding subsets $B_{i1}, B_{i2}, \ldots, B_{ik}$ of their state spaces, $P[X_{i1} \in B_{i1}, X_{i2} \in B_{i2}, \ldots, X_{ik} \in B_{ik}] = P[X_{i1} \in B_{i1}] \cdot P[X_{i2} \in B_{i2}] \cdots P[X_{ik} \in B_{ik}]$ 5

2. ***Dependent***: Random variables that are not independent are called dependent. . . 5

3. The following are equivalent:

   (a) $X_1, X_2, \ldots, X_n$ are independent random variables.

   (b) If $F(x_1, x_2, \ldots, x_n)$ is the joint c.d.f of $X_1, X_2, \ldots, X_n$ and $F_1(x_1), F_2(x_2), \ldots, F_n(x_n)$ are the marginal c.d.f.'s, then

   $$F(x_1, x_2, \ldots, x_n) = F_1(x_1) \cdot F_2(x_2) \cdots F_n(x_n).$$

   (c) If $f(x_1, x_2, \ldots, x_n)$ is the joint probability density function (mass function in the discrete) of $X_1, X_2, \ldots, X_n$ and $f_1(x_1), f_2(x_2), \ldots, f_n(x_n)$ are the marginal densities(or mass functions), then

   $$f(x_1, x_2, \ldots, x_n) = f_1(x_1) \cdot f_2(x_2) \cdots f_n(x_n).$$

   9

4. Suppose that $X_1, X_2, \ldots, X_n$ are mutually independent random variables, and suppose that $f_1, f_2, \ldots, f_n$ are functions whose domains include the state spaces of the corresponding $X_1, X_2, \ldots, X_n$. Then $f_1(X_1), f_2(X_2), \ldots, f_n(X_n)$ are mutually independent random variables. 25

5. ***Inverse image***: If A is a set in the range of a function f, define the inverse image of A as the following subset of the domain of f: $f^{-1}(A) = \{x | f(x) \in A\}$. 26

6. Suppose that $X_1, X_2, \ldots, X_n$ are independent random variables, and suppose that $h_1, h_2, \ldots, h_n$ are functions whose domains include the state spaces of the corresponding $X_1, X_2, \ldots, X_n$. Then $E[h_1(X_1) \cdot h_2(X_2) \cdots h_n(X_n)] = E[h_1(X_1)] \cdot E[h_2(X_2)] \cdots E[h_n(X_n)]$ provided the expectations exist. 27

7. ***Variance of sum of independent R.V.***: If $X_1, X_2, \ldots, X_n$ are independent random variables, then $\text{Var}\left(\sum_{i=1}^{n} c_i X_i\right) = \sum_{i=1}^{n} c_i^2 \text{Var}(X_i)$ provided the variances exist. 29

8. ***Random sample***: A random sample $X_1, \ldots, X_n$ is a collection of n independent and identically distributed (i.i.d.) random variables. 34

9. ***Conditional probability mass function***: If X and Y are discrete random variables with joint probability mass function $f(x, y)$, and f_X and f_Y are the marginal mass functions, then the conditional probability mass function of Y given $X = x$ is $f(y|x) = \frac{f(x,y)}{f_X(x)}$ provided $f_X(x) > 0$. 40

10. ***Joint conditional p.m.f.***: Consider a random vector $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ with joint p.m.f. $f(x_1, x_2, \ldots, x_n)$. The joint conditional p.m.f. of $X_{m+1}, \ldots, X_n$ given $X_1, \ldots, X_m$ is $f(x_{m+1}, \ldots, x_n | x_1, \ldots, x_m) = \frac{f(x_1, \ldots, x_n)}{f_{1,\ldots,m}(x_1, \ldots, x_m)}$. 43

11. ***Conditional probability density function***: If X and Y are continuous random variables with joint probability density function $f(x, y)$, and f_X and f_Y are the marginal density functions, then the conditional probability density function of Y given $X = x$ is $f(y|x) = \frac{f(x,y)}{f_X(x)}$ provided $f_X(x) > 0$. Similarly the conditional probability density function of X given $Y = y$ is $f(x|y) = \frac{f(x,y)}{f_Y(y)}$ provided $f_Y(y) > 0$. 47

12. ***Conditional expectation***: The conditional expectation $E[g(Y)|X = x]$ of a function of a continuous random variable given the observed value of another continuous random variable is $E[g(Y)|X = x] = \int_{-\infty}^{\infty} g(y) f(y|x)\, dy$. The integral is replaced by a sum in the discrete case. 52

13. ***Conditional mean***: The conditional mean of Y given $X = x$ is $\mu_{Y|x} = E[Y|X = x] = \int_{-\infty}^{\infty} y \cdot f(y|x)\, dy$. 53

14. ***Conditional variance***: The conditional variance of Y given $X = x$ is $\sigma^2_{Y|x} = E[(Y - \mu_{Y|x})^2 | X = x] = \int_{-\infty}^{\infty} (y - \mu_{Y|x})^2 f(y|x)\, dy$. 53

15. ***Conditional expectation formula***: If the expectations of $g(Y)$ and $h(X)$ exist, then $E[g(Y)] = E[E[g(Y)|X]]$. 55

16. ***Covariance***: The covariance of two real random variables X and Y is $\sigma_{XY} = \mathrm{Cov}(X, Y) = E[(X - \mu_X)(Y - \mu_Y)]$ provided the expectation exists. 62

17. ***Correlation***: If the covariance and the marginal variances exist, the correlation between X and Y is $\rho = \rho_{XY} = \mathrm{Corr}(X, Y) = E\left[\left(\frac{X - \mu_X}{\sigma_X}\right)\left(\frac{Y - \mu_Y}{\sigma_Y}\right)\right] = \frac{\mathrm{Cov}(X,Y)}{\sigma_X \sigma_Y}$ . . . 62

18. $\mathrm{Cov}(X, X) = \mathrm{Var}(X)$ 62

19. If X and Y are real random variables, and a, b, c, d are real constants with $b, d \neq 0$, then $\mathrm{Cov}(a + bX, c + dY) = b \cdot d \cdot \mathrm{Cov}(X, Y)$. 67

20. $\mathrm{Corr}(a + bX, c + dY) = \mathrm{Corr}(X, Y)$ (if b, d have the same sign) or $= -\mathrm{Corr}(X, Y)$ (if b, d have opposite signs). 67

21. If X and Y are real random variables with correlation ρ, then $|\rho| \leq 1$. Moreover, $|\rho| = 1$ if and only if there are constants a, b and $b \neq 0$ such that $Y = a + bX$ with probability 1. If X and Y are independent, then both $\mathrm{Cov}(X, Y) = 0$ and $\rho = 0$. . 70

22. If $X_1, X_2, \ldots, X_n$ are real random variables and $a_1, a_2, \ldots, a_n$ are real constants, then $\mathrm{Var}\left(\sum_{i=1}^n a_i X_i\right) = \sum_{i=1}^n a_i^2 \cdot \mathrm{Var}(X_i) + \sum_{j,k=1, j\neq k}^n a_j \cdot a_k \cdot \mathrm{Cov}(X_j, X_k)$ provided the variances and covariances exist. Consequently, if each pair X_j, X_k is uncorrelated, then $\mathrm{Var}\left(\sum_{i=1}^n a_i X_i\right) = \sum_{i=1}^n a_i^2 \cdot \mathrm{Var}(X_i)$. 75

23. Suppose that X and Y are random variables such that the conditional mean of Y given X is a linear function of x. Then $\mu_{Y|x} = \mu_Y + \frac{\rho \sigma_Y}{\sigma_X}(x - \mu_X)$. 85

24. **Covariance matrix**: The covariance matrix of $\mathbf{X}$ is the $n\times n$ symmetric matrix $\Sigma =$ $\mathrm{Cov}(\mathbf{X}) = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \dots & \sigma_{1n} \\ \sigma_{21} & \sigma_2^2 & \dots & \sigma_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \dots & \sigma_n^2 \end{bmatrix}$ where $\sigma_i^2 = \mathrm{Var}(X_i)$ and $\sigma_{ij} = \sigma_{ji} = \mathrm{Cov}(X_i, X_j)$. 89

25. **Correlation matrix**: The correlation matrix of $\mathbf{X}$ is the $n \times n$ symmetric matrix $\Upsilon = \mathrm{Corr}(\mathbf{X}) = \begin{bmatrix} 1 & \rho_{12} & \dots & \rho_{1n} \\ \rho_{21} & 1 & \dots & \rho_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{n1} & \rho_{n2} & \dots & 1 \end{bmatrix}$ where $\rho_{ij} = \rho_{ji} = \mathrm{Corr}(X_i, X_j)$. 90

26. Let $\mathbf{X}$ be a random vector of n components whose covariance matrix Σ exists, and let A be a constant $m \times n$ matrix. Then, $\mathrm{Cov}(A \cdot \mathbf{X}) = A \cdot \mathrm{Cov}(\mathbf{X}) \cdot A' = A\Sigma A'$. . 93

27. **Bilinear property of covariance**: Let $\mathbf{X} = [X_1, X_2, \dots, X_m]'$ and $\mathbf{Y} = [Y_1, Y_2, \dots, Y_n]'$ be random vectors, and let $\mathbf{a} = [a_1, a_2, \dots, a_m]$ and $\mathbf{b} = [b_1, b_2, \dots, b_n]$ be constant row vectors. Then $\mathrm{Cov}\left(\sum_{i=1}^m a_i X_i, \sum_{j=1}^n b_j Y_j\right) = \sum_{i=1}^m \sum_{j=1}^n a_i b_j \mathrm{Cov}(X_i, Y_j)$. 95

28. **Multivariate normal density**:

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{n/2}\sqrt{\det(\Sigma)}}\, e^{-(1/2)(\mathbf{x}-\boldsymbol{\mu})'\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})}, \quad \mathbf{x} \in \mathbb{R}^n$$

. . . . 99

29. **Bivariate normal density**: $\boldsymbol{\mu} = (\mu_x, \mu_y)$, $\Sigma = \begin{bmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{bmatrix}$

$$f(x, y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \times \exp\left[-\frac{1}{2(1-\rho^2)}\left[\frac{(x-\mu_x)^2}{\sigma_x^2} - 2\rho\frac{(x-\mu_x)(y-\mu_y)}{\sigma_x\sigma_y} + \frac{(y-\mu_y)^2}{\sigma_y^2}\right]\right]$$

. . . . 100

30. Let the random vector $\mathbf{X} = [X_1\ X_2\ \dots\ X_n]'$ have the multivariate normal distribution with mean $\boldsymbol{\mu} = [\mu_1\ \mu_2\ \dots\ \mu_n]'$ and covariance matrix Σ. Then Σ is a diagonal matrix with diagonal entries $\sigma_1^2, \sigma_2^2, \dots, \sigma_n^2$ if and only if $X_1, X_2, \dots, X_n$ are mutually independent and X_i has the $N(\mu_i, \sigma_i^2)$ distribution. 101

31. If $\mathbf{X} = [X\ Y]'$ has the bivariate normal density as described earlier, then the conditional density of Y given $X = x$ is $N(\mu_{y|x}, \sigma_{y|x}^2)$, where $\mu_{y|x} = \mu_y + \rho\frac{\sigma_y}{\sigma_x}(x - \mu_x)$; $\sigma_{y|x}^2 = \sigma_y^2(1-\rho^2)$. Similarly, the conditional density of X given $Y = y$ is $N(\mu_{x|y}, \sigma_{x|y}^2)$, where $\mu_{x|y} = \mu_x + \rho\frac{\sigma_x}{\sigma_y}(y - \mu_y)$; $\sigma_{x|y}^2 = \sigma_x^2(1-\rho^2)$. 102

32. Let $\mathbf{X} = \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{bmatrix}$, $\boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix}$, and $\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$. Then, $\mathbf{X}_1$ has the multivariate normal distribution with mean vector $\boldsymbol{\mu}_1$ and covariance matrix Σ_{11}. $\mathbf{X}_2$ has the multivariate normal distribution with mean vector $\boldsymbol{\mu}_2$ and covariance matrix Σ_{22}. Conditioned on $\mathbf{X}_1 = \mathbf{x}_1$, $\mathbf{X}_2$ has the multivariate normal distribution with mean

vector and covariance matrix $\boldsymbol{\mu}_{2|1} = \boldsymbol{\mu}_2 + \Sigma_{21}\Sigma_{11}^{-1}(\mathbf{x}_1 - \boldsymbol{\mu}_1); \Sigma_{2|1} = \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}$. Conditioned on $\mathbf{X}_2 = \mathbf{x}_2$, $\mathbf{X}_1$ has the multivariate normal distribution with mean vector and covariance matrix $\boldsymbol{\mu}_{1|2} = \boldsymbol{\mu}_1 + \Sigma_{12}\Sigma_{22}^{-1}(\mathbf{x}_2 - \boldsymbol{\mu}_2); \Sigma_{1|2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$. 110

Chapter 5

Transformations of Random Variables

Contents

5.1 Distribution function technique

- The binomial distribution, with all of its many useful applications, arises as the distribution of the sum of independent and identically distributed Bernoulli random variables.

- If a random variable X has the $N(\mu, \sigma^2)$ is distribution, then the transformed random variable $Z = (X - \mu)/\sigma$ is standard normal, which permits us to rely on the standard normal table to find arbitrary normal probabilities.

- The time T_n of the nth arrival in a Poisson process is the sum $S_1 + S_2 + \cdots + S_n$ of exponentially distributed interarrival times, and we have used without proof the fact that the distribution of T_n is of the gamma family.

- The idea of transforming random variables is not new to us, but now we would like to study transformations more systematically and with more depth.

c.d.f. technique: Find the distribution of a real-valued, continuous function of a continuous random variable.

- X be a continuous r.v. with density function $f_X(x)$.
- $Y = g(X)$
- B_y is some subset of the state space of X dependent on y.
- $F_Y(y) = P[Y \le y] = P[g(X) \le y] = P[X \in B_y] = \int_{B_y} f_X(x)\,\mathrm{d}x$
- $f_Y(y) = F_Y'(y) = \frac{\mathrm{d}}{\mathrm{d}y}\int_{B_y} f_X(x)\,\mathrm{d}x$

Example 5.1.1 *In the financial analysis of prices of commodities of various kinds, it is sometimes assumed that from one time, labeled* 0*, to the next, labeled* 1*, the logarithm of the ratio* P_1/P_0 *of prices at the two times is normally distributed with some mean* μ *and variance* σ^2*. Denote by* A *the normal random variable* $\ln(P_1/P_0) = \ln(P_1) - \ln(P_0)$*. It is easy to invert this relationship between* A *and* P_1 *to find that* $P_1 = P_0 \cdot e^A$*. Considering* P_0 *to be fixed, what is the distribution* P_1*?*

- $A \sim N(\mu, \sigma^2)$
- $F(p) = P[P_1 \le p] = P[P_0 \cdot e^A \le p] = P[A \le \ln(p/P_0) = \ln(p) - \ln(P_0)] = \int_{-\infty}^{\ln(p)-\ln(P_0)} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(a-\mu)^2}{2\sigma^2}\right) \mathrm{d}a, \quad p > 0$
- $f(p) = \frac{1}{p\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(\ln p - \ln P_0 - \mu)^2}{2\sigma^2}\right]$ □

Application of c.d.f. technique to simulation:

- There are widely available algorithms, used by many computers, that can simulate $U[0,1]$.
- Generate random sample from a discrete r.v. $P[X = x_i] = p_i, i = 1, 2, 3$.

$$X(\omega) = \begin{cases} x_1 & \text{if } 0 \le U(\omega) \le p_1 \\ x_2 & \text{if } p_1 < U(\omega) \le p_1 + p_2 \\ x_3 & \text{otherwise.} \end{cases}$$

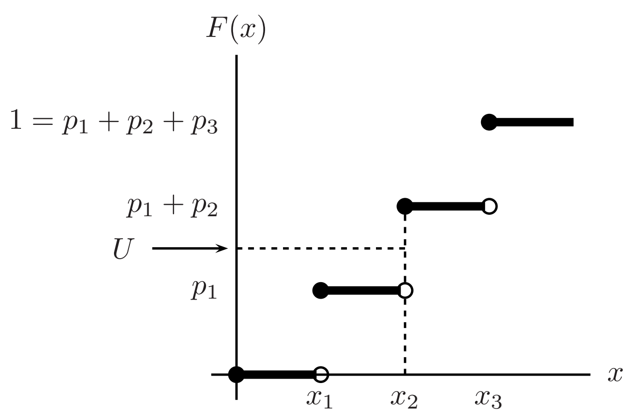


Figure 5.1: ***Simulating from a discrete distribution***

Question 5.1.1 *Try to generalize (5.5) to a formula suitable for an n-point discrete distribution.*

Ans:

- $P[X = x_1] = P[0 \leq U \leq p_1] = p_1$
- $P[X = x_2] = P[p_1 < U \leq p_1 + p_2] = p_2$
- $P[X = x_3] = P[p_1 + p_2 < U \leq p_1 + p_2 + p_3] = p_3$

$\vdots$

- $P[X = x_n] = P[p_1 + p_2 + \cdots + p_{n-1} < U \leq p_1 + p_2 + \cdots + p_n] = p_n$ □

Proposition 5.1.1 (*Inverse transformation of* $U(0,1)$) *Let F be a continuous, strictly increasing c.d.f. on a state space I that is a subinterval of the real line. Suppose that U has the uniform $(0,1)$ distribution. Then $X = F^{-1}(U)$ is a random variable with the c.d.f. F. Conversely, suppose that X is a random variable with continuous, strictly increasing c.d.f F. Then the random variable $U = F(X)$ has the uniform $(0,1)$ distribution.*

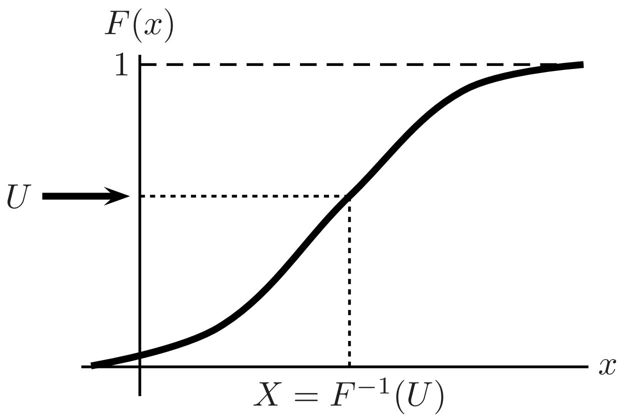


Figure 5.2: ***Simulating from a continuous distribution***

Proof:

- Beginning with the first implication, the cumulative distribution function of X is

$$P[X \leq x] = P[F^{-1}(U) \leq x] = P[U \leq F(x)] = F(x), \quad x \in I.$$

- In the third equation we take advantage of the fact that the c.d.f. of U is the identity function.
- Thus X has the c.d.f. F.
- For the converse, assume that X has c.d.f. F.
- Since F^{-1} exists and is strictly increasing (you should verify this), the c.d.f. of $U = F(X)$ is

$$P[U \leq u] = P[F(X) \leq u] = P[X \leq F^{-1}(u)] = F(F^{-1}(u)) = u.$$

- This computation is valid as long as u is in the domain of F^{-1}, which is the range of F, namely $(0, 1)$.
- (One or more of the endpoints 0 or 1 may also be in the range of F, depending on whether the state space I is bounded on one end or the other. This detail does not change the gist of argument.)
- Since the c.d.f. of U has the proper form, U is uniformly distributed on $(0, 1)$. □

Example 5.1.2 *Many practitioners of probability must simulate large, complicated service systems that are not amenable to exact analysis, in order to reach conclusions about the efficiency of the system. This often entails simulation of the interarrival times of a Poisson process. As we have learned, a typical interarrival time S of a Poisson process has the exponential(λ) distribution, whose associated c.d.f. is*

$$F(s) = 1 - e^{-\lambda s}, \quad s > 0.$$

The inverse of F, which you will check in the next question

$$F^{-1}(u) = -\frac{1}{\lambda}\ln(1-u), \quad u \in [0, 1).$$

By the last proposition, if U has the uniform $(0, 1)$ distribution, then the random variable $S = F^{-1}(U) = -\ln(1-U)/\lambda$ is exponential with parameter λ. A sequence of such uniform random variables then gives rise, by this transformation, to a sequence of interarrival times, which determines a particular outcome of a Poisson process. □

Question 5.1.2 *Verify that the function in (5.7) is the inverse of the exponential c.d.f.*

Ans:

- $F(s) = 1 - e^{-\lambda s} = u$
- $1 - u = e^{-\lambda s}$
- $\ln(1-u) = -\lambda s$
- $F^{-1}(u) = s = -\frac{1}{\lambda}\ln(1-u), \quad 0 \le u < 1$ □

The next example illustrates how the c.d.f. technique can also be used to find the distribution of a function of more than one random variable.

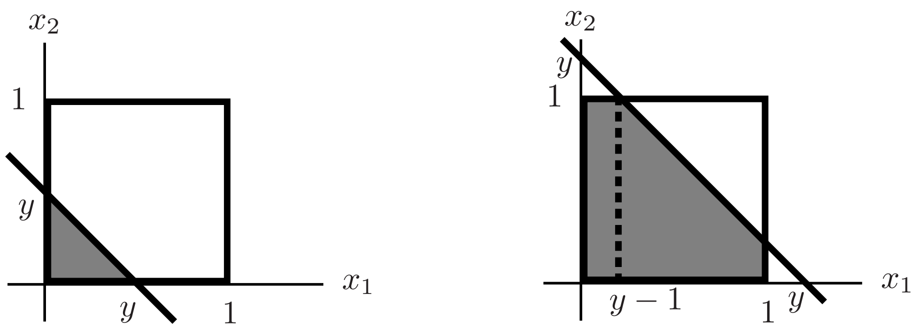
Example 5.1.3 *Let X_1 and X_2 be independent and identically distributed random variables with density $f(x) = 2x$, $x \in (0, 1)$, and let $Y = X_1 + X_2$. Find the density of Y.*

- $f(x_1, x_2) = 4x_1x_2, \quad x_1, x_2 \in (0, 1)$
- $F_Y(y) = P[Y \le y] = P[X_1 + X_2 \le y]$

$$F_Y(y) = \int_0^1 \int_0^{y-x_1} 4x_1x_2 \,\mathrm{d}x_2 \,\mathrm{d}x_1 = y^4/6, \quad \text{if } y \in [0, 1]$$

$$F_Y(y) = \int_0^{y-1} \int_0^1 4x_1x_2 \,\mathrm{d}x_2 \,\mathrm{d}x_1 + \int_{y-1}^1 \int_0^{y-x_1} 4x_1x_2 \,\mathrm{d}x_1 \,\mathrm{d}x_2$$
$$= 1 - 8y/3 + 2y^2 - y^4/6, \quad \text{if } y \in [1, 2]$$

- $f_Y(y) = \begin{cases} \frac{2}{3}y^3 & \text{if } y \in (0,1) \\ -\frac{8}{3} + 4y - \frac{2}{3}y^3 & \text{if } y \in (1,2). \end{cases}$ □



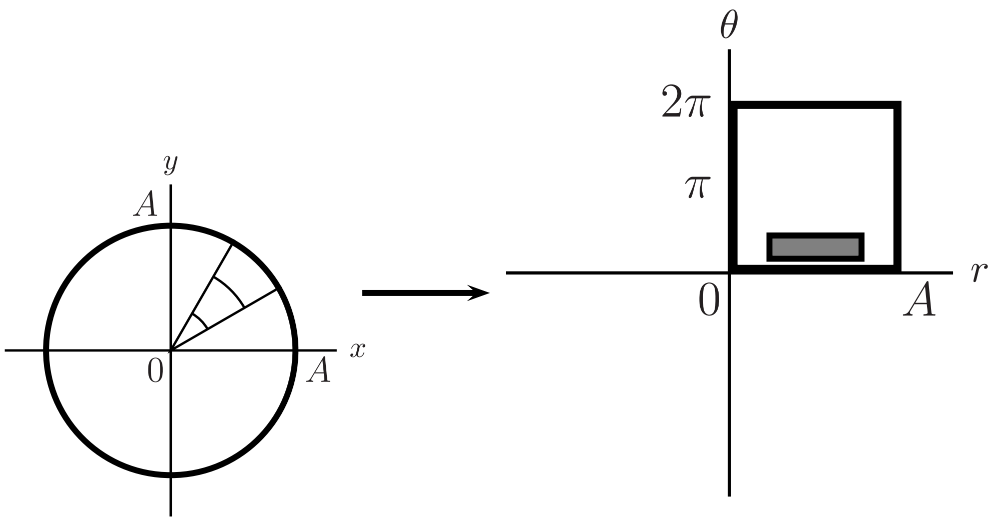
(a) Case $0 \leq y \leq 1$ (b) Case $1 < y \leq 2$

Figure 5.3: ***Finding the distribution of a sum***

- So far, we have said little about transformations on discrete random variables.
- The reason is that when they do come up, it is often in the special setting of sums of i.i.d. random variables, which can be treated by the generating function techniques introduced later.

5.2 Multivariate transformations

5.2.1 Distributions of transformed random vector



(a) (b)

Figure 5.4: ***Polar transformation***

- Consider a dart thrown randomly at a circular board such as the one shown in Fig. 5.4, which is centered at the origin and has radius A.

- Interpret the word randomly is to say the joint density of coordinates (X, Y) of the landing point of the dart is uniform over the disk:

$$f(x, y) = \frac{1}{\pi A^2}, \quad \text{if} \quad x^2 + y^2 \leq A^2.$$

- Instead of representing the point $\mathbf{X} = (X, Y)$ in rectangular coordinates, we could represent it in polar coordinates (R, Θ).
- $R^2 = X^2 + Y^2$, $\tan(\Theta) = Y/X$

A fresh look at some results from single- and multivariate calculus on changes of variables in integrals is a good idea.

- $\int_0^\infty 3x^2 \cdot e^{-x^3} \, \mathrm{d}x$
- $u = g(x) = -x^3$, $\mathrm{d}u = -3x^2 \, \mathrm{d}x$
- $u = g(x) = -x^3$ is a 1-1 function from the region of integration $[0, \infty)$ onto the interval $(-\infty, 0]$, and its inverse function is $x = g^{-1}(u) = -\sqrt[3]{u}$.
- $\int_0^\infty 3x^2 \cdot e^{-x^3} \, \mathrm{d}x = \int_0^{-\infty} 3(-\sqrt[3]{u})^2 \cdot e^{-(-\sqrt[3]{u})^3} \left(-\frac{1}{3}u^{-2/3}\right) \mathrm{d}u$
- The last integral is of the form

$$\int_{-\infty}^{0} 3(g^{-1}(u))^2 \cdot e^{-(g^{-1}(u))^3} \left| -\frac{\mathrm{d}}{\mathrm{d}u} g^{-1}(u) \right| \mathrm{d}u$$

- In general, an integral $\int_A f(x) \, \mathrm{d}x = \int_B f(g^{-1}(u)) \left| \frac{\mathrm{d}}{\mathrm{d}u} g^{-1}(u) \right| \mathrm{d}u$ where the set B is the image of set A under the transformation.
- When a transformation involves two variables in a double integral,

$$\iint_A f(x, y) \, \mathrm{d}y \, \mathrm{d}x,$$

let $u = g_1(x, y)$ and $v = g_2(x, y)$, assume the transformation is 1-1; hence the inverse transformation

$$x = h_1(u, v), \quad y = h_2(u, v).$$

- $$J = J(u, v) = \det \begin{bmatrix} \frac{\partial h_1}{\partial u} & \frac{\partial h_1}{\partial v} \\ \frac{\partial h_2}{\partial u} & \frac{\partial h_2}{\partial v} \end{bmatrix}$$
- J is called the ***Jacobian*** of the transformation.
- The two-variable transformation formula is

$$\iint_A f(x, y) \, \mathrm{d}y \, \mathrm{d}x = \iint_B f(h_1(u, v), h_2(u, v)) |J(u, v)| \, \mathrm{d}u \, \mathrm{d}v.$$

Question 5.2.1 *Using the two-variable transformation formula as a model, guess at the general n variables transformation formula.*

Ans:

$$\int_A \cdots \int f(x_1, x_2, \ldots, x_n)\, dx_1 \cdots dx_n =$$

$$\int_B \cdots \int f(h_1(y_1, \ldots, y_n), \ldots, h_n(y_1, \ldots, y_n))|J(y_1, \ldots, y_n)|\, dy_1 \cdots dy_n$$

□

Proposition 5.2.1 (*Bivariate transformation*) *Suppose that X and Y are continuous random variables with joint density $f(x, y)$. Let U and V be obtained from X and Y via the invertible transformation (5.14), whose inverse is in (5.15). Let J be the Jacobian of the transformation. Then the joint density of U and V is:*

$$f(h_1(u, v), h_2(u, v))|J(u, v)|.$$

Example 5.2.1 *We can now find the joint density of R and Θ in the dartboard example.*

- Polar transformation: $x = r\cos\theta, \quad y = r\sin\theta$.
- Jacobian of the transformation:

$$J(r, \theta) = \det \begin{bmatrix} \frac{\partial x}{\partial r} & \frac{\partial x}{\partial \theta} \\ \frac{\partial y}{\partial r} & \frac{\partial y}{\partial \theta} \end{bmatrix} = \det \begin{bmatrix} \cos\theta & -r\sin\theta \\ \sin\theta & r\cos\theta \end{bmatrix} = r$$

- Joint density:

$$f_{R,\Theta}(r, \theta) = f(r\cos\theta, r\sin\theta)|J(r, \theta)| = \frac{1}{\pi A^2} r, \quad r \in [0, A],\ \theta \in [0, 2\pi]$$

□

Question 5.2.2 *Try to explain the nonuniformity of the joint density in Example 5.2.1. What is the marginal density of Θ?*

Ans:

- $f_{R,\theta}(r, \theta) = \frac{1}{\pi A^2} r$, if r is greater, the density is greater.
- $f_\Theta(\theta) = \int_0^A \frac{1}{\pi A^2} r\, dr = 1/(2\pi)$ □

Example 5.2.2 *Let S_1 and S_2 be the first two interarrival times of a Poisson process with rate λ, and let $T_1 = S_1$ and $T_2 = S_1 + S_2$ be the associated arrival times. Compute the joint density of T_1 and T_2.*

- Since S_1 and S_2 are i.i.d. exponential (λ) random variables, their joint density is

$$f_{S_1,S_2}(s_1, s_2) = \lambda^2 e^{-\lambda(s_1+s_2)}, \quad s_1, s_2 > 0.$$

- $t_1 = s_1$ and $t_2 = s_1 + s_2$
- $s_1 = t_1$ and $s_2 = t_2 - t_1$

- $J = \det \begin{bmatrix} 1 & 0 \\ -1 & 1 \end{bmatrix} = 1$

- $f_{T_1,T_2}(t_1, t_2) = \lambda^2 e^{-\lambda t_2}, \quad 0 < t_1, t_1 < t_2$ □

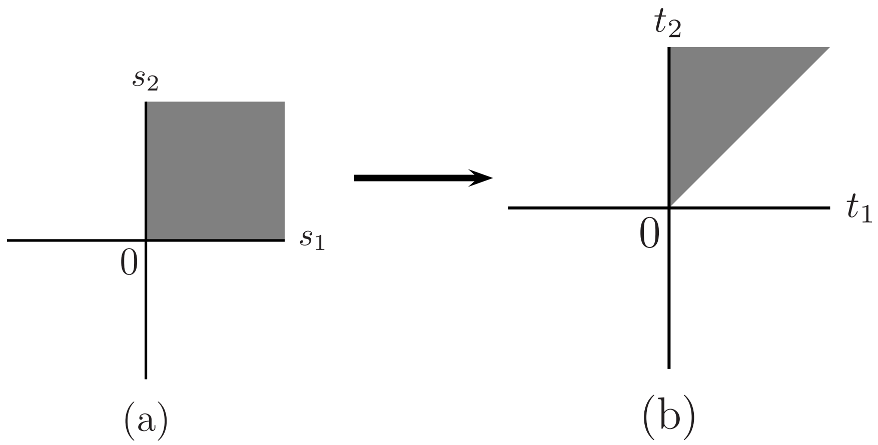


Figure 5.5: ***Transformation from interarrival to arrival times***

> **Question 5.2.3** *Devise a general theorem analogous to Proposition 5.2.1 about the density of a one-dimensional continuous random variable $Y = g(X)$, obtained as an invertible function of another one-dimensional continuous random variable.*

Ans:

- X is a continuous random variable with density $f_X(x)$.

- $Y = g(X) \Rightarrow X = g^{-1}(Y)$

- $f_Y(y) = f_X(g^{-1}(y)) \cdot \left| \frac{\mathrm{d}}{\mathrm{d}y} g^{-1}(y) \right|$ □

- Proposition 5.2.1 extends in a natural way to joint transformations of more than two random variables. Let

$$U_1 = g_1(X_1, \ldots, X_n),\ U_2 = g_2(X_1, \ldots, X_n), \ldots,\ U_n = g_n(X_1, \ldots, X_n)$$

be a 1-1 transformation with inverse

$$X_1 = h_1(U_1, \ldots, U_n),\ X_2 = h_2(U_1, \ldots, U_n), \ldots,\ X_n = h_n(U_1, \ldots, U_n).$$

- ***Jacobian of the transformation***

$$J = J(u_1, u_2, \ldots, u_n) = \det \begin{bmatrix} \frac{\partial h_1}{\partial u_1} & \frac{\partial h_1}{\partial u_2} & \cdots & \frac{\partial h_1}{\partial u_n} \\ \frac{\partial h_2}{\partial u_1} & \frac{\partial h_2}{\partial u_2} & \cdots & \frac{\partial h_2}{\partial u_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial h_n}{\partial u_1} & \frac{\partial h_n}{\partial u_2} & \cdots & \frac{\partial h_n}{\partial u_n} \end{bmatrix}$$

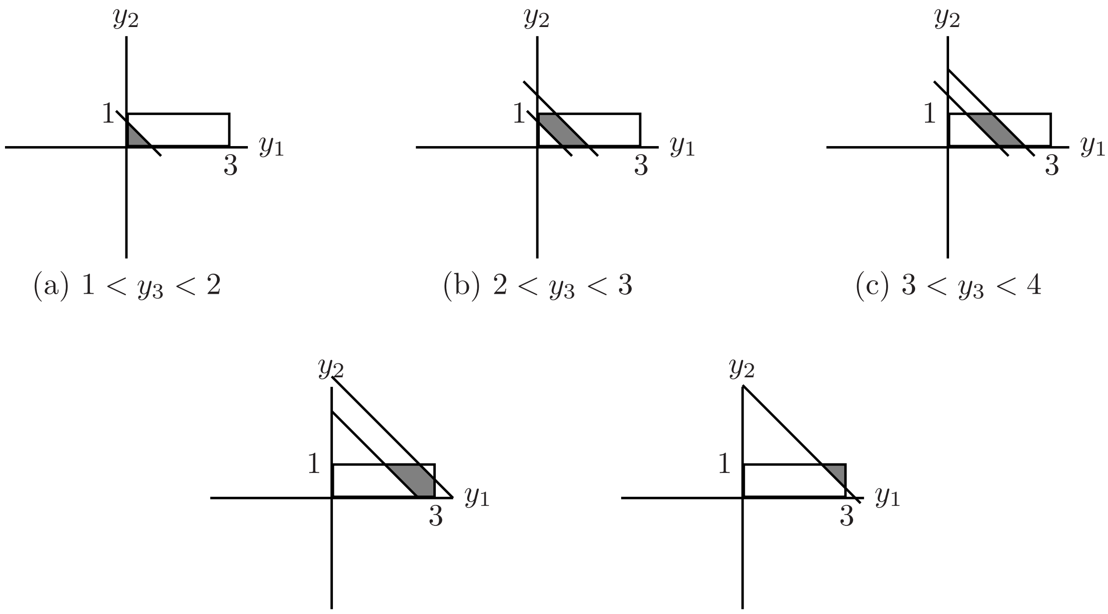
- $$\begin{aligned}&\int \cdots \int_A f(x_1, x_2, \ldots, x_n)\,\mathrm{d}x_n \ldots \mathrm{d}x_2\,\mathrm{d}x_1 \\ &= \int \cdots \int_B f(h_1, h_2, \ldots, h_n)|J|\,\mathrm{d}u_1\,\mathrm{d}u_2 \ldots \mathrm{d}u_n\end{aligned}$$
- The joint density of $U_1, U_2, \ldots, U_n$ is $f(h_1, h_2, \ldots, h_n)|J|$.

