

STAT/MATH 511
PROBABILITY

Fall, 2009

Lecture Notes

Joshua M. Tebbs

Department of Statistics

University of South Carolina

Contents

2	Probability	1
2.1	Introduction	1
2.2	Sample spaces	3
2.3	Basic set theory	3
2.4	Properties of probability	6
2.5	Discrete probability models and events	8
2.6	Tools for counting sample points	10
2.6.1	The multiplication rule	10
2.6.2	Permutations	11
2.6.3	Combinations	15
2.7	Conditional probability	17
2.8	Independence	20
2.9	Law of Total Probability and Bayes Rule	22
3	Discrete Distributions	27
3.1	Random variables	27
3.2	Probability distributions for discrete random variables	28
3.3	Mathematical expectation	32
3.4	Variance	35
3.5	Moment generating functions	37
3.6	Binomial distribution	41
3.7	Geometric distribution	45
3.8	Negative binomial distribution	48
3.9	Hypergeometric distribution	51
3.10	Poisson distribution	55

4	Continuous Distributions	62
4.1	Introduction	62
4.2	Cumulative distribution functions	62
4.3	Continuous random variables	64
4.4	Mathematical expectation	70
4.4.1	Expected value	70
4.4.2	Variance	72
4.4.3	Moment generating functions	73
4.5	Uniform distribution	74
4.6	Normal distribution	76
4.7	The gamma family of distributions	81
4.7.1	Exponential distribution	82
4.7.2	Gamma distribution	85
4.7.3	χ^2 distribution	90
4.8	Beta distribution	91
4.9	Chebyshev's Inequality	95
4.10	Expectations of piecewise functions and mixed distributions	96
4.10.1	Expectations of piecewise functions	96
4.10.2	Mixed distributions	99
5	Multivariate Distributions	101
5.1	Introduction	101
5.2	Discrete random vectors	102
5.3	Continuous random vectors	103
5.4	Marginal distributions	106
5.5	Conditional distributions	109
5.6	Independent random variables	113

5.7	Expectations of functions of random variables	117
5.8	Covariance and correlation	120
5.8.1	Covariance	120
5.8.2	Correlation	124
5.9	Expectations and variances of linear functions of random variables	126
5.10	The multinomial model	128
5.11	The bivariate normal distribution	130
5.12	Conditional expectation	131
5.12.1	Conditional means and curves of regression	131
5.12.2	Iterated means and variances	132

2 Probability

Complementary reading: Chapter 2 (WMS).

2.1 Introduction

TERMINOLOGY: The text defines **probability** as a measure of one's belief in the occurrence of a future (random) event. Probability is also known as “the mathematics of uncertainty.”

REAL LIFE EVENTS: Here are some events we may wish to assign probabilities to:

- tomorrow's temperature exceeding 80 degrees
- getting a flat tire on my way home today
- a new policy holder making a claim in the next year
- the NASDAQ losing 5 percent of its value this week
- you being diagnosed with prostate/cervical cancer in the next 20 years.

ASSIGNING PROBABILITIES: How do we assign probabilities to events? There are three general approaches.

1. *Subjective approach.*

- This approach is based on feeling and may not even be scientific.

2. *Relative frequency approach.*

- This approach can be used when some random phenomenon is observed repeatedly under identical conditions.

3. *Axiomatic/Model-based approach.* This is the approach we will take in this course.

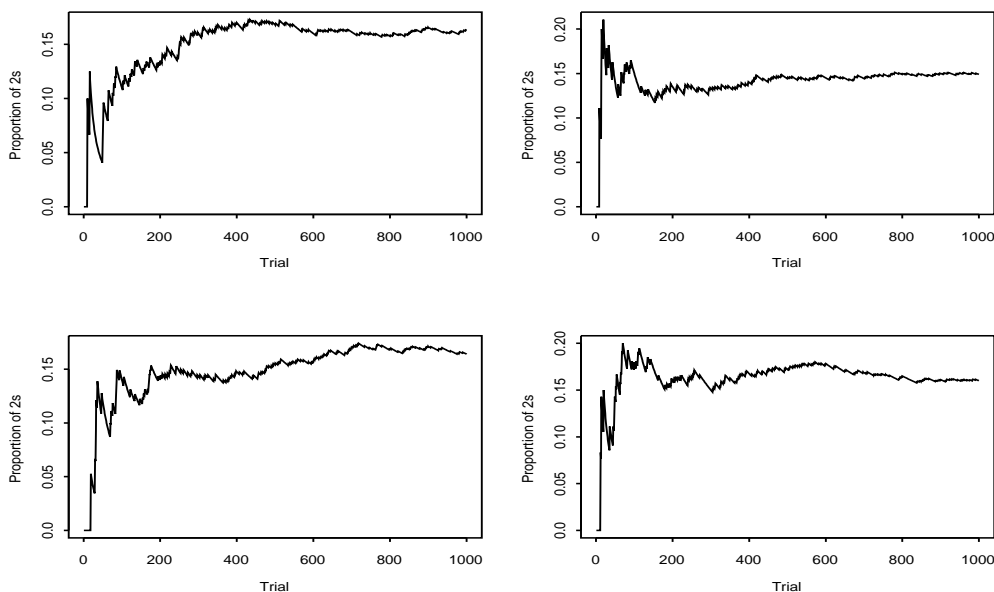


Figure 2.1: *The relative frequency of die rolls which result in a “2” ; each plot represents 1000 simulated rolls of a fair die.*

Example 2.1. *Relative frequency approach.* Suppose that we roll a die 1000 times and record the number of times we observe a “2.” Let A denote this event. The **relative frequency approach** says that

$$P(A) \approx \frac{\text{number of times } A \text{ occurs}}{\text{number of trials performed}} = \frac{n(A)}{n},$$

where $n(A)$ denotes the **frequency** of the event, and n denotes the number of trials performed. The proportion $n(A)/n$ is called the **relative frequency**. The symbol $P(A)$ is shorthand for “the probability that A occurs.”

RELATIVE FREQUENCY APPROACH: Continuing with our example, suppose that $n(A) = 158$. We would then estimate $P(A)$ by $158/1000 = 0.158$. If we performed the experiment of rolling a die repeatedly, the relative frequency approach says that

$$\frac{n(A)}{n} \rightarrow P(A),$$

as $n \rightarrow \infty$. Of course, if the die is fair, then $n(A)/n \rightarrow P(A) = 1/6$. \square

2.2 Sample spaces

TERMINOLOGY: Suppose that a **random experiment** is performed and that we observe an outcome from the experiment (e.g., rolling a die). The set of all possible outcomes for an experiment is called the **sample space** and is denoted by S .

Example 2.2. In each of the following random experiments, we write out a corresponding sample space.

(a) The Michigan state lottery calls for a three-digit integer to be selected:

$$S = \{000, 001, 002, \dots, 998, 999\}.$$

(b) A USC student is tested for chlamydia (0 = negative, 1 = positive):

$$S = \{0, 1\}.$$

(c) An industrial experiment consists of observing the lifetime of a battery, measured in hours. Different sample spaces are:

$$S_1 = \{w : w \geq 0\} \quad S_2 = \{0, 1, 2, 3, \dots\} \quad S_3 = \{\text{defective, not defective}\}.$$

Sample spaces are not unique; in fact, how we describe the sample space has a direct influence on how we assign probabilities to outcomes in this space. \square

2.3 Basic set theory

TERMINOLOGY: A **countable set** A is a set whose elements can be put into a one-to-one correspondence with $\mathcal{N} = \{1, 2, \dots\}$, the set of natural numbers. A set that is not countable is said to be **uncountable**.

TERMINOLOGY: Countable sets can be further divided up into two types.

- A **countably infinite set** has an infinite number of elements.
- A **countably finite set** has a finite number of elements.

Example 2.3. Say whether the following sets are countable (and, furthermore, finite or infinite) or uncountable.

(a) $A = \{0, 1, 2, \dots, 10\}$

(b) $B = \{1, 2, 3, \dots, \}$

(c) $C = \{x : 0 < x < 2\}$.

TERMINOLOGY: Suppose that A and B are sets (events). We say that A is a **subset** of B if every outcome in A is also in B , written $A \subset B$ or $A \subseteq B$.

- **IMPLICATION:** In a random experiment, if the event A occurs, then so does B . The converse is not necessarily true.

TERMINOLOGY: The **null set**, denoted by \emptyset , is the set that contains no elements.

TERMINOLOGY: The **union** of two sets A and B is the set of all elements in either A or B (or both), written $A \cup B$. The **intersection** of two sets A and B is the set of all elements in both A and B , written $A \cap B$. Note that $A \cap B \subseteq A \cup B$.

- **REMEMBER:** Union \longleftrightarrow “or” Intersection \longleftrightarrow “and”

EXTENSION: We extend the notion of unions and intersections to more than two sets. Suppose that A_1, A_2, \dots, A_n is a **finite** sequence of sets. The union of A_1, A_2, \dots, A_n is

$$\bigcup_{j=1}^n A_j = A_1 \cup A_2 \cup \dots \cup A_n,$$

that is, the set of all elements contained in at least one A_j . The intersection of A_1, A_2, \dots, A_n is

$$\bigcap_{j=1}^n A_j = A_1 \cap A_2 \cap \dots \cap A_n,$$

the set of all elements contained in each of the sets A_j , $j = 1, 2, \dots, n$.

EXTENSION: Suppose that A_1, A_2, \dots , is a **countable** sequence of sets. The union and intersection of this infinite collection of sets is denoted by

$$\bigcup_{j=1}^{\infty} A_j \quad \text{and} \quad \bigcap_{j=1}^{\infty} A_j,$$

respectively. The interpretation is the same as before.

Example 2.4. Define the sequence of sets $A_j = [1 - 1/j, 1 + 1/j)$, for $j = 1, 2, \dots$. Then,

$$\bigcup_{j=1}^{\infty} A_j = [0, 2) \quad \text{and} \quad \bigcap_{j=1}^{\infty} A_j = \{1\}. \quad \square$$

TERMINOLOGY: Suppose that A is a subset of S (the sample space). The **complement** of a set A is the set of all elements not in A (but still in S). We denote the complement by \bar{A} .

Distributive Laws:

1. $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$
2. $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$

DeMorgans Laws:

1. $\overline{A \cap B} = \bar{A} \cup \bar{B}$
2. $\overline{A \cup B} = \bar{A} \cap \bar{B}$

TERMINOLOGY: We call two events A and B **mutually exclusive**, or **disjoint**, if $A \cap B = \emptyset$, that is, A and B have no common elements.

Example 2.5. Suppose that a fair die is rolled. A sample space for this random experiment is $S = \{1, 2, 3, 4, 5, 6\}$.

- (a) If $A = \{1, 2, 3\}$, then $\bar{A} = \{4, 5, 6\}$.
- (b) If $A = \{1, 2, 3\}$, $B = \{4, 5\}$, and $C = \{2, 3, 6\}$, then $A \cap B = \emptyset$ and $B \cap C = \emptyset$.
Note that $A \cap C = \{2, 3\}$. Note also that $A \cap B \cap C = \emptyset$ and $A \cup B \cup C = S$. \square

2.4 Properties of probability

KOLMOLGOROV AXIOMS OF PROBABILITY: Given a nonempty sample space S , the measure $P(A)$ is a set function satisfying three axioms:

- (1) $P(A) \geq 0$, for every $A \subseteq S$
- (2) $P(S) = 1$
- (3) If A_1, A_2, \dots , is a countable sequence of **pairwise disjoint** events (i.e., $A_i \cap A_j = \emptyset$, for $i \neq j$) in S , then

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i).$$

RESULTS: The following results are important properties of the probability set function $P(\cdot)$, and each one follows from the Kolmolgorov axioms just stated. All events below are assumed to be subsets of a nonempty sample space S .

1. **Complement rule:** For any event A ,

$$P(A) = 1 - P(\bar{A}).$$

Proof. Note that $S = A \cup \bar{A}$. Thus, since A and \bar{A} are disjoint, $P(A \cup \bar{A}) = P(A) + P(\bar{A})$ by Axiom 3. By Axiom 2, $P(S) = 1$. Thus,

$$1 = P(S) = P(A \cup \bar{A}) = P(A) + P(\bar{A}). \quad \square$$

2. $P(\emptyset) = 0$.

Proof. Take $A = \emptyset$ and $\bar{A} = S$. Use the last result and Axiom 2. \square

3. **Monotonicity property:** Suppose that A and B are two events such that $A \subset B$. Then, $P(A) \leq P(B)$.

Proof. Write $B = A \cup (B \cap \bar{A})$. Clearly, A and $(B \cap \bar{A})$ are disjoint. Thus, by Axiom 3, $P(B) = P(A) + P(B \cap \bar{A})$. Because $P(B \cap \bar{A}) \geq 0$, we are done. \square

4. For any event A , $P(A) \leq 1$.

Proof. Since $A \subset S$, this follows from the monotonicity property and Axiom 2. \square

5. **Inclusion-exclusion:** Suppose that A and B are two events. Then,

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

Proof. Write $A \cup B = A \cup (\bar{A} \cap B)$. Then, since A and $(\bar{A} \cap B)$ are disjoint, by Axiom 3,

$$P(A \cup B) = P(A) + P(\bar{A} \cap B).$$

Now, write $B = (A \cap B) \cup (\bar{A} \cap B)$. Clearly, $(A \cap B)$ and $(\bar{A} \cap B)$ are disjoint. Thus, again, by Axiom 3,

$$P(B) = P(A \cap B) + P(\bar{A} \cap B).$$

Combining the last expressions for $P(A \cup B)$ and $P(B)$ gives the result. \square

Example 2.6. The probability that train 1 is on time is 0.95, and the probability that train 2 is on time is 0.93. The probability that both are on time is 0.90.

(a) What is the probability that **at least one** train is on time?

SOLUTION: Denote by A_i the event that train i is on time, for $i = 1, 2$. Then,

$$P(A_1 \cup A_2) = P(A_1) + P(A_2) - P(A_1 \cap A_2) = 0.95 + 0.93 - 0.90 = 0.98.$$

(b) What is the probability that **neither** train is on time?

SOLUTION: By DeMorgan's Law,

$$P(\bar{A}_1 \cap \bar{A}_2) = P(\overline{A_1 \cup A_2}) = 1 - P(A_1 \cup A_2) = 1 - 0.98 = 0.02. \square$$

EXTENSION: The **inclusion-exclusion** formula can be extended to any finite sequence of sets A_1, A_2, \dots, A_n . For example, if $n = 3$,

$$\begin{aligned} P(A_1 \cup A_2 \cup A_3) &= P(A_1) + P(A_2) + P(A_3) - P(A_1 \cap A_2) - P(A_1 \cap A_3) \\ &\quad - P(A_2 \cap A_3) + P(A_1 \cap A_2 \cap A_3). \end{aligned}$$

In general, the inclusion-exclusion formula can be written for any finite sequence:

$$P\left(\bigcup_{i=1}^n A_i\right) = \sum_{i=1}^n P(A_i) - \sum_{i_1 < i_2} P(A_{i_1} \cap A_{i_2}) + \sum_{i_1 < i_2 < i_3} P(A_{i_1} \cap A_{i_2} \cap A_{i_3}) - \cdots + (-1)^{n+1} P(A_1 \cap A_2 \cap \cdots \cap A_n).$$

Of course, if the sets A_1, A_2, \dots, A_n are **pairwise disjoint**, then we arrive back at

$$P\left(\bigcup_{i=1}^n A_i\right) = \sum_{i=1}^n P(A_i),$$

a result implied by Axiom 3 by taking $A_{n+1} = A_{n+2} = \cdots = \emptyset$.

2.5 Discrete probability models and events

TERMINOLOGY: If a sample space for an experiment contains a finite or countable number of sample points, we call it a **discrete sample space**.

- **Finite:** “number of sample points $< \infty$.”
- **Countable:** “number of sample points may equal ∞ , but can be counted; i.e., sample points may be put into a 1:1 correspondence with $\mathcal{N} = \{1, 2, \dots, \}$.”

Example 2.7. A standard roulette wheel contains an array of numbered compartments referred to as “pockets.” The pockets are either red, black, or green. The numbers 1 through 36 are evenly split between red and black, while 0 and 00 are green pockets. On the next play, we are interested in the following events:

$$\begin{aligned} A_1 &= \{13\} \\ A_2 &= \{\text{red}\} \\ A_3 &= \{0, 00\}. \end{aligned}$$

TERMINOLOGY: A **simple event** is an event that can not be decomposed. That is, a simple event corresponds to exactly one sample point. **Compound events** are those events that contain more than one sample point. In Example 2.7, because A_1 contains

only one sample point, it is a simple event. The events A_2 and A_3 contain more than one sample point; thus, they are compound events.

STRATEGY: Computing the probability of a compound event can be done by

- (1) counting up all sample points associated with the event (this can be very easy or very difficult)
- (2) adding up the probabilities associated with each sample point.

NOTATION: Your authors use the symbol E_i to denote the i th sample point (i.e., i th simple event). Thus, adopting the aforementioned strategy, if A denotes any compound event,

$$P(A) = \sum_{i: E_i \in A} P(E_i).$$

We simply sum up the simple event probabilities $P(E_i)$ for all i such that $E_i \in A$.

Example 2.8. *An equiprobability model.* Suppose that a discrete sample space S contains $N < \infty$ sample points, each of which are **equally likely**. If the event A consists of n_a sample points, then $P(A) = n_a/N$.

Proof. Write $S = E_1 \cup E_2 \cup \cdots \cup E_N$, where E_i corresponds to the i th sample point; $i = 1, 2, \dots, N$. Then,

$$1 = P(S) = P(E_1 \cup E_2 \cup \cdots \cup E_N) = \sum_{i=1}^N P(E_i).$$

Now, as $P(E_1) = P(E_2) = \cdots = P(E_N)$, we have that

$$1 = \sum_{i=1}^N P(E_i) = NP(E_1),$$

and, thus, $P(E_1) = \frac{1}{N} = P(E_2) = \cdots = P(E_N)$. Without loss of generality, take $A = E_1 \cup E_2 \cup \cdots \cup E_{n_a}$. Then,

$$P(A) = P(E_1 \cup E_2 \cup \cdots \cup E_{n_a}) = \sum_{i=1}^{n_a} P(E_i) = \sum_{i=1}^{n_a} \frac{1}{N} = n_a/N. \quad \square$$

Example 2.9. Two jurors are needed from a pool of 2 men and 2 women. The jurors are randomly selected from the 4 individuals. A sample space for this experiment is

$$S = \{(M1, M2), (M1, W1), (M1, W2), (M2, W1), (M2, W2), (W1, W2)\}.$$

What is the probability that the two jurors chosen consist of 1 male and 1 female?

SOLUTION. There are $N = 6$ sample points, denoted in order by E_1, E_2, \dots, E_6 . Let the event

$$A = \{\text{one male, one female}\} = \{(M1, W1), (M1, W2), (M2, W1), (M2, W2)\},$$

so that $n_A = 4$. If the sample points are equally likely (probably true if the jurors are randomly selected), then $P(A) = 4/6$. \square

2.6 Tools for counting sample points

2.6.1 The multiplication rule

MULTIPLICATION RULE: Consider an experiment consisting of $k \geq 2$ “stages,” where

$$\begin{aligned} n_1 &= \text{number of ways stage 1 can occur} \\ n_2 &= \text{number of ways stage 2 can occur} \\ &\vdots \\ n_k &= \text{number of ways stage } k \text{ can occur.} \end{aligned}$$

Then, there are

$$\prod_{i=1}^k n_i = n_1 \times n_2 \times \cdots \times n_k$$

different outcomes in the experiment.

Example 2.10. An experiment consists of rolling two dice. Envision stage 1 as rolling the first and stage 2 as rolling the second. Here, $n_1 = 6$ and $n_2 = 6$. By the multiplication rule, there are $n_1 \times n_2 = 6 \times 6 = 36$ different outcomes. \square

Example 2.11. In a controlled field experiment, I want to form all possible treatment combinations among the three factors:

Factor 1: Fertilizer (60 kg, 80 kg, 100kg: 3 levels)

Factor 2: Insects (infected/not infected: 2 levels)

Factor 3: Precipitation level (low, high: 2 levels).

Here, $n_1 = 3$, $n_2 = 2$, and $n_3 = 2$. Thus, by the multiplication rule, there are $n_1 \times n_2 \times n_3 = 12$ different treatment combinations. \square

Example 2.12. Suppose that an Iowa license plate consists of seven places; the first three are occupied by letters; the remaining four with numbers. Compute the total number of possible orderings if

- (a) there are no letter/number restrictions.
- (b) repetition of letters is prohibited.
- (c) repetition of numbers is prohibited.
- (d) repetitions of numbers and letters are prohibited.

ANSWERS:

(a) $26 \times 26 \times 26 \times 10 \times 10 \times 10 \times 10 = 175,760,000$

(b) $26 \times 25 \times 24 \times 10 \times 10 \times 10 \times 10 = 156,000,000$

(c) $26 \times 26 \times 26 \times 10 \times 9 \times 8 \times 7 = 88,583,040$

(d) $26 \times 25 \times 24 \times 10 \times 9 \times 8 \times 7 = 78,624,000$

2.6.2 Permutations

TERMINOLOGY: A **permutation** is an arrangement of distinct objects in a particular order. *Order is important.*

PROBLEM: Suppose that we have n distinct objects and we want to **order** (or **permute**) these objects. Thinking of n slots, we will put one object in each slot. There are

- n different ways to choose the object for slot 1,
- $n - 1$ different ways to choose the object for slot 2,
- $n - 2$ different ways to choose the object for slot 3,

and so on, down to

- 2 different ways to choose the object for slot $(n - 1)$, and
- 1 way to choose for the last slot.