Example 5.2.3 *Suppose that you are writing a paper for a history class, which is due in five weeks. The project requires three phases: library research, the production of a first draft, and the revision of the first draft to the final draft. One phase must completed before its successor can begin. If the time required for the research is distributed uniformly on the interval* 0 *week to* 3 *weeks, the time for the first draft is uniform on* 1 *week to* 2 *weeks, what is the distribution of the completion time of the paper? What is the probability that it will be late?*

- X_1, X_2, X_3: Times required for the research, first draft, and final draft.
- $X_1 \sim U(0,3);\ X_2 \sim U(0,1);\ X_3 \sim U(1,2)$
- $f(x_1, x_2, x_3) = 1/3, \quad x_1 \in [0,3], x_2 \in [0,1], x_3 \in [1,2]$
- $Y_1 = X_1, \quad Y_2 = X_2, \quad Y_3 = X_1 + X_2 + X_3$
- $X_1 = Y_1, \quad X_2 = Y_2, \quad X_3 = Y_3 - Y_1 - Y_2$
- $J = \det \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ -1 & -1 & 1 \end{bmatrix} = 1$
- $f_{Y_1,Y_2,Y_3}(y_1, y_2, y_3) = 1/3, \quad y_1 \in [0,3], y_2 \in [0,1], y_1 + y_2 + 1 \le y_3 \le y_1 + y_2 + 2$
- $$f_{Y_3}(y_3) = \begin{cases} (y_3^2 - 2y_3 + 1)/6 & \text{if } y_3 \in (1,2) \\ (-y_3^2 + 6y_3 - 7)/6 & \text{if } y_3 \in (2,3) \\ 1/3 & \text{if } y_3 \in (3,4) \\ (-y_3^2 + 8y_3 - 14)/6 & \text{if } y_3 \in (4,5) \\ (y_3^2 - 12y_3 + 36)/6 & \text{if } y_3 \in (5,6) \end{cases} \qquad \text{(Exercise 5.2.8)}$$
- $P[Y_3 > 5] = \int_5^6 (y_3^2 - 12y_3 + 36)/6\,\mathrm{d}y_3 = 1/18$ □

5.2.2 Order statistics

- $X_1, X_2, \ldots, X_n$: random samples
- Order statistics: $Y_1 \le Y_2 \le \cdots \le Y_n$ $(X_{(1)} \le X_{(2)} \le \cdots \le X_{(n)})$
- If the sample values are the results of a poll in which the respondents are asked to rate the effectiveness of a political officeholder on a scale of 0-100, the smallest order statistic Y_1 and the largest Y_n, give an indication of the range of public sentiment.



(a) $1 < y_3 < 2$ (b) $2 < y_3 < 3$ (c) $3 < y_3 < 4$

(d) $4 < y_4 < 5$ (e) $5 < y_5 < 6$

Figure 5.6: ***Cases for marginal of** Y_3*

- If a system of n identical, independent components with failure times $X_1, X_2, \ldots, X_n$, respectively, is built in a series structure, then the system failure time is the smallest X_i.
- More than the extreme values of a random sample are of interest here.
- In fact, in some samples the extreme values result from errors in data gathering or reporting.
- These errors can sometimes have drastic effects on the sample mean if the mean is used to measure the center point of the data.
- The middle value of the sample, however, is not affected by changes to the extremes. This statistic is called the ***median*** of the sample.
- If the components are organized in a parallel structure, then the system failure time is the largest X_i. ***Median***: $\begin{cases} Y_{(n+1)/2} & \text{if } n \text{ is odd;} \\ \frac{Y_{n/2}+Y_{n/2+1}}{2} & \text{if } n \text{ is even;} \end{cases}$
- How are the probability distributions of the order statistics found from the distribution from which the sample was taken?
- Let $\mathbf{U} = g(\mathbf{X})$ be a transformation such that the state space E_X of $\mathbf{X}$ can be partitioned into several subsets $E_1, E_2, \ldots, E_k$, as in Fig. 5.7, where g is 1-1 with an inverse transformation h_i when restricted to E_i.
- $x_1 = h_{i1}(u_1, u_2, \ldots, u_n),\ x_2 = h_{i2}(u_1, u_2, \ldots, u_n), \ldots, x_n = h_{in}(u_1, u_2, \ldots, u_n)$
- If J_i denotes the Jacobian associated with the ith inverse transformation h_i, then the generalized transformation theorem states that the joint density of $\mathbf{U} = (U_1, U_2, \ldots, U_n)$ is the sum, over all of the transformations:

$$f_U(u_1, u_2, \ldots, u_n) = \sum_{i=1}^{k} f_X(h_{i1}, h_{i2}, \ldots, h_{in}) \cdot |J_i|. \tag{5.26}$$

(Exercise 15)

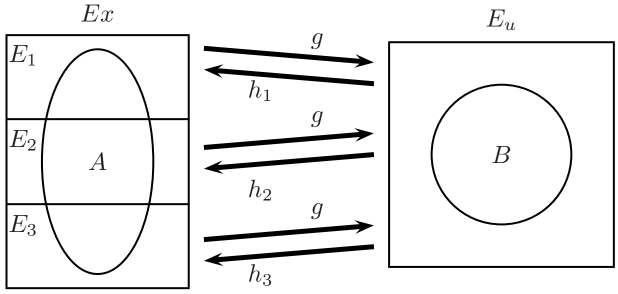


Figure 5.7: ***A noninvertible transformation***

Proposition 5.2.2 (***Joint pdf of order statistics***) *Let $Y_1 \le Y_2 \le \cdots \le Y_n$ be the order statistics of a random sample $X_1, X_2, \ldots, X_n$ taken from a continuous distribution with p.d.f f. Then the joint density of the Y's is*

$$f_Y(y_1, y_2, \ldots, y_n) = n! \cdot f(y_1) \cdot f(y_2) \cdots f(y_n), \quad y_1 < y_2 < \cdots < y_n.$$

Proof:

- Since the X's have a continuous distribution, we can safely ignore the zero probability events that sample values coincide with one another, and therefore we can break the state space of the X's into $n!$ subsets, on each of which a different ordering of the coordinates $(x_1, x_2, \ldots, x_n)$ of the state of the random vector is in effect.

- For example, here are a few such subsets, with the associated inverse transformations taking the observed x_i's back to the y_i's:

- On $E_1 = \{x_1 < x_2 < x_3 < \cdots < x_n\}$,

$$x_1 = h_{11}(\mathbf{y}) = y_1,\ x_2 = h_{12}(\mathbf{y}) = y_2,\ x_3 = h_{13}(\mathbf{y}) = y_3, \ldots,$$

$$x_n = h_{1n}(\mathbf{y}) = y_n.$$

- On $E_2 = \{x_2 < x_1 < x_3 < \cdots < x_n\}$,

$$x_1 = h_{21}(\mathbf{y}) = y_2,\ x_2 = h_{22}(\mathbf{y}) = y_1,\ x_3 = h_{23}(\mathbf{y}) = y_3, \ldots,$$

$$x_n = h_{2n}(\mathbf{y}) = y_n.$$

- On $E_3 = \{x_1 < x_3 < x_2 < \cdots < x_n\}$,

$$x_1 = h_{31}(\mathbf{y}) = y_1,\ x_2 = h_{32}(\mathbf{y}) = y_3,\ x_3 = h_{33}(\mathbf{y}) = y_2, \ldots,$$

$$x_n = h_{3n}(\mathbf{y}) = y_n.$$

- Because of the special form of each of the inverse transformations, in each of the Jacobians J_i the matrix of partials has exactly one 1 in every row and every column and the rest of the matrix entries are 0.

- It is not hard to show that the determinant of such a matrix is equal to 1 or to -1; hence the $|J_i|$ term in (5.26) is simply equal to 1.

- Also, since the $X_i's$ are a random sample, the joint density of the X's has the form $f_X(\mathbf{x}) = f(x_1) \cdot f(x_2) \cdots f(x_n)$.

- Regardless of which inverse transformation h_i you look at, when it is substituted into f_X there will be exactly one $f(y_1)$ factor, exactly one $f(y_2)$ factor, and so forth, contained in the product.

- Therefore the joint density of the order statistics simplifies to

$$\begin{aligned} f_Y(y_1, y_2, \ldots, y_n) &= \sum_{i=1}^{n!} f(y_1) f(y_2) \cdots f(y_n) \cdot 1 \\ &= n! \cdot f(y_1) \cdot f(y_2) \cdots f(y_n), \quad y_1 < y_2 < \cdots < y_n. \end{aligned}$$ □

The marginal distribution of the kth order statistic Y_k

- The infinitesimal probability $f_k(y_k)\,\mathrm{d}y_k$ that the kth smallest order statistic takes the value y_k is the probability that the n X_i values fall into three categories:
  - $k-1$ of the X_i's are less than y_k.
  - Exactly 1 X_i equals y_k.
  - The remaining $(n-k)$ X_i's are greater than y_k.

$$\begin{aligned} f_k(y_k) &= \binom{n}{k-1} [F(y_k)]^{k-1} \binom{n-k+1}{1} f(y_k) \binom{n-k}{n-k} [1 - F(y_k)]^{n-k} \\ &= \binom{n}{k-1, 1, n-k} [F(y_k)]^{k-1} f(y_k) [1 - F(y_k)]^{n-k} \end{aligned}$$

Proposition 5.2.3 (*Marginal density of ordered statistics*) *Let Y_k be the k-th order statistic in a random sample $X_1, X_2, \ldots, X_n$ taken from a continuous distribution with density function $f(x)$ and c.d.f $F(x)$. The density of Y_k is*

$$f_k(y_k) = \frac{n!}{(k-1)!(n-k)!} f(y_k) [F(y_k)]^{k-1} [1 - F(y_k)]^{n-k}.$$

Proof:

- In several of the following computations we make use of the fact that $F'(x) = f(x)$, and hence we can substitute $u = F(x)$, $\mathrm{d}u = f(x)\,\mathrm{d}x$.

- Suppose that the state space of each X_i is (a, b). (Either a or b may be infinite.)

- The marginal density of Y_k is the integral of the joint density in (5.27) with respect to all other Y_i's and after removing factors this integral becomes

$$
\begin{aligned}
f_k(y_k) &= n! f(y_k) \int_a^{y_k} f(y_1) \int_{y_1}^{y_k} f(y_2) \cdots \int_{y_{k-2}}^{y_k} f(y_{k-1}) \int_{y_k}^{b} f(y_{k+1}) \cdots \\
&\quad \int_{y_{n-2}}^{b} f(y_{n-1}) \int_{y_{n-1}}^{b} f(y_n)\, \mathrm{d}y_n\, \mathrm{d}y_{n-1} \cdots \mathrm{d}y_{k+1}\, \mathrm{d}y_{k-1} \cdots \mathrm{d}y_2\, \mathrm{d}y_1 \\
&= n! f(y_k) \int_a^{y_k} f(y_1) \int_{y_1}^{y_k} f(y_2) \cdots \int_{y_{k-2}}^{y_k} f(y_{k-1}) \int_{y_k}^{b} f(y_{k+1}) \cdots \\
&\quad \int_{y_{n-2}}^{b} f(y_{n-1})(1 - F(y_{n-1}))\, \mathrm{d}y_{n-1} \cdots \mathrm{d}y_{k+1}\, \mathrm{d}y_{k-1} \cdots \mathrm{d}y_2\, \mathrm{d}y_1
\end{aligned}
$$

(make sure you understand the limits of integration.)

- The second line evaluates the innermost integral in the first line, using the fact that $F(b) = 1$.

- To carry on the computation, a substitution $u = 1 - F(y_{n-1})$ in the innermost integral in the last line above results in the following:

$$
\int_{y_{n-2}}^{b} f(y_{n-1})(1 - F(y_{n-1}))\, \mathrm{d}y_{n-1} = \frac{1}{2 \cdot 1}(1 - F(y_{n-2}))^2.
$$

- The next integral would then have the form

$$
\int_{y_{n-3}}^{b} f(y_{n-2}) \cdot \frac{1}{2 \cdot 1}(1 - F(y_{n-2}))^2\, \mathrm{d}y_{n-2} = \frac{1}{3 \cdot 2 \cdot 1}(1 - F(y_{n-3}))^3,
$$

by a similar substitution.

- This pattern continues throughout the computation of the innermost $n - k$ integrals, with respect to the variables $y_n, \ldots, y_{k+1}$.

- At the end of this phase of computation, we have reduced the marginal density of Y_k to the form

$$
\begin{aligned}
f_k(y_k) &= n! f(y_k) \int_a^{y_k} f(y_1) \int_{y_1}^{y_k} f(y_2) \cdots \\
&\quad \cdots \int_{y_{k-2}}^{y_k} f(y_{k-1}) \cdot \frac{1}{(n-k)!}(1 - F(y_k))^{n-k}\, \mathrm{d}y_{k-1} \cdots \mathrm{d}y_2\, \mathrm{d}y_1 \\
&= \frac{n!}{(n-k)!} f(y_k)(1 - F(y_k))^{n-k} \int_a^{y_k} f(y_1) \int_{y_1}^{y_k} f(y_2) \cdots \\
&\quad \cdots \int_{y_{k-2}}^{y_k} f(y_{k-1})\, \mathrm{d}y_{k-1} \cdots \mathrm{d}y_2\, \mathrm{d}y_1.
\end{aligned}
$$

- Now the innermost integral is

$$
\int_{y_{k-2}}^{y_k} f(y_{k-1})\, \mathrm{d}y_{k-1} = F(y_k) - F(y_{k-2}).
$$

- Application of a substitution $u = F(y_k) - F(y_{k-2})$ to the next integral in line gives

$$\int_{y_{k-3}}^{y_k} (F(y_k) - F(y_{k-2}))f(y_{k-2})\,\mathrm{d}y_{k-2} = \frac{1}{2\cdot 1}\cdot (F(y_k) - F(y_{k-3}))^2.$$

- The $k-2$ integrals with respect to the variables $y_{k-1}, \ldots, y_2$ follow this pattern, leaving us with one last integral:

$$\begin{aligned} f_k(y_k) &= \frac{n!}{(n-k)!} f(y_k)(1-F(y_k))^{n-k} \\ &\quad \cdot \int_a^{y_k} f(y_1)\cdot \frac{1}{(k-2)!}\cdot (F(y_k)-F(y_1))^{k-2}\,\mathrm{d}y_1 \\ &= \frac{n!}{(n-k)!} f(y_k)(1-F(y_k))^{n-k}\left[-\frac{(F(y_k)-F(y_1))^{k-1}}{(k-1)!}\right]_a^{y_k} \\ &= \frac{n!}{(k-1)!(n-k)!}\cdot f(y_k)\cdot [F(y_k)]^{k-1}\cdot [1-F(y_k)]^{n-k}, \end{aligned}$$

which completes the derivation. □

Question 5.2.4 *Write down the formulas to which (5.28) reduces in the two cases $k = 1$ (the smallest order statistic) and $k = n$ (the largest order statistic). Interpret the expressions as multinomial probabilities.*

Ans:

$$\begin{aligned} f_1(y_1) &= \frac{n!}{0!(n-1)!}[F(y_1)]^0 f(y_1)(1-F(y_n))^{n-1} = nf(y_1)[1-F(y_1)]^{n-1} f_n(y_n) \\ &= \frac{n!}{(n-1)!0!}[F(y_n)]^{n-1} f(y_n)(1-F(y_n))^0 = n[F(y_n)]^{n-1}f(y_n) \end{aligned}$$

Alternative:

$$\begin{aligned} P[Y_n \le y_n] &= P[X_i \le y_n, i = 1, 2, \ldots, n] = F^n(y_n) \\ f_n(y_n) &= \frac{\mathrm{d}}{\mathrm{d}y_n} P[Y_n \le y_n] = n[F(y_n)]^{n-1}f(y_n) \\ P[Y_1 \ge y_1] &= P[X_i \ge y_1, i = 1, 2, \ldots, n] = (1-F(y_1))^n \\ f_1(y_1) &= \frac{\mathrm{d}}{\mathrm{d}y_1}(1 - P[Y_1 \ge y_1]) = nf(y_1)[1-F(y_1)]^{n-1} \end{aligned}$$

□

Example 5.2.4 *Suppose that the system components 1, 2, and 3 each work properly for a length of time governed by the Weibull distribution with parameters $\lambda = 2$, $\beta = 2$, and assume that their failure times are independent. Find the distribution of the system lifetime in each case.*

- $f(x) = \beta\lambda^\beta x^{\beta-1} e^{-(\lambda x)^\beta}$
- $X_1, X_2, X_3 \sim \text{Weibull}(2, 2)$
- $f(x) = 8xe^{-4x^2},\ x > 0;\quad F(x) = 1 - e^{-4x^2},\ x > 0$

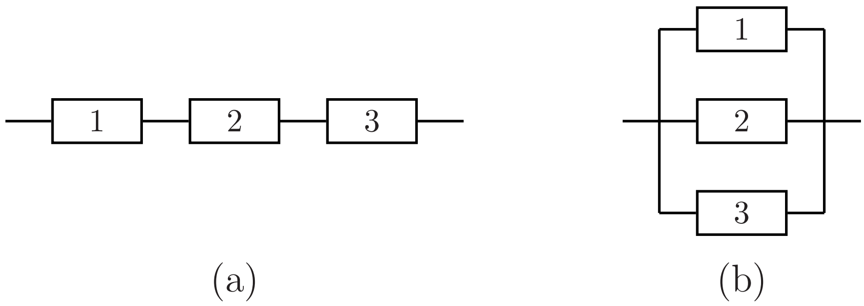


Figure 5.8: ***Series and parallel reliability structures***

- Series system:

$$\frac{3!}{0!(3-1)!}f(y_1)(F(y_1))^{1-1}(1-F(y_1))^{3-1} = 24y_1e^{-12y_1^2}, \quad y_1 > 0$$

- Parallel system:

$$\frac{3!}{(3-1)!(3-3)!}f(y_3)(F(y_3))^{3-1}(1-F(y_3))^{3-3} = 24y_3e^{-4y_3^2}\left(1-e^{-4y_3^2}\right)^2, \quad y_3 > 0$$

- $P[Y_1 > .5] = e^{-3} \approx .05$
- $P[Y_3 > .5] = 1-(1-e^{-1})^3 \approx .75$ □

Example 5.2.5 *A Pick 4 state lottery drawing claims to select four digits 0-9 randomly and independently in order to form a number ranging from 0 to 9999. We will approximate the discrete distribution of 10,000 numbers by a continuous uniform distribution [0,1] (in units of 10,000). Under uniformity, you would expect the middle of the distribution to be around 0.5 (i.e. the actual number 5000). Suppose that among seven consecutive draws, the median number drawn was 0.7637. Would you have cause to doubt the uniformity of drawing?*

- $U_i,\ i = 1, 2, \ldots, 7$: ith uniform random number.
- Y_4: Median of $U_i,\ i = 1, 2, \ldots, 7$
- $f(x) = 1$ and $F(x) = x$ for $x \in (0, 1)$.
- p.d.f. $f(y_4) = \frac{7!}{3!3!}f(y_4)F^3(y_4)(1-F(y_4))^3 = 140y_4^3(1-y_4)^3, \quad y_4 \in (0,1)$
- $P[Y_4 \geq .7637] = \int_{.7637}^{1} 140y_4^3(1-y_4)^3\, \mathrm{d}y_4 \approx .059$
- There is only about a 1/20 chance that the sample median could be so large.
- This by itself is not conclusive evidence of some kind of conspiracy, but it is enough to warrant some further investigation. □

5.3 Generating functions

- In the last section we saw how change-of-variable methods from calculus allow us to find the distribution of transformed random variables.

- Although this approach is very general, it can be difficult to apply, especially if ***more than two random variables*** are involved or if ***the transformation is complicated***.

- We now introduce a much simpler method, one that works for the special problem of ***finding the distribution of the sum of independent random variables***. This method uses a tool called the ***moment-generating function***.

- Besides its usefulness in the transformation problem, the moment-generating function has a connection to the ***moments*** of the distribution with which it is associated that can produce some computational results.

Definition 5.3.1 *The* ***moment generating function (m.g.f)*** *of a real-valued random variable X is the function*

$$M(t) = M_X(t) = E[e^{tX}],$$

which is defined for all real values of t such that the expectation is finite.

- For continuous random variables,

$$M(t) = E[e^{tX}] = \int_E e^{tx} f(x)\,\mathrm{d}x.$$

- The m.g.f. method for finding probability distributions of sums parallels the Laplace transform method for solving differential equations.

Example 5.3.1 *Consider the Poisson distribution with parameter μ. We can compute the m.g.f.*

- $M(t) = E[e^{tX}] = \sum_{n=0}^{\infty} e^{tn} \frac{e^{-\mu}\mu^n}{n!} = e^{-\mu} \sum_{n=0}^{\infty} \frac{(\mu e^t)^n}{n!}$

- $M(t) = e^{-\mu} e^{\mu e^t} = e^{\mu(e^t-1)}, \quad t \in \mathbb{R}$ □

Question 5.3.1 *Compute the derivative of the Poisson moment-generating function with respect to t, then evaluate it at $t = 0$. Try to explain what you get.*

Ans:

- $M(t) = E[e^{tX}] = e^{\mu(e^t-1)}$

- $M'(t) = E[Xe^{tX}] = \mu e^t e^{\mu(e^t-1)}$

- $M'(0) = E[X] = \mu$

Example 5.3.2 *The moment-generating function of* $\Gamma(\alpha, \lambda)$ *distribution is the following integral:*

$$
\begin{aligned}
M(t) = E[e^{tX}] &= \int_0^\infty e^{tx} \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x} \, dx \\
&= \frac{\lambda^\alpha}{\Gamma(\alpha)} \int_0^\infty x^{\alpha-1} e^{-(\lambda-t)x} \, dx
\end{aligned}
$$

The substitution $u = (\lambda - t)x, \;\; du = (\lambda - t)\, dx$ produces the following integral, which converges only for $t < \lambda$:

$$
\begin{aligned}
M(t) &= \frac{\lambda^\alpha}{\Gamma(\alpha)} \int_0^\infty \left(\frac{u}{\lambda - t} \right)^{\alpha-1} e^{-u} \cdot \frac{1}{\lambda - t} \, du \\
&= \left(\frac{\lambda}{\lambda - t} \right)^\alpha \frac{1}{\Gamma(\alpha)} \int_0^\infty u^{\alpha-1} e^{-u} \, du \\
&= \left(\frac{\lambda}{\lambda - t} \right)^\alpha \frac{1}{\Gamma(\alpha)} \Gamma(\alpha) = \left(\frac{\lambda}{\lambda - t} \right)^\alpha, \quad t < \lambda.
\end{aligned}
$$

□

Question 5.3.2 *Use (5.33) to write expressions for the m.g.f's of the important subcases of the gamma family: the exponential and* χ^2 *distributions.*

Ans:

- Exponential$(\lambda) = \Gamma(1, \lambda)$, $M(t) = \frac{\lambda}{\lambda - t}, \quad t < \lambda$.
- $\chi^2(n) = \Gamma(n/2, 1/2)$, $M(t) = \left(\frac{1}{1-2t}\right)^{\frac{n}{2}}$. □

m.g.f. of the $N(\mu, \sigma^2)$ distribution is (Exercise 5.3.4)

$$M(t) = \exp(\mu t + \sigma^2 t^2 / 2).$$

- The main reason why we are concerned with moment-generating functions of distribution is that m.g.f. of the sum $Y = X_1 + X_2 + \cdots + X_n$ of independent random variables can be expressed easily in terms of the m.g.f.'s of the X_i's.
- There is a result in probability theory (see, e.g., Parzen, 1960, p. 400) that states that a probability distribution is uniquely characterized by its moment-generating function; that is no two distributions share the same m.g.f.
- Thus, if the m.g.f. of the sum matches a moment generating function of a distribution with which we are familiar, then Y must have that distribution.
- This gives us a simple and powerful method of finding the distribution of a sum of independent random variables.

Let $X_1, X_2, \ldots, X_n$ be independent random variables having m.g.f. $M_{X_i}(t)$, respectively. Then the random variable $Y = X_1 + X_2 + \cdots + X_n$ has the m.g.f. $\prod_{i=1}^n M_{X_i}(t)$.

Proposition 5.3.1 *Let $X_1, X_2, \ldots, X_n$ be independent random variables having the Poisson distribution with parameter $\mu_1, \mu_2, \ldots, \mu_n$, respectively. Then the random variable $Y = X_1 + X_2 + \cdots + X_n$ has the Poisson distribution with parameter $\mu_1 + \mu_2 + \cdots + \mu_n$.*

Proof:

- From formula (5.32), the moment-generating function of X_i is

$$M_{X_i}(t) = E[e^{tX_i}] = e^{\mu_i(e^t-1)}.$$

- Thus the m.g.f. of Y is

$$\begin{aligned} M_Y(t) &= E[e^{tY}] \\ &= E[e^{t(X_1+\cdots+X_n)}] \\ &= E[e^{tX_1}]E[e^{tX_2}]\cdots E[e^{tX_n}] \\ &= M_{X_1}(t)M_{X_2}(t)\cdots M_{X_n}(t) \\ &= e^{\mu_1(e^t-1)}e^{\mu_2(e^t-1)}\cdots e^{\mu_n(e^t-1)} \\ &= e^{(\mu_1+\cdots+\mu_n)(e^t-1)} \end{aligned}$$

- Since this is the moment-generating function of the Poisson $(\mu_1 + \mu_2 + \cdots + \mu_n)$ distribution, the proposition is proved. □

Proposition 5.3.2 *Let $X_1, X_2, \ldots, X_n$ be independent random variables having gamma distributions with alpha parameters $\alpha_1, \alpha_2, \ldots, \alpha_n$, respectively, and common lambda parameter λ. Then the random variable $Y = X_1 + X_2 + \cdots + X_n$ has the gamma distribution with parameters $\alpha_1 + \alpha_2 + \cdots + \alpha_n$ and λ.*

- $M_{X_i}(t) = \left(\frac{\lambda}{\lambda-t}\right)^{\alpha_i}$
- $M_Y(t) = E\left[e^{t(X_1+\cdots+X_n)}\right] = E\left[e^{tX_1}\right]\cdots E\left[e^{tX_n}\right] = \left(\frac{\lambda}{\lambda-t}\right)^{\alpha_1+\cdots+\alpha_n}$ □

Recall that the $\chi^2(r)$ density is the special case of the gamma density for which $\alpha = r/2$ and $\lambda = 1/2$.

Proposition 5.3.3 *Let $X_1, X_2, \ldots, X_n$ be independent random variables having χ^2 distributions with parameters $r_1, r_2, \ldots, r_n$, respectively. Then the random variable $Y = X_1 + X_2 + \cdots + X_n$ has the $\chi^2(r_1 + r_2 + \cdots + r_n)$ distribution.*

- $M_{X_i}(t) = \left(\frac{1}{1-2t}\right)^{r_i/2}$
- $M_Y(t) = E[e^{t(X_1+\cdots+X_n)}] = E[e^{tX_1}]\cdots E[e^{tX_n}] = \left(\frac{1}{1-2t}\right)^{(r_1+\cdots+r_n)/2}$ □

- Proposition 5.3.2 also tells us that the distribution of the sum of n i.i.d. $\exp(\lambda)$ [i.e., $\Gamma(1,\lambda)$] random variables is $\Gamma(n,\lambda)$.
- Because of this result, the time of the nth arrival $T_n = S_1 + S_2 + \cdots + S_n$, in a Poisson process has the $\Gamma(n,\lambda)$ distribution.

- The m.g.f. method produces some very powerful results just by writing the m.g.f. of the sum random variable Y as the product $M_{X_1}(t)M_{X_2}(t)\cdots M_{X_n}(t)$ of the m.g.f.'s of the X's in the sum, and then simplifying and recognizing that product.

Proposition 5.3.4 *Suppose that $X_1, X_2, \ldots, X_n$ are independent normal random variables with mean $\mu_1, \mu_2, \ldots, \mu_n$ and variance $\sigma_1^2, \sigma_2^2, \ldots, \sigma_n^2$, respectively. Let $a_1, a_2, \ldots, a_n$ be constants. Then $Y = \sum_{i=1}^n a_i X_i$ has the $N(\sum_{i=1}^n a_i\mu_i, \sum_{i=1}^n a_i^2\sigma_i^2)$ distribution.*

- $M_{a_iX_i}(t) = \exp(a_i\mu_i t + a_i^2\sigma_i^2 t^2/2)$
- $M_Y(t) = E[e^{t(X_1+\cdots+X_n)}] = E[e^{tX_1}]\cdots E[e^{tX_n}] = \exp((a_1\mu_1 + \cdots + a_n\mu_n)\mu t + (a_1^2\sigma_1^2 + \cdots + a_n^2\sigma_n^2)t^2/2)$ □

Proposition 5.3.1 has implications for superpositions of Poisson processes.

Example 5.3.3 *Suppose that calls to a vendor for technical support, sales, and account questions arrive to a central access number according to three independent Poisson processes with rates λ_1, λ_2, and λ_3, respectively.*

- N_t^i denotes the total number of calls of the ith type that have arrived by time t, $i = 1, 2, 3$.
- N_t^i is Poisson with parameter $\lambda_i t$.
- m.g.f. of N_t^i is $\exp(\lambda_i t(e^s - 1))$.
- m.g.f. of $N_t = N_t^1 + N_t^2 + N_t^3$ is $\exp((\lambda_1 + \lambda_2 + \lambda_3)t(e^s - 1))$. □

Example 5.3.4 *Suppose that scores on a mathematics placement exam are normally distributed with a mean of 18 and a variance of 4. Let $\bar{X} = (X_1 + X_2 + \cdots + X_n)/100$ be the mean of a random sample of 100 of the scores.*

- By Proposition 5.3.4 $\bar{X}$ is also normally distributed, with mean and variance

$$\mu_{\bar{X}} = \frac{1}{100}(18 + 18 + \cdots + 18) = 18, \quad \sigma_{\bar{X}}^2 = \frac{1}{100^2}(4 + \cdots + 4) = \frac{4}{100}.$$

- The standard deviation of the sample mean is .2. □

Use m.g.f. to compute the moments:

- $\frac{\mathrm{d}}{\mathrm{d}t}M(t) = E\left[\frac{\mathrm{d}}{\mathrm{d}t}e^{tX}\right] = E[Xe^{tX}]$
- $\frac{\mathrm{d}}{\mathrm{d}t}M(t)\Big|_{t=0} = E[X]$
- $\frac{\mathrm{d}^2}{\mathrm{d}t^2}M(t) = E\left[\frac{\mathrm{d}^2}{\mathrm{d}t^2}e^{tX}\right] = E[X^2e^{tX}]$
- $\frac{\mathrm{d}^2}{\mathrm{d}t^2}M(t)\Big|_{t=0} = E[X^2]$
- $\frac{\mathrm{d}^n}{\mathrm{d}t^n}M(t) = E\left[\frac{\mathrm{d}^n}{\mathrm{d}t^n}e^{tX}\right] = E[X^ne^{tX}]$

- $\frac{d^n}{dt^n} M(t)\big|_{t=0} = E[X^n]$

Question 5.3.3 *How could you use the m.g.f. to compute the moments about the mean?*

Ans:

- Let $Y = X - \mu$.
- $M_Y(t) = E[e^{tY}] = E[e^{t(X-\mu)}] = E[e^{tX-t\mu}] = e^{-t\mu} E[e^{tX}]$
- $M_Y^{(n)}(0) = E[Y^n]$ □

If $Y = aX + b$, then $M_Y(t) = E\left[e^{t(aX+b)}\right] = e^{bt} E\left[e^{(at)X}\right] = e^{bt} M_X(at)$.

Example 5.3.5 *Use the moment generating function to calculate the mean and variance of the $\Gamma(\alpha, \lambda)$ distribution.*

- $M(t) = \lambda^\alpha (\lambda - t)^{-\alpha}$
- $M'(t) = \frac{\alpha \lambda^\alpha}{(\lambda - t)^{\alpha+1}}$
- $M'(0) = E[X] = \frac{\alpha}{\lambda}$
- $M''(t) = \frac{\alpha(\alpha+1)\lambda^\alpha}{(\lambda-t)^{\alpha+2}}$
- $M''(0) = \frac{\alpha(\alpha+1)}{\lambda^2}$
- $\text{Var}(X) = E[X^2] - (E[X])^2 = \frac{\alpha}{\lambda^2}$ □

We can learn more about the multivariate normal distribution if we generalize the concept of moment-generating function to a random vector.

Definition 5.3.2 *The **moment generating function of a random vector** $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ is following real-valued function of a vector variable $\mathbf{t} = (t_1, \ldots, t_n)$:*

$$M_{\mathbf{X}}(\mathbf{t}) = E\left[e^{\mathbf{t}'\mathbf{X}}\right] = E\left[\exp\left(\sum_{i=1}^{n} t_i X_i\right)\right],$$

which is defined for all $\mathbf{t} \in \mathbb{R}^n$ such that the expectation is finite.