IMPLICATION: By the multiplication rule, there are $n(n - 1)(n - 2) \cdots (2)(1) = n!$ different ways to order (permute) the n distinct objects.

Example 2.13. My bookshelf has 10 books on it. How many ways can I permute the 10 books on the shelf? **ANSWER:** $10! = 3,628,800$. \square

Example 2.14. Now, suppose that in Example 2.13 there are 4 math books, 2 chemistry books, 3 physics books, and 1 statistics book. I want to order the 10 books so that all books of the same subject are together. How many ways can I do this?

SOLUTION: Use the multiplication rule.

Stage 1	Permute the 4 math books	4!
Stage 2	Permute the 2 chemistry books	2!
Stage 3	Permute the 3 physics books	3!
Stage 4	Permute the 1 statistics book	1!
Stage 5	Permute the 4 subjects $\{m, c, p, s\}$	4!

Thus, there are $4! \times 2! \times 3! \times 1! \times 4! = 6912$ different orderings. \square

PERMUTATIONS: With a collection of n distinct objects, we now want to choose and **permute** r of them ($r \leq n$). The number of ways to do this is

$$P_{n,r} \equiv \frac{n!}{(n-r)!}.$$

The symbol $P_{n,r}$ is read “the permutation of n things taken r at a time.”

Proof. Envision r slots. There are n ways to fill the first slot, $n-1$ ways to fill the second slot, and so on, until we get to the r th slot, in which case there are $n-r+1$ ways to fill it. Thus, by the multiplication rule, there are

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

different permutations. \square

Example 2.15. With a group of 5 people, I want to choose a committee with three members: a president, a vice-president, and a secretary. There are

$$P_{5,3} = \frac{5!}{(5-3)!} = \frac{120}{2} = 60$$

different committees possible. **Here, note that order is important.** \square

Example 2.16. What happens if the objects to permute are **not distinct**? Consider the word *PEPPER*. How many permutations of the letters are possible?

TRICK: Initially, treat all letters as distinct objects by writing, say,

$$P_1 E_1 P_2 P_3 E_2 R.$$

There are $6! = 720$ different orderings of these distinct objects. Now, there are

3! ways to permute the P s

2! ways to permute the E s

1! ways to permute the R s.

So, $6!$ is actually $3! \times 2! \times 1!$ times too large. That is, there are

$$\frac{6!}{3! 2! 1!} = 60 \text{ possible permutations. } \square$$

MULTINOMIAL COEFFICIENTS: Suppose that in a set of n objects, there are n_1 that are similar, n_2 that are similar, ..., n_k that are similar, where $n_1 + n_2 + \cdots + n_k = n$. The number of permutations (i.e., distinguishable permutations) of the n objects is given by the **multinomial coefficient**

$$\binom{n}{n_1 n_2 \cdots n_k} \equiv \frac{n!}{n_1! n_2! \cdots n_k!}.$$

NOTE: Multinomial coefficients arise in the algebraic expansion of the multinomial expression $(x_1 + x_2 + \cdots + x_k)^n$; i.e.,

$$(x_1 + x_2 + \cdots + x_k)^n = \sum_D \binom{n}{n_1 n_2 \cdots n_k} x_1^{n_1} x_2^{n_2} \cdots x_k^{n_k},$$

where

$$D = \left\{ (n_1, n_2, \dots, n_k) : \sum_{j=1}^k n_j = n \right\}.$$

Example 2.17. How many signals, each consisting of 9 flags in a line, can be made from 4 white flags, 2 blue flags, and 3 yellow flags?

ANSWER:

$$\frac{9!}{4! 2! 3!} = 1260. \quad \square$$

Example 2.18. In Example 2.17, assuming all permutations are equally likely, what is the probability that all of the white flags are grouped together? We offer two solutions. The solutions differ in the way we construct the sample space. Define

$$A = \{\text{all four white flags are grouped together}\}.$$

SOLUTION 1. Work with a sample space that does **not** treat the flags as distinct objects, but merely considers color. Then, we know from Example 2.17 that there are 1260 different orderings. Thus,

$$N = \text{number of sample points in } S = 1260.$$

Let n_a denote the number of ways that A can occur. We find n_a by using the multiplication rule.

Stage 1	Pick four adjacent slots	$n_1 = 6$
Stage 2	With the remaining 5 slots, permute the 2 blues and 3 yellows	$n_2 = \frac{5!}{2!3!} = 10$

Thus, $n_a = 6 \times 10 = 60$. Finally, since we have equally likely outcomes, $P(A) = n_a/N = 60/1260 \approx 0.0476$. \square

SOLUTION 2. Initially, treat all 9 flags as **distinct objects**; i.e.,

$$W_1W_2W_3W_4B_1B_2Y_1Y_2Y_3,$$

and consider the sample space consisting of the $9!$ different permutations of these 9 distinct objects. Then,

$$N = \text{number of sample points in } S = 9!$$

Let n_a denote the number of ways that A can occur. We find n_a , again, by using the multiplication rule.

Stage 1	Pick adjacent slots for W_1, W_2, W_3, W_4	$n_1 = 6$
Stage 2	With the four chosen slots, permute W_1, W_2, W_3, W_4	$n_2 = 4!$
Stage 3	With remaining 5 slots, permute B_1, B_2, Y_1, Y_2, Y_3	$n_3 = 5!$

Thus, $n_a = 6 \times 4! \times 5! = 17280$. Finally, since we have equally likely outcomes, $P(A) = n_a/N = 17280/9! \approx 0.0476$. \square

2.6.3 Combinations

COMBINATIONS: Given n distinct objects, the number of ways to choose r of them ($r \leq n$), *without regard to order*, is given by

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

The symbol $C_{n,r}$ is read “the combination of n things taken r at a time.” By convention, we take $0! = 1$.

Proof: Choosing r objects is equivalent to breaking the n objects into two distinguishable groups:

Group 1 r chosen

Group 2 $(n - r)$ not chosen.

There are $C_{n,r} = \frac{n!}{r!(n-r)!}$ ways to do this. \square

REMARK: We will adopt the notation $\binom{n}{r}$, read “ n choose r ,” as the symbol for $C_{n,r}$. The terms $\binom{n}{r}$ are called **binomial coefficients** since they arise in the algebraic expansion of a binomial; viz.,

$$(x + y)^n = \sum_{r=0}^n \binom{n}{r} x^{n-r} y^r.$$

Example 2.19. Return to Example 2.15. Now, suppose that we only want to choose 3 committee members from 5 (without designations for president, vice-president, and secretary). Then, there are

$$\binom{5}{3} = \frac{5!}{3!(5-3)!} = \frac{5 \times 4 \times 3!}{3! \times 2!} = 10$$

different committees. \square

NOTE: From Examples 2.15 and 2.19, one should note that

$$P_{n,r} = r! \times C_{n,r}.$$

Recall that combinations do not regard order as important. Thus, once we have chosen our r objects (there are $C_{n,r}$ ways to do this), there are then $r!$ ways to permute those r chosen objects. Thus, we can think of a permutation as simply a combination times the number of ways to permute the r chosen objects.

Example 2.20. A company receives 20 hard drives. Five of the drives will be randomly selected and tested. If all five are satisfactory, the entire lot will be accepted. Otherwise, the entire lot is rejected. If there are really 3 defectives in the lot, what is the probability of accepting the lot?

SOLUTION: First, the number of sample points in S is given by

$$N = \binom{20}{5} = \frac{20!}{5!(20-5)!} = 15504.$$

Let A denote the event that the lot is accepted. How many ways can A occur? Use the multiplication rule.

Stage 1 Choose 5 good drives from 17 $\binom{17}{5}$

Stage 2 Choose 0 bad drives from 3 $\binom{3}{0}$

By the multiplication rule, there are $n_a = \binom{17}{5} \times \binom{3}{0} = 6188$ different ways A can occur. Assuming an equiprobability model (i.e., each outcome is equally likely), $P(A) = n_a/N = 6188/15504 \approx 0.399$. \square

2.7 Conditional probability

MOTIVATION: In some problems, we may be fortunate enough to have prior knowledge about the likelihood of events related to the event of interest. We may want to incorporate this information into a probability calculation.

TERMINOLOGY: Let A and B be events in a nonempty sample space S . The **conditional probability** of A , given that B has occurred, is given by

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

provided that $P(B) > 0$.

Example 2.21. A couple has two children.

- (a) What is the probability that both are girls?
- (b) What is the probability that both are girls, if the eldest is a girl?

SOLUTION: (a) The sample space is given by

$$S = \{(M, M), (M, F), (F, M), (F, F)\}$$

and $N = 4$, the number of sample points in S . Define

$$\begin{aligned} A_1 &= \{\text{1st born child is a girl}\}, \\ A_2 &= \{\text{2nd born child is a girl}\}. \end{aligned}$$

Clearly, $A_1 \cap A_2 = \{(F, F)\}$ and $P(A_1 \cap A_2) = 1/4$, assuming that the four outcomes in S are equally likely.

SOLUTION: (b) Now, we want $P(A_2|A_1)$. Applying the definition of conditional probability, we get

$$P(A_2|A_1) = \frac{P(A_1 \cap A_2)}{P(A_1)} = \frac{1/4}{2/4} = 1/2. \quad \square$$

Example 2.22. In a certain community, 36 percent of the families own a dog, 22 percent of the families that own a dog also own a cat, and 30 percent of the families own a cat. A family is selected at random.

- (a) Compute the probability that the family owns both a cat and dog.
- (b) Compute the probability that the family owns a dog, given that it owns a cat.

SOLUTION: Let $C = \{\text{family owns a cat}\}$ and $D = \{\text{family owns a dog}\}$. From the problem, we are given that $P(D) = 0.36$, $P(C|D) = 0.22$ and $P(C) = 0.30$. In (a), we want $P(C \cap D)$. We have

$$0.22 = P(C|D) = \frac{P(C \cap D)}{P(D)} = \frac{P(C \cap D)}{0.36}.$$

Thus,

$$P(C \cap D) = 0.36 \times 0.22 = 0.0792.$$

For (b), we want $P(D|C)$. Simply use the definition of conditional probability:

$$P(D|C) = \frac{P(C \cap D)}{P(C)} = \frac{0.0792}{0.30} = 0.264. \quad \square$$

RESULTS: It is interesting to note that conditional probability $P(\cdot|B)$ satisfies the axioms for a probability set function when $P(B) > 0$. In particular,

1. $P(A|B) \geq 0$
2. $P(B|B) = 1$
3. If A_1, A_2, \dots is a countable sequence of **pairwise mutually exclusive** events (i.e., $A_i \cap A_j = \emptyset$, for $i \neq j$) in S , then

$$P\left(\bigcup_{i=1}^{\infty} A_i \middle| B\right) = \sum_{i=1}^{\infty} P(A_i|B).$$

EXERCISE. Show that the measure $P(\cdot|B)$ satisfies the Kolmogorov axioms when $P(B) > 0$; i.e., establish the results above.

MULTIPLICATION LAW OF PROBABILITY: Suppose A and B are events in a non-empty sample space S . Then,

$$\begin{aligned} P(A \cap B) &= P(B|A)P(A) \\ &= P(A|B)P(B). \end{aligned}$$

Proof. As long as $P(A)$ and $P(B)$ are strictly positive, this follows directly from the definition of conditional probability. \square

EXTENSION: The multiplication law of probability can be extended to more than 2 events. For example,

$$\begin{aligned} P(A_1 \cap A_2 \cap A_3) &= P[(A_1 \cap A_2) \cap A_3] \\ &= P(A_3|A_1 \cap A_2) \times P(A_1 \cap A_2) \\ &= P(A_3|A_1 \cap A_2) \times P(A_2|A_1) \times P(A_1). \end{aligned}$$

NOTE: This suggests that we can compute probabilities like $P(A_1 \cap A_2 \cap A_3)$ “sequentially” by first computing $P(A_1)$, then $P(A_2|A_1)$, then $P(A_3|A_1 \cap A_2)$. The probability of a k -fold intersection can be computed similarly; i.e.,

$$P\left(\bigcap_{i=1}^k A_i\right) = P(A_1) \times P(A_2|A_1) \times P(A_3|A_1 \cap A_2) \times \cdots \times P\left(A_k \middle| \bigcap_{i=1}^{k-1} A_i\right).$$

Example 2.23. I am dealt a hand of 5 cards. What is the probability that they are all spades?

SOLUTION. Define A_i to be the event that card i is a spade ($i = 1, 2, 3, 4, 5$). Then,

$$\begin{aligned} P(A_1) &= \frac{13}{52} \\ P(A_2|A_1) &= \frac{12}{51} \\ P(A_3|A_1 \cap A_2) &= \frac{11}{50} \\ P(A_4|A_1 \cap A_2 \cap A_3) &= \frac{10}{49} \\ P(A_5|A_1 \cap A_2 \cap A_3 \cap A_4) &= \frac{9}{48}, \end{aligned}$$

so that

$$P(A_1 \cap A_2 \cap A_3 \cap A_4 \cap A_5) = \frac{13}{52} \times \frac{12}{51} \times \frac{11}{50} \times \frac{10}{49} \times \frac{9}{48} \approx 0.0005.$$

NOTE: As another way to solve this problem, a student recently pointed out that we could simply regard the cards as belonging to two groups: spades and non-spades. There are $\binom{13}{5}$ ways to draw 5 spades from 13. There are $\binom{52}{5}$ possible hands. Thus, the probability of drawing 5 spades (assuming that each hand is equally likely) is $\binom{13}{5} / \binom{52}{5} \approx 0.0005$. \square

2.8 Independence

TERMINOLOGY: When the occurrence or non-occurrence of A has no effect on whether or not B occurs, and vice versa, we say that the events A and B are **independent**. Mathematically, we define A and B to be independent iff

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

Otherwise, A and B are called **dependent** events. Note that if A and B are independent,

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

and

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

Example 2.24. A red die and a white die are rolled. Let $A = \{4 \text{ on red die}\}$ and $B = \{\text{sum is odd}\}$. Of the 36 outcomes in S , 6 are favorable to A , 18 are favorable to B , and 3 are favorable to $A \cap B$. Assuming the outcomes are equally likely,

$$\frac{3}{36} = P(A \cap B) = P(A)P(B) = \frac{6}{36} \times \frac{18}{36},$$

and the events A and B are independent. \square

Example 2.25. In an engineering system, two components are placed in a **series**; that is, the system is functional as long as both components are. Let A_i ; $i = 1, 2$, denote the event that component i is functional. Assuming independence, the probability the system is functional is then $P(A_1 \cap A_2) = P(A_1)P(A_2)$. If $P(A_i) = 0.95$, for example, then $P(A_1 \cap A_2) = 0.95 \times 0.95 = 0.9025$. If the events A_1 and A_2 are not independent, we do not have enough information to compute $P(A_1 \cap A_2)$. \square

INDEPENDENCE OF COMPLEMENTS: If A and B are independent events, so are

- (a) \bar{A} and B
- (b) A and \bar{B}
- (c) \bar{A} and \bar{B} .

Proof. We will only prove (a). The other parts follow similarly.

$$P(\bar{A} \cap B) = P(\bar{A}|B)P(B) = [1 - P(A|B)]P(B) = [1 - P(A)]P(B) = P(\bar{A})P(B). \quad \square$$

EXTENSION: The concept of independence (and independence of complements) can be extended to any finite number of events in S .

TERMINOLOGY: Let A_1, A_2, \dots, A_n denote a collection of $n \geq 2$ events in a nonempty sample space S . The events A_1, A_2, \dots, A_n are said to be **mutually independent** if for any subcollection of events, say, $A_{i_1}, A_{i_2}, \dots, A_{i_k}$, $2 \leq k \leq n$, we have

$$P\left(\bigcap_{j=1}^k A_{i_j}\right) = \prod_{j=1}^k P(A_{i_j}).$$

CHALLENGE: Come up with a random experiment and three events which are **pairwise independent**, but not mutually independent.

COMMON SETTING: Many experiments consist of a sequence of n trials that are viewed as independent (e.g., flipping a coin 10 times). If A_i denotes the event associated with the i th trial, and the trials are **independent**, then

$$P\left(\bigcap_{i=1}^n A_i\right) = \prod_{i=1}^n P(A_i).$$

Example 2.26. An unbiased die is rolled six times. Let $A_i = \{i \text{ appears on roll } i\}$, for $i = 1, 2, \dots, 6$. Then, $P(A_i) = 1/6$, and assuming independence,

$$P(A_1 \cap A_2 \cap A_3 \cap A_4 \cap A_5 \cap A_6) = \prod_{i=1}^6 P(A_i) = \left(\frac{1}{6}\right)^6.$$

Suppose that if A_i occurs, we will call it “a match.” What is the probability of at least one match in the six rolls?

SOLUTION: Let B denote the event that there is at least one match. Then, \bar{B} denotes the event that there are no matches. Now,

$$P(\bar{B}) = P(\bar{A}_1 \cap \bar{A}_2 \cap \bar{A}_3 \cap \bar{A}_4 \cap \bar{A}_5 \cap \bar{A}_6) = \prod_{i=1}^6 P(\bar{A}_i) = \left(\frac{5}{6}\right)^6 = 0.335.$$

Thus, $P(B) = 1 - P(\bar{B}) = 1 - 0.335 = 0.665$, by the complement rule.

EXERCISE: Generalize this result to an n sided die. What does this probability converge to as $n \rightarrow \infty$? \square

2.9 Law of Total Probability and Bayes Rule

SETTING: Suppose A and B are events in a nonempty sample space S . We can express the event A as follows

$$A = \underbrace{(A \cap B) \cup (A \cap \bar{B})}_{\text{union of disjoint events}}.$$

By the third Kolmogorov axiom,

$$\begin{aligned} P(A) &= P(A \cap B) + P(A \cap \overline{B}) \\ &= P(A|B)P(B) + P(A|\overline{B})P(\overline{B}), \end{aligned}$$

where the last step follows from the multiplication law of probability. This is called the **Law of Total Probability** (LOTP). The LOTP is helpful. Sometimes $P(A|B)$, $P(A|\overline{B})$, and $P(B)$ may be easily computed with available information whereas computing $P(A)$ directly may be difficult.

NOTE: The LOTP follows from the fact that B and \overline{B} **partition** S ; that is,

- (a) B and \overline{B} are disjoint, and
- (b) $B \cup \overline{B} = S$.

Example 2.27. An insurance company classifies people as “accident-prone” and “non-accident-prone.” For a fixed year, the probability that an accident-prone person has an accident is 0.4, and the probability that a non-accident-prone person has an accident is 0.2. The population is estimated to be 30 percent accident-prone. (a) What is the probability that a new policy-holder will have an accident?

SOLUTION:

Define $A = \{\text{policy holder has an accident}\}$ and $B = \{\text{policy holder is accident-prone}\}$. Then, $P(B) = 0.3$, $P(A|B) = 0.4$, $P(\overline{B}) = 0.7$, and $P(A|\overline{B}) = 0.2$. By the LOTP,

$$\begin{aligned} P(A) &= P(A|B)P(B) + P(A|\overline{B})P(\overline{B}) \\ &= (0.4)(0.3) + (0.2)(0.7) = 0.26. \quad \square \end{aligned}$$

(b) Now suppose that the policy-holder does have an accident. What is the probability that he was “accident-prone?”

SOLUTION: We want $P(B|A)$. Note that

$$P(B|A) = \frac{P(A \cap B)}{P(A)} = \frac{P(A|B)P(B)}{P(A)} = \frac{(0.4)(0.3)}{0.26} = 0.46. \quad \square$$

NOTE: From this last part, we see that, in general,

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

This is a form of **Bayes Rule**.

Example 2.28. A lab test is 95 percent effective at detecting a certain disease when it is present (sensitivity). When the disease is not present, the test is 99 percent effective at declaring the subject negative (specificity). If 8 percent of the population has the disease (prevalence), what is the probability that a subject has the disease given that (a) his test is positive? (b) his test is negative?

SOLUTION: Let $D = \{\text{disease is present}\}$ and $\mathfrak{X} = \{\text{test is positive}\}$. We are given that $P(D) = 0.08$ (prevalence), $P(\mathfrak{X}|D) = 0.95$ (sensitivity), and $P(\bar{\mathfrak{X}}|\bar{D}) = 0.99$ (specificity). In part (a), we want to compute $P(D|\mathfrak{X})$. By Bayes Rule,

$$\begin{aligned} P(D|\mathfrak{X}) &= \frac{P(\mathfrak{X}|D)P(D)}{P(\mathfrak{X}|D)P(D) + P(\mathfrak{X}|\bar{D})P(\bar{D})} \\ &= \frac{(0.95)(0.08)}{(0.95)(0.08) + (0.01)(0.92)} \approx 0.892. \end{aligned}$$

In part (b), we want $P(D|\bar{\mathfrak{X}})$. By Bayes Rule,

$$\begin{aligned} P(D|\bar{\mathfrak{X}}) &= \frac{P(\bar{\mathfrak{X}}|D)P(D)}{P(\bar{\mathfrak{X}}|D)P(D) + P(\bar{\mathfrak{X}}|\bar{D})P(\bar{D})} \\ &= \frac{(0.05)(0.08)}{(0.05)(0.08) + (0.99)(0.92)} \approx 0.004. \end{aligned}$$

Table 2.1: *The general Bayesian scheme.*

Measure before test		Result		Updated measure
$P(D)$		F		$P(D F)$
0.08	→	\mathfrak{X}	→	0.892
0.08	→	$\bar{\mathfrak{X}}$	→	0.004

NOTE: We have discussed the LOTP and Bayes Rule in the case of the partition $\{B, \bar{B}\}$. However, these rules hold for any partition of S .