Example 5.3.6 *Compute the moment generating function of the multivariate normal distribution.*

- $M(\mathbf{t}) = E[e^{\mathbf{t}'\mathbf{X}}] = \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} e^{\mathbf{t}'\cdot\mathbf{x}} \frac{1}{(2\pi)^{n/2}\sqrt{\det(\Sigma)}} e^{(-1/2)(\mathbf{x}-\boldsymbol{\mu})'\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})}\, d\mathbf{x}$
- $f(\mathbf{x}) = Ce^{(-1/2)(\mathbf{x}-\boldsymbol{\mu})'\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})}$
- $-\frac{1}{2}[(\mathbf{x}-\boldsymbol{\mu})'\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu}) - 2\mathbf{t}'\mathbf{x}] = -\frac{1}{2}[(\mathbf{x}-(\boldsymbol{\mu}+\Sigma\mathbf{t}))'\Sigma^{-1}(\mathbf{x}-(\boldsymbol{\mu}+\Sigma\mathbf{t}))] + \boldsymbol{\mu}'\mathbf{t} + \frac{1}{2}\mathbf{t}'\Sigma\mathbf{t}$

- Since the integrand is a multivariate normal density with mean vector $\boldsymbol{\mu} + \Sigma\mathbf{t}$ and covariance matrix Σ, the multiple integral equals 1. Hence $M(\mathbf{t}) = e^{\boldsymbol{\mu}'\mathbf{t}+\frac{1}{2}\mathbf{t}'\Sigma\mathbf{t}}$, $\mathbf{t} \in \mathbb{R}^n$. □

Example 5.3.7 *The mixed partials of the joint moment generating function (5.41) of a random vector are useful in finding expected products of the components of the vector. To illustrate, let us use the m.g.f. to verify that the covariance of bivariate normal random variables X_1 and X_2 is $\rho\sigma_1\sigma_2$.*

- $\frac{\partial M}{\partial t_1} = \frac{\partial}{\partial t_1} E[e^{t_1X_1+t_2X_2}] = E[X_1 e^{t_1X_1+t_2X_2}]$
- $\left.\frac{\partial M}{\partial t_1}\right|_{(0,0)} = E[X_1]$
- $\frac{\partial M}{\partial t_2} = \frac{\partial}{\partial t_2} E[e^{t_1X_1+t_2X_2}] = E[X_2 e^{t_1X_1+t_2X_2}]$
- $\left.\frac{\partial M}{\partial t_2}\right|_{(0,0)} = E[X_2]$
- $\frac{\partial^2 M}{\partial t_1 \partial t_2} = \frac{\partial^2}{\partial t_1 \partial t_2} E[e^{t_1X_1+t_2X_2}] = E[X_1X_2 e^{t_1X_1+t_2X_2}]$
- $\left.\frac{\partial^2 M}{\partial t_1 \partial t_2}\right|_{(0,0)} = E[X_1X_2]$
- $M(t_1, t_2) = \exp[\mu_1 t_1 + \mu_2 t_2 + \sigma_1^2 t_1^2/2 + \rho\sigma_1\sigma_2 t_1 t_2 + \sigma_2^2 t_2^2/2]$
- $E[X_1] = (\mu_1 + \sigma_1^2 t_1 + \rho\sigma_1\sigma_2 t_2) \cdot M(t_1, t_2)|_{(0,0)} = \mu_1$ and $E[X_2] = (\mu_2 + \sigma_2^2 t_2 + \rho\sigma_1\sigma_2 t_1) \cdot M(t_1, t_2)|_{(0,0)} = \mu_2$.
- $E[X_1X_2] = (\rho\sigma_1\sigma_2 M(t_1, t_2) + (\mu_2 + \sigma_2^2 t_2 + \rho\sigma_1\sigma_2 t_1) \cdot (\mu_1 \sigma_1^2 t_1 + \rho\sigma_1\sigma_2 t_2) \cdot M(t_1, t_2)|_{(0,0)} = \rho\sigma_1\sigma_2 + \mu_1\mu_2$
- $\text{Cov}(X_1, X_2) = E[X_1X_2] - E[X_1]E[X_2] = \rho\sigma_1\sigma_2$ □

Question 5.3.4 *What moments can be determined from second-order partials $\partial^2 M/\partial t_1^2$,$\partial^2 M/\partial t_2^2$ for a general bivariate random vector? Can you generalize?*

Ans:

- $\frac{\partial^2}{\partial t_1^2} M|_{(0,0)} = E[X_1^2]$
- $\frac{\partial^2}{\partial t_2^2} M|_{(0,0)} = E[X_2^2]$
- $\frac{\partial^n}{\partial t_i^n} M|_{(0,\ldots,0)} = E[X_i^n]$

Other types of generating functions:

- ***Probability generating function***: (discrete distribution)

$$P(t) = E[t^X] = \sum_{i=1}^{\infty} t^{x_i} P[X = x_i]$$

which is particularly useful for discrete distributions on the nonnegative integers. In such a case it would take the form

$$P(t) = E[t^X] = \sum_{k=0}^{\infty} t^k P[X = k].$$

Queueing theorists sometimes use the probability generating function to derive queue length probabilities and moments.

- ***Characteristic function***:

  $$\phi(t) = E[e^{itX}]$$

  where i is the complex number $i = \sqrt{-1}$ and the complex exponential function is defined by $e^{iu} = \cos(u) + i\sin(u)$.

  - The complex exponential function has similar algebraic and calculus properties to the real exponential function.
  - The characteristic function can be used to do the same things as the moment-generating function.
  - The characteristic function is actually better, however, in the sense that it exists for all real t and is bounded in magnitude by 1.

5.4 Transformations of normal random variables

5.4.1 Basic results

Random samples $X_1, X_2, \ldots, X_n$:

- ***Sample mean***: Estimates the central tendency of the distribution, measured by μ.

  $$\bar{X} = \sum_{i=1}^{n} \frac{X_i}{n}$$

- ***Sample variance***: Estimates the spread of the distribution, measured by σ^2, by the average squared distance of data points from $\bar{X}$.

  $$S^2 = \sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{n-1}$$

- How likely are they to be close to the underlying parameters μ and σ^2? To answer these questions, we must find the probability distributions of these particular transformations of normal random variables.

- The sample mean is a linear transformation of independent normal random variables, and the sample variance is a combination of squares and first powers of normals.

- Proposition 3.3.6, says that

  $$X \sim N(\mu, \sigma^2) \Rightarrow Z = \frac{X - \mu}{\sigma} \sim N(0, 1).$$

Proposition 5.4.1 *If X has the $N(\mu, \sigma^2)$ distribution, then the random variable $Y = cX + d$ has the $N(c\mu + d, c^2\sigma^2)$ distribution.*

Proof:

- $M_X(t) = e^{\mu t + \sigma^2 t^2/2}$
- $M_Y(t) = E\left[e^{tY}\right] = E\left[e^{t(cX+d)}\right] = e^{td}E\left[e^{ctX}\right] = e^{td}M_X(ct)$
- $M_X(ct) = e^{\mu ct + \sigma^2 c^2 t^2/2}$
- $M_Y(t) = e^{td}e^{\mu ct + \sigma^2 c^2 t^2/2} = e^{(c\mu+d)t + (c^2\sigma^2)t^2/2}$
- $Y \sim N(c\mu + d, c^2\sigma^2)$ $\square$

- $X_1, X_2, \ldots, X_n$ are independent $N(\mu_i, \sigma_i^2) \Rightarrow Y = \sum_{i=1}^n c_i X_i \sim N(\sum_{i=1}^n c_i\mu_i, \sum_{i=1}^n c_i^2\sigma_i^2)$.
- If all of the means $\mu_i = \mu$ and all of the variances $\sigma_i^2 = \sigma^2$, we can set all of the coefficients $c_i = 1/n$ to obtain the following result about the distribution of the sample mean.

Proposition 5.4.2 *Let $X_1, X_2, \ldots, X_n$ be a random sample from the $N(\mu, \sigma^2)$ distribution. Then*

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right) \quad \textit{and} \quad Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1).$$

Example 5.4.1 *Suppose that past evidence suggests that the calorie count in a cup of a brand of low-fat yogurt is normally distributed with mean 150 calories and standard deviation 10 calories. An improved formulation of the yogurt has been concocted, for which the standard deviation has not changed but the mean may have been reduced. If 100 cups are measured and the sample average calorie content is 145, is there strong evidence to believe that the true mean calorie content has been reduced?*

- We can quantify the question by asking how likely is the event that the sample mean is 145 or less ***if the true mean is still 150***.
- $P[\bar{X} \leq 145] = P\left[Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \leq \frac{145-150}{10/\sqrt{100}}\right] = P[Z \leq -5] \approx 0$
- The observed deviation of five standard errors from the old mean of 150 is a very rare event, which is indeed strong evidence that the population mean has been reduced. $\square$

Proposition 5.4.3 *If Z has the $N(0, 1)$ distribution, then the transformed random variable $Y = Z^2$ has the $\chi^2(1)$ distribution. Hence, if X is $N(\mu, \sigma^2)$, then*

$$Y = \left(\frac{X - \mu}{\sigma}\right)^2 \sim \chi^2(1). \tag{5.56}$$

Proof:

- The c.d.f. of Y is

$$\begin{aligned} G_Y(y) = P[Y \le y] &= P[Z^2 \le y] \\ &= P[-\sqrt{y} \le Z \le \sqrt{y}] \\ &= 2 \cdot P[0 \le Z \le \sqrt{y}], \quad y > 0. \end{aligned}$$

- The final equation follows from the symmetry of the $N(0, 1)$ density about 0.
- Therefore the density of Y is

$$\begin{aligned} g_Y(y) &= \frac{\mathrm{d}}{\mathrm{d}y} G_Y(y) \\ &= \frac{\mathrm{d}}{\mathrm{d}y}\left(2 \cdot \int_0^{\sqrt{y}} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz\right) \\ &= 2 \cdot \frac{1}{2} y^{-1/2} \cdot \frac{1}{\sqrt{2\pi}} e^{-(\sqrt{y})^2/2} \\ &= \frac{1}{2^{1/2}\Gamma(1/2)} y^{1/2-1} e^{-y/2}, \quad y > 0. \end{aligned}$$

- In the last line, we have used the fact that $\Gamma(1/2) = \sqrt{\pi}$.
- The density of Y is of the $\Gamma(1/2, 1/2)$ form, which is the same as the $\chi^2(1)$ density; hence the first statement is proved.
- Statement (5.56) follows from the fact that the random variable in parentheses has the standard normal distribution.
- Alternative: $Y = Z^2$ and $Z \sim N(0, 1)$, $M_Z(t) = e^{t^2/2}$
- $M_Y(t) = M_{Z^2}(t) = \int_{-\infty}^{\infty} e^{tz^2} \frac{1}{\sqrt{2\pi}} e^{-z^2/2}\,\mathrm{d}z = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-z^2(1-2t)/2}\,\mathrm{d}z$
- $M_Y(t) = (1 - 2t)^{-1/2}$, $t < 1/2$
- $Y \sim \chi^2(1)$ □

Proposition 5.4.4 *Let $Z_1, Z_2, \ldots, Z_n$ be independent and identically distributed $N(0, 1)$ random variables. Then*

$$Y = Z_1^2 + Z_2^2 + \cdots + Z_n^2 \sim \chi^2(n).$$

In particular, if $X_1, X_2, \ldots, X_n$ are normal random variables with mean $\mu_1, \ldots, \mu_n$ and variance $\sigma_1^2, \ldots, \sigma_n^2$, respectively, then

$$Y = \left(\frac{X_1 - \mu_1}{\sigma_1}\right)^2 + \cdots + \left(\frac{X_n - \mu_n}{\sigma_n}\right)^2 \sim \chi^2(n).$$

Proof:

- $M_{Z_i^2}(t) = (1 - 2t)^{-1/2}$, $t < 1/2$
- $M_Y(t) = \prod_{i=1}^n M_{Z_i^2}(t) = \prod_{i=1}^n (1 - 2t)^{-1/2} = (1 - 2t)^{-n/2}$, $t < 1/2$

- $Y \sim \chi^2(n)$ □

Question 5.4.1 *How does Proposition 5.4.4 follow from previous results?*

Ans:

- It follows from Proposition 5.3.3.
- $X_1, X_2, \ldots, X_n$ be independent random variables having χ^2 distributions with parameters $r_1, r_2, \ldots, r_n$, then $Y = X_1 + X_2 + \cdots + X_n$ has $\chi^2(r_1 + r_2 + \cdots + r_n)$ distribution.
- Now, $r_1 = r_2 = \cdots = r_n = 1$. □

Notice how closely the form of this random variable resembles the form of a simple function of the sample variance:

$$\frac{(n-1)S^2}{\sigma^2} = \frac{(n-1)}{\sigma^2} \cdot \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2 = \sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{\sigma^2}$$

The only difference is that the true mean μ is estimated by the sample mean $\bar{X}$.

Proposition 5.4.5 *Let $X_1, X_2, \ldots, X_n$ be a random sample from the $N(\mu, \sigma^2)$ distribution. Then the random variable*

$$\frac{(n-1)S^2}{\sigma^2} = \sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{\sigma^2}$$

has the $\chi^2(n-1)$ distribution and furthermore is independent of $\bar{X}$.

Example 5.4.2 *A cutting machine saws two-by-four wood boards to a length X that is normally distributed with mean 8 feet. What is the probability that, among 16 such boards, the sample variance of the board lengths is at least twice the population variance?*

- $Y = (n-1)S^2/\sigma^2 = 15S^2/\sigma^2$ has the $\chi^2(15)$ distribution.
- $S^2 \geq 2\sigma^2 \Leftrightarrow 15S^2/\sigma^2 \geq 2 \cdot 15$ iff $Y \geq 30$.
- $P[Y \geq 30] \approx .01$ □

5.4.2 Multivariate transformations of normal random vectors

Two reasons underline our need to carry the development further.

- A proof of Proposition 5.4.5.
- There is an important topic in statistics that relies heavily on stronger theorems than we have at present: The design and analysis of experiments, whose purpose is to gain information about the influence of one or more factors on the mean of a normally distributed response variable.

We will now extend our study to linear and quadratic functions of multivariate normal data.

Proposition 5.4.6 *Let A be an $m \times n$ matrix of row rank m, where $m \leq n$, and let* $\mathbf{b}$ *an $m \times 1$ column vector. If $\mathbf{X} = [X_1, X_2, \ldots, X_n]'$ is a multivariate normal random vector with mean vector $\boldsymbol{\mu}_X = [\mu_1, \ldots, \mu_n]'$ and covariance matrix $\sum_X$, then the random vector*

$$\mathbf{Y} = A\mathbf{X} + \mathbf{b}$$

has the multivariate normal distribution with mean vector $\boldsymbol{\mu}_Y = A\boldsymbol{\mu}_X + \mathbf{b}$ and covariance $\Sigma_Y = A\Sigma_X A'$.

Proof:

- We will first reduce to the case where A is an $n \times n$ invertible matrix.
- If the number of rows m of A is strictly less than n, then since A has row rank m, its rows comprise m linearly independent vectors in $\mathbb{R}^n$.
- These rows may be supplemented by an additional $n - m$ vectors in $\mathbb{R}^n$ to form a basis.
- Let the new vectors be the rows of an $(n-m) \times n$ matrix B, which we can adjoin to the bottom of A to form an $n \times n$ matrix:

$$C = \begin{bmatrix} A \\ \cdots \\ B \end{bmatrix}.$$

- Since the rows of C are linearly independent, C is an invertible matrix.
- Now adjoin $(n - m)$ 0's to the bottom of the constant vector $\mathbf{b}$ to obtain a new constant vector $\mathbf{c}$ and a new transformation:

$$\mathbf{W} = C\mathbf{X} + \mathbf{c} = \begin{bmatrix} A \\ \cdots \\ B \end{bmatrix} \begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix} + \begin{bmatrix} \mathbf{b} \\ \cdots \\ \mathbf{0} \end{bmatrix} = \begin{bmatrix} A\mathbf{X} + \mathbf{b} \\ \cdots \\ B\mathbf{X} \end{bmatrix}.$$

- Since the joint marginal distributions of multivariate normal vectors are also multivariate normal [see Proposition 4.4.3(a)], if we can show that the entire vector $\mathbf{W}$ is multivariate normal, then its first block $A\mathbf{X} + \mathbf{b}$ is also multivariate normal.
- Thus it suffices to prove the proposition in the case that the coefficient matrix is an invertible $n \times n$ matrix.
- From this point, we return to the original notation and abandon $\mathbf{W}$, C and $\mathbf{c}$.
- The inverse transformation of $\mathbf{Y} = A\mathbf{X} + \mathbf{b}$ is $\mathbf{X} = A^{-1}(\mathbf{Y} - \mathbf{b})$.
- You can check (see Question 5.4.2) that the Jacobian is just A^{-1}.
- The determinant of A^{-1} is the same as $(\det A)^{-1}$.
- By formula (5.23), the density of $\mathbf{Y} = A\mathbf{X} + \mathbf{b}$ is therefore

$$f_Y(\mathbf{y}) = f_X(A^{-1}(\mathbf{y} - \mathbf{b})) \cdot |\det A|^{-1},$$

and since $\mathbf{X}$ is multivariate normal, this density is

$$\begin{aligned} f_Y(\mathbf{y}) &= \frac{1}{(2\pi)^{n/2}\sqrt{\det\Sigma}} \cdot \frac{1}{|\det A|} \\ &\quad \cdot \exp\left[-\frac{1}{2}(A^{-1}(\mathbf{y}-\mathbf{b})-\boldsymbol{\mu}_X)'\Sigma^{-1}(A^{-1}(\mathbf{y}-\mathbf{b})-\boldsymbol{\mu}_X)\right] \\ &= \frac{1}{(2\pi)^{n/2}\sqrt{(\det\Sigma)(\det A)(\det A)}} \\ &\quad \cdot \exp\left[-\frac{1}{2}(A^{-1}(\mathbf{y}-(A\boldsymbol{\mu}_X+\mathbf{b})))'\Sigma^{-1}(A^{-1}(\mathbf{y}-(A\boldsymbol{\mu}_X+\mathbf{b})))\right] \\ &= \frac{1}{(2\pi)^{n/2}\sqrt{(\det\Sigma)(\det A)(\det A)}} \\ &\quad \cdot \exp\left[-\frac{1}{2}(\mathbf{y}-(A\boldsymbol{\mu}_X+\mathbf{b}))'(A^{-1})'\Sigma^{-1}(A^{-1})'(\mathbf{y}-(A\boldsymbol{\mu}_X+\mathbf{b}))\right]. \end{aligned}$$

- Since $(\det A)(\det\Sigma)(\det A') = \det(A\Sigma A')$ and $(A^{-1})'\Sigma^{-1}A^{-1} = (A\Sigma A')^{-1}$, we have reduced the density of $\mathbf{Y}$ to the bivariate normal form, and in the process we have obtained the correct formulas for the mean vector and covariance matrix. □

Question 5.4.2 *Confirm that the Jacobian of* $\mathbf{Y} = A\mathbf{X}+b, \mathbf{X} = A^{-1}(\mathbf{Y}-\mathbf{b})$ *is* A^{-1}.

Ans:

- $\mathbf{X} = A^{-1}(\mathbf{Y}-\mathbf{b}) \Rightarrow A^{-1}\begin{bmatrix} Y_1-b_1 \\ Y_2-b_2 \\ \vdots \\ Y_n-b_n \end{bmatrix} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix}$

- Let $A^{-1} = \begin{bmatrix} B_{11} & B_{12} & \dots & B_{1n} \\ B_{21} & B_{22} & \dots & B_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ B_{n1} & B_{n2} & \dots & B_{nn} \end{bmatrix}$.

- $X_1 = B_{11}(Y_1-b_1) + B_{12}(Y_2-b_2) + \cdots + B_{1n}(Y_n-b_n)$

$\vdots$

- $X_n = B_{n1}(Y_1-b_1) + B_{n2}(Y_2-b_2) + \cdots + B_{nn}(Y_n-b_n)$

- Jacobian is $\begin{bmatrix} \frac{\partial X_1}{\partial Y_1} & \frac{\partial X_1}{\partial Y_2} & \dots & \frac{\partial X_1}{\partial Y_n} \\ \frac{\partial X_2}{\partial Y_1} & \frac{\partial X_2}{\partial Y_2} & \dots & \frac{\partial X_2}{\partial Y_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial X_n}{\partial Y_1} & \frac{\partial X_n}{\partial Y_2} & \dots & \frac{\partial X_n}{\partial Y_n} \end{bmatrix} = \begin{bmatrix} B_{11} & B_{12} & \dots & B_{1n} \\ B_{21} & B_{22} & \dots & B_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ B_{n1} & B_{n2} & \dots & B_{nn} \end{bmatrix} = A^{-1}$. □

Example 5.4.3 *Let* X_1, X_2, X_3 *be i.i.d* $N(\mu, \sigma^2)$ *random variables, so that the vector* $\mathbf{X} = [X_1, X_2, X_3]'$ *is multivariate normal with mean vector* $\boldsymbol{\mu}_X = [\mu, \mu, \mu]'$ *and covariance matrix* $\sigma^2 I$*, where* I *is a* 3×3 *identity matrix.*

- $A = \begin{bmatrix} 1/3 & 1/3 & 1/3 \\ -1 & 0 & 1 \\ 0 & 1 & -1 \end{bmatrix}$
- $\mathbf{Y} = A\mathbf{X}$ has the form

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \end{bmatrix} = \begin{bmatrix} (X_1 + X_2 + X_3)/3 \\ -X_1 + X_3 \\ X_2 - X_3 \end{bmatrix}.$$

- $\boldsymbol{\mu}_Y = A\boldsymbol{\mu}_X = \begin{bmatrix} \mu \\ 0 \\ 0 \end{bmatrix}$ and $\sum_Y = A \sum_X A' = \begin{bmatrix} \sigma^2/3 & 0 & 0 \\ 0 & 2\sigma^2 & -\sigma^2 \\ 0 & -\sigma^2 & 2\sigma^2 \end{bmatrix}$
- $Y_1 = \bar{X} \sim N(\mu, \sigma^2/3)$
- $Y_2 = -X_1 + X_3$ and $Y_3 = X_2 - X_3 \sim N(0, 2\sigma^2)$ □

Question 5.4.3 *What is the correlation between Y_2 and Y_3 in Example 5.4.3?*

Ans:

- $\mathrm{Corr}(Y_2, Y_3) = \frac{\mathrm{Cov}(Y_2,Y_3)}{\sqrt{\mathrm{Var}(Y_2)}\sqrt{\mathrm{Var}(Y_3)}} = \frac{-\sigma^2}{2\sigma^2} = -\frac{1}{2}$ □

Distribution of quadratic form:

$$\mathbf{Y} = \mathbf{X}'Q\mathbf{X},$$

where Q is a symmetric matrix and $\mathbf{X} = [X_1\, X_2\, \ldots\, X_n]'$ has the multivariate normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix Σ.

- Consider a random sample $X_1, X_2, \ldots, X_n$ from the $N(\mu, \sigma^2)$.
- $\sum_{i=1}^n X_i^2 = \sum_{i=1}^n (X_i - \bar{X})^2 + n\bar{X}^2 \Rightarrow \frac{\sum_{i=1}^n X_i^2}{\sigma^2} = \frac{(n-1)S^2}{\sigma^2} + \frac{n\bar{X}^2}{\sigma^2}$
- $\sum_{i=1}^n X_i^2 = \mathbf{X}' \begin{bmatrix} 1 & 0 & \ldots & 0 \\ 0 & 1 & \ldots & 0 \\ & & \ddots & \\ 0 & 0 & \ldots & 1 \end{bmatrix} \mathbf{X} = \mathbf{X}' I_n \mathbf{X}$
- $n\bar{X}^2 = \mathbf{X}' \begin{bmatrix} 1/n & 1/n & \cdots & 1/n \\ 1/n & 1/n & \cdots & 1/n \\ \vdots & & \ddots & \vdots \\ 1/n & 1/n & \cdots & 1/n \end{bmatrix} \mathbf{X} = \mathbf{X}' Q_1 \mathbf{X}$
- $\sum_{i=1}^n (X_i - \bar{X})^2 = \mathbf{X}' Q_2 \mathbf{X}$ where $Q_2 = I_n - Q_1$. □

Idempotent matrix: Symmetric and $Q^2 = Q$

- Eigenvalues are 0 or 1 only and the number of eigenvalues equal to 1 is the rank r of the matrix.
- Diagonal decomposition: $Q = N \begin{bmatrix} I_r & 0 \\ 0 & 0 \end{bmatrix} N' = N \Lambda N'$, where N is an orthogonal matrix ($NN' = N'N = I$) whose columns are the eigenvectors of Q.

Proposition 5.4.7 *Suppose that* $\mathbf{X} = [X_1, X_2, \ldots, X_n]'$ *is a random vector that has the multivariate normal distribution with mean vector* $\boldsymbol{\mu} = [\mu, \mu, \ldots, \mu]'$ *and covariance matrix* $\sigma^2 I_n$. *Let* Q *be an* $n \times n$ *symmetric, idempotent matrix of rank* r. *If either* $\mu = 0$ *or all row sums of* Q *are* 0, *then the random variable* $Y = \mathbf{X}'Q\mathbf{X}/\sigma^2$ *has the* $\chi^2(r)$ *distribution.*

Proof:

- We will show that the moment-generating function of Y agrees with the moment-generating function of the $\chi^2(r)$ distribution.

- Because $\boldsymbol{\mu}$ has identical entries, in either the case that the row sums of Q are zero or $\boldsymbol{\mu} = \mathbf{0}$, we have
$$\mathbf{X}'Q\boldsymbol{\mu} = \boldsymbol{\mu}'Q\mathbf{X} = 0 \quad \text{and} \quad \boldsymbol{\mu}'Q\boldsymbol{\mu} = 0$$
$$\Rightarrow \mathbf{X}'Q\mathbf{X} = (\mathbf{X} - \boldsymbol{\mu})'Q(\mathbf{X} - \boldsymbol{\mu}).$$

- Because the two quadratic forms that are equated in the last line are the same, their moment-generating functions are the same; hence we may as well proceed as if we had a random vector $\mathbf{Z} = \mathbf{X} - \boldsymbol{\mu}$ with mean vector $\mathbf{0}$.

- By the definition of the m.g.f.,
$$\begin{aligned} M_Y(t) &= E[e^{t\mathbf{Z}'Q\mathbf{Z}/\sigma^2}] \\ &= \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} e^{t\mathbf{z}'Q\mathbf{z}/\sigma^2} \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2}\mathbf{z}'I_n\mathbf{z}\right) \, d\mathbf{z} \\ &= \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left[-\frac{1}{2\sigma^2}\mathbf{z}'(I_n - 2tQ)\mathbf{z}\right] \, d\mathbf{z}. \end{aligned} \tag{5.66}$$

- In the last integral, multiply and divide by the quantity
$$\sqrt{|\det(I_n - 2tQ)^{-1}|} = |\det(I_n - 2tQ)|^{-1/2}.$$

- Then
$$M_Y(t) = |\det(I_n - 2tQ)|^{-1/2} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} \frac{1}{(2\pi)^{n/2}\sqrt{(\sigma^2)^n |\det(I_n - 2tQ)^{-1}|}} \exp\left[-\frac{1}{2\sigma^2}\mathbf{z}'(I_n - 2tQ)\mathbf{z}\right] \, d\mathbf{z}.$$

- The integrand is now a multivariate normal density with mean vector $\mathbf{0}$ and covariance matrix $\sigma^2(I_n - 2tQ)^{-1}$ (which exists if t is small enough).

- Therefore the integral reduces to 1, and we have the following formula for the m.g.f. of Y (which is true even when Q is not idempotent):
$$M_Y(t) = E\left[e^{t\mathbf{Z}'Q\mathbf{Z}/\sigma^2}\right] = |\det(I_n - 2tQ)|^{-1/2}. \tag{5.67}$$

- Now, since Q is idempotent, matrices N and A that satisfy the relations in (5.65) exist.

- Since $NN' = NI_nN' = I_n$, we can write the m.g.f. of Y as

$$\begin{aligned} M_Y(t) &= |\det(I_n - 2tQ)|^{-1/2} \\ &= |\det(NI_nN' - 2tNAN')|^{-1/2} \\ &= |\det(N(I_n - 2tA)N')|^{-1/2} \\ &= |\det(N)|^{-1/2}|\det(I_n - 2tA)|^{-1/2}|\det(N')|^{-1/2} \\ &= |\det(I_n - 2tA)|^{-1/2} \\ &= (1-2t)^{-r/2}, \quad t < 1/2. \end{aligned} \tag{5.68}$$

- In the fourth line of (5.68) we use the fact that the determinant of a product is the product of the determinants.
- In the fifth line we observe that since N is orthogonal, the determinant of its transpose is the reciprocal of its determinant.
- You should verify the last equation yourself by answering Question 5.4.4.
- Since we have now put the m.g.f. of Y into the form of the $\chi^2(r)$ m.g.f., the proof is complete. □

Question 5.4.4 *Verify that* $\det(I_n - 2tA) = (1-2t)^r$ *in formula (5.68) of the proof of Proposition 5.4.7.*

Ans:

- $$\begin{aligned} \det(I_n - 2tA) &= \det\left(\begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ & & \ddots & \\ 0 & 0 & \cdots & 1 \end{bmatrix} - 2t \begin{bmatrix} I_r & 0 \\ 0 & 0 \end{bmatrix} \right) \\ &= \det \begin{bmatrix} 1-2t & 0 & \cdots & \cdots & \cdots & \cdots & 0 \\ 0 & 1-2t & \cdots & \cdots & \cdots & \cdots & 0 \\ 0 & 0 & \ddots & \ddots & \ddots & \ddots & 0 \\ 0 & 0 & \cdots & 1-2t & \cdots & \cdots & 0 \\ 0 & 0 & \cdots & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & \cdots & \cdots & \cdots & \cdots & 1 \end{bmatrix} \\ &= (1-2t)^r \end{aligned}$$ □

- $Q_2^2 = (I_n - Q_1)(I_n - Q_1) = I_n - 2Q_1 + Q_1^2 = I_n - 2Q_1 + Q_1 = I_n - Q_1 = Q_2$ (idempotent matrix) with rank $n-1$.
- $\frac{(n-1)S^2}{\sigma^2} = \frac{\mathbf{X}'Q_2\mathbf{X}}{\sigma^2} \sim \chi^2(n-1)$ □

Proposition 5.4.8 *Suppose that* $\mathbf{X} = [X_1, X_1, \dots, X_n]'$ *is a random vector that has the multivariate normal distribution with mean vector* $\boldsymbol{\mu} = [\mu, \dots, \mu]'$ *and covariance matrix* $\sigma^2 I_n$. *Let* Q_1 *and* Q_2 *be symmetric* $n \times n$ *matrices such that* $Q_1 Q_2 = 0$. *If either* $\boldsymbol{\mu} = \mathbf{0}$ *or the row sums of both* Q_1 *and* Q_2 *are all zero, then the random variables* $Y_1 = \mathbf{X}'Q_1\mathbf{X}/\sigma^2$ *and* $Y_2 = \mathbf{X}'Q_2\mathbf{X}/\sigma^2$ *are independent.*

Proof:

- We will show that the joint moment-generating function of Y_1 and Y_2 factors into the product of the two marginal m.g.f.'s which is sufficient for independence.
- By the same derivation as the one in Proposition 5.4.7 that led to formula (5.67), this joint m.g.f. is

$$\begin{aligned} M(t_1, t_2) &= E[e^{t_1 Y_1 + t_2 Y_2}] \\ &= E[\exp(t_1 \mathbf{X}'Q_1\mathbf{X}/\sigma^2 + t_2\mathbf{X}'Q_2\mathbf{X}/\sigma^2)] \\ &= E[\exp(\mathbf{X}'(t_1 Q_1 + t_2 Q_2)\mathbf{X}/\sigma^2)] \\ &= |\det(I_n - 2(t_1 Q_1 + t_2 Q_2))|^{-1/2}. \end{aligned}$$

- The hypothesis that $Q_1 Q_2 = 0$ allows us to write

$$\begin{aligned} I - 2(t_1 Q_1 + t_2 Q_2) &= I_n - 2(t_1 Q_1 - 2t_1 t_2 Q_1 Q_2 + t_2 Q_2) \\ &= (I_n - 2t_1 Q_1)(I_n - 2t_2 Q_2). \end{aligned}$$

- Thus the determinant factors:

$$M(t_1, t_2) = |\det(I_n - 2t_1 Q_1)|^{-1/2} |\det(I_n - 2t_2 Q_2)|^{-1/2}.$$

- By (5.67) again, the two factors are the marginal m.g.f.'s $M_{Y_1}(t_1)$ and $M_{Y_2}(t_2)$, which finishes the proof.
- You should note that we did not as assume that the matrices were idempotent here. Independence of the forms follows strictly from the zero product hypothesis. The fact that each has a χ^2 distribution follows strictly from idempotency. □

Apply Proposition 5.4.8 to $\mathbf{X}'Q_1\mathbf{X}/\sigma^2 = n\bar{X}^2/\sigma^2$ and $\mathbf{X}'Q_2\mathbf{X}/\sigma^2 = (n-1)S^2/\sigma^2$:

- $Q_1 Q_2 = Q_1(I - Q_1) = 0$ since $Q_1 = Q_1^2$
- By proposition 5.4.8, $\bar{X}$ and $(n-1)S^2/\sigma^2$ as well, are independent.

5.5 t- and F-distributions

Recall Example 5.4.1, in which we were to decide whether the mean calorie content of a cup of yogurt could he 150 if the sample mean of 100 cups was 145 calories. We assumed a known population standard deviation $\sigma = 10$ in order to answer the question. But it is really not reasonable to assume that we know the standard deviation a priori. In place of the random variable that we used in that example,

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}}$$

In this random variable, the population standard deviation σ is estimated by the sample standard deviation S.

Question 5.5.1 *Recall the distribution of $\bar{X}$ and $(n-1)S^2/\sigma^2$ derived in the last section. Try to put these random variables together in such a way that the random variable in (5.72) results.*

Ans: $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} / \sqrt{\frac{(n-1)S^2}{\sigma^2}/(n-1)} = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} / \frac{S}{\sigma} = \frac{\bar{X}-\mu}{S/\sqrt{n}}$ □

Definition 5.5.1 *The* ***t-distribution with parameter r*** *(called the* ***degrees of freedom*** *of the distribution) is the distribution of the random variable,*

$$T = \frac{Z}{\sqrt{V/r}},$$

where $Z \sim N(0,1), V \sim \chi^2(r)$, and Z and V are independent. We use $t(r)$ as a shorthand notation for this distribution.

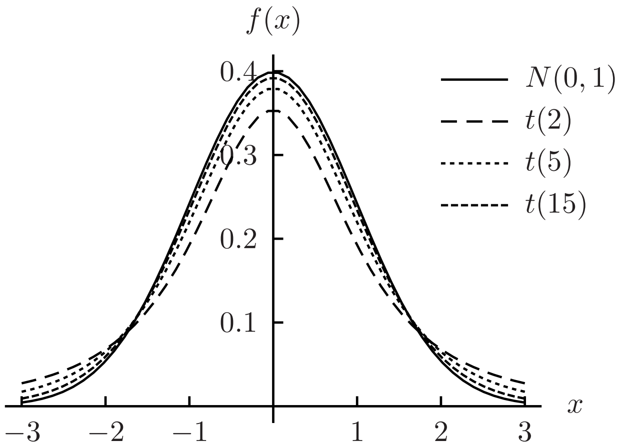


Figure 5.9: ***t-densities***

- $T = \frac{Z}{\sqrt{V/r}}$, $U = V$ then $f(t) = \frac{\Gamma((r+1)/2)}{\Gamma(r/2)\sqrt{\pi r}} \left(1 + \frac{t^2}{r}\right)^{-(r+1)/2}, \quad t \in \mathbb{R}$
- Symmetric about 0.
- $T \equiv \text{Cauchy}(0,1)$ if $r = 1$.
- $T \to N(0,1)$ as $r \to \infty$.
- $P[T \leq 1.44] = .90$ and $P[T > 1.94] = .05$ when $r = 6$.

Question 5.5.2 *For $r = 10$, what is $P[T > 2.76]$? For what value of t, $P[T \leq t] = 0.975$?*

Ans: $P[T > 2.76] = 0.01$, $P[T \leq 2.23] = 0.975$. □

Proposition 5.5.1 *Let $\bar{X}$ and S^2 be the sample mean and variance of a random sample of size n from the $N(\mu, \sigma^2)$ distribution. Then*

$$T = \frac{\bar{X}-\mu}{S/\sqrt{n}} \sim t(n-1). \tag{5.76}$$

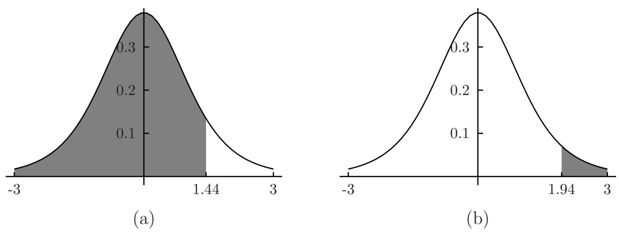


Figure 5.10: ***Some t-percentiles***

Proof:

- We have from Propositions 5.4.2 and 5.4.5 that

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1), \quad V = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1),$$

  and Z and V are independent.

- By the definition of the t-distribution, it follows that

$$T = \frac{Z}{\sqrt{V/(n-1)}} = \frac{\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}}{\sqrt{\frac{(n-1)S^2}{\sigma^2}/(n-1)}} = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t(n-1). \qquad \square$$

Example 5.5.1 *We recently opened a small computer lab in my building. A sign-in system was instituted that allowed the computer center to monitor the number of users in a particular day. Following are the results from 29 days in January 1995. Suppose that the center will decide to staff the lab with an assistant if the true average number of daily users is more than 30, but clear evidence is needed to justify the expense. Should the assistant be hired?*

$$30, 32, 29, 9, 25, 25, 32, 41, 49, 44, 37, 12, 13, 44, 31,$$
$$45, 55, 52, 16, 26, 39, 40, 46, 58, 38, 9, 15, 39, 46$$

- $\bar{x} = 33.69$ and $s = 13.98$.
- $t = \frac{\bar{x}-\mu}{s/\sqrt{n}} = 1.42$
- $P[T \geq t = 1.42] = .0833$
- Strong evidence that $\mu > 30$ and should hire the assistant.
- If the administration is particularly concerned about service, then they probably should hire the assistant. $\square$

Definition 5.5.2 *The* ***F-density*** *with parameters r_1, r_2 (both called* ***degrees of freedom*** *is the p.d.f of the random variable*

$$F = \frac{U/r_1}{V/r_2}$$

where $U \sim \chi^2(r_1), V \sim \chi^2(r_2)$, and U and V are independent random variables. We use the shorthand notation $f(r_1, r_2)$ when referring to the distribution.

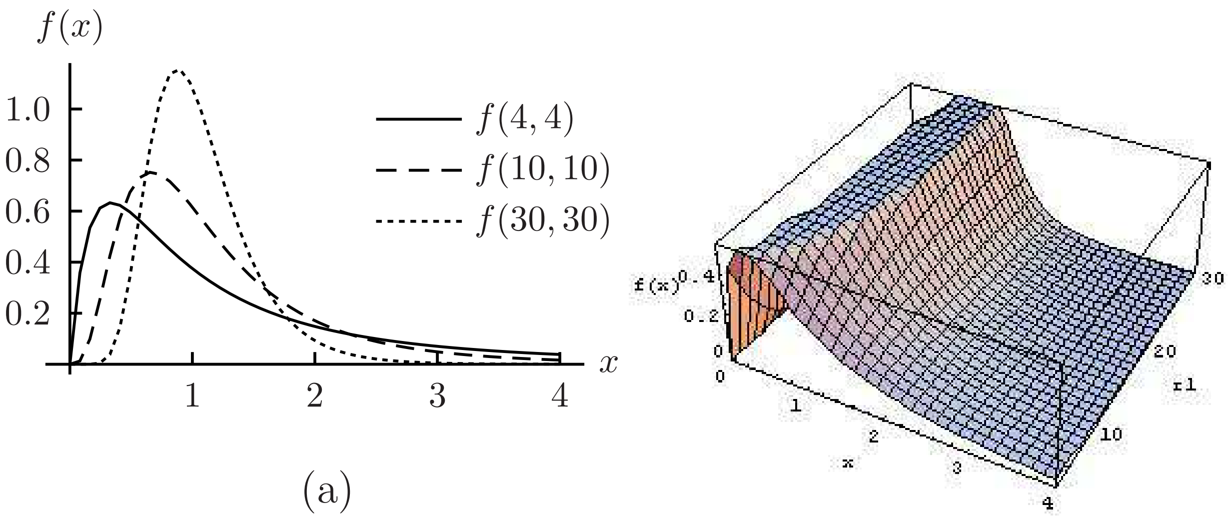


Figure 5.11: ***F*-densities**

- Set $F = \frac{U/r_1}{V/r_2}$ and $Y = V$, then

$$g(x) = \frac{\Gamma((r_1+r_2)/2)(r_1/r_2)^{r_1/2}}{\Gamma(r_1/2)\Gamma(r_2/2)} \frac{x^{r_1/2-1}}{(1+r_1x/r_2)^{(r_1+r_2)/2}}, \quad 0 < x < \infty.$$

- $F_{r_1,r_2} = 1/F_{r_2,r_1}$

Question 5.5.3 *If* $F \sim f(8, 10)$, *find* $P[F \leq 5.06]$, *and find a constant* c *such that* $P[F > c] = 0.05$.

Ans: $P[F \leq 5.06] = 0.99$, $P[F > 3.07] = 0.05$ □

- The F-table is not quite as incomplete as it might seem.
- $P\left[F_{r_1,r_2} = \frac{1}{F_{r_2,r_1}} < \frac{1}{f_\alpha(r_1,r_2)}\right] = P[F_{r_1,r_2} > f_\alpha(r_1, r_2)]$

Proposition 5.5.2 *Let* $X_1, X_2, \ldots, X_m$ *and* $Y_1, Y_2, \ldots, Y_n$ *be independent, random samples from normal populations with variances* σ_X^2 *and* σ_Y^2, *respectively. Denote the sample variances by* S_X^2 *and* S_Y^2. *Then*

$$F = \frac{S_X^2/\sigma_X^2}{S_Y^2/\sigma_Y^2} \sim f(m-1, n-1). \tag{5.81}$$

Proof:

- Since the samples are independent, by Proposition 5.4.5 we can form two independent χ^2 random variables:

$$U = \frac{(m-1)S_X^2}{\sigma_X^2} \sim \chi^2(m-1), \quad V = \frac{(n-1)S_Y^2}{\sigma_Y^2} \sim \chi^2(n-1).$$

- By (5.77),

$$F = \frac{(m-1)S_X^2/\sigma_X^2(m-1)}{(n-1)S_Y^2/\sigma_Y^2(n-1)} = \frac{S_X^2/\sigma_X^2}{X_Y^2/\sigma_Y^2} \sim f(m-1, n-1).$$ □

- Note that if $\sigma_X^2 = \sigma_y^2$, then the ratio S_X^2/S_Y^2 itself will have the $f(m-l, n-1)$ distribution.

- Extreme values of this ratio, which are either too close to zero or too large, furnish evidence against the equality of the population variances.

Example 5.5.2 *A compilation of average faculty salaries by rank for 25 institutions in our consortium recently came across my desk. I begin to wonder whether there was any difference in variability between salaries of assistant and full professors, hypothesizing that institutions might complete strongly to hire new Ph.D.s by matching other offers, which would tend to level out the assistant professor salaries, but they have different practices on merit raises, which would make salaries of senior people tend to vary more. If our consortium is typical of all colleges, we can extrapolate from it to the universe of all colleges (a risky proposition, admittedly).*

- The data follow.

Assistant
42725, 41602, 40965, 40547, 39959, 38585, 38214, 38131, 37385,
37167, 37141, 36499, 36458, 36224, 36186, 35927, 35764, 35703,
34613, 34474, 34469, 33067, 32994, 31618, 31451

Full
69306, 68411, 65136, 650102 63736, 63041, 62741, 59813, 59520,
58000, 57495, 57487, 56728, 56484, 55564, 55004, 54116, 53984,
53802, 53496, 52740, 50207, 48610, 45509, 42100

- Assume that the salaries are normally distributed and the samples are independent.

- If there is no difference, then $\sigma_x^2 = \sigma_y^2$.

- $F = \frac{S_x^2}{S_y^2} \sim f(24, 24)$

- $(2961)^2/(6700)^2 = .195$

- $P[F \leq .195] = .0001$ □

Question 5.5.4 *Plot histograms of the salary data for both ranks in the last example. Are there any striking departures from normality?*

5.6 Summary

1. The distribution of the sum of independent and identically distributed Bernoulli random variables is binomial. 3

2. If a random variable X has the $N(\mu, \sigma^2)$ is distribution, then the transformed random variable $Z = (X - \mu)/\sigma$ is standard normal. 3

3. The time T_n of the nth arrival in a Poisson process is the sum $S_1 + S_2 + \cdots + S_n$ of exponentially distributed interarrival times which is of the gamma family. 3

4. ***c.d.f. technique***: Find the distribution of a real-valued, continuous function of a continuous random variable. 3

5. ***Inverse transformation of $U(0,1)$***: Let F be a continuous, strictly increasing c.d.f. on a state space I that is a subinterval of the real line. Suppose that U has the uniform $(0,1)$ distribution. Then $X = F^{-1}(U)$ is a random variable with the c.d.f. F. Conversely, suppose that X is a random variable with continuous, strictly increasing c.d.f F. Then the random variable $U = F(X)$ has the uniform $(0,1)$ distribution. . 5

6. If U has the uniform $(0,1)$ distribution, then the random variable $S = F^{-1}(U) = -\ln(1-U)/\lambda = -\ln(U)/\lambda$ is exponential with parameter λ. 6

7. ***Polar transformation***: $R^2 = X^2 + Y^2,\ \tan(\Theta) = Y/X$ 8

8. ***One-variable transformation formula***:

$$\int_A f(x)\,\mathrm{d}x = \int_B f(g^{-1}(u)) \left|\frac{\mathrm{d}}{\mathrm{d}u} g^{-1}(u)\right| \mathrm{d}u.$$

. . . . 8

9. ***Two-variable transformation formula***:

$$\iint\limits_A f(x,y)\,\mathrm{d}y\,\mathrm{d}x = \iint\limits_B f(h_1(u,v), h_2(u,v))|J(u,v)|\,\mathrm{d}u\,\mathrm{d}v.$$

. . . . 8

10. ***Bivariate transformation***: Suppose that X and Y are continuous random variables with joint density $f(x,y)$. Let U and V be obtained from X and Y via the invertible transformation (5.14), whose inverse is in (5.15). Let J be the Jacobian of the transformation. Then the joint density of U and V is: $f(h_1(u,v), h_2(u,v))|J(u,v)|$.9

11. ***Joint transformation of n random variables***: Let

$$U_1 = g_1(X_1, \ldots, X_n),\ U_2 = g_2(X_1, \ldots, X_n), \ldots,\ U_n = g_n(X_1, \ldots, X_n)$$

be a 1-1 transformation with inverse

$$X_1 = h_1(U_1, \ldots, U_n),\ X_2 = h_2(U_1, \ldots, U_n), \ldots,\ X_n = h_n(U_1, \ldots, U_n).$$

Jacobian of the transformation

$$J = J(u_1, u_2, \ldots, u_n) = \det \begin{bmatrix} \frac{\partial h_1}{\partial u_1} & \frac{\partial h_1}{\partial u_2} & \cdots & \frac{\partial h_1}{\partial u_n} \\ \frac{\partial h_2}{\partial u_1} & \frac{\partial h_2}{\partial u_2} & \cdots & \frac{\partial h_2}{\partial u_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial h_n}{\partial u_1} & \frac{\partial h_n}{\partial u_2} & \cdots & \frac{\partial h_n}{\partial u_n} \end{bmatrix}$$

$$\begin{aligned} &\int\limits_A \cdots \int f(x_1, x_2, \ldots, x_n)\,\mathrm{d}x_n \ldots \mathrm{d}x_2\,\mathrm{d}x_1 \\ &= \int\limits_B \cdots \int f(h_1, h_2, \ldots, h_n)|J|\,\mathrm{d}u_1\,\mathrm{d}u_2 \ldots \mathrm{d}u_n \end{aligned}$$

. . . . 11

12. ***Order statistics of the sample*** $X_1, X_2, \ldots, X_n$: $Y_1 \leq Y_2 \leq \cdots \leq Y_n$ (some authors use the notation $X_{(1)}, \ldots, X_{(n)}$) 11

13. ***Joint pdf of order statistics***: Let $Y_1 \leq Y_2 \leq \cdots \leq Y_n$ be the order statistics of a random sample $X_1, X_2, \ldots, X_n$ taken from a continuous distribution with p.d.f f. Then the joint density of the Y's is

$$f_Y(y_1, y_2, \ldots, y_n) = n! \cdot f(y_1) \cdot f(y_2) \cdots f(y_n), \quad y_1 < y_2 < \cdots < y_n.$$

. . . . 13

14. ***Marginal density of ordered statistics***: Let Y_k be the k-th order statistic in a random sample $X_1, X_2, \ldots, X_n$ taken from a continuous distribution with density function $f(x)$ and c.d.f $F(x)$. The density of Y_k is $f_k(y_k) = \frac{n!}{(k-1)!(n-k)!} f(y_k)[F(y_k)]^{k-1}[1 - F(y_k)]^{n-k}$ 14

15. ***pdf of smallest order statistic***: $f_1(y_1) n f(y_1)[1 - F(y_1)]^{n-1}$ 16

16. ***pdf of largest order statistic***: $f_n(y_n) = n[F(y_n)]^{n-1} f(y_n)$ 16

17. ***Moment generating function (m.g.f)***: The moment generating function (m.g.f) of a real-valued random variable X is the function $M(t) = M_X(t) = E[e^{tX}]$ which is defined for all real values of t such that the expectation is finite. 18

18. ***m.g.f. of Poisson(μ)***: $M(t) = e^{\mu(e^t - 1)}, \quad t \in \mathbb{R}$ 18

19. ***m.g.f. of gamma(α, λ) distribution***: $M(t) = \left(\frac{\lambda}{\lambda - t}\right)^{\alpha}, \quad t < \lambda$ 19

20. ***m.g.f. of Exponential(λ) = $\Gamma(1, \lambda)$***: $M(t) = \frac{\lambda}{\lambda - t}, \quad t < \lambda$ 19

21. ***m.g.f. of $\chi^2(n)$ distribution***: $M(t) = \left(\frac{1}{1-2t}\right)^{\frac{n}{2}}, \quad t < 1/2$ 19

22. ***m.g.f. of $N(\mu, \sigma^2)$ distribution***: $M(t) = \exp(\mu t + \sigma^2 t^2/2), \quad t \in \mathbb{R}$. 19

23. Let $X_1, X_2, \ldots, X_n$ be independent random variables having m.g.f. $M_{X_i}(t)$, respectively. Then the random variable $Y = X_1 + X_2 + \cdots + X_n$ has the m.g.f. $\prod_{i=1}^{n} M_{X_i}(t)$. 20

24. ***Distribution of sum of i.i.d. Poisson r.v.***: Let $X_1, X_2, \ldots, X_n$ be independent random variables having the Poisson distribution with parameter $\mu_1, \mu_2, \ldots, \mu_n$, respectively. Then the random variable $Y = X_1 + X_2 + \cdots + X_n$ has the Poisson distribution with parameter $\mu_1 + \mu_2 + \cdots + \mu_n$. 20

25. ***Distribution of sum i.i.d. gamma r.v.***: Let $X_1, X_2, \ldots, X_n$ be independent random variables having gamma distributions with alpha parameters $\alpha_1, \alpha_2, \ldots, \alpha_n$, respectively, and common lambda parameter λ. Then the random variable $Y = X_1 + X_2 + \cdots + X_n$ has the gamma distribution with parameters $\alpha_1 + \alpha_2 + \cdots + \alpha_n$ and λ. 20

26. ***Distribution of sum i.i.d. χ^2 r.v.***: Let $X_1, X_2, \ldots, X_n$ be independent random variables having χ^2 distributions with parameters $r_1, r_2, \ldots, r_n$, respectively. Then the random variable $Y = X_1 + X_2 + \cdots + X_n$ has the $\chi^2(r_1 + r_2 + \cdots + r_n)$ distribution. 20

27. ***Distribution of sum i.i.d.*** $\exp(\lambda)$ ***r.v.***: Let $X_1, X_2, \ldots, X_n$ be independent random variables having $\exp(\lambda)$ distributions. Then the random variable $Y = X_1 + X_2 + \cdots + X_n$ has the $\Gamma(n, \lambda)$ distribution. 20

28. ***Distribution of sum i.i.d. normal r.v.***: Suppose that $X_1, X_2, \ldots, X_n$ are independent normal random variables with mean $\mu_1, \mu_2, \ldots, \mu_n$ and variance $\sigma_1^2, \sigma_2^2, \ldots, \sigma_n^2$, respectively. Let $a_1, a_2, \ldots, a_n$ be constants. Then $Y = \sum_{i=1}^n a_i X_i$ has the $N(\sum_{i=1}^n a_i\mu_i, \sum_{i=1}^n a_i^2\sigma_i^2)$ distribution. 21

29. ***Use m.g.f. to compute the moments***: (1) $\frac{\mathrm{d}}{\mathrm{d}t}M(t) = E\left[\frac{\mathrm{d}}{\mathrm{d}t}e^{tX}\right] = E[Xe^{tX}]$. (2) $\frac{\mathrm{d}}{\mathrm{d}t}M(t)\big|_{t=0} = E[X]$. (3) $\frac{\mathrm{d}^2}{\mathrm{d}t^2}M(t) = E\left[\frac{\mathrm{d}^2}{\mathrm{d}t^2}e^{tX}\right] = E[X^2e^{tX}]$. (4) $\frac{\mathrm{d}^2}{\mathrm{d}t^2}M(t)\Big|_{t=0} = E[X^2]$. (5) $\frac{\mathrm{d}^n}{\mathrm{d}t^n}M(t) = E\left[\frac{\mathrm{d}^n}{\mathrm{d}t^n}e^{tX}\right] = E[X^ne^{tX}]$. (6) $\frac{\mathrm{d}^n}{\mathrm{d}t^n}M(t)\big|_{t=0} = E[X^n]$. . . . 22

30. ***Use m.g.f. to compute variance***: $\mathrm{Var}(X) = M''(0) - (M'(0))^2$ 22

31. If $Y = aX + b$, then $M_Y(t) = E\left[e^{t(aX+b)}\right] = e^{bt}E\left[e^{(at)X}\right] = e^{bt}M_X(at)$. 22

32. The mean and variance of the $\Gamma(\alpha, \lambda)$ distribution are $\frac{\alpha}{\lambda}$ and $\frac{\alpha}{\lambda^2}$, respectively. 22

33. ***Moment generating function of a random vector***: The moment generating function of a random vector $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ is following real-valued function of a vector variable $\mathbf{t} = (t_1, \ldots, t_n)$: $M_{\mathbf{X}}(\mathbf{t}) = E[e^{\mathbf{t}'\mathbf{X}}] = E\left[\exp\left(\sum_{i=1}^n t_iX_i\right)\right]$ which is defined for all $\mathbf{t} \in \mathbb{R}^n$ such that the expectation is finite. 22

34. ***m.g.f. of multivariate normal distribution***: $M(\mathbf{t}) = e^{\boldsymbol{\mu}'\mathbf{t} + \frac{1}{2}\mathbf{t}'\Sigma\mathbf{t}}$, $\mathbf{t} \in \mathbb{R}^n$. . . . 23

35. ***Probability generating function***: (Discrete distribution) $P(t) = E[t^X] = \sum_{i=1}^{\infty} t^{x_i}P[X = x_i]$. 24

36. ***Characteristic function***: $\phi(t) = E[e^{itX}]$ 24

37. ***Sample mean***: Estimates the central tendency of the distribution, measured by μ. $\bar{X} = \sum_{i=1}^n \frac{X_i}{n}$. 24

38. ***Sample variance***: Estimates the spread of the distribution, measured by σ^2, by the average squared distance of data points from $\bar{X}$. $S^2 = \sum_{i=1}^n \frac{(X_i - \bar{X})^2}{n-1}$. 24

39. If X has the $N(\mu, \sigma^2)$ distribution, then the random variable $Y = cX + d$ has the $N(c\mu + d, c^2\sigma^2)$ distribution. 25

40. Let $X_1, X_2, \ldots, X_n$ be a random sample from the $N(\mu, \sigma^2)$ distribution. Then $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$ and $Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$. 25

41. If Z has the $N(0, 1)$ distribution, then the transformed random variable $Y = Z^2$ has the $\chi^2(1)$ distribution. Hence, if X is $N(\mu, \sigma^2)$, then $Y = \left(\frac{X-\mu}{\sigma}\right)^2 \sim \chi^2(1)$. 25

42. Let $Z_1, Z_2, \ldots, Z_n$ be independent and identically distributed $N(0, 1)$ random variables. Then $Y = Z_1^2 + Z_2^2 + \cdots + Z_n^2 \sim \chi^2(n)$. In particular, if $X_1, X_2, \ldots, X_n$ are normal random variables with mean $\mu_1, \ldots, \mu_n$ and variance $\sigma_1^2, \ldots, \sigma_n^2$, respectively, then $Y = \left(\frac{X_1 - \mu_1}{\sigma_1}\right)^2 + \cdots + \left(\frac{X_n - \mu_n}{\sigma_n}\right)^2 \sim \chi^2(n)$. 26

43. Let $X_1, X_2, \ldots, X_n$ be a random sample from the $N(\mu, \sigma^2)$ distribution. Then the random variable $\frac{(n-1)S^2}{\sigma^2} = \sum_{i=1}^n \frac{(X_i - \bar{X})^2}{\sigma^2}$ has the $\chi^2(n-1)$ distribution and furthermore is independent of $\bar{X}$. 27

44. Let A be an $m \times n$ matrix of row rank m, where $m \leq n$, and let b an $m \times 1$ column vector. If $\mathbf{X} = [X_1, X_2, \ldots, X_n]'$ is a multivariate normal random vector with mean vector $\boldsymbol{\mu}_X = [\mu_1, \ldots, \mu_n]'$ and covariance matrix $\sum_X$, then the random vector $\mathbf{Y} = A\mathbf{X} + \mathbf{b}$ has the multivariate normal distribution with mean vector $\boldsymbol{\mu}_Y = A\boldsymbol{\mu}_X + \mathbf{b}$ and covariance $\Sigma_Y = A\Sigma_X A'$. 28

45. $\sum_{i=1}^n X_i^2 = \sum_{i=1}^n (X_i - \bar{X})^2 + n\bar{X}^2 = \mathbf{X}'(I - \mathbf{1}_n\mathbf{1}_n'/n)\mathbf{X} + X'(\mathbf{1}_n\mathbf{1}_n'/n)\mathbf{X}$. 30

46. ***Idempotent matrix***: A matrix Q is symmetric and $Q^2 = Q$. It's eigenvalues are 0 or 1 only and the number of eigenvalues equal to 1 is the rank r of the matrix and it can be diagonal decomposed. 31

47. Suppose that $\mathbf{X} = [X_1, X_2, \ldots, X_n]'$ is a random vector that has the multivariate normal distribution with mean vector $\boldsymbol{\mu} = [\mu, \mu, \ldots, \mu]'$ and covariance matrix $\sigma^2 I_n$. Let Q be an $n \times n$ symmetric, idempotent matrix of rank r. If either $\mu = 0$ or all row sums of Q are 0, then the random variable $Y = \mathbf{X}'Q\mathbf{X}/\sigma^2$ has the $\chi^2(r)$ distribution. 31

48. Suppose that $\mathbf{X} = [X_1, X_1, \ldots, X_n]'$ is a random vector that has the multivariate normal distribution with mean vector $\boldsymbol{\mu} = [\mu, \ldots, \mu]'$ and covariance matrix $\sigma^2 I_n$. Let Q_1 and Q_2 be symmetric $n \times n$ matrices such that $Q_1 Q_2 = 0$. If either $\boldsymbol{\mu} = \mathbf{0}$ or the row sums of both Q_1 and Q_2 are all zero, then the random variables $Y_1 = \mathbf{X}'Q_1\mathbf{X}/\sigma^2$ and $Y_2 = \mathbf{X}'Q_2\mathbf{X}/\sigma^2$ are independent. 33

49. ***t-distribution***: The t-distribution with parameter r (called the degrees of freedom of the distribution) is the distribution of the random variable, $T = \frac{Z}{\sqrt{V/r}}$ where $Z \sim N(0,1), V \sim \chi^2(r)$, and Z and V are independent. We use $t(r)$ as a shorthand notation for this distribution. 34

50. Let $\bar{X}$ and S^2 be the sample mean and variance of a random sample of size n from the $N(\mu, \sigma^2)$ distribution. Then $T = \frac{\bar{X}-\mu}{S/\sqrt{n}} \sim t(n-1)$. 34

51. ***F-distribution***: The F-density with parameters r_1, r_2 (both called degrees of freedom is the p.d.f of the random variable $F = \frac{U/r_1}{V/r_2}$ where $U \sim \chi^2(r_1), V \sim \chi^2(r_2)$, and U and V are independent random variables. We use the shorthand notation $f(r_1, r_2)$ when referring to the distribution. 35

52. ***Variance comparison***: Let $X_1, X_2, \ldots, X_m$ and $Y_1, \ldots, Y_n$ be independent, random samples from normal populations with variances σ_X^2 and σ_Y^2, respectively. Denote the sample variances by S_X^2 and S_Y^2. Then $F = \frac{S_X^2/\sigma_X^2}{S_Y^2/\sigma_Y^2} \sim f(m-1, n-1)$. 36

Chapter 6

Summary

6.1 Summary

1. ***Random***: Pertain to an experiment whose result remains uncertain until the experiment is performed or phenomenon is observed. 3

2. ***Sample space***: The set of all possible outcomes of the experiment. 4

3. ***Event***: Subset of the sample space. 4

4. ***Probability***: A measure of the likelihood of events. 4

5. ***Random variable***: A function gives a numerical value to each outcome of a random experiment. 5

6. ***Sample space***: The sample space Ω of a random experiment is the collection of all possible outcomes. 5

7. ***Axioms of probability***: A probability measure on a sample space Ω of a random experiment is a function $P[\cdot]$ that maps events in Ω to real numbers such that $P[\Omega] = 1, P[F] \geq 0$ for all events $F, P[E_1 \cup E_2 \cup \cdots] = P[E_1] + P[E_2] + \cdots$ where E_i are disjoint events. 5

8. ***Probability of empty set***: $P[\emptyset] = 0$. 7

9. For any event E, denote the complement of E by E^c. Then $P[E] + P[E^c] = 1$. 7

10. For any event $E, P[E] \leq 1$. 7

11. For any event A and B, $P[A] = P[A \cap B] + P[A \cap B^c]$. 7

12. If $E \subseteq F$ are event, then $P[E] \leq P[F]$. 8

13. If A and B are events, then $P[A \cup B] = P[A] + P[B] - P[A \cap B]$. 8

14. ***Principle of inclusion-exclusion***: If $A_i, i = 1, 2, \ldots, n$, are events, then $P\left[\bigcup_{i=1}^{n} A_i\right] = \sum_{k=1}^{n} (-1)^{k-1} \sum_{1 \leq i_1 < i_2 < \cdots < i_k \leq n} P[A_{i_1} A_{i_2} \cdots A_{i_k}]$. 9

15. ***Random variable***: Let Ω be a sample space, and let E be a subset of $\mathbb{R}^n$. A ***random variable*** X is a function from Ω into E. 9

16. ***State space***: Let Ω be a sample space, and let E be a subset of $\mathbb{R}^n$. We call E the state space of the random variable X. 9

17. ***Probability distribution***: The probability distribution Q of a random variable X is the probability measure on E defined by $Q(A) = P[X \in A]$, for $A \subseteq E$. 10

18. ***Probability mass function (p.m.f.)***: Suppose that a random variable X has a discrete (i.e., finite or countable) state space. The function $q : E \to [0, 1]$ is called the probability mass function (p.m.f.) of X if $q(x) = Q(\{x\}) = P[X = x]$. 12

19. ***Probability density function (p.d.f.)***: A random variable X is said to have probability density function (p.d.f.) f if, for all subsets A of the state space, $Q(A) = P[X \in A] = \int_A f(x)\,dx$. 12

20. ***Cumulative distribution function (c.d.f.)***: The cumulative distribution function (c.d.f.) is defined as $F(x) = P[X \le x] = \int_{-\infty}^{x} f(t)\,dt$ where $f(x) = F'(x)$. 12

21. ***Conditional probability***: If B is an event such that $P[B] > 0$, then the conditional probability of an event A given B is defined as $P[A|B] = \frac{P[A \cap B]}{P[B]}$. 13

22. ***Alternative definition of conditional probability***: $P[A \cap B] = P[B] \cdot P[A|B]$ 15

23. ***Generalized multiplication rule***: If $A_1, \dots, A_n$ are events such that the following conditional probabilities are defined, then $P[A_1 \cap A_2 \cap \dots \cap A_n] = P[A_1] \cdot P[A_2|A_1] \cdots P[A_n|A_1 \cap \dots \cap A_{n-1}]$. 15

24. ***Law of Total Probability***: Let A be an event, and let $B_1, \dots, B_n$ be mutually exclusive events of nonzero probability whose union is the sample space Ω. Then $P[A] = P[B_1] \cdot P[A|B_1] + P[B_2] \cdot P[A|B_2] + \dots + P[B_n] \cdot P[A|B_n]$. 16

25. ***Bayes' Theorem***: Let the sets A and $B_i, i = 1, 2, \dots, n$ satisfy the hypothesis of Proposition 1.4.2. Then for each $i = 1, 2, \dots, n$, $P[B_i|A] = \frac{P[A|B_i] \cdot P[B_i]}{P[A|B_1] \cdot P[B_1] + \dots + P[A|B_n] \cdot P[B_n]}$. 19

26. ***Independent events***: Events A and B are said to be independent of one another if $P[A|B] = P[A]$ provided $P[B] > 0$. 20

27. Let A and B be events of positive probability. Then A and B are independent of each other if and only if $P[A \cap B] = P[A] \cdot P[B]$. 21

28. ***Mutually independent events***: Events $A_1, \dots, A_n$ are said to be ***mutually independent*** if for any subcollection $A_{i_1}, \dots, A_{i_k}, 1 \le i_1 < \dots < i_k \le n$, of the events, $P[A_{i_1} \cap A_{i_2} \cap \dots \cap A_{i_k}] = P[A_{i_1}] \cdot P[A_{i_2}] \cdots P[A_{i_k}]$. 22

29. End of Chap. 1 ———————————————— 25

30. ***Fundamental counting principle***:

    (a) ***Rule of product***: Suppose that an experiment has two stages. For the first stage, there are m possible outcomes, and for each of these, the second stage has n possible outcomes. Then the two-stage experiment has $m \cdot n$ outcomes.

(b) ***Rule of sum***: For a more general two-stage experiment, let the first-stage outcomes be labeled $i = 1, 2, \ldots, m$. Assume that if the first-stage outcome is i, then there are n_i possible outcomes for stage 2. Then the two-stage experiment has

$$\sum_{i=1}^{m} n_i$$

possible outcomes.

(c) ***Rule of product***: Suppose that an experiment consists of k stages such that the first stage has m_1 possible outcomes, for each outcome of stage 1 there are m_2 possible outcomes of stage 2, for each combined outcome of the first two stages there are m_3 possible outcomes of stages 3, and so on. Then there are $m_1 \cdot m_2 \cdots m_k$ outcomes of the entire experiment.

. . . . 28

31. ***Permutation***: A permutation of n objects $\{y_1, \ldots, y_n\}$, taken r at a time is an ordered list $(x_1, \ldots, x_r)$ selected from the original n objects, such that $x_i \neq x_j, \forall i \neq j$. We denote the number of such permutations by $P_{n,r}$. 30

32. ***Combination***: A combination of n items $\{y_1, \ldots, y_n\}, r$ at a time, is a subset $\{x_1, \ldots, x_r\}$ selected from the original n items, such that $x_i \neq x_j, \forall i \neq j$. We denote the number of such combinations by $C_{n,r}$, or $\binom{n}{r}$. The latter is read "n choose r." 31

33. $P_{n,r} = n(n-1)\cdots(n-r+1) = \frac{n!}{(n-r)!}$ 31

34. $C_{n,r} = \binom{n}{r} = \frac{n!}{r!\cdot(n-r)!}$ 32

35. ***Probability mass function (p.m.f.)***: $q(x) = P[X = x] \quad \text{for } x \in E$. 33

36. ***Cumulative distribution function (c.d.f.)***: (Discrete case) $F(x) = P[X \leq x] = \sum_{t \leq x} q(t)$. 33

37. ***Discrete uniform distribution***: $q(x) = \frac{1}{n}, x \in \{x_1, x_2, \ldots, x_n\}$ 34

38. ***Empirical probability mass function (emf)***: The empirical cumulative distribution function (edf) of the sample is then the c.d.f. associated with $\hat{q}$ through equation (2.7), alternatively, $\hat{F}(w_j) = \frac{\text{number of } X_i = w_j}{n}$. 34

39. ***Empirical cumulative distribution function (edf)***: The empirical probability mass function (emf) of the sample is $\hat{q}(w_j) = \frac{\text{number of } X_i \leq w_j}{n}$. 34

40. ***Hypergeometric distribution***: The experiments involving sampling without replacement discussed in the last section give rise to a frequently observed distribution. 35

41. ***Marginal mass function***: The marginal mass function of X_i is $q_i(x_i) = P[X_i = x_i] = \sum_{x_1} \cdots \sum_{x_{i-1}} \sum_{x_{i+1}} \cdots \sum_{x_n} q(x_1, \ldots, x_n)$. 37

42. ***Joint marginal mass function***: The joint marginal mass function of a subcollection $X_{i_1}, \ldots, X_{i_k}$ of the random variables is $q(x_{i_1}, \ldots, x_{i_k}) = P[X_{i_1} = x_{i_1}, \ldots, X_{i_k} = x_{i_k}] = \sum \cdots \sum q(x_1, \ldots, x_n)$. 37

43. ***Binomial probability mass function***: Let X be the total number of successes in n Bernoulli trials. Then X has the binomial probability mass function: $q(k) = P[X = k] = \binom{n}{k} p^k (1-p)^{n-k}$. 39

44. ***Geometric distribution***: Let T_1 be the random variable that returns the trial on which the first success occurs in a sequence of Bernoulli trials. Then T_1 has the geometric distribution $P[T_1 = n] = (1-p)^{n-1} p, n = 1, 2, \ldots$. 41

45. ***Geometric distribution without memory***:

$$P[X > n + k | X > n] = P[n > k] = (1-p)^k.$$

. . . . 42

46. ***Negative binomial distribution***: Let T_r be the trial on which the rth success occurs in a sequence of Bernoulli trials. Then T_r has the negative binomial distribution: $P[T_r = n] = \binom{n-1}{r-1} p^r (1-p)^{n-r}, n = r, r+1, \ldots$. 42

47. If $X_1, \ldots, X_k$ have the $m(n, p_1, \ldots, p_k)$ distribution, then for each i, X_i has the $b(n, p_i) = m(n, p_1, 1 - p_1)$ distribution. 44

48. If $X_1, \ldots, X_k$ have the $m(n, p_1, \ldots, p_k)$ distribution, then for $X_{i_1}, X_{i_2}, \ldots, X_{i_m}$ has the $m(n, p_{i_1}, p_{i_2}, \ldots, p_{i_m}, 1 - \sum_{j=1}^{m} p_{i_j})$ distribution. 45

49. ***Poisson probability mass function***: The Poisson probability mass function with parameter λ is $q(k) = P[X = k] = \frac{e^{-\lambda}\lambda^k}{k!}, k = 0, 1, 2, \ldots$. 47

50. ***Poisson process***: A Poisson process is a family $(N_t)_{t\geq 0}$ of random variables whose paths are step functions beginning at state 0 at time 0, which jump by 1 at a sequence of random times $T_1, T_2, \ldots$. Additionally, the some conditions are assumed. If $(N_t)_{t\geq 0}$ is a Poisson process with rate parameter λ, then $P[N_t = n] = \frac{e^{-\lambda t}(\lambda t)^n}{n!}$. 50

51. ***Poisson(λt) distribution***: If $(N_t)_{t\geq 0}$ is a Poisson process with rate parameter λ, then $P[N_t = n] = \frac{e^{-\lambda t}(\lambda t)^n}{n!}$, that is, the number of arrivals up through time t has the Poisson(λt) distribution. 51

52. ***Expected value or expectation***: Let X be a discrete, real-value random variable with state space $E = \{e_1, e_2, \ldots\}$ and probability mass function q. Then the expected value or expectation of X is $\mu = E[X] = \sum_i e_i P[X = e_i] = \sum_i e_i q(e_i)$ provided the series converges. 53

53. ***Mean***: $\mu = E[X]$ 53

54. If c is a real constant, then $E[c] = c$. 58

55. ***Linearity of expectation***: If X and Y are random variables with finite expectation, then $E[aX + bY] = aE[X] + bE[Y]$. 59

56. ***Variance***: The variance of a real random variable X is $\sigma^2 = \text{Var}(X) = E[(X-\mu)^2]$ provided the expectation is finite. 60

57. ***Standard deviation***: The square root of the variance, σ, is referred to as the standard deviation of the random variable. 60

58. $\sigma^2 = E[X^2] - \mu^2$. 61

59. If X is a real random variable with finite variance and a and b are real constants, then $\mathrm{Var}(aX+b) = a^2\mathrm{Var}(X)$. 61

60. ***Sample mean***: $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ 61

61. ***Variance of $\overline{X}$***: $\mathrm{Var}(\overline{X}) = \frac{1}{n}\sigma^2$ 62

62. ***Moment***: The rth moment of the distribution of a real random variable X is $E[X^r]$, provided the expectation exists. 62