TERMINOLOGY: A sequence of sets B_1, B_2, \dots, B_k is said to form a **partition** of the sample space S if

- (a) $B_1 \cup B_2 \cup \dots \cup B_k = S$ (exhaustive condition), and
- (b) $B_i \cap B_j = \emptyset$, for all $i \neq j$ (disjoint condition).

LAW OF TOTAL PROBABILITY (restated): Suppose that B_1, B_2, \dots, B_k form a partition of S , and suppose $P(B_i) > 0$ for all $i = 1, 2, \dots, k$. Then,

$$P(A) = \sum_{i=1}^k P(A|B_i)P(B_i).$$

Proof. Write

$$A = A \cap S = A \cap (B_1 \cup B_2 \cup \dots \cup B_k) = \bigcup_{i=1}^k (A \cap B_i).$$

Thus,

$$P(A) = P\left[\bigcup_{i=1}^k (A \cap B_i)\right] = \sum_{i=1}^k P(A \cap B_i) = \sum_{i=1}^k P(A|B_i)P(B_i). \quad \square$$

BAYES RULE (restated): Suppose that B_1, B_2, \dots, B_k form a partition of S , and suppose that $P(A) > 0$ and $P(B_i) > 0$ for all $i = 1, 2, \dots, k$. Then,

$$P(B_j|A) = \frac{P(A|B_j)P(B_j)}{\sum_{i=1}^k P(A|B_i)P(B_i)}.$$

Proof. Simply apply the definition of conditional probability and the multiplication law of probability to get

$$P(B_j|A) = \frac{P(A|B_j)P(B_j)}{P(A)}.$$

Then, just apply LOTP to $P(A)$ in the denominator to get the result. \square

REMARK: Bayesians will call $P(B_j)$ the **prior probability** for the event B_j ; they call $P(B_j|A)$ the **posterior probability** of B_j , given the information in A .

Example 2.29. Suppose that a manufacturer buys approximately 60 percent of a raw material (in boxes) from Supplier 1, 30 percent from Supplier 2, and 10 percent from

Supplier 3. For each supplier, defective rates are as follows: Supplier 1: 0.01, Supplier 2: 0.02, and Supplier 3: 0.03. The manufacturer observes a defective box of raw material.

- (a) What is the probability that it came from Supplier 2?
- (b) What is the probability that the defective did not come from Supplier 3?

SOLUTION: (a) Let $A = \{\text{observe defective box}\}$. Let B_1 , B_2 , and B_3 , respectively, denote the events that the box comes from Supplier 1, 2, and 3. The prior probabilities (ignoring the status of the box) are

$$P(B_1) = 0.6$$

$$P(B_2) = 0.3$$

$$P(B_3) = 0.1.$$

Note that $\{B_1, B_2, B_3\}$ partitions the space of possible suppliers. Thus, by Bayes Rule,

$$\begin{aligned} P(B_2|A) &= \frac{P(A|B_2)P(B_2)}{P(A|B_1)P(B_1) + P(A|B_2)P(B_2) + P(A|B_3)P(B_3)} \\ &= \frac{(0.02)(0.3)}{(0.01)(0.6) + (0.02)(0.3) + (0.03)(0.1)} = 0.40. \end{aligned}$$

This is the updated (posterior) probability that the box came from Supplier 2 (updated to include the information that the box was defective).

SOLUTION: (b) First, compute the posterior probability $P(B_3|A)$. By Bayes Rule,

$$\begin{aligned} P(B_3|A) &= \frac{P(A|B_3)P(B_3)}{P(A|B_1)P(B_1) + P(A|B_2)P(B_2) + P(A|B_3)P(B_3)} \\ &= \frac{(0.03)(0.1)}{(0.01)(0.6) + (0.02)(0.3) + (0.03)(0.1)} = 0.20. \end{aligned}$$

Thus,

$$P(\overline{B_3}|A) = 1 - P(B_3|A) = 1 - 0.20 = 0.80,$$

by the complement rule. \square

NOTE: Read Sections 2.11 (Numerical Events and Random Variables) and 2.12 (Random Sampling) in WMS.

3 Discrete Distributions

Complementary reading: Chapter 3 (WMS), except § 3.10 and § 3.11.

3.1 Random variables

PROBABILISTIC DEFINITION: A **random variable** Y is a function whose domain is the sample space S and whose range is the set of real numbers $\mathcal{R} = \{y : -\infty < y < \infty\}$. That is, $Y : S \rightarrow \mathcal{R}$ takes sample points in S and assigns them a real number.

WORKING DEFINITION: In simpler terms, a random variable is a variable whose observed value is determined by chance.

Example 3.1. Suppose that an experiment consists of flipping two fair coins. The sample space is

$$S = \{(H, H), (H, T), (T, H), (T, T)\}.$$

Let Y denote the number of heads observed. Before we perform the experiment, we do not know, with certainty, the value of Y . We can, however, list out the possible values of Y corresponding to each sample point:

E_i	$Y(E_i) = y$	E_i	$Y(E_i) = y$
(H, H)	2	(T, H)	1
(H, T)	1	(T, T)	0

For each sample point E_i , Y takes on a numerical value specific to E_i . This is precisely why we can think of Y as a function; i.e.,

$$Y[(H, H)] = 2 \quad Y[(H, T)] = 1 \quad Y[(T, H)] = 1 \quad Y[(T, T)] = 0,$$

so that

$$\begin{aligned} P(Y = 2) &= P[(H, H)] = 1/4 \\ P(Y = 1) &= P[(H, T)] + P[(T, H)] = 1/4 + 1/4 = 1/2 \\ P(Y = 0) &= P[(T, T)] = 1/4. \end{aligned}$$

NOTE: From these probability calculations; note that we can

- work on the sample space S and compute probabilities from S , or
- work on \mathcal{R} and compute probabilities for events $\{Y \in B\}$, where $B \subset \mathcal{R}$.

NOTATION: We denote a random variable Y using a capital letter. We denote an observed value of Y by y , a lowercase letter. **This is standard notation.** For example, if Y denotes the weight (in ounces) of the next newborn boy in Columbia, SC, then Y is a random variable. After the baby is born, we observe that the baby weighs $y = 128$ oz.

3.2 Probability distributions for discrete random variables

TERMINOLOGY: The **support** of a random variable Y is the set of all possible values that Y can assume. We will denote the support set by R .

TERMINOLOGY: If the random variable Y has a support set R that is countable (finitely or infinitely), we call Y a **discrete** random variable.

Example 3.2. An experiment consists of rolling an unbiased die. Consider the two random variables:

X = face value on the first roll

Y = number of rolls needed to observe a six.

The support of X is $R_X = \{1, 2, 3, 4, 5, 6\}$. The support of Y is $R_Y = \{1, 2, 3, \dots\}$. R_X is finitely countable and R_Y is infinitely countable; thus, both X and Y are discrete. \square

GOAL: For a discrete random variable Y , we would like to find $P(Y = y)$ for any $y \in R$. Mathematically,

$$p_Y(y) \equiv P(Y = y) = \sum P[E_i \in S : Y(E_i) = y],$$

for all $y \in R$.

TERMINOLOGY: Suppose that Y is a discrete random variable. The function $p_Y(y) = P(Y = y)$ is called the **probability mass function (pmf)** for Y . The pmf $p_Y(y)$ consists of two parts:

- (a) R , the support set of Y
- (b) a probability assignment $P(Y = y)$, for all $y \in R$.

PROPERTIES: A pmf $p_Y(y)$ for a discrete random variable Y satisfies the following:

- (1) $p_Y(y) > 0$, for all $y \in R$ [NOTE: if $y \notin R$, then $p_Y(y) = 0$]
- (2) The sum of the probabilities, taken over all support points, must equal one; i.e.,

$$\sum_{y \in R} p_Y(y) = 1.$$

IMPORTANT: Suppose that Y is a **discrete** random variable. The probability of an event $\{Y \in B\}$ is computed by adding the probabilities $p_Y(y)$ for all $y \in B$; i.e.,

$$P(Y \in B) = \sum_{y \in B} p_Y(y).$$

Example 3.3. An experiment consists of rolling two fair dice and observing the face on each. The sample space consists of $6 \times 6 = 36$ sample points:

$$\begin{aligned} S = & \{(1, 1), (1, 2), (1, 3), (1, 4), (1, 5), (1, 6), (2, 1), (2, 2), (2, 3), (2, 4), (2, 5), (2, 6), \\ & (3, 1), (3, 2), (3, 3), (3, 4), (3, 5), (3, 6), (4, 1), (4, 2), (4, 3), (4, 4), (4, 5), (4, 6), \\ & (5, 1), (5, 2), (5, 3), (5, 4), (5, 5), (5, 6), (6, 1), (6, 2), (6, 3), (6, 4), (6, 5), (6, 6)\}. \end{aligned}$$

Let the random variable Y record the sum of the two faces. Note that $R = \{2, 3, \dots, 12\}$.

We now compute the probability associated with each support point $y \in R$:

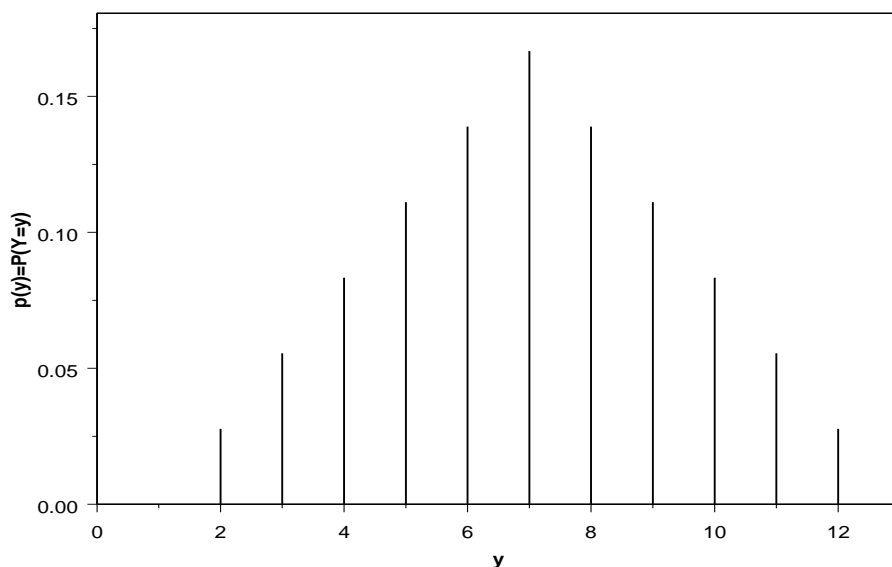
$$\begin{aligned} P(Y = 2) &= P(\{\text{all } E_i \in S \text{ where } Y(E_i) = y = 2\}) \\ &= P[(1, 1)] = 1/36. \end{aligned}$$

$$\begin{aligned}
 P(Y = 3) &= P(\{\text{all } E_i \in S \text{ where } Y(E_i) = y = 3\}) \\
 &= P[(1, 2)] + P[(2, 1)] = 2/36.
 \end{aligned}$$