63. ***Central moment or moment about the mean***: The rth central moment or moment about the mean is $E[(X-\mu)^r]$. 62

64. ***Skewness***: The third moment about the mean $E[(X-\mu)^3]$. 62

65. ***Skewed to the right***: A distribution has a long right tail. 62

66. ***Skewed to the left***: A distribution has a long left tail. 62

67. If X has the Bernoulli distribution with success parameter p, then $E[X] = p$; $\mathrm{Var}(X) = p(1-p)$. 62

68. If X has the $b(n,p)$ distribution, then $E[X] = np$ and $\mathrm{Var}(X) = np(1-p)$. . . 63

69. If X has the Poisson(λ) distribution, then $E[X] = \lambda$; $\mathrm{Var}(X) = \lambda$. 65

70. If X has the geometric distribution with parameter p, then $E[X] = \frac{1}{p}$ and $\mathrm{Var}(X) = \frac{1-p}{p^2}$. If X has the negative binomial distribution with parameters r and p, then $E[X] = \frac{r}{p}$ and $\mathrm{Var}(X) = r \cdot \frac{1-p}{p^2}$. 66

71. If $\mathbf{X} = [X_1\ X_2\ \ldots\ X_n]'$ is a random vector, then the expected value of $\mathbf{X}$ is the vector $E[\mathbf{X}] = \begin{pmatrix} E[X_1] \\ E[X_2] \\ \vdots \\ E[X_n] \end{pmatrix}$ provided the individual component expectations exist. . 67

72. ***Linearity of expectation***: Let $\mathbf{X}$ and $\mathbf{Y}$ be random vectors of the same dimension n, and let a and b be scalar constants. Then, $E[a\mathbf{X}+b\mathbf{Y}] = aE[\mathbf{X}]+bE[\mathbf{Y}]$ provided the expectations of $\mathbf{X}$ and $\mathbf{Y}$ exist. 68

73. Let $\mathbf{X}$ be an n-dimensional random vector and let A be an $m \times n$ matrix of constants. Then, $E[A\mathbf{X}] = AE[\mathbf{X}]$ provided the expectation of $\mathbf{X}$ exists. 68

74. End of Chap. 6 ———————— 72

75. ***Probability density function (p.d.f.)***: $Q(A) = P[X \in A] = \int_A f(x)\,\mathrm{d}x$. (a) $f(x) \geq 0, \forall x \in E$; (b) $\int_E f(x)\,\mathrm{d}x = 1$ 75

76. ***Cumulative distribution function (c.d.f.)***: $F(x) = P[X \leq x] = \int_{-\infty}^{x} f(t)\,\mathrm{d}t$. (a) $F'(x) = f(x)$; (b) F is a nondecreasing, nonnegative function; (c) $\lim_{x\to-\infty} F(x) = 0$; (d) $\lim_{x\to\infty} F(x) = 1$; (e) $P[a < X \leq b] = F(b) - F(a)$. 75

94. ***Mean and variance of exponential distribution***: If X has the exponential distribution with parametric λ, then $E[X] = \frac{1}{\lambda}$ and $\text{Var}(X) = \frac{1}{\lambda^2}$. 89

95. ***Distribution of interarrival times***: Let $T_1, T_2, T_3, \ldots$ be the arrival times of a Poisson process with rate λ. Then the interarrival time random variables $T_1, T_2 - T_1, T_3 - T_2, \ldots$ each have the $\exp(\lambda)$ distribution. Furthermore, the interarrival times are independent. 89

96. ***Gamma density***: The gamma density is $f(t) = \frac{\lambda^\alpha}{\Gamma(\alpha)} t^{\alpha-1} e^{-\lambda t}, t > 0$. 89

97. ***Gamma function***: The gamma function is defined by $\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x}\, dx$. $\Gamma(\alpha+1) = \alpha\Gamma(\alpha),\ \alpha > 0,\ \Gamma(1/2) = \sqrt{\pi}$ and $\Gamma(n) = (n-1)!,\ n \in \mathbb{N}$. 89

98. ***Erlang density***: is $\Gamma(n, \lambda)$ with $n \in \mathbb{N}$. 91

99. ***Distribution of nth arrival time***: Let $T_1, T_2, T_3, \ldots$ be the arrival times of a Poisson process with rate λ. Then T_n has $\Gamma(n, \lambda)$ distribution. Furthermore, the time between the mth and $(m+n)$th arrivals $T_{m+n} - T_m$ has the $\Gamma(n, \lambda)$ distribution for $m, n > 0$. 91

100. ***Mean and variance of Gamma distribution***: If T has the $\Gamma(\alpha, \lambda)$ distribution, then $E[T] = \frac{\alpha}{\lambda}$ and $\text{Var}(T) = \frac{\alpha}{\lambda^2}$. 91

101. ***Chi-square density***: The chi-square density is $f(x) = \frac{1}{2^{n/2}\Gamma(n/2)} x^{n/2-1} e^{-x/2}, x > 0$. It is $\Gamma(n/2, 1/2)$ with mean n and variance $2n$. 92

102. ***Normal density***: The normal density is $f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), -\infty < x < \infty$. 93

103. ***Mean and variance of normal distribution***: If X has the $N(\mu, \sigma^2)$ distribution, then $E[X] = \mu$ and $\text{Var}(X) = \sigma^2$. 95

104. ***Standardizing***: If a random variable X has the $N(\mu, \sigma^2)$ distribution, then the random variable Z defined by $Z = \frac{X-\mu}{\sigma}$ has the standard normal distribution. Therefore $P[X \le b] = P\left[Z = \frac{X-\mu}{\sigma} \le \frac{b-\mu}{\sigma}\right] = \int_{-\infty}^{(b-\mu)/\sigma} \frac{1}{\sqrt{2\pi}} e^{-z^2/2}\, dz$. 96

105. ***Weibull(λ, β) distribution***: $f(t) = \beta\lambda^\beta t^{\beta-1} e^{-(\lambda t)^\beta},\ F(t) = 1 - e^{-(\lambda t)^\beta}, \lambda > 0, \beta > 0, t > 0$. $\text{Weibull}(\lambda, 1) \equiv \exp(\lambda)$ 99

106. ***Lognormal density***: $f(y) = \frac{1}{y}\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(\ln(y)-\mu)^2}{2\sigma^2}\right], \mu \in \mathbb{R}, \sigma^2 > 0, y > 0$. $X = \log Y \sim N(\mu, \sigma^2)$. 99

107. ***Pareto density***: $f(x) = \frac{\theta}{(1+x)^{\theta+1}},\ x \ge 0, \theta > 0$ 100

108. ***Cauchy density***: $f(x) = \frac{1}{\pi[1+(x-\mu)^2]}, -\infty < x < \infty$. 100

109. ***Beta density***: $f(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}, \alpha, \beta > 0, 0 < x < 1$. 100

110. ***Multivariate normal distribution***: A random vector $\mathbf{X} = (X_1, X_2, \ldots, X_n)'$ is said to have the multivariate normal distribution with parameters $\boldsymbol{\mu}$ and Σ if its density is $f(\mathbf{x}) = \frac{1}{(2\pi)^{n/2}\sqrt{\det(\Sigma)}} e^{-(1/2)(\mathbf{x}-\boldsymbol{\mu})'\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})}, \mathbf{x} \in \mathbb{R}^n$. 103

111. Let $\mathbf{X} = (X_1, X_2, \ldots, X_n)'$ have the multivariate normal distribution with mean $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_n)'$ and covariance Σ. If Σ is a diagonal matrix with entries $\sigma_1^2, \ldots, \sigma_n^2$ on its diagonal, then the joint density is $f(\mathbf{x}) = f_1(x_1) \cdot f_2(x_2) \cdots f_n(x_n)$ where f_i is the $N(\mu_i, {\sigma_i}^2)$ density. Consequently, under these assumptions the marginal distribution of each X_i is $N(\mu_i, {\sigma_i}^2)$. 104

112. ***Bivariate normal distribution***: $N(\boldsymbol{\mu}, \Sigma)$ where $\Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{bmatrix}$

$$f(\mathbf{x}) = \frac{1}{2\pi\sqrt{\det \Sigma}} \cdot \exp\left[-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})'\Sigma(\mathbf{x} - \boldsymbol{\mu})\right]$$

. . . . 104

113. ***Marginal distribution of bivariate normal distribution***: Let $\mathbf{X} = (X, Y)$ have the bivariate normal distribution $(\mu_x, \mu_y, \sigma_x^2, \sigma_y^2, \rho)$. Then X has the $N(\mu_x, {\sigma_x}^2)$ distribution, and Y has the $N(\mu_y, {\sigma_y}^2)$ distribution. 105

114. End of Chap. 3 ——————————————————————————————————— 109

115. ***Mutually independent***: Random variables $X_1, X_2, \ldots, X_n$ are called mutually independent if for any subcollection of them $X_{i1}, X_{i2}, \ldots, X_{ik}, k \leq n$, and corresponding subsets $B_{i1}, B_{i2}, \ldots, B_{ik}$ of their state spaces, $P[X_{i1} \in B_{i1}, X_{i2} \in B_{i2}, \ldots, X_{ik} \in B_{ik}] = P[X_{i1} \in B_{i1}] \cdot P[X_{i2} \in B_{i2}] \cdots P[X_{ik} \in B_{ik}]$. 112

116. ***Dependent***: Random variables that are not independent are called dependent. 112

117. The following are equivalent:

    (a) $X_1, X_2, \ldots, X_n$ are independent random variables.

    (b) If $F(x_1, x_2, \ldots, x_n)$ is the joint c.d.f of $X_1, X_2, \ldots, X_n$ and $F_1(x_1), F_2(x_2), \ldots, F_n(x_n)$ are the marginal c.d.f.'s, then

    $$F(x_1, x_2, \ldots, x_n) = F_1(x_1) \cdot F_2(x_2) \cdots F_n(x_n).$$

    (c) If $f(x_1, x_2, \ldots, x_n)$ is the joint probability density function (mass function in the discrete) of $X_1, X_2, \ldots, X_n$ and $f_1(x_1), f_2(x_2), \ldots, f_n(x_n)$ are the marginal densities(or mass functions), then

    $$f(x_1, x_2, \ldots, x_n) = f_1(x_1) \cdot f_2(x_2) \cdots f_n(x_n).$$

. . . . 113

118. Suppose that $X_1, X_2, \ldots, X_n$ are mutually independent random variables, and suppose that $f_1, f_2, \ldots, f_n$ are functions whose domains include the state spaces of the corresponding $X_1, X_2, \ldots, X_n$. Then $f_1(X_1), f_2(X_2), \ldots, f_n(X_n)$ are mutually independent random variables. 118

119. ***Inverse image***: If A is a set in the range of a function f, define the inverse image of A as the following subset of the domain of f: $f^{-1}(A) = \{x | f(x) \in A\}$. 118

120. Suppose that $X_1, X_2, \ldots, X_n$ are independent random variables, and suppose that $h_1, h_2, \ldots, h_n$ are functions whose domains include the state spaces of the corresponding $X_1, X_2, \ldots, X_n$. Then $E[h_1(X_1)\cdot h_2(X_2)\cdots h_n(X_n)] = E[h_1(X_1)]\cdot E[h_2(X_2)]\cdots E[h_n(X_n)]$ provided the expectations exist. 118

121. ***Variance of sum of independent R.V.***: If $X_1, X_2, \ldots, X_n$ are independent random variables, then $\text{Var}\left(\sum_{i=1}^n c_i X_i\right) = \sum_{i=1}^n c_i^2 \text{Var}(X_i)$ provided the variances exist. 119

122. ***Random sample***: A random sample $X_1, \ldots, X_n$ is a collection of n independent and identically distributed (i.i.d.) random variables. 120

123. ***Conditional probability mass function***: If X and Y are discrete random variables with joint probability mass function $f(x, y)$, and f_X and f_Y are the marginal mass functions, then the conditional probability mass function of Y given $X = x$ is $f(y|x) = \frac{f(x,y)}{f_X(x)}$ provided $f_X(x) > 0$. 122

124. ***Joint conditional p.m.f.***: Consider a random vector $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ with joint p.m.f. $f(x_1, x_2, \ldots, x_n)$. The joint conditional p.m.f. of $X_{m+1}, \ldots, X_n$ given $X_1, \ldots, X_m$ is $f(x_{m+1}, \ldots, x_n | x_1, \ldots, x_m) = \frac{f(x_1,\ldots,x_n)}{f_{1,\ldots,m}(x_1,\ldots,x_m)}$. 122

125. ***Conditional probability density function***: If X and Y are continuous random variables with joint probability density function $f(x, y)$, and f_X and f_Y are the marginal density functions, then the conditional probability density function of Y given $X = x$ is $f(y|x) = \frac{f(x,y)}{f_X(x)}$ provided $f_X(x) > 0$. Similarly the conditional probability density function of X given $Y = y$ is $f(x|y) = \frac{f(x,y)}{f_Y(y)}$ provided $f_Y(y) > 0$. 124

126. ***Conditional expectation***: The conditional expectation $E[g(Y)|X = x]$ of a function of a continuous random variable given the observed value of another continuous random variable is $E[g(Y)|X = x] = \int_{-\infty}^{\infty} g(y) f(y|x)\, \mathrm{d}y$. The integral is replaced by a sum in the discrete case. 125

127. ***Conditional mean***: The conditional mean of Y given $X = x$ is $\mu_{Y|x} = E[Y|X = x] = \int_{-\infty}^{\infty} y \cdot f(y|x)\, \mathrm{d}y$. 126

128. ***Conditional variance***: The conditional variance of Y given $X = x$ is $\sigma^2_{Y|x} = E[(Y - \mu_{Y|x})^2 | X = x] = \int_{-\infty}^{\infty} (y - \mu_{Y|x})^2 f(y|x)\, \mathrm{d}y$. 126

129. ***Conditional expectation formula***: If the expectations of $g(Y)$ and $h(X)$ exist, then $E[g(Y)] = E[E[g(Y)|X]]$. 126

130. ***Covariance***: The covariance of two real random variables X and Y is $\sigma_{XY} = \text{Cov}(X, Y) = E[(X - \mu_X)(Y - \mu_Y)]$ provided the expectation exists. 128

131. ***Correlation***: If the covariance and the marginal variances exist, the correlation between X and Y is $\rho = \rho_{XY} = \text{Corr}(X, Y) = E\left[\left(\frac{X-\mu_X}{\sigma_X}\right)\left(\frac{Y-\mu_Y}{\sigma_Y}\right)\right] = \frac{\text{Cov}(X,Y)}{\sigma_X \sigma_Y}$. . 128

132. $\text{Cov}(X, X) = \text{Var}(X)$ 128

133. If X and Y are real random variables, and a, b, c, d are real constants with $b, d \neq 0$, then $\text{Cov}(a + bX, c + dY) = b \cdot d \cdot \text{Cov}(X, Y)$. 130

134. $\text{Corr}(a + bX, c + dY) = \text{Corr}(X, Y)$ (if b, d have the same sign) or $= -\text{Corr}(X, Y)$ (if b, d have opposite signs). 130

135. If X and Y are real random variables with correlation ρ, then $|\rho| \le 1$. Moreover, $|\rho| = 1$ if and only if there are constants a, b and $b \ne 0$ such that $Y = a + bX$ with probability 1. If X and Y are independent, then both $\operatorname{Cov}(X, Y) = 0$ and $\rho = 0$.131

136. If $X_1, X_2, \ldots, X_n$ are real random variables and $a_1, a_2, \ldots, a_n$ are real constants, then $\operatorname{Var}\left(\sum_{i=1}^n a_i X_i\right) = \sum_{i=1}^n a_i^2 \cdot \operatorname{Var}(X_i) + \sum_{j,k=1, j\ne k}^n a_j \cdot a_k \cdot \operatorname{Cov}(X_j, X_k)$ provided the variances and covariances exist. Consequently, if each pair X_j, X_k is uncorrelated, then $\operatorname{Var}\left(\sum_{i=1}^n a_i X_i\right) = \sum_{i=1}^n a_i^2 \cdot \operatorname{Var}(X_i)$. 133

137. Suppose that X and Y are random variables such that the conditional mean of Y given X is a linear function of x. Then $\mu_{Y|x} = \mu_Y + \frac{\rho\sigma_Y}{\sigma_X}(x - \mu_X)$. 135

138. ***Covariance matrix***: The covariance matrix of $\mathbf{X}$ is the $n \times n$ symmetric matrix $\Sigma =$ $\operatorname{Cov}(\mathbf{X}) = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \ldots & \sigma_{1n} \\ \sigma_{21} & \sigma_2^2 & \ldots & \sigma_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \ldots & \sigma_n^2 \end{bmatrix}$ where $\sigma_i^2 = \operatorname{Var}(X_i)$ and $\sigma_{ij} = \sigma_{ji} = \operatorname{Cov}(X_i, X_j)$. 136

139. ***Correlation matrix***: The correlation matrix of $\mathbf{X}$ is the $n \times n$ symmetric matrix $\Upsilon = \operatorname{Corr}(\mathbf{X}) = \begin{bmatrix} 1 & \rho_{12} & \ldots & \rho_{1n} \\ \rho_{21} & 1 & \ldots & \rho_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{n1} & \rho_{n2} & \ldots & 1 \end{bmatrix}$ where $\rho_{ij} = \rho_{ji} = \operatorname{Corr}(X_i, X_j)$. 137

140. Let $\mathbf{X}$ be a random vector of n components whose covariance matrix Σ exists, and let A be a constant $m \times n$ matrix. Then, $\operatorname{Cov}(A \cdot \mathbf{X}) = A \cdot \operatorname{Cov}(\mathbf{X}) \cdot A' = A\Sigma A'$.137

141. ***Bilinear property of covariance***: Let $\mathbf{X} = [X_1, X_2, \ldots, X_m]'$ and $\mathbf{Y} = [Y_1, Y_2, \ldots, Y_n]'$ be random vectors, and let $\mathbf{a} = [a_1, a_2, \ldots, a_m]$ and $\mathbf{b} = [b_1, b_2, \ldots, b_n]$ be constant row vectors. Then $\operatorname{Cov}\left(\sum_{i=1}^m a_i X_i, \sum_{j=1}^n b_j Y_j\right) = \sum_{i=1}^m \sum_{j=1}^n a_i b_j \operatorname{Cov}(X_i, Y_j)$. . . . 138

142. ***Multivariate normal density***:

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{n/2}\sqrt{\det(\Sigma)}} e^{-(1/2)(\mathbf{X}-\boldsymbol{\mu})'\Sigma^{-1}(\mathbf{X}-\boldsymbol{\mu})}, \quad \mathbf{x} \in \mathbb{R}^n$$

. . . . 139

143. ***Bivariate normal density***: $\boldsymbol{\mu} = (\mu_x, \mu_y)$, $\Sigma = \begin{bmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{bmatrix}$

$$f(x, y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \times \exp\left[-\frac{1}{2(1-\rho^2)}\left[\frac{(x-\mu_x)^2}{\sigma_x^2} - 2\rho\frac{(x-\mu_x)(y-\mu_y)}{\sigma_x\sigma_y} + \frac{(y-\mu_y)^2}{\sigma_y^2}\right]\right]$$

. . . . 139

144. Let the random vector $\mathbf{X} = [X_1\ X_2\ \ldots\ X_n]'$ have the multivariate normal distribution with mean $\boldsymbol{\mu} = [\mu_1\ \mu_2\ \ldots\ \mu_n]'$ and covariance matrix Σ. Then Σ is a diagonal matrix with diagonal entries $\sigma_1^2, \sigma_2^2, \ldots, \sigma_n^2$ if and only if $X_1, X_2, \ldots, X_n$ are mutually independent and X_i has the $N(\mu_i, \sigma_i^2)$ distribution. 140

145. If $\mathbf{X} = [X\ Y]'$ has the bivariate normal density as described earlier, then the conditional density of Y given $X = x$ is $N(\mu_{y|x}, \sigma^2_{y|x})$, where $\mu_{y|x} = \mu_y + \rho\frac{\sigma_y}{\sigma_x}(x - \mu_x); \sigma^2_{y|x} = \sigma^2_y(1-\rho^2)$. Similarly, the conditional density of X given $Y = y$ is $N(\mu_{x|y}, \sigma^2_{x|y})$, where $\mu_{x|y} = \mu_x + \rho\frac{\sigma_x}{\sigma_y}(y - \mu_y); \sigma^2_{x|y} = \sigma^2_x(1-\rho^2)$. 140

146. Let $\mathbf{X} = \begin{bmatrix}\mathbf{X}_1\\ \mathbf{X}_2\end{bmatrix}$, $\boldsymbol{\mu} = \begin{bmatrix}\boldsymbol{\mu}_1\\ \boldsymbol{\mu}_2\end{bmatrix}$, and $\Sigma = \begin{bmatrix}\Sigma_{11} & \Sigma_{12}\\ \Sigma_{21} & \Sigma_{22}\end{bmatrix}$. Then, $\mathbf{X}_1$ has the multivariate normal distribution with mean vector $\boldsymbol{\mu}_1$ and covariance matrix Σ_{11}. $\mathbf{X}_2$ has the multivariate normal distribution with mean vector $\boldsymbol{\mu}_2$ and covariance matrix Σ_{22}. Conditioned on $\mathbf{X}_1 = \mathbf{x}_1$, $\mathbf{X}_2$ has the multivariate normal distribution with mean vector and covariance matrix $\boldsymbol{\mu}_{2|1} = \boldsymbol{\mu}_2 + \Sigma_{21}\Sigma_{11}^{-1}(\mathbf{x}_1 - \boldsymbol{\mu}_1); \Sigma_{2|1} = \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}$. Conditioned on $\mathbf{X}_2 = \mathbf{x}_2$, $\mathbf{X}_1$ has the multivariate normal distribution with mean vector and covariance matrix $\boldsymbol{\mu}_{1|2} = \boldsymbol{\mu}_1 + \Sigma_{12}\Sigma_{22}^{-1}(\mathbf{x}_2 - \boldsymbol{\mu}_2); \Sigma_{1|2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$. 142

147. End of Chap. 4 ———————— 148

148. The distribution of the sum of independent and identically distributed Bernoulli random variables is binomial. 149

149. If a random variable X has the $N(\mu, \sigma^2)$ is distribution, then the transformed random variable $Z = (X - \mu)/\sigma$ is standard normal. 149

150. The time T_n of the nth arrival in a Poisson process is the sum $S_1 + S_2 + \cdots + S_n$ of exponentially distributed interarrival times which is of the gamma family. 149

151. ***c.d.f. technique***: Find the distribution of a real-valued, continuous function of a continuous random variable. 149

152. ***Inverse transformation of $U(0,1)$***: Let F be a continuous, strictly increasing c.d.f. on a state space I that is a subinterval of the real line. Suppose that U has the uniform $(0,1)$ distribution. Then $X = F^{-1}(U)$ is a random variable with the c.d.f. F. Conversely, suppose that X is a random variable with continuous, strictly increasing c.d.f F. Then the random variable $U = F(X)$ has the uniform $(0,1)$ distribution.151

153. If U has the uniform $(0,1)$ distribution, then the random variable $S = F^{-1}(U) = -\ln(1-U)/\lambda = -\ln(U)/\lambda$ is exponential with parameter λ. 152

154. ***Polar transformation***: $R^2 = X^2 + Y^2$, $\tan(\Theta) = Y/X$ 154

155. ***One-variable transformation formula***:

$$\int_A f(x)\,\mathrm{d}x = \int_B f(g^{-1}(u)) \left|\frac{\mathrm{d}}{\mathrm{d}u} g^{-1}(u)\right| \mathrm{d}u.$$

. . . . 154

156. ***Two-variable transformation formula***:

$$\iint_A f(x,y)\,\mathrm{d}y\,\mathrm{d}x = \iint_B f(h_1(u,v), h_2(u,v))|J(u,v)|\,\mathrm{d}u\,\mathrm{d}v.$$

. . . . 154

157. ***Bivariate transformation***: Suppose that X and Y are continuous random variables with joint density $f(x,y)$. Let U and V be obtained from X and Y via the invertible transformation (5.14), whose inverse is in (5.15). Let J be the Jacobian of the transformation. Then the joint density of U and V is: $f(h_1(u,v),h_2(u,v))|J(u,v)|$. 155

158. ***Joint transformation of n random variables***: Let

$$U_1 = g_1(X_1,\dots,X_n),\; U_2 = g_2(X_1,\dots,X_n),\dots,\; U_n = g_n(X_1,\dots,X_n)$$

be a 1-1 transformation with inverse

$$X_1 = h_1(U_1,\dots,U_n),\; X_2 = h_2(U_1,\dots,U_n),\dots,\; X_n = h_n(U_1,\dots,U_n).$$

Jacobian of the transformation

$$J = J(u_1,u_2,\dots,u_n) = \det\begin{bmatrix} \frac{\partial h_1}{\partial u_1} & \frac{\partial h_1}{\partial u_2} & \cdots & \frac{\partial h_1}{\partial u_n} \\ \frac{\partial h_2}{\partial u_1} & \frac{\partial h_2}{\partial u_2} & \cdots & \frac{\partial h_2}{\partial u_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial h_n}{\partial u_1} & \frac{\partial h_n}{\partial u_2} & \cdots & \frac{\partial h_n}{\partial u_n} \end{bmatrix}$$

$$\int \cdots \int_A f(x_1,x_2,\dots,x_n)\,\mathrm{d}x_n \dots \mathrm{d}x_2\,\mathrm{d}x_1$$
$$= \int \cdots \int_B f(h_1,h_2,\dots,h_n)|J|\,\mathrm{d}u_1\,\mathrm{d}u_2 \dots \mathrm{d}u_n$$

. . . . 157

159. ***Order statistics of the sample $X_1,X_2,\dots,X_n$***: $Y_1 \le Y_2 \le \cdots \le Y_n$ (some authors use the notation $X_{(1)},\dots,X_{(n)}$) 157

160. ***Joint pdf of order statistics***: Let $Y_1 \le Y_2 \le \cdots \le Y_n$ be the order statistics of a random sample $X_1,X_2,\dots,X_n$ taken from a continuous distribution with p.d.f f. Then the joint density of the Y's is

$$f_Y(y_1,y_2,\dots,y_n) = n!\cdot f(y_1)\cdot f(y_2)\cdots f(y_n), \quad y_1 < y_2 < \cdots < y_n.$$

. . . . 159

161. ***Marginal density of ordered statistics***: Let Y_k be the k-th order statistic in a random sample $X_1,X_2,\dots,X_n$ taken from a continuous distribution with density function $f(x)$ and c.d.f $F(x)$. The density of Y_k is $f_k(y_k) = \frac{n!}{(k-1)!(n-k)!}f(y_k)[F(y_k)]^{k-1}[1-F(y_k)]^{n-k}$. 160

162. ***pdf of smallest order statistic***: $f_1(y_1)nf(y_1)[1-F(y_1)]^{n-1}$ 162

163. ***pdf of largest order statistic***: $f_n(y_n) = n[F(y_n)]^{n-1}f(y_n)$ 162

164. ***Moment generating function (m.g.f)***: The moment generating function (m.g.f) of a real-valued random variable X is the function $M(t) = M_X(t) = E[e^{tX}]$ which is defined for all real values of t such that the expectation is finite. 164

165. ***m.g.f. of Poisson***(μ): $M(t) = e^{\mu(e^t-1)}, \quad t \in \mathbb{R}$ 164

166. ***m.g.f. of gamma***(α, λ) ***distribution***: $M(t) = \left(\frac{\lambda}{\lambda-t}\right)^{\alpha}, \quad t < \lambda$ 165

167. ***m.g.f. of Exponential***$(\lambda) = \Gamma(1, \lambda)$: $M(t) = \frac{\lambda}{\lambda-t}, \quad t < \lambda$ 165

168. ***m.g.f. of*** $\chi^2(n)$ ***distribution***: $M(t) = \left(\frac{1}{1-2t}\right)^{\frac{n}{2}}, \quad t < 1/2$ 165

169. ***m.g.f. of*** $N(\mu, \sigma^2)$ ***distribution***: $M(t) = \exp(\mu t + \sigma^2 t^2/2), \quad t \in \mathbb{R}$. 165

170. Let $X_1, X_2, \ldots, X_n$ be independent random variables having m.g.f. $M_{X_i}(t)$, respectively. Then the random variable $Y = X_1 + X_2 + \cdots + X_n$ has the m.g.f. $\prod_{i=1}^{n} M_{X_i}(t)$. 166

171. ***Distribution of sum of i.i.d. Poisson r.v.***: Let $X_1, X_2, \ldots, X_n$ be independent random variables having the Poisson distribution with parameter $\mu_1, \mu_2, \ldots, \mu_n$, respectively. Then the random variable $Y = X_1 + X_2 + \cdots + X_n$ has the Poisson distribution with parameter $\mu_1 + \mu_2 + \cdots + \mu_n$. 166

172. ***Distribution of sum i.i.d. gamma r.v.***: Let $X_1, X_2, \ldots, X_n$ be independent random variables having gamma distributions with alpha parameters $\alpha_1, \alpha_2, \ldots, \alpha_n$, respectively, and common lambda parameter λ. Then the random variable $Y = X_1 + X_2 + \cdots + X_n$ has the gamma distribution with parameters $\alpha_1 + \alpha_2 + \cdots + \alpha_n$ and λ. 166

173. ***Distribution of sum i.i.d.*** χ^2 ***r.v.***: Let $X_1, X_2, \ldots, X_n$ be independent random variables having χ^2 distributions with parameters $r_1, r_2, \ldots, r_n$, respectively. Then the random variable $Y = X_1 + X_2 + \cdots + X_n$ has the $\chi^2(r_1 + r_2 + \cdots + r_n)$ distribution. 166

174. ***Distribution of sum i.i.d.*** $\exp(\lambda)$ ***r.v.***: Let $X_1, X_2, \ldots, X_n$ be independent random variables having $\exp(\lambda)$ distributions. Then the random variable $Y = X_1 + X_2 + \cdots + X_n$ has the $\Gamma(n, \lambda)$ distribution. 166

175. ***Distribution of sum i.i.d. normal r.v.***: Suppose that $X_1, X_2, \ldots, X_n$ are independent normal random variables with mean $\mu_1, \mu_2, \ldots, \mu_n$ and variance $\sigma_1^2, \sigma_2^2, \ldots, \sigma_n^2$, respectively. Let $a_1, a_2, \ldots, a_n$ be constants. Then $Y = \sum_{i=1}^{n} a_i X_i$ has the $N(\sum_{i=1}^{n} a_i \mu_i, \sum_{i=1}^{n} a_i^2 \sigma_i^2)$ distribution. 167

176. ***Use m.g.f. to compute the moments***: (1) $\frac{d}{dt} M(t) = E\left[\frac{d}{dt} e^{tX}\right] = E[X e^{tX}]$. (2) $\frac{d}{dt} M(t)\big|_{t=0} = E[X]$. (3) $\frac{d^2}{dt^2} M(t) = E\left[\frac{d^2}{dt^2} e^{tX}\right] = E[X^2 e^{tX}]$. (4) $\frac{d^2}{dt^2} M(t)\Big|_{t=0} = E[X^2]$. (5) $\frac{d^n}{dt^n} M(t) = E\left[\frac{d^n}{dt^n} e^{tX}\right] = E[X^n e^{tX}]$. (6) $\frac{d^n}{dt^n} M(t)\big|_{t=0} = E[X^n]$. . . 168

177. ***Use m.g.f. to compute variance***: $\mathrm{Var}(X) = M''(0) - (M'(0))^2$ 168

178. If $Y = aX + b$, then $M_Y(t) = E\left[e^{t(aX+b)}\right] = e^{bt} E\left[e^{(at)X}\right] = e^{bt} M_X(at)$. 168

179. The mean and variance of the $\Gamma(\alpha, \lambda)$ distribution are $\frac{\alpha}{\lambda}$ and $\frac{\alpha}{\lambda^2}$, respectively. . . . 168

180. ***Moment generating function of a random vector***: The moment generating function of a random vector $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ is following real-valued function of a vector variable $\mathbf{t} = (t_1, \ldots, t_n)$: $M_{\mathbf{X}}(\mathbf{t}) = E[e^{\mathbf{t}'\mathbf{X}}] = E\left[\exp\left(\sum_{i=1}^{n} t_i X_i\right)\right]$ which is defined for all $\mathbf{t} \in \mathbb{R}^n$ such that the expectation is finite. 168

181. ***m.g.f. of multivariate normal distribution***: $M(\mathbf{t}) = e^{\boldsymbol{\mu}'\mathbf{t} + \frac{1}{2}\mathbf{t}'\Sigma\mathbf{t}}, \ \mathbf{t} \in \mathbb{R}^n$. . . 169

182. ***Probability generating function***: (Discrete distribution) $P(t) = E[t^X] = \sum_{i=1}^{\infty} t^{x_i} P[X = x_i]$. 170

183. ***Characteristic function***: $\phi(t) = E[e^{itX}]$ 170

184. ***Sample mean***: Estimates the central tendency of the distribution, measured by μ. $\bar{X} = \sum_{i=1}^{n} \frac{X_i}{n}$. 170

185. ***Sample variance***: Estimates the spread of the distribution, measured by σ^2, by the average squared distance of data points from $\bar{X}$. $S^2 = \sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{n-1}$. 170

186. If X has the $N(\mu, \sigma^2)$ distribution, then the random variable $Y = cX + d$ has the $N(c\mu + d, c^2\sigma^2)$ distribution. 171

187. Let $X_1, X_2, \ldots, X_n$ be a random sample from the $N(\mu, \sigma^2)$ distribution. Then $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$ and $Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$. 171

188. If Z has the $N(0, 1)$ distribution, then the transformed random variable $Y = Z^2$ has the $\chi^2(1)$ distribution. Hence, if X is $N(\mu, \sigma^2)$, then $Y = \left(\frac{X-\mu}{\sigma}\right)^2 \sim \chi^2(1)$. 171

189. Let $Z_1, Z_2, \ldots, Z_n$ be independent and identically distributed $N(0, 1)$ random variables. Then $Y = Z_1^2 + Z_2^2 + \cdots + Z_n^2 \sim \chi^2(n)$. In particular, if $X_1, X_2, \ldots, X_n$ are normal random variables with mean $\mu_1, \ldots, \mu_n$ and variance $\sigma_1^2, \ldots, \sigma_n^2$, respectively, then $Y = \left(\frac{X_1 - \mu_1}{\sigma_1}\right)^2 + \cdots + \left(\frac{X_n - \mu_n}{\sigma_n}\right)^2 \sim \chi^2(n)$. 172

190. Let $X_1, X_2, \ldots, X_n$ be a random sample from the $N(\mu, \sigma^2)$ distribution. Then the random variable $\frac{(n-1)S^2}{\sigma^2} = \sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{\sigma^2}$ has the $\chi^2(n-1)$ distribution and furthermore is independent of $\bar{X}$. 173

191. Let A be an $m \times n$ matrix of row rank m, where $m \leq n$, and let b an $m \times 1$ column vector. If $\mathbf{X} = [X_1, X_2, \ldots, X_n]'$ is a multivariate normal random vector with mean vector $\boldsymbol{\mu}_X = [\mu_1, \ldots, \mu_n]'$ and covariance matrix $\sum_X$, then the random vector $\mathbf{Y} = A\mathbf{X} + \mathbf{b}$ has the multivariate normal distribution with mean vector $\boldsymbol{\mu}_Y = A\boldsymbol{\mu}_X + \mathbf{b}$ and covariance $\Sigma_Y = A\Sigma_X A'$. 174

192. $\sum_{i=1}^{n} X_i^2 = \sum_{i=1}^{n} (X_i - \bar{X})^2 + n\bar{X}^2 = \mathbf{X}'(I - \mathbf{1}_n\mathbf{1}_n'/n)\mathbf{X} + X'(\mathbf{1}_n\mathbf{1}_n'/n)\mathbf{X}$ 176

193. ***Idempotent matrix***: A matrix Q is symmetric and $Q^2 = Q$. It's eigenvalues are 0 or 1 only and the number of eigenvalues equal to 1 is the rank r of the matrix and it can be diagonal decomposed. 177

194. Suppose that $\mathbf{X} = [X_1, X_2, \ldots, X_n]'$ is a random vector that has the multivariate normal distribution with mean vector $\boldsymbol{\mu} = [\mu, \mu, \ldots, \mu]'$ and covariance matrix $\sigma^2 I_n$. Let Q be an $n \times n$ symmetric, idempotent matrix of rank r. If either $\mu = 0$ or all row sums of Q are 0, then the random variable $Y = \mathbf{X}'Q\mathbf{X}/\sigma^2$ has the $\chi^2(r)$ distribution. 177

195. Suppose that $\mathbf{X} = [X_1, X_1, \ldots, X_n]'$ is a random vector that has the multivariate normal distribution with mean vector $\boldsymbol{\mu} = [\mu, \ldots, \mu]'$ and covariance matrix $\sigma^2 I_n$. Let Q_1 and Q_2 be symmetric $n \times n$ matrices such that $Q_1 Q_2 = 0$. If either $\boldsymbol{\mu} = \mathbf{0}$ or the row sums of both Q_1 and Q_2 are all zero, then the random variables $Y_1 = \mathbf{X}'Q_1\mathbf{X}/\sigma^2$ and $Y_2 = \mathbf{X}'Q_2\mathbf{X}/\sigma^2$ are independent. 179

196. ***t-distribution***: The t-distribution with parameter r (called the degrees of freedom of the distribution) is the distribution of the random variable, $T = \frac{Z}{\sqrt{V/r}}$ where $Z \sim N(0,1), V \sim \chi^2(r)$, and Z and V are independent. We use $t(r)$ as a shorthand notation for this distribution. 180

197. Let $\bar{X}$ and S^2 be the sample mean and variance of a random sample of size n from the $N(\mu, \sigma^2)$ distribution. Then $T = \frac{\bar{X}-\mu}{S/\sqrt{n}} \sim t(n-1)$. 180

198. ***F-distribution***: The F-density with parameters r_1, r_2 (both called degrees of freedom is the p.d.f of the random variable $F = \frac{U/r_1}{V/r_2}$ where $U \sim \chi^2(r_1), V \sim \chi^2(r_2)$, and U and V are independent random variables. We use the shorthand notation $f(r_1, r_2)$ when referring to the distribution. 181

199. ***Variance comparison***: Let $X_1, X_2, \ldots, X_m$ and $Y_1, \ldots, Y_n$ be independent, random samples from normal populations with variances σ_X^2 and σ_Y^2, respectively. Denote the sample variances by S_X^2 and S_Y^2. Then $F = \frac{S_X^2/\sigma_X^2}{S_Y^2/\sigma_Y^2} \sim f(m-1, n-1)$. 182

200. End of Chap. 5 187

Part II

Appendices

Appendix A 統計學（一）報告

A.1 報告寫作的注意事項

請自己從各大學圖書館的考古題網站：如

1. 臺閩地區圖書館暨資料單位名錄
   http://wwwsrch.ncl.edu.tw/libdir/

   碩士班入學考古題

   國立大學

2. 國立臺灣大學圖書館
   http://www.lib.ntu.edu.tw/exam/graduate/college.htm

3. 國立臺灣師範大學圖書館
   http://www.lib.ntnu.edu.tw/libweb/qlink/exam.php

4. 國立政治大學圖書館
   http://www.lib.nccu.edu.tw/exam/index.htm

5. 國立交通大學浩然圖書館
   http://www.lib.nctu.edu.tw/n_exam/index.html

6. 國立清華大學圖書館
   http://www.lib.nthu.edu.tw/library/department/ref/exam/index.htm

7. 國立中央大學圖書館
   http://www.lib.ncu.edu.tw/cexamn.html

8. 國立中興大學圖書館
   http://recruit.nchu.edu.tw/

9. 國立中正大學圖書館
   http://www.lib.ccu.edu.tw/gradexam/kind.htm

10. 國立成功大學圖書館
    http://eserv.lib.ncku.edu.tw/exam/index.php

11. 國立中山大學圖書館
    http://www.lib.nsysu.edu.tw/exam

12. 國立高雄師範大學圖書館
http://www.nknu.edu.tw/~math/mathweb/frame.asp
13. 國立東華大學圖書館
http://www.lib.ndhu.edu.tw/index.phtml?path=,175,164&language=zh_tw
14. 國立海洋大學圖書館
http://www.lib.ntou.edu.tw/exam/exam.htm
15. 國立台北大學圖書館
http://www.ntpu.edu.tw/library/lib/ntpulib_exam.htm
16. 國立暨南大學圖書館
http://www.library.ncnu.edu.tw/download/old_exam.htm
17. 國立中正理工學院
http://www.lib.ccit.edu.tw/search/search6.htm

私立大學

18. 私立文化大學圖書館
http://www.lib.pccu.edu.tw/exam.html
19. 私立淡江大學圖書館
http://www.lib.tku.edu.tw/exam/exam-tku.shtml
20. 私立中原大學圖書館
http://www.lib.cycu.edu.tw/exams_new/exams_new.html
21. 私立輔仁大學圖書館
http://lib.fju.edu.tw/collection/examine.htm
22. 私立逢甲大學圖書館
http://www.admission.fcu.edu.tw/test_question.htm
23. 私立元智大學圖書館
http://www.yzu.edu.tw/library/index.php/content/view/152/253/
24. 私立銘傳大學圖書館
http://140.131.66.3/
25. 私立靜宜大學圖書館
http://www.lib.pu.edu.tw/new/exam/
26. 私立東吳大學
http://www.scu.edu.tw/entrance/exam92/index.htm
27. 私立中山醫學大學圖書館
http://www.lib.csmu.edu.tw/overlib/2206.php
28. 私立高雄醫學大學圖書館
http://www.kmu.edu.tw/%7Elib/kmul/exam.htm
29. 私立大同大學圖書館
http://www.library.ttu.edu.tw/eresource/exam.htm

30. 私立義守大學
 http://www1.isu.edu.tw/exam/exam/
31. 私立世新大學圖書館
 http://lib.shu.edu.tw/search_taskpaper.asp
32. 私立南華大學圖書館
 http://libserver2.nhu.edu.tw/20.htm
33. 私立華梵大學圖書館
 http://huafan.hfu.edu.tw/~lib/srvc/exam2/exam2.htm
34. 私立玄奘大學
 http://www.hcu.edu.tw/hcu2/old/old.asp

或研究所升學的統計或機率參考書，找出十題研究所入學考題與統計學（一）所教授的觀念相關的題目，每章題目找兩題。寫作的注意事項：

1. 第一行標明出處及關鍵詞，第二行題目所在的網址或參考書的作者、年代、版次及書名。
2. 題目不變，中文就用中文，英文就用英文，不用翻譯成中文。
3. 解題過程詳細講解每一個解題步驟，請參考範例。
4. 預計第三、五章結束需要上台報告自己每階段的作品（五題）。
5. 使用提供的範本作報告，檔名：m962040002蔡仲信.doc。
6. 學期末繳交m962040002蔡仲信.doc。

A.2 報告範例

統計學（一）報告

蔡仲信
m962040002
國立中山大學應用數學系
tsaijs1@gmail.com
2008-06-10

1. 【94中山應數統計組，隨機變數分解求期望值】
 http://www.lib.nsysu.edu.tw/exam/master/sci/math/94.pdf
 An urn contains $n+m$ balls, of which n are red and m are black. They are withdrawn from the urn, one at a time and without replacement. Let Y denote the number of red balls chosen after the first but before the second black ball has been chosen. Number the red balls from 1 to n. Find $E[Y]$.

 Ans: 在第一個黑球被選中後且第二個黑球被選中前，若紅球i 被選中，則令$Y_i = 1, i = 1, \ldots, n$。所以$Y = \sum_{i=1}^n Y_i$。

$$\begin{aligned} E[Y_i] &= P[Y_i = 1] \\ &= P[\text{從}m+1\text{ 個球中選中紅球}i\ ] \\ &= 1/(m+1)\ \text{因為}m+1\text{ 個球被選中的機率皆相等} \end{aligned}$$

因此，

$$E[Y] = n/(m+1)$$

2. 【94中山應數統計組，非單調隨機變數變換的機率密度函數】
http://www.lib.nsysu.edu.tw/exam/master/sci/math/94.pdf
Let X have pdf $f_X(x) = \frac{2}{9}(x+1),\ -1 < x < 2$. Find the pdf of $Y = X^2$.

Ans: 這轉換是非單調的。所以我們利用

$$\begin{aligned} f_Y(y) &= [f_X(\sqrt{y}) + f_X(-\sqrt{y})] \left| \frac{1}{2\sqrt{y}} \right| \\ &= \begin{cases} \frac{1}{2\sqrt{y}}[\frac{2}{9}(\sqrt{y}+1) + \frac{2}{9}(-\sqrt{y}+1)] = \frac{2}{9}\frac{1}{\sqrt{y}}, & \text{對}0 < y \leq 1, \\ \frac{1}{2\sqrt{y}}[\frac{2}{9}(\sqrt{y}+1) + 0] = \frac{1}{9} + \frac{1}{9\sqrt{y}}, & \text{對}1 < y \leq 4. \end{cases} \end{aligned}$$

$f_Y(y) = 0$ 對所有$y < 0$ 且對$y > 4$。

3. 【95中央經濟，柴比雪夫不等式】
http://www.lib.ncu.edu.tw/exam/MA07_95_01.pdf
假設台積電股價報酬率X為一隨機變數，若已知台積電股價報酬率期望值為2%，且$X \leq -2\%$的機率為0.3, $X \geq 6\%$的機率為0.1，則購買台積電股票的風險（以標準差衡量）不能小於多少?

Ans: 已知$P[X \leq -0.02] = 0.3, P[X \geq 0.06] = 0.1$，因此

$$\begin{aligned} P[X \leq -0.02, X \geq 0.06] &= 0.3 + 0.1 \\ &= 0.4 \end{aligned}$$

即$P[|X - 0.02| \geq 0.04] = 0.4$。利用柴比雪夫不等式

$$P[|X - 0.02| \geq k] \leq \frac{\sigma^2}{k^2}$$

當$k = 0.04$時得$\sigma^2 \geq k^2 P[|X - 0.02| \geq k] = 0.04^2 \times 0.4$，所以$\sigma \geq 0.04 \times \sqrt{0.4} = 0.025298$。

4. 【95高應大國企、商務經營，隨機變數和的期望值】
http://academic.kuas.edu.tw/recruit/Files//2007q2k8l15y35h49p.pdf
有一骰子出現1點及2點的機率均為$\frac{1}{12}$，出現3點的機率為$\frac{1}{3}$，其餘點數出現的機率均為$\frac{1}{6}$。將此骰子連續擲兩次，求此兩次投擲結果點數和之期望值。

Ans: 令X_1為骰子第一次出現點數的隨機變數:

X_1	1	2	3	4	5	6
$f(x_1)$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{3}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

令X_2為骰子第二次出現點數的隨機變數:

X_2	1	2	3	4	5	6
$f(x_2)$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{3}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

令$Y = X_1 + X_2$爲骰子兩次點數和的隨機變數。因爲

$$\begin{aligned} E[X_1] = E[X_2] &= \sum x_2 f(x_2) \\ &= 1 \times \frac{1}{12} + 2 \times \frac{1}{12} + 3 \times \frac{1}{3} + 4 \times \frac{1}{6} + 5 \times \frac{1}{6} + 6 \times \frac{1}{6} \\ &= 3.75 \end{aligned}$$

所以

$$\begin{aligned} E[Y] = E[X_1 + X_2] &= E[X_1] + E[X_2] \\ &= 2E[X_1] \\ &= 2 \times 3.75 \\ &= 7.5 \end{aligned}$$

Appendix **B**

What is Probability?

Contents

Probability: A number expressing the likelihood that a specific event will occur, expressed as the ratio of the number of actual occurrences to the number of possible occurrences. (The American Heritage® Dictionary of the English Language: Fourth Edition. 2000.)

B.1 Introduction

Probability, also theory of probability, branch of mathematics that deals with measuring or determining quantitatively the likelihood that an event or experiment will have a particular outcome. Probability is based on the study of permutations and combinations and is the necessary foundation for statistics.

The foundation of probability is usually ascribed to the 17th-century French mathematicians Blaise Pascal and Pierre de Fermat, but mathematicians as early as Gerolamo Cardano had made important contributions to its development. Mathematical probability began in an attempt to answer certain questions arising in games of chance, such as how many times a pair of dice must be thrown before the chance that a six will appear is 50-50. Or, in another example, if two players of equal ability, in a match to be won by the first to win ten games, are obliged to suspend play when one player has won five games, and the other seven, how should the stakes be divided?

The probability of an outcome is represented by a number between 0 and 1, inclusive, with "probability 0"indicating certainty that an event will not occur and "probability 1"indicating certainty that it will occur. The simplest problems are concerned with the probability of a specified "favorable"result of an event that has a finite number of equally likely outcomes. If an event has n equally likely outcomes and f of them are termed favorable, the probability, p, of a favorable outcome is f/n. For example, a fair die can be cast in six equally likely ways; therefore, the probability of throwing a 5 or a 6 is $2/6$. More involved problems are concerned with events in which the various possible outcomes are not equally likely. For example, in finding the probability of throwing a 5 or 6 with a pair of dice, the various outcomes $(2, 3, \ldots, 12)$ are not all equally likely. Some events may have infinitely many outcomes, such as the probability that a chord drawn at random in a circle will be longer than the radius.

Problems involving repeated trials form one of the connections between probability and statistics. To illustrate, what is the probability that exactly five 3s and at least four 6s will occur in 50 tosses of a fair die? Or, a person, tossing a fair coin twice, takes a step to the north, east, south, or west, according to whether the coin falls head, head; head, tail; tail, head; or tail, tail. What is the probability that at the end of 50 steps the person will be within 10 steps of the starting point?
In probability problems, two outcomes of an event are mutually exclusive if the probability of their joint occurrence is zero; two outcomes are independent if the probability of their joint occurrence is given as the product of the probability of their separate occurrences. Two outcomes are mutually exclusive if the occurrence of one precludes the occurrence of the other; two outcomes are independent if the occurrence or nonoccurrence of one does not alter the probability that the other will or will not occur. Compound probability is the probability of all outcomes of a certain set occurring jointly; total probability is the probability that at least one of a certain set of outcomes will occur. Conditional probability is the probability of an outcome when it is known that some other outcome has occurred or will occur.
If the probability that an outcome will occur is p, the probability that it will not occur is $q = 1 - p$. The odds in favor of the occurrence are given by the ratio $p : q$, and the odds against the occurrence are given by the ratio $q : p$. If the probabilities of two mutually exclusive outcomes X and Y are p and P, respectively, the odds in favor of X and against Y are p to P. If an event must result in one of the mutually exclusive outcomes $O_1, O_2, \ldots, O_n$, with probabilities $p_1, p_2, \ldots, p_n$, respectively, and if $v_1, v_2, \ldots, v_n$ are numerical values attached to the respective outcomes, the expectation of the event is $E = p_1v_1 + p_2v_2 + \cdots + p_nv_n$. For example, a person throws a die and wins 40 cents if it falls $1, 2$, or 3; 30 cents for 4 or 5; but loses \$1.20 if it falls 6. The expectation on a single throw is $(3/6) \times .40 + (2/6) \times .30 - (1/6) \times 1.20 = .10$.
The most common interpretation of probability is used in statistical analysis. For example, the probability of throwing a 7 in one throw of two dice is $1/6$, and this answer is interpreted to mean that if two fair dice are randomly thrown a very large number of times, about one-sixth of the throws will be 7s. This concept is frequently used to statistically determine the probability of an outcome that cannot readily be tested or is impossible to obtain. Thus, if long-range statistics show that out of every 100 people between 20 and 30 years of age, 42 will be alive at age 70, the assumption is that a person between those ages has a 42 percent probability of surviving to the age of 70.
Mathematical probability is widely used in the physical, biological, and social sciences and in industry and commerce. It is applied in such diverse areas as genetics, quantum mechanics, and insurance. It also involves deep and important theoretical problems in pure mathematics and has strong connections with the theory, known as mathematical analysis, that developed out of calculus.

Contributed By: James Singer

Reviewed By: J. Lennart Berggren Microsoft® Encarta® 2006.

B.2 Difference between statistics and probability

(Image by MIT OpenCourseWare. Based on Gilbert, Norma. Statistics. W.B. Saunders Co., 1976.)

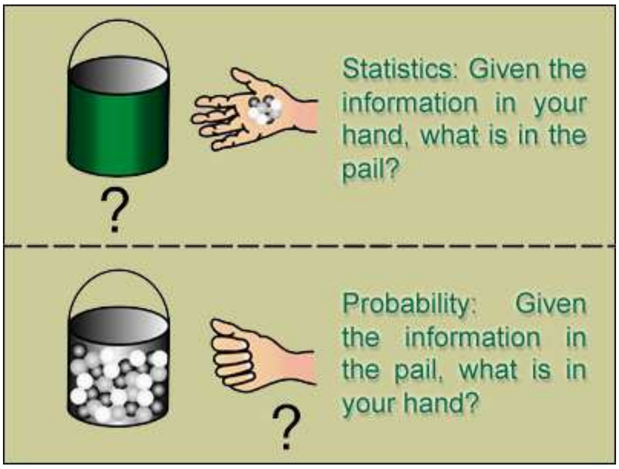


Figure B.1: Diagram showing the difference between statistics and probability

Appendix C

What is Statistics?

Contents

Statistics: The mathematics of the collection, organization, and interpretation of numerical data, especially the analysis of population characteristics by inference from sampling. (***American Heritage Dictionary***®)

C.1 Introduction

Statistics, branch of mathematics that deals with the collection, organization, and analysis of numerical data and with such problems as experiment design and decision making.

C.2 History

Simple forms of statistics have been used since the beginning of civilization, when pictorial representations or other symbols were used to record numbers of people, animals, and inanimate objects on skins, slabs, or sticks of wood and the walls of caves. Before 3000 BC the Babylonians used small clay tablets to record tabulations of agricultural yields and of commodities bartered or sold. The Egyptians analyzed the population and material wealth of their country before beginning to build the pyramids in the 31st century BC. The biblical books of Numbers and 1 Chronicles are primarily statistical works, the former containing two separate censuses of the Israelites and the latter describing the material wealth of various Jewish tribes. Similar numerical records existed in China before 2000

BC. The ancient Greeks held censuses to be used as bases for taxation as early as 594 BC. *See* Census.

The Roman Empire was the first government to gather extensive data about the population, area, and wealth of the territories that it controlled. During the Middle Ages in Europe few comprehensive censuses were made. The Carolingian kings Pepin the Short and Charlemagne ordered surveys of ecclesiastical holdings: Pepin in 758 and Charlemagne in 762. Following the Norman Conquest of England in 1066, William I, king of England, ordered a census to be taken; the information gathered in this census, conducted in 1086, was recorded in the Domesday Book. Registration of deaths and births was begun in England in the early 16th century, and in 1662 the first noteworthy statistical study of population, *Observations on the London Bills of Mortality*, was written. A similar study of mortality made in Breslau, Germany, in 1691 was used by the English astronomer Edmond Halley as a basis for the earliest mortality table. In the 19th century, with the application of the scientific method to all phenomena in the natural and social sciences, investigators recognized the need to reduce information to numerical values to avoid the ambiguity of verbal description.

At present, statistics is a reliable means of describing accurately the values of economic, political, social, psychological, biological, and physical data and serves as a tool to correlate and analyze such data. The work of the statistician is no longer confined to gathering and tabulating data, but is chiefly a process of interpreting the information. The development of the theory of probability increased the scope of statistical applications. Much data can be approximated accurately by certain probability distributions, and the results of probability distributions can be used in analyzing statistical data. Probability can be used to test the reliability of statistical inferences and to indicate the kind and amount of data required for a particular problem.

C.3 Statistical methods

The raw materials of statistics are sets of numbers obtained from enumerations or measurements. In collecting statistical data, adequate precautions must be taken to secure complete and accurate information.

The first problem of the statistician is to determine what and how much data to collect. Actually, the problem of the census taker in obtaining an accurate and complete count of the population, like the problem of the physicist who wishes to count the number of molecule collisions per second in a given volume of gas under given conditions, is to decide the precise nature of the items to be counted. The statistician faces a complex problem when, for example, he or she wishes to take a sample poll or straw vote. It is no simple matter to gauge the size and constitution of the sample that will yield reasonably accurate predictions concerning the action of the total population.

In protracted studies to establish a physical, biological, or social law, the statistician may start with one set of data and gradually modify it in light of experience. For example, in early studies of the growth of populations, future change in size of population was predicted by calculating the excess of births over deaths in any given period. Population statisticians soon recognized that rate of increase ultimately depends on the number of births, regardless of the number of deaths, so they began to calculate future population growth on the basis of the number of births each year per 1000 population. When predictions based on this method yielded inaccurate results, statisticians realized that other limiting factors exist in population growth. Because the number of births possible depends on the number of women rather than the total population, and because women bear children during only part of their total lifetime, the basic datum used to calculate future population size is now the number of live births per 1000 females of childbearing age. The predictive value of this basic datum can be further refined by combining it with other data on the percentage of women who remain childless because of choice or circumstance, sterility, contraception, death before the end of the childbearing period, and other limiting factors. The excess of births over deaths, therefore, is meaningful only as an indication of gross population growth over a definite period in the past; the number of births per 1000 population is meaningful only as an expression of the proportion of increase during a similar period; and the number of live births per 1000 women of childbearing age is meaningful for predicting future size of populations.

C.4 Tabulation and presentation of data

INTERVALS	INTERVAL MIDPOINTS	FREQUENCY	RELATIVE FREQUENCY	CUMULATIVE FREUQENCY	CUMULATIVE RELATIVE FREUQENCY
0-10	5	20	0.017	20	0.017
10-20	15	15	0.012	35	0.029
20-30	25	18	0.015	53	0.044
30-40	35	25	0.021	78	0.065
40-50	45	44	0.037	122	0.102
50-60	55	88	0.073	210	0.175
60-70	65	222	0.185	432	0.360
70-80	75	335	0.279	767	0.639
80-90	85	218	0.182	985	0.821
90-100	95	215	0.179	1200	1.000

The collected data must be arranged, tabulated, and presented to permit ready and meaningful analysis and interpretation. To study and interpret the examination-grade distribution in a class of 30 pupils, for instance, the grades are arranged in ascending order: $30, 35, 43, 52, 61, 65, 65, 65, 68, 70, 72, 72, 73, 75, 75, 76, 77, 78, 78, 80, 83, 85, 88, 88, 90, 91, 96, 97, 100, 100$. This progression shows at a glance that the maximum is 100, the minimum 30, and the range, or difference, between the maximum and minimum is 70.

In a cumulative-frequency graph, such as Fig. 1, the grades are marked on the horizontal axis and double marked on the vertical axis with the cumulative number of the grades on the left and the corresponding percentage of the total number on the right. Each dot represents the accumulated number of students who have attained a particular grade or less. For example, the dot A corresponds to the second 72; reading on the vertical axis, it is evident that there are 12, or 40 percent, of the grades equal to or less than 72.

In analyzing the grades received by 10 sections of 30 pupils each on four examinations, a total of 1200 grades, the amount of data is too large to be exhibited conveniently as in Fig. 1. The statistician separates the data into suitably chosen groups, or intervals. For example, ten intervals might be used to tabulate the 1200 grades, as in column (a) of the accompanying frequency-distribution table; the actual number in an interval, called the frequency of the interval, is entered in column (c). The numbers that define the interval range are called the interval boundaries. It is convenient to choose the interval boundaries so that the interval ranges are equal to each other; the interval midpoints, half the sum of the interval boundaries, are simple numbers, because they are used in many calculations. A grade such as 87 will be tallied in the 80-90 interval; a boundary grade such as 90 may be tallied uniformly throughout the groups in either the lower or upper intervals. The

relative frequency, column (d), is the ratio of the frequency of an interval to the total count; the relative frequency is multiplied by 100 to obtain the percent relative frequency. The cumulative frequency, column (e), represents the number of students receiving grades equal to or less than the range in each succeeding interval; thus, the number of students with grades of 30 or less is obtained by adding the frequencies in column (c) for the first three intervals, which total 53. The cumulative relative frequency, column (f), is the ratio of the cumulative frequency to the total number of grades.
The data of a frequency-distribution table can be presented graphically in a frequency histogram, as in Fig. 2, or a cumulative-frequency polygon, as in Fig. 3. The histogram is a series of rectangles with bases equal to the interval ranges and areas proportional to the frequencies. The polygon in Fig. 3 is drawn by connecting with straight lines the interval midpoints of a cumulative frequency histogram.
Newspapers and other printed media frequently present statistical data pictorially by using different lengths or sizes of various symbols to indicate different values.

C.5 Measures of central tendency

After data have been collected and tabulated, analysis begins with the calculation of a single number, which will summarize or represent all the data. Because data often exhibit a cluster or central point, this number is called a measure of central tendency.
Let $x_1, x_2, \dots, x_n$ be the n tabulated (but ungrouped) numbers of some statistic; the most frequently used measure is the simple arithmetic average, or mean, written $\overline{x}$, which is the sum of the numbers divided by n:

$$\overline{x} = \frac{\sum x}{n}$$

If the x's are grouped into k intervals, with midpoints $m_1, m_2, \dots, m_k$ and frequencies $f_1, f_2, \dots, f_k$, respectively, the simple arithmetic average is given by

$$\frac{\sum f_1 m_1}{\sum f_1} \qquad \text{with } i = 1, 2, \dots, k.$$

The median and the mode are two other measures of central tendency. Let the x's be arranged in numerical order; if n is odd, the median is the middle x; if n is even, the median is the average of the two middle x's. The mode is the x that occurs most frequently. If two or more distinct x's occur with equal frequencies, but none with greater frequency, the set of x's may be said not to have a mode or to be bimodal, with modes at the two most frequent x's, or trimodal, with modes at the three most frequent x's.

C.6 Measures of variability

The investigator frequently is concerned with the variability of the distribution, that is, whether the measurements are clustered tightly around the mean or spread over the range. One measure of this variability is the difference between two percentiles, usually the 25th and the 75th percentiles. The pth percentile is a number such that p percent of the measurements are less than or equal to it; in particular, the 25th and the 75th percentiles are called the lower and upper quartiles, respectively. The pth percentile is readily found from the cumulative-frequency graph, (Fig. 1) by running a horizontal line through the p percent mark on the vertical axis on the graph, then a vertical line from this point on

the graph to the horizontal axis; the abscissa of the intersection is the value of the pth percentile.
The standard deviation is a measure of variability that is more convenient than percentile differences for further investigation and analysis of statistical data. The standard deviation of a set of measurements $x_1, x_2, \ldots, x_n$, with the mean $\overline{x}$ is defined as the square root of the mean of the squares of the deviations; it is usually designated by the Greek letter sigma (σ). In symbols

$$\sigma = \sqrt{\frac{1}{n}[(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + \cdots + (x_n - \overline{x})^2]} = \sqrt{\frac{1}{n}\sum(x_i - \overline{x})^2}$$

The square, σ^2, of the standard deviation is called the variance. If the standard deviation is small, the measurements are tightly clustered around the mean; if it is large, they are widely scattered.

C.7 Correlation

When two social, physical, or biological phenomena increase or decrease proportionately and simultaneously because of identical external factors, the phenomena are correlated positively; under the same conditions, if one increases in the same proportion that the other decreases, the two phenomena are negatively correlated. Investigators calculate the degree of correlation by applying a coefficient of correlation to data concerning the two phenomena. The most common correlation coefficient is expressed as

$$\frac{\sum \left(\frac{x}{\sigma^x} \cdot \frac{y}{\sigma^y}\right)}{N}$$

in which x is the deviation of one variable from its mean, y is the deviation of the other variable from its mean, and N is the total number of cases in the series. A perfect positive correlation between the two variables results in a coefficient of +1, a perfect negative correlation in a coefficient of -1, and a total absence of correlation in a coefficient of 0. Intermediate values between +1 and 0 or -1 are interpreted by degree of correlation. Thus, .89 indicates high positive correlation, -.76 high negative correlation, and .13 low positive correlation.

C.8 Mathematical models

A mathematical model is a mathematical idealization in the form of a system, proposition, formula, or equation of a physical, biological, or social phenomenon. Thus, a theoretical, perfectly balanced die that can be tossed in a purely random fashion is a mathematical model for an actual physical die. The probability that in n throws of a mathematical die a throw of 6 will occur k times is

$$p(k) = \binom{n}{k}\left(\frac{1}{6}\right)^n (\frac{5}{6})^{n-k}$$

in which $\binom{n}{k}$ is the symbol for the binomial coefficient

$$\frac{n(n-1)\cdots(n-k+1)}{1 \cdot 2 \cdot \cdots \cdot k} \cdot \left(\binom{n}{0} = 1\right)$$

The statistician confronted with a real physical die will devise an experiment, such as tossing the die n times repeatedly, for a total of Nn tosses, and then determine from the observed throws the likelihood that the die is balanced and that it was thrown in a random way.

In a related but more involved example of a mathematical model, many sets of measurements have been found to have the same type of frequency distribution. For example, let $x_1, x_2, \ldots, x_N$ be the number of 6's cast in the N respective runs of n tosses of a die and assume N to be moderately large. Let $y_1, y_2, \ldots, y_N$ be the weights, correct to the nearest 1/100 g, of N lima beans chosen haphazardly from a 100-kg bag of lima beans. Let $z_1, z_2, \ldots, z_N$ be the barometric pressures recorded to the nearest 1/1000 cm by N students in succession, reading the same barometer. It will be observed that the x's, y's, and z's have amazingly similar frequency patterns. The statistician adopts a model that is a mathematical prototype or idealization of all these patterns or distributions. One form of the mathematical model is an equation for the frequency distribution, in which N is assumed to be infinite:

$$y = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

in which e (approximately 2.7) is the base for natural logarithms (see Logarithm). The graph of this equation (Fig. 4) is the bell-shaped curve called the normal, or Gaussian, probability curve. If a variate x is normally distributed, the probability that its value lies between a and b is given by

$$\frac{1}{\sqrt{2\pi}} \int_d^b e^{(-(x^2)/2)} dx$$

The mean of the x's is 0, and the standard deviation is 1. In practice, if N is large, the error is exceedingly small.

C.9 Tests of reliability

The statistician is often called upon to decide whether an assumed hypothesis for some phenomenon is valid or not. The assumed hypothesis leads to a mathematical model; the model, in turn, yields certain predicted or expected values, for example, 10, 15, 25. The corresponding actually observed values are 12, 16, 21. To determine whether the hypothesis is to be kept or rejected, these deviations must be judged as normal fluctuations caused by sampling techniques or as significant discrepancies. Statisticians have devised several tests for the significance or reliability of data. One is the chi-square (χ^2) test. The deviations (observed values minus expected values) are squared, divided by the expected values, and summed:

$$x^2 = \frac{(12-10)^2}{10} + \frac{(16-15)^2}{15} + \frac{(21-25)^2}{25} = 1.11$$

The value of χ^2 is then compared with values in a statistical table to determine the significance of the deviations.

C.10 Higher statistics

The statistical methods described above are the simpler, more commonly used methods in the physical, biological, and social sciences. More advanced methods, often involving advanced mathematics, are used in further statistical studies, such as sampling theory, inference and estimation theory, and design of experiments.

Contributed By: James Singer

C.11 Difference between statistics and probability

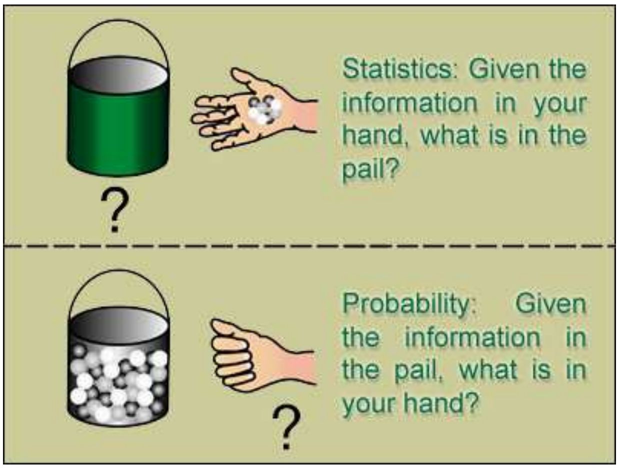


Figure C.1: Diagram showing the difference between statistics and probability

(Image by MIT OpenCourseWare. Based on Gilbert, Norma. Statistics. W.B. Saunders Co., 1976.)

Appendix D

COMMON UNIVARIATE DISTRIBUTIONS

Contents

D.1 Common discrete distributions

1. Bernoulli$(p) \equiv$ Binomial$(1, p)$
   (a) $P(X = x|p) = p^x(1-p)^{1-x}, \quad x = 0, 1;\ 0 < p < 1$
   (b) $E[X] = p$, $\text{Var}(X) = p(1-p)$
   (c) $M_X(t) = (1-p) + pe^t; \quad E[X^k] = p$
   (d) $Y = X_1 + \cdots + X_n \sim$ Binomial(n, p) if $X_1, \ldots, X_n \overset{\text{iid}}{\sim}$ Bernoulli(p).
2. Binomial(n, p)
   (a) $P(X = x|n, p) = \binom{n}{x} p^x(1-p)^{n-x}, \quad x = 0, 1, \ldots, n;\ 0 < p < 1$
   (b) $E[X] = np$, $\text{Var}(X) = np(1-p)$
   (c) $M_X(t) = [pe^t + (1-p)]^n; \quad E[X^k] = npE[(Y+1)^{k-1}]$ where $Y \sim$ Binomial$(n-1, p)$.
   (d) Let $X \sim$ Binomial(n, p).
      i. If $n = 1$, then $X \sim$ Bernoulli(p).
      ii. As $n \to \infty$ if $np \geq 5$ and $n(1-p) \geq 5$, then $X \approx N(np, np(1-p))$.
      iii. As $n \to \infty$ if $p \geq .1$ and $np < 10$, then $X \approx$ Poisson(np).
      iv. Let $X_1, \ldots, X_k$ be independent, binomial random variables with parameters n_i and p, respectively. The random variable $Y = X_1 + \cdots + X_k$ has a Binomial$(\sum_{i=1}^{k} n_i, p)$.