The calculation $P(Y = y)$ is performed similarly for $y = 4, 5, \dots, 12$. The pmf for Y can be given as a formula, a table, or a graph. In tabular form, the pmf of Y is given by

y	2	3	4	5	6	7	8	9	10	11	12
$p_Y(y)$	1/36	2/36	3/36	4/36	5/36	6/36	5/36	4/36	3/36	2/36	1/36

A **probability histogram** is a display which depicts a pmf in graphical form. In this example, the probability histogram looks like



A closed-form formula for the pmf exists and is given by

$$p_Y(y) = \begin{cases} \frac{1}{36} (6 - |7 - y|), & y = 2, 3, \dots, 12 \\ 0, & \text{otherwise.} \end{cases}$$

Define the event $B = \{3, 5, 7, 9, 11\}$; i.e., the sum Y is odd. We have

$$\begin{aligned}
 P(Y \in B) &= \sum_{y \in B} p_Y(y) = p_Y(3) + p_Y(5) + p_Y(7) + p_Y(9) + p_Y(11) \\
 &= 2/36 + 4/36 + 6/36 + 4/36 + 2/36 = 1/2. \quad \square
 \end{aligned}$$

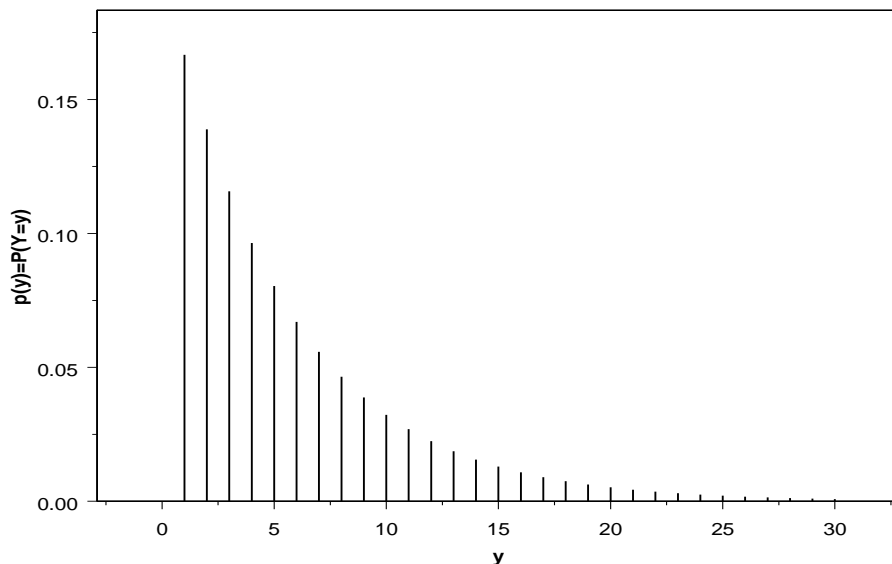
Example 3.4. An experiment consists of rolling an unbiased die until the first “6” is observed. Let Y denote the number of rolls needed. The support is $R = \{1, 2, \dots\}$. Assuming independent trials, we have

$$\begin{aligned} P(Y = 1) &= \frac{1}{6} \\ P(Y = 2) &= \frac{5}{6} \times \frac{1}{6} \\ P(Y = 3) &= \frac{5}{6} \times \frac{5}{6} \times \frac{1}{6}; \end{aligned}$$

Recognizing the pattern, we see that the pmf for Y is given by

$$p_Y(y) = \begin{cases} \frac{1}{6} \left(\frac{5}{6}\right)^{y-1}, & y = 1, 2, \dots \\ 0, & \text{otherwise.} \end{cases}$$

This pmf is depicted in a probability histogram below:



QUESTION: Is this a **valid** pmf; i.e., do the probabilities $p_Y(y)$ sum to one? Note that

$$\begin{aligned} \sum_{y \in R} p_Y(y) &= \sum_{y=1}^{\infty} \frac{1}{6} \left(\frac{5}{6}\right)^{y-1} \\ &= \sum_{x=0}^{\infty} \frac{1}{6} \left(\frac{5}{6}\right)^x \\ &= \left(\frac{\frac{1}{6}}{1 - \frac{5}{6}}\right) = 1. \quad \square \end{aligned}$$

IMPORTANT: In the last calculation, we have used an important fact concerning **infinite geometric series**; namely, if a is any real number and $|r| < 1$. Then,

$$\sum_{x=0}^{\infty} ar^x = \frac{a}{1-r}.$$

We will use this fact many times in this course!

EXERCISE: Find the probability that the first “6” is observed on (a) an odd-numbered roll (b) an even-numbered roll. Which event is more likely? \square

3.3 Mathematical expectation

TERMINOLOGY: Let Y be a discrete random variable with pmf $p_Y(y)$ and support R . The **expected value** of Y is given by

$$E(Y) = \sum_{y \in R} yp_Y(y).$$

The expected value for discrete random variable Y is simply a weighted average of the possible values of Y . Each support point y is weighted by the probability $p_Y(y)$.

ASIDE: When R is a countably infinite set, then the sum $\sum_{y \in R} yp_Y(y)$ may not exist (not surprising since sometimes infinite series do diverge). Mathematically, we require the sum above to be **absolutely convergent**; i.e.,

$$\sum_{y \in R} |y|p_Y(y) < \infty.$$

If this is true, we say that $E(Y)$ exists. If this is not true, then we say that $E(Y)$ does not exist. **NOTE:** If R is a finite set, then $E(Y)$ always exists, because a finite sum of finite quantities is always finite.

Example 3.5. Let the random variable Y have pmf

$$p_Y(y) = \begin{cases} \frac{1}{10}(5-y), & y = 1, 2, 3, 4 \\ 0, & \text{otherwise.} \end{cases}$$

The expected value of Y is given by

$$E(Y) = \sum_{y \in R} yp_Y(y) = \sum_{y=1}^4 y \left[\frac{1}{10}(5-y) \right] = 1(4/10) + 2(3/10) + 3(2/10) + 4(1/10) = 2. \quad \square$$

INTERPRETATION: The quantity $E(Y)$ has many interpretations:

- (a) the “center of gravity” of a probability distribution
- (b) a long-run average
- (c) the **first moment** of the random variable
- (d) the **mean** of a population.

FUNCTIONS OF Y: Let Y be a discrete random variable with pmf $p_Y(y)$ and support R . Suppose that g is a real-valued function. Then, $g(Y)$ is a random variable and

$$E[g(Y)] = \sum_{y \in R} g(y)p_Y(y).$$

The proof of this result is given on pp 93 (WMS). Again, we require that

$$\sum_{y \in R} |g(y)|p_Y(y) < \infty.$$

If this is not true, then $E[g(Y)]$ does not exist.

Example 3.6. In Example 3.5, find $E(Y^2)$ and $E(e^Y)$.

SOLUTION: The functions $g_1(Y) = Y^2$ and $g_2(Y) = e^Y$ are real functions of Y . From the definition, we have

$$\begin{aligned} E(Y^2) &= \sum_{y \in R} y^2 p_Y(y) \\ &= \sum_{y=1}^4 y^2 \left[\frac{1}{10}(5-y) \right] = 1^2(4/10) + 2^2(3/10) + 3^2(2/10) + 4^2(1/10) = 5. \end{aligned}$$

Also,

$$\begin{aligned} E(e^Y) &= \sum_{y \in R} e^y p_Y(y) \\ &= \sum_{y=1}^4 e^y \left[\frac{1}{10}(5-y) \right] = e^1(4/10) + e^2(3/10) + e^3(2/10) + e^4(1/10) \approx 12.78. \quad \square \end{aligned}$$

Example 3.7. *The discrete uniform distribution.* Suppose that the random variable X has pmf

$$p_X(x) = \begin{cases} 1/m, & x = 1, 2, \dots, m \\ 0, & \text{otherwise,} \end{cases}$$

where m is a positive integer larger than 1. Find the expected value of X .

SOLUTION. The expected value of X is given by

$$E(X) = \sum_{x \in R} xp_X(x) = \sum_{x=1}^m x \left(\frac{1}{m} \right) = \frac{1}{m} \sum_{x=1}^m x = \frac{1}{m} \left[\frac{m(m+1)}{2} \right] = \frac{m+1}{2}.$$

We have used the well-known fact that $\sum_{x=1}^m x = m(m+1)/2$; this can be proven by induction. If $m = 6$, then the discrete uniform distribution serves as a probability model for the outcome of an unbiased die:

x	1	2	3	4	5	6
$p_X(x)$	1/6	1/6	1/6	1/6	1/6	1/6

The expected value of X is $E(X) = (6 + 1)/2 = 3.5$. \square

PROPERTIES OF EXPECTATIONS: Let Y be a discrete random variable with pmf $p_Y(y)$ and support R . Suppose that g, g_1, g_2, \dots, g_k are real-valued functions, and let c be any real constant. Expectations satisfy the following (linearity) properties:

- (a) $E(c) = c$
- (b) $E[cg(Y)] = cE[g(Y)]$
- (c) $E[\sum_{j=1}^k g_j(Y)] = \sum_{j=1}^k E[g_j(Y)]$.

Example 3.8. In a one-hour period, the number of gallons of a certain toxic chemical that is produced at a local plant, say Y , has the following pmf:

y	0	1	2	3
$p_Y(y)$	0.2	0.3	0.3	0.2

- (a) Compute the expected number of gallons produced during a one-hour period.
- (b) The cost (in hundreds of dollars) to produce Y gallons is given by the cost function $C(Y) = 3 + 12Y + 2Y^2$. What is the expected cost in a one-hour period?

SOLUTION: (a) The expected value of Y is

$$E(Y) = \sum_{y \in R} yp_Y(y) = 0(0.2) + 1(0.3) + 2(0.3) + 3(0.2) = 1.5.$$

That is, we would expect 1.5 gallons of the toxic chemical to be produced per hour. For (b), we first compute $E(Y^2)$:

$$E(Y^2) = \sum_{y \in R} y^2 p_Y(y) = 0^2(0.2) + 1^2(0.3) + 2^2(0.3) + 3^2(0.2) = 3.3.$$

Finally,

$$\begin{aligned} E[C(Y)] &= E(3 + 12Y + 2Y^2) \\ &= 3 + 12E(Y) + 2E(Y^2) = 3 + 12(1.5) + 2(3.3) = 27.6. \end{aligned}$$

The expected hourly cost is \$2,760.00. \square

3.4 Variance

TERMINOLOGY: Let Y be a discrete random variable with pmf $p_Y(y)$, support R , and expected value $E(Y) = \mu$. The **variance** of Y is given by

$$\sigma^2 \equiv V(Y) \equiv E[(Y - \mu)^2] = \sum_{y \in R} (y - \mu)^2 p_Y(y).$$

The **standard deviation** of Y is given by the positive square root of the variance; i.e.,

$$\sigma = \sqrt{\sigma^2} = \sqrt{V(Y)}.$$

FACTS: The variance σ^2 satisfies the following:

- (a) $\sigma^2 \geq 0$.