3. Discrete uniform(m, n)
   (a) $P(X = x|m, n) = \frac{1}{n-m+1}, \quad x = m, m+1, \ldots, n$
   (b) $E[X] = \frac{m+n}{2}$, $\text{Var}(X) = \frac{(n-m)(n-m+2)}{12}$
   (c) $M_X(t) = \frac{1}{n-m+1} \sum_{k=m}^{n} e^{kt}$
4. Geometric$(p) \equiv$ Negative binomial$(1, p)$
   (a) $P(X = x|p) = p(1-p)^{x-1}, \quad x = 1, 2, \ldots;\ 0 < p < 1$
   (b) $E[X] = \frac{1}{p}$, $\text{Var}(X) = \frac{1-p}{p^2}$
   (c) $M_X(t) = \frac{pe^t}{1-(1-p)e^t}$, $t < -\log(1-p)$; $E[X^k] = \frac{1}{p}E[(Y-1)^{k-1}]$ where $Y \sim$ Negative binomial$(2, p)$.
   (d) Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim}$ Geometric(p).
      i. $Y = X_1 + \cdots + X_n \sim$ Negative Binomial(n, p)
      ii. $Y = \min(X_1, \ldots, X_n) \sim$ Geometric(p)
5. Hypergeometric(N, m, n)
   (a) $P(X = x|N, m, n) = \frac{\binom{m}{x}\binom{N-m}{n-x}}{\binom{N}{n}}, \quad x = 0, 1, \ldots, n;\ m - (N-n) \le x \le m$
   (b) $E[X] = \frac{mn}{N} = np$, $\text{Var}(X) = \frac{mn}{N}\frac{(N-m)(N-n)}{N(N-1)} = \frac{N-n}{N-1}np(1-p)$ where $p = \frac{m}{N}$.
   (c) $M_X(t)$ is messy! $E[X^k] = \frac{mn}{N}E[(Y+1)^{k-1}]$ where $Y \sim$ Hypergeometric$(N-1, m-1, n-1)$.
   (d) Let $X \sim$ Hypergeometric(N, m, n).
      i. As $N \to \infty$ if $n/N < 0.1$, then $X \approx$ Binomial$(n, m/N)$.
      ii. As n, m, and N all tend to infinity, if m/N is small, then $X \approx$ Poisson(nm/N).
6. Multinomial$(n, k, p_1, \ldots, p_k)$
   (a) $P(\boldsymbol{X} = (x_1, \ldots, x_k)|n, k, p_1, \ldots, p_k) = n! \prod_{i=1}^{k} \frac{p_i^{x_i}}{x_i!}, \quad \sum_{i=1}^{k} x_i = n;\ 0 \le p \le 1$
   (b) $E[X_i] = np_i$, $\text{Var}(X_i) = np_i(1-p_i)$, $\text{Cov}(X_i, X_j) = -np_ip_j, i \ne j$
   (c) $M_{\boldsymbol{X}}(\boldsymbol{t}) = (p_1e^{t_1} + \cdots + p_ke^{t_k})^n$
   (d) Let $\boldsymbol{X} \sim$ Multinomial$(n, k, p_1, \ldots, p_k)$.
      i. $X_i \sim$ Binomial(n, p_i)
      ii. If $k = 2$ and $p_1 = p$, then $X_1 \sim$ Binomial(n, p)
7. Negative binomial(r, p)
   (a) $P(X = x|r, p) = \binom{x-1}{r-1}p^r(1-p)^{x-r}, \quad x = r, r+1, \ldots;\ 0 \le p \le 1$
   (b) $E[X] = \frac{r}{p}$, $\text{Var}(X) = \frac{r(1-p)}{p^2}$
   (c) $M_X(t) = \left(\frac{pe^t}{1-(1-p)e^t}\right)^r$, $t < -\log(1-p)$; $E[X^k] = \frac{r}{p}E[(Y-1)^{k-1}]$ where $Y \sim$ Negative binomial$(r+1, p)$.
   (d) Let $X \sim$ Negative binomial(r, p).
      i. If $r = 1$, then $X \sim$ Geometric(p).
      ii. As $r \to \infty$ and $p \to 1$ with $r(1-p)$ held constant, $X \approx$ Poisson$(r(1-p))$.

iii. Let $X_1, \ldots, X_k$ be independent, negative binomial random variables with parameters r_i and p, respectively. The random variable $Y = X_1 + \cdots + X_k$ has an negative binomial($\sum_{i=1}^{k} r_i, p$).

8. Poisson(λ)

(a) $P(X = x|\lambda) = \frac{e^{-\lambda}\lambda^x}{x!}, \quad x = 0, 1, \ldots;\ 0 \leq \lambda < \infty$

(b) $E[X] = \lambda,\ \mathrm{Var}(X) = \lambda$

(c) $M_X(t) = e^{\lambda(e^t-1)};\quad E[X^k] = \lambda E[(X+1)^{k-1}]$

(d) Let $X \sim$ Poisson(λ).

i. As $\lambda \to \infty$, $X \approx$ N(λ, λ).

ii. Let $X_1, \ldots, X_k$ be independent, Poisson random variables with parameters λ_i, respectively. The random variable $Y = X_1 + \cdots + X_k$ has a Poisson($\sum_{i=1}^{k} \lambda_i$).

D.2 Common continuous distributions

1. Beta(α, β)

(a) $f(x|\alpha, \beta) = \frac{1}{B(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1}, \quad 0 \leq x \leq 1;\ \alpha > 0,\ \beta > 0$

(b) $E[X] = \frac{\alpha}{\alpha+\beta},\ \mathrm{Var}(X) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$

(c) $M_X(t) = 1 + \sum_{k=1}^{\infty} \left(\prod_{\gamma=0}^{k-1} \frac{\alpha+\gamma}{\alpha+\beta+\gamma} \right) \frac{t^k}{k!}$

(d) Related distributions
Let $X \sim$ Beta(α, β).

i. If $\alpha = \beta = 1/2$,then X is an arcsin random variable.

ii. If $\alpha = \beta = 1$, then X is a uniform random variable with parameters $a = 0$ and $b = 1$.

iii. If $\beta = 1$, then X is pa power function random variable with parameters $b = 1$ and $c = \alpha$.

iv. As α and β tend to infinity such that α/β is constant, X tends to a standard normal random variable.

2. Cauchy(θ, σ) Cauchy$(0, 1) \equiv t_1$

(a) $f(x|\theta, \sigma) = \frac{1}{\pi\sigma} \frac{1}{1+\left(\frac{x-\theta}{\sigma}\right)^2}, \quad -\infty < x < \infty;\ -\infty < \theta < \infty,\ \sigma > 0$

(b) $E[X] =$ DO NOT EXIST, $\mathrm{Var}(X) =$ DO NOT EXIST

(c) $M_X(t) =$ DO NOT EXIST

(d) Related distributions
Let $X \sim$ Cauchy(a, b).

i. If $a = 0.b = 1$ then X is a standard Cauchy random variable.

ii. The random variable $1/X$ is also a Cauchy random variable with parameters $a/(a^2 + b^2)$ and $b/(a^2 + b^2)$.

iii. Let X_i (for $i = 1, 2, \ldots, n$) be independent, Cauchy random variables with parameters a_i and b_i, respectively. The random variable $Y = X_1 + X_2 + \cdots + X_n$ has a Cauchy distribution with parameters $a = a_1 + a_2 + \cdots + a_n$ and $b = b_1 + b_2 + \cdots + b_n$.

3. Chi squared χ_p^2

   (a) $f(x|p) = \frac{1}{\Gamma(\frac{p}{2})2^{p/2}} x^{(p/2)-1} e^{-x/2}, \quad 0 \le x < \infty;\ p = 1, 2, \dots$

   (b) $E[X] = p,\ \mathrm{Var}(X) = 2p$

   (c) $M_X(t) = (\frac{1}{1-2t})^{p/2},\ t < \frac{1}{2}$

   (d) Related distributions

      i. If X is a chi-square random variable with $v = 2$, then X is an exponential random variable with $\lambda = 1/2$.

      ii. If X_1 and X_2 are independent chi-square random variables with parameters ν_1 and ν_2, then random variable $(X_1/\nu_1)/(X_2/\nu_2)$ has an F distribution with ν_1 and ν_2 degree of freedom.

      iii. If X_1 and X_2 are independent chi-square random variables with parameters $\nu_1 = \nu_2 \nu$, the random variable

$$Y = \frac{\sqrt{\nu}}{2} \frac{X_1 - X_2}{\sqrt{X_1 X_2}}$$

      has a t distribution with ν degrees of freedom.

      iv. Let X_i (for $i = 1, 2, \dots, n$) be independent chi-square random variables ν_i. The random variable $Y = X_1 + X_2 + \cdots + X_n$ has a chi-square distribution with $\nu = \nu_1 + \nu_2 + \cdots + \nu_n$ degrees of freedom.

      v. If X is a chi-square random variable with ν degrees of freedom, the random variable $\sqrt{X}$ has a **chi distribution** with parameter ν.
      Properties of a chi random variable:

pdf	$f(x)$	$=$	$\frac{x^{n-1}e^{-x^2/2}}{2^{(n/2)-1}\Gamma(n/2)}$
mean	μ	$=$	$\frac{\Gamma(\frac{n+1}{2})}{\Gamma(\frac{n}{2})}$
variance	σ^2	$=$	$\frac{\Gamma(\frac{n+2}{2})}{\Gamma(\frac{n}{2})} - \left[\frac{\Gamma(\frac{n+1}{2})}{\Gamma(\frac{n}{2})}\right]^2$

      where $\Gamma(x)$ is the gamma function.
      If X is a chi random variable with parameter $n = 2$, then X is a Rayleigh random variable with $\sigma = 1$.

4. Double exponential(μ, σ)

   (a) $f(x|\mu, \sigma) = \frac{1}{2\sigma} e^{-|x-\mu|/\sigma}, \quad -\infty < x < \infty;\ -\infty < \mu < \infty,\ \sigma > 0$

   (b) $E[X] = \mu,\ \mathrm{Var}(X) = 2\sigma^2$

   (c) $M_X(t) = \frac{e^{\mu t}}{1-(\sigma t)^2},\ |t| < \frac{1}{\sigma}$

   (d) Related distributions

      i. Let X be a Laplace random variable with parameters α and β. The random variable $Y = |x - \alpha|$ has an exponential distribution with parameter $\lambda = \beta$. The random variable $W = |X - \alpha|/\beta$ has an exponential distribution with parameter $\lambda = 1$.

ii. Let X_1 and X_2 be independent Laplace random variables with parameters $\alpha = 0$, and β_1 and β_2, respectively. The random variable $Y = |X_1/X_2|$ has an F distribution with parameters $\nu_1 = \nu_2 = 2$.

5. Exponential(β) $\equiv$ Gamma($1, \beta$)

(a) $f(x|\beta) = \frac{1}{\beta} e^{-x/\beta}, \quad 0 \leq x < \infty;\ \beta > 0$

(b) $E[X] = \beta,\ \mathrm{Var}(X) = \beta^2$

(c) $M_X(t) = \frac{1}{1-\beta t},\ t < \frac{1}{\beta}$

(d) Related distributions
Let $X \sim$Exp(λ).

i. If $\lambda = 1/2$, then X is a chi-square random variable with $\nu = 2$.

ii. The random variable $\sqrt{X}$ has Rayleigh distribution with parameter $\sigma = \sqrt{1/(2\lambda)}$.

iii. The random variable $Y = X^{1/\alpha}$ has a weibull distribution with parameters α and $\lambda^{-1/\alpha}$.

iv. The random variable $Y = e^{-X}$ has a power function distribution with parameters $b = 1$ and $c = \lambda$.

v. The random variable $Y = ae^X$ has a Pareto distribution with parameters a and $\theta = \lambda$.

vi. The random variable $Y = \alpha - \ln X$ has an extreme-value distribution with parameters α and $\beta = 1/\lambda$.

vii. Let $X_1, X_2, \ldots, X_n$ be independent exponential random variables each with parameter λ.

(1) The random variable $Y = \min(X_1, X_2, \ldots, X_n)$ has an exponential distribution with parameter $n\lambda$.

(2) The random variable $Y = X_1 + X_2 + \cdots + X_n$ has an Erlang distribution with parameter $\beta = 1/\lambda$ and n.

viii. Let X_1 and X_2 be independent exponential random variables each with parameter λ. The random variable $Y = X_1 - X_2$ has a Laplace distribution with parameters 0 and $1/\lambda$.

ix. Let X be an exponential random variable with parameter $\lambda = 1$. The random variable $Y == -\ln[e^{-X}/(1+e^{-X})]$ has a (standard) logistic distribution with parameters $\alpha = 0$ and $\beta = 1$.

x. Let X_1 and X_2 be independent exponential random variables with parameter $\lambda = 1$.

(1) The random variable $Y = X_1/(X_1 + X_2)$ has a (standard) uniform distribution with parameters $a = 0$ and $b = 1$.

(2) The random variable $W = -\ln(X_1/X_2)$ has a (standard) logistic distribution with parameters $\alpha = 0$ and $\beta = 1$.

6. $F_{v_1,v_2} \equiv \frac{\chi^2_{v_1}/v_1}{\chi^2_{v_2}/v_2}$

(a) $f(x|v_1, v_2) = \frac{\Gamma\left(\frac{v_1+v_2}{2}\right)}{\Gamma\left(\frac{v_1}{2}\right)\Gamma\left(\frac{v_2}{2}\right)} \left(\frac{v_1}{v_2}\right)^{v_1/2} \frac{x^{(v_1-2)/2}}{\left(1+ \frac{v_1}{v_2} x\right)^{(v_1+v_2)/2}}, \quad 0 \leq x < \infty;\ v_1, v_2 = 1, \ldots$

(b) $E[X] = \frac{v_2}{v_2-2},\ v_2 > 2,\ \mathrm{Var}(X) = 2\left(\frac{v_2}{v_2-2}\right)^2 \frac{(v_1+v_2-2)}{v_1(v_2-4)},\ v_2 > 4$

(c) $M_X(t) =$ DO NOT EXIST

(d) Related distributions

i. The standard extreme-value distribution has $\alpha = 0$ and $\beta = 1$.

ii. If X is an extreme-value random variable with parameters α and β, then the random variable $Y = (X-\alpha)/\beta$ has a (standard) extreme-value distribution with parameter 0 and 1.

iii. If X is a (standard) extreme-value random variable with parameters $\alpha = 0$ and $\beta = 1$, then the random variable $Y = e^{(-e^{-X/c})}$ has a power function distribution with parameters $b = 0$ and c.

iv. If X is a extreme-value random variable with parameters $\alpha = 0$ and $\beta = 1$, then the random variable $Y = a[1 - e^{(-e^{-X})}]^{1/\theta}$ has a Pareto distribution with parameters a and θ.

7. Gamma(α, β)

(a) $f(x|\alpha, \beta) = \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-x/\beta}, \quad 0 \leq x < \infty;\ \alpha, \beta > 0$

(b) $E[X] = \alpha\beta,\ \mathrm{Var}(X) = \alpha\beta^2$

(c) $M_X(t) = \left(\frac{1}{1-\beta t}\right)^\alpha,\ t < \frac{1}{\beta}$

(d) Related distributions
Let $X \sim \Gamma(\alpha, \beta)$

i. The random variable X has a standard gamma distribution if $\alpha = 1$.

ii. If $\alpha = 1$ and $\beta = 1/\lambda$, then X has an exponential distribution with parameter λ.

iii. If $\alpha = \nu/2$ and $\beta = 2$, then X has a chi-square distribution with ν degrees of freedom.

iv. If $\alpha = n$ is an inter, then X has an Erlang distribution with parameters β and n.

v. If $\alpha = \nu/2$ and $\beta = 1$, then the random variable $Y = 2X$ has a chi-square distribution with ν degrees of freedom.

vi. As $\alpha \to \infty$, X tends to a normal distribution with parameters $\mu = \alpha\beta$ and $\sigma^2 = \alpha\beta^2$

vii. Suppose X_1 is a gamma random variable with parameter $\alpha = \alpha_1, \beta$, X_2 is a gamma random variable with parameter $\alpha = \alpha_2, \beta$, and X_1 and X_2 are independent. The random variable $Y = X_1/(X_1 + X_2)$ has a beta distribution with parameters α_1 and α_2.

viii. Let $X_1, X_2, \ldots, X_n$ be independent gamma random variables with parameters α and β_i for $i = 1, 2, \ldots, n$. The random variable $Y = X_1 + X_2 + \cdots + X_n$ has a gamma distribution with parameters α and $\beta = \beta_1 + \beta_2 + \cdots + \beta_n$

8. Logistic(μ, β)

(a) $f(x|\mu, \beta) = \frac{1}{\beta} \frac{e^{-(x-\mu)/\beta}}{[1+e^{-(x-\mu)/\beta}]^2}, \quad -\infty < x < \infty;\ -\infty < \mu < \infty,\ \beta > 0$

(b) $E[X] = \mu,\ \mathrm{Var}(X) = \frac{\pi^2\beta^2}{3}$

(c) $M_X(t) = e^{\mu t}\Gamma(1 - \beta t)\Gamma(1 + \beta t),\ |t| < \frac{1}{\beta}$

9. Lognormal$(\mu, \sigma^2) \equiv e^{N(\mu, \sigma^2)}$

(a) $f(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma}\frac{e^{-(\log x-\mu)^2/(2\sigma^2)}}{x}, \quad 0 \leq x < \infty;\ -\infty < \mu < \infty,\ \sigma > 0$

(b) $E[X] = e^{\mu+(\sigma^2/2)},\ \mathrm{Var}(X) = e^{2(\mu+\sigma^2)} - e^{2\mu+\sigma^2}$

(c) $M_X(t) =$ DO NOT EXIST

(d) Related distributions

i. If X is a lognormal random variable with parameters μ and σ, then the random variable $Y = \ln X$ has a normal distribution with mean μ and variance σ^2.

ii. If X is a lognormal random variable with parameters μ and σ and a and b are constants, then the random variable $Y = e^a X^b$ has a lognormal distribution with parameters $a + b\mu$ and $b\sigma$.

iii. Let X_1 and X_2 be independent lognormal random variables with parameters μ_1, σ_1 and μ_2, σ_2, respectively. The random variable $Y = X_1/X_2$ has a lognormal distribution with parameters $\mu_1 - \mu_2$ and $\sigma_1 + \sigma_2$.

iv. Let $X_1, X_2, \ldots, X_n$ be independent lognormal random variables with parameters μ and σ. The random variable $Y = \sqrt[n]{X_1 \cdots X_n}$ has a lognormal distribution with parameters μ and σ/n.

10. Normal $N(\mu,\sigma^2)$

(a) $f(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad -\infty < x < \infty;\ -\infty < \mu < \infty,\ \sigma > 0$

(b) $E[X] = \mu,\ \mathrm{Var}(X) = \sigma^2$

(c) $M_X(t) = e^{\mu t+(\sigma^2/2)t^2}$

(d) Related distributions

i. The random variable X has a standard normal distribution if $\mu = 0$ and $\sigma = 1$.

ii. If $X \sim N(\mu,\sigma^2)$, then $Y = (X-\mu)/\sigma \sim N(0,1)$.

iii. If $X \sim N(\mu,\sigma^2)$, then $Y = e^X \sim$ lognormal(μ,σ^2).

iv. If $X \sim N(0,1)$, then $Y = e^{\mu+\sigma X} \sim$ lognormal(μ,σ^2).

v. If $X \sim N(\mu,\sigma^2)$ and a and b are constants, then $Y = ax + b \sim N(a + b\mu, b^2\sigma^2)$.

vi. If X and Y are iid $N(0,1)$, then $Y = X_1/X_2 \sim$ Cauchy$(0,1)$.

vii. If X and Y are iid $N(0,\sigma^2)$, then $Y = \sqrt{X_1^2 + X_2^2} \sim$ Rayleigh(σ^2).

viii. If X_i be independent $N(\mu_i,\sigma_i^2)$ and c_i be any constants, then $Y = \sum_{i=1}^n c_i X_i \sim N(\sum_{i=1}^n c_i\mu_i, \sum_{i=1}^n c_i^2\sigma_i^2)$.

ix. If $X_i \overset{\text{iid}}{\sim} N(0,1)$, then $Y = \sum_{i=1}^n X_i^2 \sim \chi_n^2$.

x. If X_i be independent $N(\mu_i,1)$, then $Y = \sum_{i=1}^n X_i^2 \sim \chi_n^2(\sum_{i=1}^n \mu_i^2)$.

11. Pareto(α,β)

(a) $f(x|\alpha,\beta) = \frac{\beta\alpha^\beta}{x^{\beta+1}}, \quad \alpha < x < \infty;\ \alpha > 0,\ \beta > 0$

(b) $E[X] = \frac{\beta\alpha}{\beta-1},\ \beta > 1,\ \mathrm{Var}(X) = \frac{\beta\alpha^2}{(\beta-1)^2(\beta-2)},\ \beta > 2$

(c) $M_X(t) =$ DO NOT EXIST

(d) Related distributions

i. Let X be a Pareto random variable with parameters a and θ.

(1) The random variable $Y = \ln(X/a)$ has an exponential distribution with parameter $\lambda = 1/\theta$.

(2) The random variable $Y = 1/X$ has a power function distribution with parameters $1/a$ and θ.

(3) The random variable $Y = -\ln[(X/a)^\theta - 1]$ has a logistic distribution with parameters $\alpha = 0$ and $\beta = 1$.

ii. Let X_i (for $i = 1, 2, \ldots, n$) be independent Pareto random variables with parameters a and θ. The random variable $Y = 2a\sum_{i=1}^n \ln(X_i/\theta)$ has a chi-square distribution with $\nu = 2n$.

12. Student-t $\quad t_v \equiv \frac{N(0,1)}{\sqrt{\chi_v^2/v}}$

(a) $f(x|v) = \frac{\Gamma\left(\frac{v+1}{2}\right)}{\Gamma\left(\frac{v}{2}\right)} \frac{1}{\sqrt{v\pi}} \frac{1}{1+\frac{x^2}{v}}^{(v+1)/2}, \quad -\infty < x < \infty,\ v = 1, \ldots$

(b) $E[X] = 0,\ v > 1,\ \text{Var}(X) = \frac{v}{v-2},\ v > 2$

(c) $M_X(t) =$ DO NOT EXIST

(d) Related distributions

i. If X is a t random variable with parameter ν, then the random variable $Y = X^2$ has an F distribution with 1 and ν degree of freedom.

ii. If X is a t random variable with parameter $\nu = 1$, then X has a Cauchy distribution with parameters $a = 0$ and $b = 1$.

iii. If X is a t random variable with parameter ν, as ν tends to infinity X tends to a standard normal distribution. The approximation is reasonable for $\nu \geq 30$.

13. Uniform $\quad U(a, b) \equiv a + (b - a)$ Beta$(1, 1)$

(a) $f(x|a, b) = \frac{1}{b-a}, \quad a \leq x \leq b$

(b) $E[X] = \frac{a+b}{2},\ \text{Var}(X) = \frac{(b-a)^2}{12}$

(c) $M_X(t) = \frac{e^{bt} - e^{at}}{(b-a)t}$

(d) Related distributions

i. The random variable X has a standard uniform distribution if $a = 0$ and $b = 1$.

ii. If X is a uniform random variable with parameters $a = 0$ and $b = 1$, the random variable $Y = -(\ln X)/\lambda$ has an exponential distribution with parameter λ.

iii. Let X_1 and X_2 be independent uniform random variables with parameters $a = 0$ and $b = 1$. The random variable $Y = (X_1 + X_2)/2$ has a triangular distribution with parameters 0 and 1.

iv. If X is a uniform random variable with parameters $a = -\pi/2$ and $b = \pi/2$, then the random variable $Y = \tan X$ has a Cauchy distribution with parameters $a = 0$ and $b = 1$.

14. Weibull$(\gamma, \beta) \equiv$ Exponential$(\beta)^{1/\gamma}$

(a) $f(x|\gamma, \beta) = \frac{\gamma}{\beta} x^{\gamma-1} e^{-x^\gamma/\beta}, \quad 0 \leq x < \infty;\ \gamma > 0,\ \beta > 0$

(b) $E[X] = \beta^{\frac{1}{\gamma}} \Gamma\left(1 + \frac{1}{\gamma}\right),\ \text{Var}(X) = \beta^{\frac{2}{\gamma}}\left[\Gamma\left(1 + \frac{2}{\gamma}\right) - \Gamma^2\left(1 + \frac{1}{\gamma}\right)\right]$

(c) $M_X(t) =$ DO NOT EXIST

(d) Related distributions
Let $X \sim \text{Weibull}(\alpha, \beta)$

i. The random variable X has a standard Weibull distribution α and β.
ii. If $\alpha = 1$ then X has an exponential distribution with parameter $\lambda = 1/\beta$.
iii. The random variable $Y = X^\alpha$ has an exponential distribution with parameter $\lambda = \beta$.
iv. If $\alpha = 2$ then X has a Rayleigh distribution with parameter $\sigma = \beta/\sqrt{2}$.
v. The random variable $Y = -\alpha \ln(X/\beta)$ has a (standard) extreme-value distribution with parameters $\alpha = 0$ and $\beta = 1$.

D.3 Distributions related to normal distribution: chi-square, t, F distribution

D.3.1 chi-square distribution

1. Let Z be a standard normal random variable, then Z^2 has a chi-square distribution with 1 degree of freedom.

2. Let $Z_1, Z_2, \ldots, Z_n$ be independent standard normal random variables. The random variable $Y = \sum_{i=1}^{n} Z_i^2$ has a chi–square distribution with n degrees of freedom.

3. Let $X_1, X_2, \ldots, X_n$ be independent random variables such that X_i has a chi–square distribution with ν_i degrees of freedom. The random variable $Y = \sum_{i=1}^{n} X_i$ has a chi–square distribution with $\nu = \nu_1 + \nu_2 + \cdots + \nu_n$ degrees of freedom.

4. Let U have a chi–square distribution with ν_1 degrees of freedom, U and V be independent, and $U + V$ have a chi–square distribution with $\nu > \nu_1$ degrees of freedom. The random variable V has a chi–sqaure distribution with $\nu - \nu_1$ degrees of freedom.

5. Let $X_1, X_2, \ldots, X_n$ be a random sample from a normal population with mean μ and variance σ_2. Then

   (a) The sample mean, $\bar{X}$, and the sample variance, S^2, are independent.

   (b) The random variable $\frac{(n-1)S^2}{\sigma^2}$ has a chi–square distribution with $n - 1$ degrees of freedom.

D.3.2 t distribution

1. Let Z have a standard normal distribution, X have a chi–square distribution with ν degrees of freedom, and X and Z be independent. The random variable

$$T = \frac{Z}{\sqrt{X/\nu}} \tag{D.1}$$

has a t distribution with ν degrees of freedom.

2. Let $X_1, X_2, \ldots, X_n$ be a random sample from a normal population with mean μ and variance σ^2. The random variable

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}} \tag{D.2}$$

has a t distribution with $n-1$ degrees of freedom.

D.3.3 F distribution

1. Let U have a chi–square distribution with ν_1 degrees of freedom, V have a chi–square distribution with ν_2 degrees of freedom, and U and V be independent. The random variable

$$F = \frac{U/\nu_1}{V/\nu_2} \tag{D.3}$$

has an F distribution with ν_1 and ν_2 degrees of freedom.

2. Let $X_1, X_2, \ldots, X_m$ and $Y_1, Y_2, \ldots, Y_n$ be random samples from normal populations with variances σ_X^2 and σ_Y^2, respectively. The random variable

$$F = \frac{S_X^2/\sigma_X^2}{S_Y^2/\sigma_Y^2} \tag{D.4}$$

has an F distribution with $m-1$ and $n-1$ degrees of freedom.

3. Let F_{α,ν_1,ν_2} be a critical value for the F distribution defined by $P\{F \geq F_{\alpha,\nu_1,\nu_2}\} = \alpha$. Then $F_{1-\alpha,\nu_1,\nu_2} = 1/F_{\alpha,\nu_2,\nu_1}$.

D.4 Relationship among common univariate distributions

Figure (D.1) presents some of the relationships among common univariate distributions. The first line of each box is the name of the distribution and the second line lists the parameters that characterize the distribution. The random variable X is used to represent each distribution. The three types of relationships presented in the figure are transformations (independent random variables are assumed) and special cases (both indicated with a solid arrow), and limiting distributions (indicated with a dashed arrow).

1. If X_1 has a standard normal distribution, X_2 has a chi-square distribution with ν degrees of freedom, and X_1 and X_2 are independent, then the random variable

$$Y = \frac{X_1}{\sqrt{X_2/\nu}} \tag{D.5}$$

has a t distribution with ν degrees of freedom.

2. Let $X_1, X_2, \ldots, X_n$ be independent normal random variables with parameters μ and σ, and define

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \quad \text{and} \quad S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2. \tag{D.6}$$

(a) The random variable $Y = nS^2/\sigma^2$ has a chi-square distribution with $n-1$ degrees of freedom.

(b) The random variable

$$W = \frac{\bar{X} - \mu}{S/\sqrt{n}} \tag{D.7}$$

has a t distribution with $n-1$ degrees of freedom.

3. Let $X_1, X_2, \ldots, X_n$ be independent normal random variables with parameters μ and σ, and define

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \quad \text{and} \quad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2. \tag{D.8}$$

The random variable

$$Y = \frac{\bar{X} - \mu}{S/\sqrt{n}} \tag{D.9}$$

has a t distribution with $n-1$ degrees of freedom.

4. Let $X_1, X_2, \ldots, X_{n_1}$ be independent normal random variables with parameters μ_1 and σ, and $Y_1, Y_2, \ldots, Y_{n_2}$ be independent normal random variables with parameters μ_2 and σ. Define

$$\begin{aligned} \bar{X} &= \tfrac{1}{n_1}\textstyle\sum_{i=1}^{n_1} X_i & S_1^2 &= \tfrac{1}{n_1}\textstyle\sum_{i=1}^{n_1}(X_i - \bar{X})^2 \\ \bar{Y} &= \tfrac{1}{n_2}\textstyle\sum_{i=1}^{n_2} Y_i & S_2^2 &= \tfrac{1}{n_2}\textstyle\sum_{i=1}^{n_2}(Y_i - \bar{Y})^2 \end{aligned} \tag{D.10}$$

(a) The random variable $Y = (n_1S_1^2 + n_2S_2^2)/\sigma^2$ has a chi-square distribution with $n_1 + n_2 - 2$ degrees of freedom.

(b) The random variable

$$W = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\sqrt{\frac{n_1S_1^2 + n_2S_2^2}{n_1+n_2-2}}} \tag{D.11}$$

has a t distribution with $n_1 + n_2 - 2$ degrees of freedom.

5. Let $X_1, X_2, \ldots, X_{n_1}$ be independent normal random variables with parameters μ_1 and σ_1, and $Y_1, Y_2, \ldots, Y_{n_2}$ be independent normal random variables with parameters μ_2 and σ_2. Define

$$\begin{aligned} \bar{X} &= \tfrac{1}{n_1}\textstyle\sum_{i=1}^{n_1} X_i & S_1^2 &= \tfrac{1}{n_1}\textstyle\sum_{i=1}^{n_1}(X_i - \bar{X})^2 \\ \bar{Y} &= \tfrac{1}{n_2}\textstyle\sum_{i=1}^{n_2} Y_i & S_2^2 &= \tfrac{1}{n_2}\textstyle\sum_{i=1}^{n_2}(Y_i - \bar{Y})^2 \end{aligned} \tag{D.12}$$

The random variable

$$Y = \frac{n_1S_1^2}{(n_1-1)\sigma_1^2}\bigg/\frac{n_2S_2^2}{(n_2-1)\sigma_2^2} \tag{D.13}$$

has an F distribution with n_1 and n_2 degrees of freedom.

6. Let X_1 be a normal random variable with parameters $\mu = \lambda$ and $\sigma = 1$, X_2 a chi-square random variable with parameter ν, and X_1 and X_2 be independent. The random variable $Y = X_1/\sqrt{X_2/\nu}$ has a noncentral t distribution with parameters ν and λ.

7. Let X be a continuous random variable with cumulative distribution function $F(x)$.

   (a) The random variable $Y = F(X)$ has a (standard) uniform distribution with parameters $a = 0$ and $b = 1$.

   (b) The random variable $Y = -\ln[1 - F(X)]$ has a (standard) exponential distribution with parameter $\lambda = 1$.

   Let X be a continuous random variable with probability density function $f(x)$. The random variable $Y = |X|$ has probability density function $g(y)$ given by

$$g(y) = \begin{cases} f(y) + f(-y), & \text{if } y > 0; \\ 0, & \text{elsewhere.} \end{cases} \tag{D.14}$$

   If X has a standard normal distribution ($\mu = 0$, $\sigma = 1$) then $g(y) = 2f(y)$.