- (b) $\sigma^2 = 0$ if and only if the random variable Y has a **degenerate distribution**; i.e., all the probability mass is located at one support point.
- (c) The larger (smaller) σ^2 is, the more (less) spread in the possible values of Y about the mean $\mu = E(Y)$.
- (d) σ^2 is measured in (units)² and σ is measured in the original units.

VARIANCE COMPUTING FORMULA: Let Y be a random variable with (finite) mean $E(Y) = \mu$. Then

$$V(Y) = E[(Y - \mu)^2] = E(Y^2) - [E(Y)]^2.$$

Proof. Expand the $(Y - \mu)^2$ term and distribute the expectation operator as follows:

$$\begin{aligned} E[(Y - \mu)^2] &= E(Y^2 - 2\mu Y + \mu^2) \\ &= E(Y^2) - 2\mu E(Y) + \mu^2 \\ &= E(Y^2) - 2\mu^2 + \mu^2 \\ &= E(Y^2) - \mu^2. \quad \square \end{aligned}$$

Example 3.9. *The discrete uniform distribution.* Suppose that the random variable X has pmf

$$p_X(x) = \begin{cases} 1/m, & x = 1, 2, \dots, m \\ 0, & \text{otherwise,} \end{cases}$$

where m is a positive integer larger than 1. Find the variance of X .

SOLUTION. We find $\sigma^2 = V(X)$ using the variance computing formula. In Example 3.7, we computed

$$\mu = E(X) = \frac{m+1}{2}.$$

We first find $E(X^2)$:

$$\begin{aligned} E(X^2) &= \sum_{x \in R} x^2 p_X(x) = \sum_{x=1}^m x^2 \left(\frac{1}{m}\right) = \frac{1}{m} \sum_{x=1}^m x^2 = \frac{1}{m} \left[\frac{m(m+1)(2m+1)}{6} \right] \\ &= \frac{(m+1)(2m+1)}{6}. \end{aligned}$$

We have used the well-known fact that $\sum_{x=1}^m x^2 = m(m+1)(2m+1)/6$; this can be proven by induction. The variance of X is equal to

$$\begin{aligned}\sigma^2 &= E(X^2) - [E(X)]^2 \\ &= \frac{(m+1)(2m+1)}{6} - \left(\frac{m+1}{2}\right)^2 = \frac{m^2-1}{12}. \quad \square\end{aligned}$$

EXERCISE: Find $\sigma^2 = V(Y)$ in Examples 3.5 and 3.8 (notes).

IMPORTANT RESULT: Let Y be a random variable (not necessarily a discrete random variable). Suppose that a and b are fixed constants. Then

$$V(a + bY) = b^2V(Y).$$

REMARK: Taking $b = 0$ above, we see that $V(a) = 0$, for any constant a . This makes sense intuitively. The variance is a measure of variability for a random variable; a constant (such as a) does not vary. Also, by taking $a = 0$, we see that $V(bY) = b^2V(Y)$.

3.5 Moment generating functions

TERMINOLOGY: Let Y be a discrete random variable with pmf $p_Y(y)$ and support R . The **moment generating function (mgf)** for Y , denoted by $m_Y(t)$, is given by

$$m_Y(t) = E(e^{tY}) = \sum_{y \in R} e^{ty} p_Y(y),$$

provided $E(e^{tY}) < \infty$ for all t in an open neighborhood about 0; i.e., there exists some $h > 0$ such that $E(e^{tY}) < \infty$ for all $t \in (-h, h)$. If $E(e^{tY})$ does not exist in an open neighborhood of 0, we say that the moment generating function does not exist.

TERMINOLOGY: We call $\mu'_k \equiv E(Y^k)$ the **k th moment** of the random variable Y :

$$\begin{array}{ll} E(Y) & \text{1st moment (mean!)} \\ E(Y^2) & \text{2nd moment} \\ E(Y^3) & \text{3rd moment} \\ E(Y^4) & \text{4th moment} \\ \vdots & \vdots \end{array}$$

REMARK: The moment generating function (mgf) can be used to generate moments. In fact, from the theory of Laplace transforms, it follows that if the mgf exists, it characterizes an infinite set of moments. So, how do we generate moments?

RESULT: Let Y denote a random variable (not necessarily a discrete random variable) with support R and mgf $m_Y(t)$. Then,

$$E(Y^k) = \left. \frac{d^k m_Y(t)}{dt^k} \right|_{t=0}.$$

Note that derivatives are taken with respect to t .

Proof. Assume, without loss, that Y is discrete. With $k = 1$, we have

$$\frac{d}{dt} m_Y(t) = \frac{d}{dt} \sum_{y \in R} e^{ty} p_Y(y) = \sum_{y \in R} \frac{d}{dt} e^{ty} p_Y(y) = \sum_{y \in R} y e^{ty} p_Y(y) = E(Y e^{tY}).$$

Thus,

$$\left. \frac{dm_Y(t)}{dt} \right|_{t=0} = E(Y e^{tY}) \Big|_{t=0} = E(Y).$$

Continuing to take higher-order derivatives, we can prove that

$$\left. \frac{d^k m_Y(t)}{dt^k} \right|_{t=0} = E(Y^k),$$

for any integer $k \geq 1$. See pp 139-140 (WMS) for a slightly different proof. \square

ASIDE: In the proof of the last result, we interchanged the derivative and (possibly infinite) sum. This is permitted as long as $m_Y(t) = E(e^{tY})$ exists.

MEANS AND VARIANCES: Suppose that Y is a random variable (not necessarily a discrete random variable) with mgf $m_Y(t)$. We know that

$$E(Y) = \left. \frac{dm_Y(t)}{dt} \right|_{t=0}$$

and

$$E(Y^2) = \left. \frac{d^2 m_Y(t)}{dt^2} \right|_{t=0}.$$

We can get $V(Y)$ using $V(Y) = E(Y^2) - [E(Y)]^2$.

REMARK: Being able to find means and variances is important in mathematical statistics. **Thus, we can use the mgf as a tool to do this.** This is helpful because sometimes computing

$$E(Y) = \sum_{y \in R} yp_Y(y)$$

directly (or even higher order moments) may be extremely difficult, depending on the form of $p_Y(y)$.

Example 3.10. Suppose that Y is a random variable with pmf

$$p_Y(y) = \begin{cases} \left(\frac{1}{2}\right)^y, & y = 1, 2, 3, \dots \\ 0, & \text{otherwise.} \end{cases}$$

Find the mean of Y .

SOLUTION. Using the definition of expected value, the mean of Y is given by

$$E(Y) = \sum_{y \in R} yp_Y(y) = \sum_{y=1}^{\infty} y \left(\frac{1}{2}\right)^y.$$

Finding this infinite sum is not obvious (at least, this sum is not a geometric sum).

Another option is to use moment generating functions! The mgf of Y is given by

$$\begin{aligned} m_Y(t) = E(e^{tY}) &= \sum_{y \in R} e^{ty} p_Y(y) \\ &= \sum_{y=1}^{\infty} e^{ty} \left(\frac{1}{2}\right)^y = \sum_{y=1}^{\infty} \left(\frac{e^t}{2}\right)^y = \left[\sum_{y=0}^{\infty} \left(\frac{e^t}{2}\right)^y \right] - 1. \end{aligned}$$

The series $\sum_{y=0}^{\infty} (e^t/2)^y$ is an infinite geometric sum with common ratio $r = e^t/2$. This series converges as long as $e^t/2 < 1$, in which case

$$m_Y(t) = \frac{1}{1 - \frac{e^t}{2}} - 1 = \frac{e^t}{2 - e^t},$$

for $e^t/2 < 1 \iff t < \ln 2$. Note that $(-h, h)$ with $h = \ln 2$ is an open neighborhood around zero for which $m_Y(t)$ exists. Now,

$$\begin{aligned} E(Y) &= \left. \frac{dm_Y(t)}{dt} \right|_{t=0} = \left. \frac{d}{dt} \left(\frac{e^t}{2 - e^t} \right) \right|_{t=0} \\ &= \left. \frac{e^t(2 - e^t) - e^t(-e^t)}{(2 - e^t)^2} \right|_{t=0} = 2. \quad \square \end{aligned}$$

Example 3.11. Let the random variable Y have pmf $p_Y(y)$ given by

$$p_Y(y) = \begin{cases} \frac{1}{6}(3-y), & y = 0, 1, 2 \\ 0, & \text{otherwise.} \end{cases}$$

Simple calculations show that $E(Y) = 2/3$ and $V(Y) = 5/9$ (verify!). Let's "check" these calculations using the mgf of Y . It is given by

$$\begin{aligned} m_Y(t) = E(e^{tY}) &= \sum_{y \in R} e^{ty} p_Y(y) \\ &= e^{t(0)} \frac{3}{6} + e^{t(1)} \frac{2}{6} + e^{t(2)} \frac{1}{6} \\ &= \frac{3}{6} + \frac{2}{6}e^t + \frac{1}{6}e^{2t}. \end{aligned}$$

Taking derivatives of $m_Y(t)$ with respect to t , we get

$$\begin{aligned} \frac{d}{dt} m_Y(t) &= \frac{2}{6}e^t + \frac{2}{6}e^{2t} \\ \frac{d^2}{dt^2} m_Y(t) &= \frac{2}{6}e^t + \frac{4}{6}e^{2t}. \end{aligned}$$

Thus,

$$\begin{aligned} E(Y) &= \left. \frac{dm_Y(t)}{dt} \right|_{t=0} = \frac{2}{6}e^0 + \frac{2}{6}e^{2(0)} = 4/6 = 2/3 \\ E(Y^2) &= \left. \frac{d^2 m_Y(t)}{dt^2} \right|_{t=0} = \frac{2}{6}e^0 + \frac{4}{6}e^{2(0)} = 1 \end{aligned}$$

so that

$$V(Y) = E(Y^2) - [E(Y)]^2 = 1 - (2/3)^2 = 5/9.$$

In this example, it is easier to compute $E(Y)$ and $V(Y)$ directly (using the definition). However, it nice to see that we get the same answer using the mgf approach. \square

REMARK: Not only is the mgf a tool for computing moments, but it also helps us to characterize a probability distribution. How? When an mgf exists, it happens to be unique. This means that if two random variables have same mgf, then they have the same probability distribution! This is called the **uniqueness property** of mgfs (it is based on the uniqueness of Laplace transforms). For now, however, it suffices to envision the mgf as a "special expectation" that generates moments. This, in turn, helps us to compute means and variances of random variables.