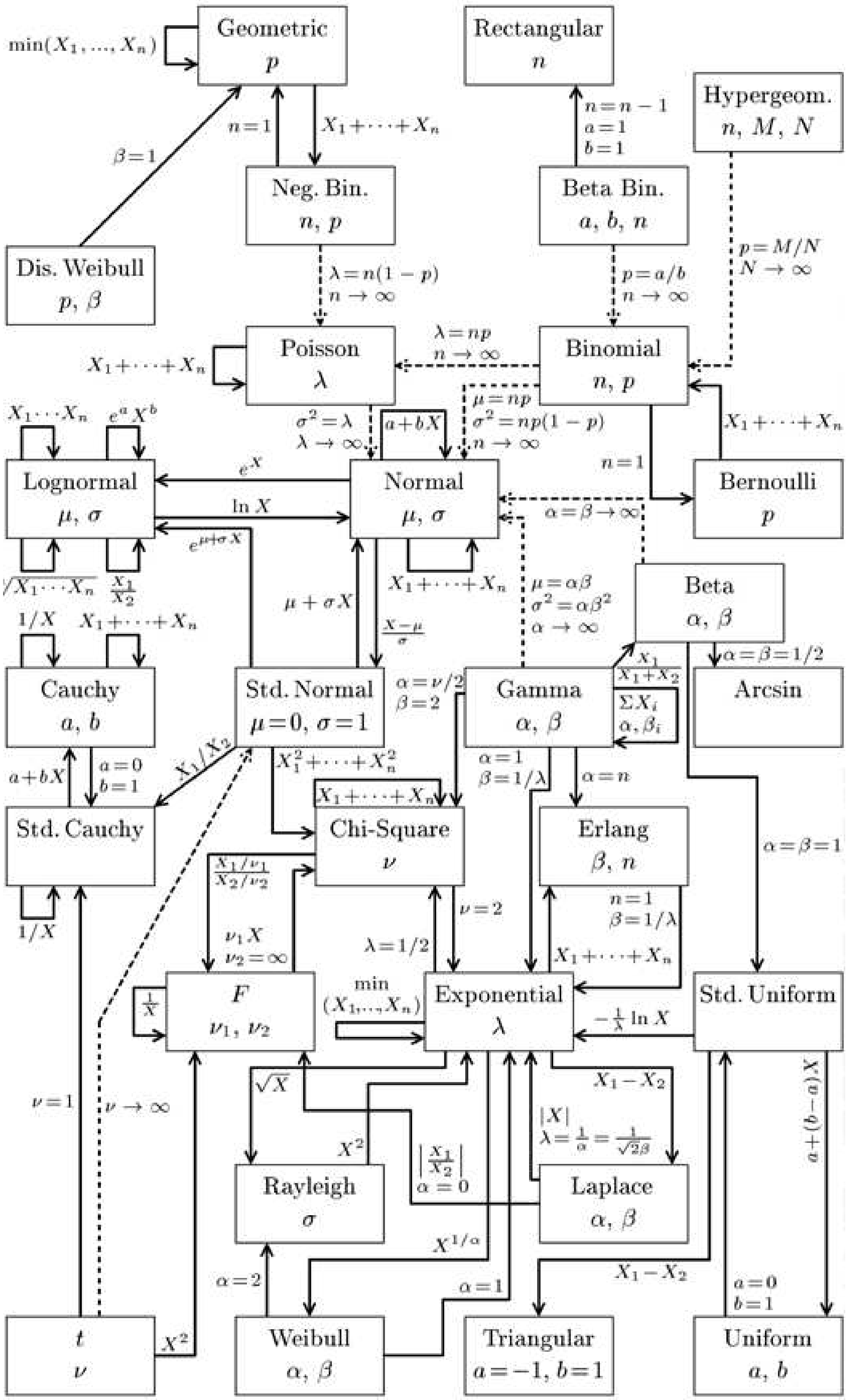


Figure D.1: Relationship among common univariate distributions

Appendix E

Tables of Common Univariate Distributions

Contents

E.1 Tables of common discrete distributions

Distribution	Notation	pmf $f(x)$	cdf $F(x)$	$E[X]$	$\text{Var}(X)$	MGF $M_X(t)$
Bernoulli	$X \sim \text{Bernoulli}(p)$	$p^x(1-p)^{1-x}$	$\sum_{i=0}^{[x]} f(i)$	p	$p(1-p)$	$(1-p)+pe^t$
	$p \in (0,1)$	$x = 0, 1$	$0 \le x \le 1$			$t \in \mathbb{R}$
Binomial	$X \sim \text{Binomial}(n,p)$	$\binom{n}{x}p^x(1-p)^{n-x}$	$\sum_{i=0}^{[x]} f(i)$	np	$np(1-p)$	$((1-p)+pe^t)^n$
	$n \in \mathbb{N},\ p \in (0,1)$	$x = 0, 1, \ldots, n$	$0 \le x \le n$			$t \in \mathbb{R}$
Discrete	$X \sim \text{DUniform}(a,b)$	$\frac{1}{b-a+1}$	$\frac{[x-a]+1}{b-a+1}$	$\frac{a+b}{2}$	$\frac{(b-a+1)^2-1}{12}$	$\frac{e^{at}(1-e^{(b-a+1)t})}{(b-a+1)(1-e^t)}$
Uniform	$a \le b,\ b-a \in \mathbb{N}_0$	$x = a, a+1, \ldots, b$	$a \le x \le b$			$t \in \mathbb{R}$
Geometric	$X \sim \text{Geometric}(p)$	$p(1-p)^{x-1}$	$1-(1-p)^{[x]}$	$\frac{1}{p}$	$\frac{1-p}{p^2}$	$\frac{pe^t}{1-(1-p)e^t}$
# Trials	$p \in (0,1)$	$x \in \mathbb{N}$	$x \ge 1$			$t < -\ln(1-p)$
Geometric	$X \sim \text{Geometric}(p)$	$p(1-p)^x$	$1-(1-p)^{[x+1]}$	$\frac{1-p}{p}$	$\frac{1-p}{p^2}$	$\frac{p}{1-(1-p)e^t}$
# Failures	$p \in (0,1)$	$x \in \mathbb{N}_0$	$x \ge 0$			$t < -\ln(1-p)$
Hypergeometric	$X \sim \text{HypG}(N,M,n)$	$\frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}}$	$\sum_{i=0}^{[x]} f(i)$	np	$np(1-p)$	NU
	$N \in \mathbb{N},\ n \in \mathbb{N}$	$x = 0, 1, \ldots, b$	$0 \le x \le b$	$p = \frac{M}{N}$	$\times\frac{(N-n)}{(N-1)}$	$t \in \mathbb{R}$
	$M = 0, 1, \ldots, N$	$b = \min(n, M)$				
Negative	$X \sim \text{NBinomial}(r,p)$	$\binom{x-1}{r-1}p^r(1-p)^{x-r}$	$\sum_{i=r}^{[x]} f(i)$	$\frac{r}{p}$	$\frac{r(1-p)}{p^2}$	$\left(\frac{pe^t}{1-(1-p)e^t}\right)^r$
Binomial	$r \in \mathbb{N},\ p \in (0,1)$	$x = r, r+1, \ldots$	$x \ge r$			$t < -\ln(1-p)$
# Trials						
Negative	$X \sim \text{NBinomial}(r,p)$	$\binom{x-r-1}{x}p^r(1-p)^x$	$\sum_{i=0}^{[x]} f(i)$	$\frac{r(1-p)}{p}$	$\frac{r(1-p)}{p^2}$	$\left(\frac{p}{1-(1-p)e^t}\right)^r$
Binomial	$r \in \mathbb{N},\ p \in (0,1)$	$x \in \mathbb{N}_0$	$x \ge 0$			$t < -\ln(1-p)$
# Failures						
Poisson	$X \sim \text{Poisson}(\lambda)$	$\frac{e^{-\lambda}\lambda^x}{x!}$	$\sum_{i=0}^{[x]} f(i)$	λ	λ	$e^{\lambda(e^t-1)}$
	$\lambda \ge 0$	$x \in \mathbb{N}_0$	$x \ge 0$			$t \in \mathbb{R}$

pmf: probability mass function, cdf: cumulative distribution function, E: expectation, Var: variance, MGF: moment generating function
$\mathbb{R} = \{x \mid -\infty < x < \infty\}$, $\mathbb{N} = \{1, 2, \ldots\}$, $\mathbb{N}_0 = \{0, 1, \ldots\}$, $\binom{n}{x} = \frac{n!}{x!(n-x)!}$, $[x]$: greatest integer smaller than or equal to x, NU: Not useful

E.2 Tables of common continuous distributions

Distribution	Notation	pdf $f(x)$	cdf $F(x)$	$E[X]$	$\mathrm{Var}(X)$	MGF $M_X(t)$
Beta	$X \sim \mathrm{Beta}(\alpha, \beta)$ $\alpha > 0,\ \beta > 0$	$\frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)}$ $0 < x < 1$	$\int_0^x f(x)\,dx$ $0 < x < 1$	$\frac{\alpha}{\alpha+\beta}$	$\frac{\alpha\beta}{(\alpha+\beta+1)}\frac{1}{(\alpha+\beta)^2}$	NU $t \in \mathbb{R}$
Cauchy	$X \sim \mathrm{Cauchy}(\mu, \sigma)$ $\mu \in \mathbb{R},\ \sigma > 0$	$\frac{1}{\pi\sigma\left[1+\left(\frac{x-\mu}{\sigma}\right)^2\right]}$ $x \in \mathbb{R}$	$\frac{1}{2} + \frac{\arctan\left(\frac{x-\mu}{\sigma}\right)}{\pi}$ $x \in \mathbb{R}$	DNE	DNE	DNE $t \neq 0$
Chi-Squared	$X \sim \chi^2_\nu$ $\nu > 0$	$\frac{x^{\frac{\nu}{2}-1}e^{-\frac{x}{2}}}{\Gamma(\frac{\nu}{2})2^{\frac{\nu}{2}}}$ $x > 0$	$\int_0^x f(u)\,du$ $x > 0$	ν	2ν	$(1-2t)^{-\frac{\nu}{2}}$ $t < \frac{1}{2}$
Exponential	$X \sim \mathrm{Exponential}(\beta)$ $\beta > 0$	$\frac{e^{-x/\beta}}{\beta}$ $x > 0$	$1 - e^{-x/\beta}$ $x > 0$	β	β^2	$(1-\beta t)^{-1}$ $t < \frac{1}{\beta}$
Exponential	$X \sim \mathrm{Exponential}(\beta)$ $\beta > 0$	$\beta e^{-\beta x}$ $x > 0$	$1 - e^{-\beta x}$ $x > 0$	$\frac{1}{\beta}$	$\frac{1}{\beta^2}$	$\frac{\beta}{\beta-t}$ $t < \beta$
Exponential: Two-parameter	$X \sim \mathrm{Exponential}(\mu, \sigma)$ $\mu \in \mathbb{R},\ \sigma > 0$	$\frac{e^{-\frac{(x-\mu)}{\sigma}}}{\sigma}$ $x > \mu$	$1 - e^{-\frac{(x-\mu)}{\sigma}}$ $x > \mu$	$\mu + \sigma$	σ^2	$\frac{e^{\mu t}}{1-\sigma t}$ $t < \frac{1}{\sigma}$
F	$X \sim F_{\nu_1,\nu_2}$ $\nu_1 > 0,\ \nu_2 > 0$	$\left(\frac{v_1}{v_2}\right)^{v_1/2}\frac{x^{\nu_1/2-1}}{B(\nu_1/2,\nu_2/2)}$ $\times\left(1+\frac{\nu_1 x}{\nu_2}\right)^{-(\nu_1+\nu_2)/2}$ $x > 0$	$\int_0^x f(u)\,du$ $x > 0$	$\frac{\nu_2}{\nu_2-2}$ $\nu_2 > 2$	$2\left(\frac{v_2}{v_2-2}\right)^2$ $\times\frac{(v_1+v_2-2)}{v_1(v_2-4)}$ $\nu_2 > 4$	DNE $t \neq 0$
Gamma	$X \sim \mathrm{Gamma}(\alpha, \beta)$ $\alpha > 0,\ \beta > 0$	$\frac{1}{\Gamma(\alpha)\beta^\alpha}x^{\alpha-1}e^{-x/\beta}$ $x > 0$	$\int_0^x f(u)\,du$ $x > 0$	$\alpha\beta$	$\alpha\beta^2$	$(1-\beta t)^{-\alpha}$ $t < \frac{1}{\beta}$
Gamma	$X \sim \mathrm{Gamma}(\alpha, \beta)$ $\alpha > 0,\ \beta > 0$	$\frac{\beta^\alpha}{\Gamma(\alpha)}x^{\alpha-1}e^{-\beta x}$ $x > 0$	$\int_0^x f(u)\,du$ $x > 0$	$\frac{\alpha}{\beta}$	$\frac{\alpha}{\beta^2}$	$\left(\frac{\beta}{\beta-t}\right)^\alpha$ $t < \beta$

pdf: probability density function, cdf: cumulative distribution function, E: expectation, Var: variance, MGF: moment generating function
$\mathbb{R} = \{x \mid -\infty < x < \infty\}$, $\mathbb{N} = \{1, 2, \ldots\}$, $\Gamma(\alpha) = \int_0^\infty x^{\alpha-1}e^{-x}\,dx$ $(\alpha > 0)$, $B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} = \int_0^1 x^{\alpha-1}(1-x)^{\beta-1}\,dx$ $(\alpha > 0, \beta > 0)$
DNE: Does not exist, NU: Not useful

Distribution	Notation	pdf $f(x)$	cdf $F(x)$	$E[X]$	Var(X)	MGF $M_X(t)$
Laplace (Double Exponential)	$X \sim \text{Laplace}(\mu, \sigma)$ $\mu \in \mathbb{R},\ \sigma > 0$	$\frac{1}{2\sigma} e^{-\lvert x-\mu \rvert/\sigma}$ $x \in \mathbb{R}$	$\frac{e^{(x-\mu)/\sigma}}{2},\ x \leq \mu$ $1 - \frac{e^{-(x-\mu)/\sigma}}{2},\ x > \mu$	μ	$2\sigma^2$	$\frac{e^{\mu t}}{1-(\sigma t)^2}$ $\lvert t \rvert < \frac{1}{\sigma}$
Logistic	$X \sim \text{Logistic}(\mu, \beta)$ $\mu \in \mathbb{R},\ \beta > 0$	$\frac{1}{\beta} \frac{e^{-(x-\mu)/\beta}}{[1+e^{-(x-\mu)/\beta}]^2}$ $x \in \mathbb{R}$	$\frac{1}{1+e^{-(x-\mu)/\beta}}$ $x \in \mathbb{R}$	μ	$\frac{\pi^2\beta^2}{3}$	$e^{\mu t}\Gamma(1-\beta t)\times$ $\Gamma(1+\beta t),\ t < \frac{1}{\beta}$
Lognormal	$X \sim \text{Lognormal}(\mu, \sigma^2)$ $\mu \in \mathbb{R},\ \sigma > 0$	$\frac{1}{\sqrt{2\pi}\sigma} \frac{e^{-(\log x-\mu)^2/(2\sigma^2)}}{x}$ $x > 0$	$\int_0^x f(u)\,du$ $x > 0$	$e^{\mu+(\sigma^2/2)}$	$e^{2(\mu+\sigma^2)}$ $-e^{2\mu+\sigma^2}$	DNE $t \neq 0$
Normal (Gaussian)	$X \sim \text{N}(\mu, \sigma^2)$ $\mu \in \mathbb{R},\ \sigma > 0$	$\frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/(2\sigma^2)}$ $x \in \mathbb{R}$	$\Phi(x) = \int_{-\infty}^x f(u)\,du$ $x \in \mathbb{R}$	μ	σ^2	$e^{\mu t+\sigma^2 t^2/2}$ $t \in \mathbb{R}$
Pareto	$X \sim \text{Pareto}(\alpha, \beta)$ $\alpha > 0,\ \beta > 0$	$\frac{\beta\alpha^\beta}{x^{\beta+1}}$ $x > \alpha$	$1 - \left(\frac{\alpha}{x}\right)^\beta$ $x > \alpha$	$\frac{\alpha\beta}{\beta-1}$ $\beta > 1$	$\frac{\alpha^2\beta}{(\beta-1)^2(\beta-2)}$ $\beta > 2$	DNE $t \neq 0$
t	$X \sim t_\nu$ $\nu > 0$	$\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)} \frac{1}{\sqrt{\nu\pi}}$ $\times \frac{1}{\left(1+\frac{x^2}{\nu}\right)^{(\nu+1)/2}}$	$\int_0^x f(u)\,du$ $x > 0$	0 $\nu > 1$	$\frac{\nu}{\nu-2}$ $\nu > 2$	DNE $t \neq 0$
Uniform (Rectangular)	$X \sim U(a, b)$ $b > a$	$\frac{1}{b-a}$ $a < x < b$	$\frac{x-a}{b-a}$ $a < x < b$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$	$\frac{e^{bt}-e^{at}}{(b-a)t}$ $t \in \mathbb{R}$
Weibull	$X \sim \text{Weibull}(\gamma, \beta)$ $\gamma > 0,\ \beta > 0$	$\frac{\gamma}{\beta} x^{\gamma-1} e^{-x^\gamma/\beta}$ $x > 0$	$1 - e^{-x^\gamma/\beta}$ $x > 0$	$\beta^{\frac{1}{\gamma}}\Gamma\left(1+\frac{1}{\gamma}\right)$	$\beta^{\frac{2}{\gamma}}\left[\Gamma\left(1+\frac{2}{\gamma}\right)\right.$ $\left.-\Gamma^2\left(1+\frac{1}{\gamma}\right)\right]$	NU $\gamma \geq 1$

pdf: probability density function, cdf: cumulative distribution function, E: expectation, Var: variance, MGF: moment generating function
$\mathbb{R} = \{x \mid -\infty < x < \infty\}$, $\mathbb{N} = \{1, 2, \ldots\}$, $\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x}\,dx\ (\alpha > 0)$, $B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} = \int_0^1 x^{\alpha-1}(1-x)^{\beta-1}\,dx\ (\alpha > 0, \beta > 0)$
DNE: Does not exist, NU: Not useful

Appendix F

Glossary

1 ***Alternative definition of conditional probability***: $P[A \cap B] = P[B] \cdot P[A|B]$ 15

2 ***Bayes' Theorem***: Let the sets A and $B_i, i = 1, 2, \ldots, n$ satisfy the hypothesis of Proposition 1.4.2. Then for each $i = 1, 2, \ldots, n$, $P[B_i|A] = \frac{P[A|B_i] \cdot P[B_i]}{P[A|B_1] \cdot P[B_1] + \cdots + P[A|B_n] \cdot P[B_n]}$. 18

3 ***Beta density***: $f(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}, \alpha, \beta > 0, 0 < x < 1$. 84

4 ***Binomial probability mass function***: Let X be the total number of successes in n Bernoulli trials. Then X has the binomial probability mass function: $q(k) = P[X = k] = \binom{n}{k} p^k (1-p)^{n-k}$. 36

5 ***Cauchy density***: $f(x) = \frac{1}{\pi[1+(x-\mu)^2]}, -\infty < x < \infty$. 84

6 ***c.d.f. technique***: Find the distribution of a real-valued, continuous function of a continuous random variable. 125

7 ***Central Limit Theorem***: Let $X_1, X_2, \ldots$ be a sequence of i.i.d random variables with finite mean μ and variance σ^2. Let $\bar{X}_n = \sum_{i=1}^{n} X_i/n$. Then the c.d.f. $F_n(x)$ of the random variable $Z_n = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}$ converges to the standard normal c.d.f. at all points x. 160

8 ***Central moment or moment about the mean***: The rth central moment or moment about the mean is $E[(X-\mu)^r]$. 51

9 ***Characteristic function***: $\phi(t) = E[e^{itX}]$. 140

10 ***Chebyshev's Inequality***: If a random variable X has a finite mean μ and variance σ^2, then for all $h > 0$, $P[|X - \mu| \geq h\sigma] \leq \frac{1}{h^2}$ and hence$P[|X - \mu| < h\sigma] \geq 1 - \frac{1}{h^2}$.155

11 ***Chi-square density***: The chi-square density is $f(x) = \frac{1}{2^{n/2}\Gamma(n/2)} x^{n/2-1} e^{-x/2}, x > 0$. 78

12 ***Combination***: A combination of n items $\{y_1, \ldots, y_n\}, r$ at a time, is a subset $\{x_1, \ldots, x_r\}$ selected from the original n items, such that $x_i \neq x_j, \forall i \neq j$. We denote the number of such combinations by $C_{n,r}$, or $\binom{n}{r}$. The latter is read "n choose r." 29

13 ***Conditional expectation***: The conditional expectation $E[g(Y)|X = x]$ of a function of a continuous random variable given the observed value of another continuous random variable is $E[g(Y)|X = x] = \int_{-\infty}^{\infty} g(y)f(y|x)\,dy$. The integral is replaced by a sum in the discrete case. 105

14 ***Conditional mean***: The conditional mean of Y given $X = x$ is $\mu_{Y|x} = E[Y|X = x] = \int_{-\infty}^{\infty} y \cdot f(y|x)dy$. 106

15 ***Conditional probability density function***: X and Y are continuous random variables with joint probability density function $f(x, y)$, and f_X and f_Y are the marginal density functions, then the conditional probability density function of Y given $X = x$ is $f(y|x) = \frac{f(x,y)}{f_X(x)}$ provided $f_X(x) > 0$. Similarly the conditional probability density function of X given $Y = y$ is $f(x|y) = \frac{f(x,y)}{f_Y(y)}$ provided $f_Y(y) > 0$. 104

16 ***Conditional probability***: If B is an event such that $P[B] > 0$, then the conditional probability of an event A given B is defined as $P[A|B] = \frac{P[A\cap B]}{P[B]}$. 13

17 ***Conditional probability mass function***: If X and Y are discrete random variables with joint probability mass function $f(x, y)$, and f_X and f_Y are the marginal mass functions, then the conditional probability mass function of Y given $X = x$ is $f(y|x) = \frac{f(x,y)}{f_X(x)}$ provided $f_X(x) > 0$. 102

18 ***Conditional variance***: The conditional variance of Y given $X = x$ is $\sigma^2_{Y|x} = E[(Y - \mu_{Y|x})^2|X = x] = \int_{-\infty}^{\infty}(y - \mu_{Y|x})^2 f(y|x)dy$. 106

19 ***Converge in probability***: A sequence of random variables (Y_n) is said to converge in probability to a constant a if, for any $\epsilon > 0$, $P[|Y_n - a| < \epsilon] \to 1, \quad \text{as } n \to \infty$ or equivalently, $P[|Y_n - a| \geq \epsilon] \to 0, \text{ as } n \to \infty$. 158

20 ***Converges almost surely***: A sequence of random variables (Y_n) is said to converges almost surely to a constant a if $\lim_{n\to\infty} Y_n(\omega) = a$ for all outcomes ω except possibly those in some event $N \subseteq \Omega$ of the probability 0. 159

21 ***Correlation***: If the covariance and the marginal variances exist, the correlation between X and Y is $\rho = \rho_{XY} = Corr(X, Y) = E\left[\left(\frac{X-\mu_X}{\sigma_X}\right)\left(\frac{Y-\mu_Y}{\sigma_Y}\right)\right] = \frac{\text{Cov}(X,Y)}{\sigma_X\sigma_Y}$. 107

22 ***Correlation matrix***: The correlation matrix of $\mathbf{X}$ is the $n \times n$ symmetric matrix $\Upsilon = \text{Corr}(\mathbf{X}) = \begin{bmatrix} 1 & \rho_{12} & \dots & \rho_{1n} \\ \rho_{21} & 1 & \dots & \rho_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{n1} & \rho_{n2} & \dots & 1 \end{bmatrix}$ where $\rho_{ij} = \rho_{ji} = \text{Corr}(X_i, X_j)$. 115

23 ***Covariance matrix***: The covariance matrix of $\mathbf{X}$ is the $n\times n$ symmetric matrix $\Sigma = \text{Cov}(\mathbf{X}) = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \dots & \sigma_{1n} \\ \sigma_{21} & \sigma_2^2 & \dots & \sigma_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \dots & \sigma_n^2 \end{bmatrix}$ where $\sigma_i^2 = \text{Var}(X_i)$ and $\sigma_{ij} = \sigma_{ji} = \text{Cov}(X_i, X_j)$. 115

24 ***Covariance***: The covariance of two real random variables X and Y is $\sigma_{XY} = \text{Cov}(X, Y) = E[(X - \mu_X)(Y - \mu_Y)]$ provided the expectation exists. 107

25 ***Cumulative distribution function (c.d.f.)***: (Discrete case) $F(x) = P[X \leq x] = \sum_{t \leq x} q(t)$. 31

26 ***Cumulative distribution function (c.d.f.)***: The cumulative distribution function (c.d.f.) is defined as $F(x) = P[X \leq x] = \int_{-\infty}^{x} f(t)\, dt$ where $f(x) = F'(x)$. 12

27 ***Dependent***: Random variables that are not independent are called dependent. 93

28 ***Discrete uniform distribution***: $q(x) = \frac{1}{n}, x \in \{x_1, x_2, \ldots, x_n\}$ 31

29 ***Empirical cumulative distribution function (edf)***: The empirical probability mass function (emf) of the sample is $\hat{q}(w_j) = \frac{\text{number of } x_i = w_j}{n}$. 32

30 ***Empirical probability mass function (emf)***: The empirical cumulative distribution function (edf) of the sample is then the c.d.f. associated with $\hat{q}$ through equation (2.7), alternatively, $\hat{F}(w_j) = \frac{\text{number of } x_i \leq w_j}{n}$. 32

31 ***Event***: Subset of the sample space. 4

32 ***Expected value or expectation***: Let X be a discrete, real-value random variable with state space $E = \{e_1, e_2, \ldots\}$ and probability mass function q. Then the expected value or expectation of X is $\mu = E[X] = \sum_i e_i P[X = e_i] = \sum_i e_i q(e_i)$ provided the series converges. 46

33 ***Expected value***: The expected value of a real-valued, continuous random variable X with p.d.f. f and state space E is $E[X] = \int_E x \cdot f(x)\, dx$ provided the integral exists. 71

34 ***Exponential density*** The exponential density is $f(t) = \lambda e^{-\lambda t},\ t > 0$. 74

35 ***F-distribution***: The F-density with parameters r_1, r_2 (both called degrees of freedom is the p.d.f of the random variable $F = \frac{U/r_1}{V/r_2}$ where $U \sim \chi^2(r_1), V \sim \chi^2(r_2)$, and U and V are independent random variables. We use the shorthand notation $f(r_1, r_2)$ when referring to the distribution. 150

36 ***Fundamental counting principle***:

(a) Suppose that an experiment has two stages. For the first stage, there are m possible outcomes, and for each of these, the second stage has n possible outcomes. Then the two-stage experiment has $m \cdot n$ outcomes.

(b) For a more general two-stage experiment, let the first-stage outcomes be labeled $i = 1, 2, \ldots, m$. Assume that if the first-stage outcome is i, then there are n_i possible outcomes for stage 2. Then the two-stage experiment has

$$\sum_{i=1}^{m} n_i$$

possible outcomes.

(c) Suppose that an experiment consists of k stages such that the first stage has m_1 possible outcomes, for each outcome of stage 1 there are m_2 possible outcomes of stage 2, for each combined outcome of the first two stages there are m_3 possible outcomes of stages 3, and so on. Then there are $m_1 \cdot m_2 \cdots m_k$ outcomes of the entire experiment.

. . . . 26

37 ***Gamma density***: The gamma density is $f(t) = \frac{\lambda^\alpha}{\Gamma(\alpha)} t^{\alpha-1} e^{-\lambda t}, t > 0$. 76

38 ***Gamma function***: The gamma function is defined by $\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x}\, dx$. . 76

39 ***Generalized multiplication rule***: If $A_1, \ldots, A_n$ are events such that the following conditional probabilities are defined, then $P[A_1 \cap A_2 \cap \cdots \cap A_n] = P[A_1] \cdot P[A_2|A_1] \cdots P[A_n|A_1 \cap \cdots \cap A_{n-1}]$. 15

40 ***Geometric distribution***: Let T_1 be the random variable that returns the trial on which the first success occurs in a sequence of Bernoulli trials. Then T_1 has the geometric distribution $P[T_1 = n] = (1-p)^{n-1} p, n = 1, 2, \ldots$. 37

41 ***Hypergeometric distribution***: The experiments involving sampling without replacement discussed in the last section give rise to a frequently observed distribution. 32

42 ***Idempotent matrix***: A matrix Q is symmetric and $Q^2 = Q$. It's eigenvalues are 0 or 1 only and the number of eigenvalues equal to 1 is the rank r of the matrix and it can be diagonal decomposed. 146

43 ***Independent***: Events A and B are said to be independent of one another if $P[A|B] = P[A]$ provided $P[B] > 0$. 19

44 ***Inverse image***: If A is a set in the range of a function f, define the inverse image of A as the following subset of the domain of f: $f^{-1}(A) = \{x | f(x) \in A\}$. 99

45 ***Joint conditional p.m.f.***: Consider a random vector $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ with joint p.m.f. $f(x_1, x_2, \ldots, x_n)$. The joint conditional p.m.f. of $X_{m+1}, \ldots, X_n$ given $X_1, \ldots, X_m$ is $f(x_{m+1}, \ldots, x_n | x_1, \ldots, x_m) = \frac{f(x_1, \ldots, x_n)}{f_{1,\ldots,m}(x_1, \ldots, x_m)}$. 103

46 ***Joint marginal mass function***: The joint marginal mass function of a subcollection $X_{i_1}, \ldots, X_{i_k}$ of the random variables is $q(x_{i_1}, \ldots, x_{i_k}) = P[X_{i_1} = x_{i_1}, \ldots, X_{i_k} = x_{i_k}] = \sum \cdots \sum q(x_1, \ldots, x_n)$. 34

47 ***Law of Total Probability***: Let A be an event, and let $B_1, \ldots, B_n$ be mutually exclusive events of nonzero probability whose union is the sample space Ω. Then $P[A] = P[B_1] \cdot P[A|B_1] + P[B_2] \cdot P[A|B_2] + \cdots + P[B_n] \cdot P[A|B_n]$. 15

48 ***Lognormal density***: $f(y) = \frac{1}{y} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(\ln(y)-\mu)^2}{2\sigma^2}\right], \mu \in \mathbb{R}, \sigma^2 > 0, y > 0$ 84

49 ***Marginal mass function***: The marginal mass function of X_i is $q_i(x_i) = P[X_i = x_i] = \sum_{x_1} \cdots \sum_{x_{i-1}} \sum_{x_{i+1}} \cdots \sum_{x_n} q(x_1, \ldots, x_n)$. 34

50 ***Mean***: The mean of X is $\mu = E[X]$. 71

51 ***Moment about mean***: The rth moment about the mean of X is $\mu'_r = E[(X-\mu)^r] = \int_E (x-\mu)^r f(x)\, dx$. 71

52 ***Moment generating function (m.g.f)***: The moment generating function (m.g.f) of a real-valued random variable X is the function $M(t) = M_X(t) = E[e^{tX}]$ which is defined for all real values of t such that the expectation is finite. 136

53 ***Moment generating function***: The moment generating function of a random vector $\mathbf{X} = (X_1, \ldots, X_n)$ is following real-valued function of a vector variable $\mathbf{t} = (t_1, \ldots, t_n)$: $M_{\mathbf{X}}(\mathbf{t}) = E[e^{\mathbf{t}'\mathbf{X}}] = E\left[\exp\left(\sum_{i=1}^n t_i X_i\right)\right]$ which is defined for all $t \in \mathbb{R}^n$ such that the expectation is finite. 139

54 ***Moment***: The rth moment of the distribution of a real random variable X is $E[X^r]$, provided the expectation exists. 51

55 ***Moment***: The rth moment of X is $\mu_r = E[X^r] = \int_E x^r f(x)\, dx$. 71

56 ***Multivariate normal distribution***: A random vector $\mathbf{X} = (X_1, X_2, \ldots, X_n)'$ is said to have the multivariate normal distribution with parameters $\boldsymbol{\mu}$ and Σ if its density is $f(\mathbf{x}) = \frac{1}{(2\pi)^{n/2}\sqrt{\det(\Sigma)}} e^{-1/2(\mathbf{x}-\boldsymbol{\mu})'\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})}, \mathbf{x} \in \mathbb{R}^n$. 86

57 ***Mutually independent***: Events $A_1, \ldots, A_n$ are said to be ***mutually independent*** if for any subcollection $A_{i_1}, \ldots, A_{i_k}$ of the events, $P[A_{i_1} \cap A_{i_2} \cap \cdots \cap A_{i_k}] = P[A_{i_1}] \cdot P[A_{i_2}] \cdots P[A_{i_k}]$ 21

58 ***Mutually independent***: Random variables $X_1, X_2, \ldots, X_n$ are called mutually independent if for any subcollection of them $X_{i1}, X_{i2}, \ldots, X_{ik}, k \leq n$, and corresponding subsets $B_{i1}, B_{i2}, \ldots, B_{ik}$ of their state spaces, $P[X_{i1} \in B_{i1}, X_{i2} \in B_{i2}, \ldots, X_{ik} \in B_{ik}] = P[X_{i1} \in B_{i1}] \cdot P[X_{i2} \in B_{i2}] \cdots P[X_{ik} \in B_{ik}]$. 93

59 ***Negative binomial distribution***: Let T_r be the trial on which the rth success occurs in a sequence of Bernoulli trials. Then T_r has the negative binomial distribution: $P[T_r = n] = \binom{n-1}{r-1} p^r (1-p)^{n-r}, n = r, r+1, \ldots$. 38

60 ***Normal density***: The normal density is $f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), -\infty < x < \infty$. 79

61 ***Order statistics***: We denote these ordered sample values by $Y_1 \leq Y_2 \leq \cdots \leq Y_n$ (some authors use the notation $X_{(1)}, \ldots, X_{(n)}$), and we refer to them as the order statistics of the sample. 132

62 ***Pareto density***: $f(x) = \frac{\theta}{(1+x)^{\theta+1}}, \theta > 0$ 84

63 ***Permutation***: A permutation of n objects $\{y_1, \ldots, y_n\}$, taken r at a time is an ordered list $(x_1, \ldots, x_r)$ selected from the original n objects, such that $x_i \neq x_j, \forall i \neq j$. We denote the number of such permutations by $P_{n,r}$. 28

64 ***Poisson probability mass function***: The Poisson probability mass function with parameter λ is $q(k) = P[X = k] = \frac{e^{-\lambda}\lambda^k}{k!}, k = 0, 1, 2, \ldots$. 41

65 ***Poisson process***: A Poisson process is a family $(N_t)_{t\geq 0}$ of random variables whose paths are step functions beginning at state 0 at time 0, which jump by 1 at a sequence of random times $T_1, T_2, \ldots$. Additionally, the some conditions are assumed. If $(N_t)_{t\geq 0}$ is a Poisson process with rate parameter λ, then $P[N_t = n] = \frac{e^{-\lambda t}(\lambda t)^n}{n!}$. 43

66 ***Poisson(λt) distribution***: If $(N_t)_{t\geq 0}$ is a Poisson process with rate parameter λ, then $P[N_t = n] = \frac{e^{-\lambda t}(\lambda t)^n}{n!}$, that is, the number of arrivals up through time t has the Poisson(λt) distribution. 44

67 ***Probability***: A measure of the likelihood of events. 4

68 ***Probability density function (p.d.f.)***: A random variable X is said to have probability density function (p.d.f.) f if, for all subsets A of the state space, $Q(A) = P[X \in A] = \int_A f(x)\,dx$. 12

69 ***Probability distribution***: The probability distribution Q of a random variable X is the probability measure on E defined by $Q(A) = P[X \in A]$, for $A \subseteq E$. 10

70 ***Probability generating function***: (Discrete distribution) $P(t) = E[t^X] = \sum_{i=1}^{\infty} t^{x_i} P[X = x_i]$. 140

71 ***Probability mass function (p.m.f.)***: Suppose that a random variable X has a discrete (i.e., finite or countable) state space. The function $q : E \to [0,1]$ is called the probability mass function (p.m.f.) of X if $q(x) = Q(\{x\}) = P[X = x]$. 11

72 ***Probability mass function (p.m.f.)***: $q(x) = P[X = x] \quad$ for $x \in E$. 31

73 ***Probability measure***: A probability measure on a sample space Ω of a random experiment is a function $P[\cdot]$ that maps events in Ω to real numbers such that $P[\Omega] = 1, P[F] \geq 0$ for all events $F, P[E_1 \cup E_2 \cup \cdots] = P[E_1] + P[E_2] + \cdots$ where E_i are disjoint events. 5

74 ***Random***: Pertain to an experiment whose result remains uncertain until the experiment is performed or phenomenon is observed. 3

75 ***Random sample***: A random sample $X_1, \ldots, X_n$ is a collection of n independent and identically distributed (i.i.d.) random variables. 101

76 ***Random variable***: A function gives a numerical value to each outcome of a random experiment. 5

77 ***Sample mean***: Estimates the central tendency of the distribution, measured by μ. $\bar{X} = \sum_{i=1}^{n} \frac{X_i}{n}$. 141

78 ***Sample space***: The sample space Ω of a random experiment is the collection of all possible outcomes. 5

79 ***Sample space***: The set of all possible outcomes of the experiment. 4

80 ***Sample variance***: Estimates the spread of the distribution, measured by σ^2, by the average squared distance of data points from $\bar{X}$. $S^2 = \sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{n-1}$. 141

81 ***Standard deviation***: The square root of the variance, σ, is referred to as the standard deviation of the random variable. 50

82 ***Standardizing***If a random variable X has the $N(\mu, \sigma^2)$ distribution, then the random variable Z defined by $Z = \frac{X-\mu}{\sigma}$ has the standard normal distribution. Therefore $P[X \leq b] = P\left[Z = \frac{X-\mu}{\sigma} \leq \frac{b-\mu}{\sigma}\right] = \int_{-\infty}^{(b-\mu)/\sigma} \frac{1}{\sqrt{2\pi}} e^{-z^2/2}\,dz$ (The algebraic operation of subtracting μ and dividing by σ is known as standardizing. 81

83 ***State space***: Let Ω be a sample space, and let E be a subset of $\mathbb{R}^n$. We call E the state space of the random variable X. 9

84 ***Strong law of large numbers***: Let $X_1, X_2, X_3, \ldots$ be a sequence of i.i.d. random variables with finite mean μ and variance σ^2. Let $\bar{X}_n = \sum_{i=1}^{n} X_i/n$ be the sample mean of the first n $X_i's$. Then the sequence $(\bar{X}_n)$ converges almost surely to μ. 159

85 ***t-distribution***: The t-distribution with parameter r (called the degrees of freedom of the distribution) is the distribution of the random variable, $T = \frac{Z}{\sqrt{V/r}}$ where $Z \sim N(0,1), V \sim \chi^2(r)$, and Z and V are independent. We use $t(r)$ as a shorthand notation for this distribution. 149

86 ***Uniform distribution on*** $[0,c]$: Median: $1/2 = \int_0^m 1/c\,dx = m/c$ gives $m = c/2$. pth percentile: $p = \int_0^{x_p} 1/c\,dx = x_p/c$ gives $x_p = cp$. 67

87 ***Use m.g.f to compute the moments***: (1) $\frac{d}{dt}M(t) = E\left[\frac{d}{dt}e^{tX}\right] = E[Xe^{tX}]$. (2) $\frac{d}{dt}M(t)\big|_{t=0} = E[X]$. (3) $\frac{d^n}{dt^n}M(t) = E\left[\frac{d^n}{dt^n}e^{tX}\right] = E[X^n e^{tX}]$. (4) $\frac{d^n}{dt^n}M(t)\big|_{t=0} = E[X^n]$. 139

88 ***Variance***: The variance of a real random variable X is $\sigma^2 = \mathrm{Var}(X) = E[(X-\mu)^2]$ provided the expectation is finite. 50

89 ***Variance***: The variance of X is $\mathrm{Var}(X) = \sigma^2 = E[(X-\mu)^2] = \int_E (x-\mu)^2 f(x)\,dx$. 71

90 ***Weak law of large numbers***: Let $X_1, X_2, X_3, \ldots$ be a sequence of i.i.d. random variables with finite mean μ and variance σ^2. Let $\bar{X}_n = \sum_{i=1}^n X_i/n$ be the sample mean of the first n $X_i's$. Then the sequence $(\bar{X}_n)$ converges in probability to μ. 158

91 ***Weibull(λ,β) distribution***: $f(t) = \beta\lambda^\beta t^{\beta-1} e^{-(\lambda t)^\beta}, \lambda > 0, \beta > 0, t > 0$ 84

Index