3.6 Binomial distribution

BERNOULLI TRIALS: Many processes can be envisioned as consisting of a sequence of “trials,” where

- (i) each trial results in a “success” or a “failure,”
- (ii) the trials are **independent**, and
- (iii) the probability of “success,” denoted by p , $0 < p < 1$, is the same on every trial.

TERMINOLOGY: In a sequence of n Bernoulli trials, denote by Y the number of successes out of n (where n is fixed). We say that Y has a **binomial distribution** with number of trials n and success probability p . Shorthand notation is $Y \sim b(n, p)$.

Example 3.12. Each of the following situations could be conceptualized as a binomial experiment. Are you satisfied with the Bernoulli assumptions in each instance?

- (a) We flip a fair coin 10 times and let Y denote the number of tails in 10 flips. Here, $Y \sim b(n = 10, p = 0.5)$.
- (b) Forty percent of all plots of land respond to a certain treatment. I have four plots to be treated. If Y is the number of plots that respond to the treatment, then $Y \sim b(n = 4, p = 0.4)$.
- (c) In rural Kenya, the prevalence rate for HIV is estimated to be around 8 percent. Let Y denote the number of HIV infecteds in a sample of 740 individuals. Here, $Y \sim b(n = 740, p = 0.08)$.
- (d) Parts produced by a certain company do not meet specifications (i.e., are defective) with probability 0.001. Let Y denote the number of defective parts in a package of 40. Then, $Y \sim b(n = 40, p = 0.001)$. \square

DERIVATION: We now derive the pmf of a binomial random variable. The support of Y is $R = \{y : y = 0, 1, 2, \dots, n\}$. We need to find an expression for $p_Y(y) = P(Y = y)$ for each value of $y \in R$.

QUESTION: In a sequence of n trials, how can we get exactly y successes? Denoting “success” and “failure” by S and F , respectively, one possible sample point might be

$$SSFSFSFFS \cdots FSF.$$

Because the trials are **independent**, the probability that we get a particular ordering of y successes and $n - y$ failures is $p^y(1 - p)^{n-y}$. Furthermore, there are $\binom{n}{y}$ sample points that contain exactly y successes. Thus, we add the term $p^y(1 - p)^{n-y}$ a total of $\binom{n}{y}$ times to get $P(Y = y)$. The pmf for Y is, for $0 < p < 1$,

$$p_Y(y) = \begin{cases} \binom{n}{y} p^y (1 - p)^{n-y}, & y = 0, 1, 2, \dots, n \\ 0, & \text{otherwise.} \end{cases}$$

Example 3.13. In Example 3.12(b), assume that $Y \sim b(n = 4, p = 0.4)$. Here are the probability calculations for this binomial model:

$$P(Y = 0) = p_Y(0) = \binom{4}{0} (0.4)^0 (1 - 0.4)^{4-0} = 1 \times (0.4)^0 \times (0.6)^4 = 0.1296$$

$$P(Y = 1) = p_Y(1) = \binom{4}{1} (0.4)^1 (1 - 0.4)^{4-1} = 4 \times (0.4)^1 \times (0.6)^3 = 0.3456$$

$$P(Y = 2) = p_Y(2) = \binom{4}{2} (0.4)^2 (1 - 0.4)^{4-2} = 6 \times (0.4)^2 \times (0.6)^2 = 0.3456$$

$$P(Y = 3) = p_Y(3) = \binom{4}{3} (0.4)^3 (1 - 0.4)^{4-3} = 4 \times (0.4)^3 \times (0.6)^1 = 0.1536$$

$$P(Y = 4) = p_Y(4) = \binom{4}{4} (0.4)^4 (1 - 0.4)^{4-4} = 1 \times (0.4)^4 \times (0.6)^0 = 0.0256.$$

EXERCISE: What is the probability that at least 2 plots respond? at most one? What are $E(Y)$ and $V(Y)$? \square

Example 3.14. In a small clinical trial with 20 patients, let Y denote the number of patients that respond to a new skin rash treatment. The physicians assume that a binomial model is appropriate and that $Y \sim b(n = 20, p = 0.4)$. Under this model, compute (a) $P(Y = 5)$, (b) $P(Y \geq 5)$, and (c) $P(Y < 10)$.

$$(a) P(Y = 5) = p_Y(5) = \binom{20}{5} (0.4)^5 (0.6)^{20-5} = 0.0746.$$

(b)

$$P(Y \geq 5) = \sum_{y=5}^{20} P(Y = y) = \sum_{y=5}^{20} \binom{20}{y} (0.4)^y (0.6)^{20-y}.$$

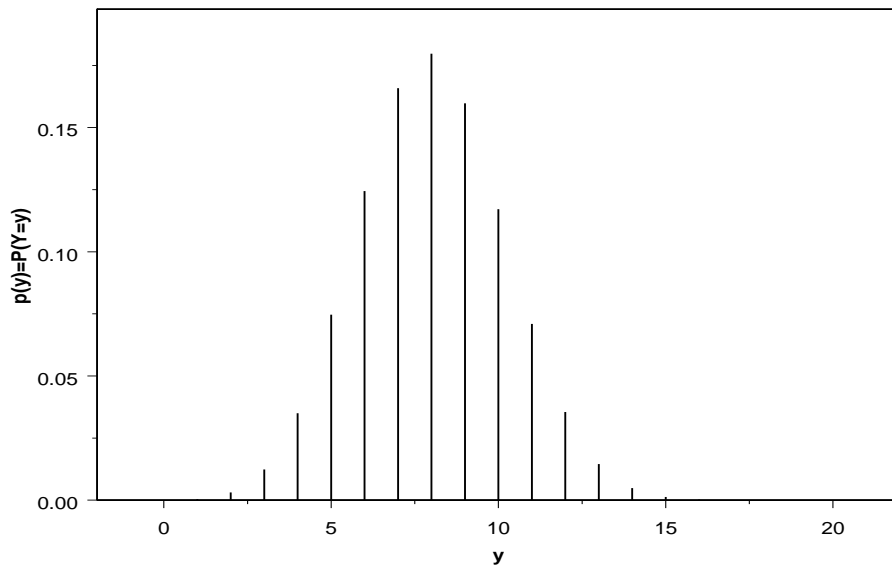


Figure 3.2: Probability histogram for the number of patients responding to treatment. This represents the $b(n = 20, p = 0.4)$ model in Example 3.14.

This calculation involves using the binomial pmf 16 times and adding the results!

TRICK: Instead of computing the sum $\sum_{y=5}^{20} \binom{20}{y} (0.4)^y (0.6)^{20-y}$ directly, we can write

$$P(Y \geq 5) = 1 - P(Y \leq 4),$$

by the complement rule. We do this because WMS's Appendix III (Table 1, pp 839-841) contains binomial probability calculations of the form

$$P(Y \leq a) = \sum_{y=0}^a \binom{n}{y} p^y (1-p)^{n-y},$$

for different n and p . With $n = 20$ and $p = 0.4$, we see from Table 1 that

$$P(Y \leq 4) = 0.051.$$

Thus, $P(Y \geq 5) = 1 - 0.051 = 0.949$.

(c) $P(Y < 10) = P(Y \leq 9) = 0.755$, from Table 1. \square

REMARK: The function $P(Y \leq y)$ is called the **cumulative distribution function** of a random variable Y ; we'll talk more about this function in the next chapter.

RECALL: The **binomial expansion** of $(a + b)^n$ is given by

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

CURIOSITY: Is the binomial pmf a **valid** pmf? Clearly $p_Y(y) > 0$ for all y . To check that the pmf sums to one, consider the binomial expansion

$$[(1 - p) + p]^n = \sum_{y=0}^n \binom{n}{y} p^y (1 - p)^{n-y}.$$

The LHS clearly equals 1, and the RHS is the $b(n, p)$ pmf. Thus, $p_Y(y)$ is valid. \square

BINOMIAL MGF: Suppose that $Y \sim b(n, p)$. The mgf of Y is given by

$$m_Y(t) = E(e^{tY}) = \sum_{y=0}^n e^{ty} \binom{n}{y} p^y (1 - p)^{n-y} = \sum_{y=0}^n \binom{n}{y} (pe^t)^y (1 - p)^{n-y} = (q + pe^t)^n,$$

where $q = 1 - p$. The last step follows from noting that $\sum_{y=0}^n \binom{n}{y} (pe^t)^y (1 - p)^{n-y}$ is the binomial expansion of $(q + pe^t)^n$. \square

MEAN AND VARIANCE: We want to compute $E(Y)$ and $V(Y)$ where $Y \sim b(n, p)$. We will use the mgf. Taking the derivative of $m_Y(t)$ with respect t , we get

$$m'_Y(t) \equiv \frac{d}{dt} m_Y(t) = \frac{d}{dt} (q + pe^t)^n = n(q + pe^t)^{n-1} pe^t.$$

Thus,

$$E(Y) = \left. \frac{d}{dt} m_Y(t) \right|_{t=0} = n(q + pe^0)^{n-1} pe^0 = n(q + p)^{n-1} p = np,$$

since $q + p = 1$. Now, we need to find the second moment. By using the product rule for derivatives, we have

$$\frac{d^2}{dt^2} m_Y(t) = \frac{d}{dt} \underbrace{n(q + pe^t)^{n-1} pe^t}_{m'_Y(t)} = n(n-1)(q + pe^t)^{n-2} (pe^t)^2 + n(q + pe^t)^{n-1} pe^t.$$

Thus,

$$E(Y^2) = \left. \frac{d^2}{dt^2} m_Y(t) \right|_{t=0} = n(n-1)(q + pe^0)^{n-2} (pe^0)^2 + n(q + pe^0)^{n-1} pe^0 = n(n-1)p^2 + np.$$

Appealing to the variance computing formula, we have

$$V(Y) = E(Y^2) - [E(Y)]^2 = n(n-1)p^2 + np - (np)^2 = np(1-p).$$

NOTE: WMS derive the binomial mean and variance using a different approach (not using the mgf). See pp 107-108. \square

Example 3.15. Artichokes are a marine climate vegetable and thrive in the cooler coastal climates. Most will grow in a wide range of soils, but produce best on a deep, fertile, well-drained soil. Suppose that 15 artichoke seeds are planted in identical soils and temperatures, and let Y denote the number of seeds that germinate. If 60 percent of all seeds germinate (on average) and we assume a $b(15, 0.6)$ probability model for Y , the mean number of seeds that will germinate is

$$E(Y) = \mu = np = 15(0.6) = 9.$$

The variance of Y is

$$V(Y) = \sigma^2 = np(1-p) = 15(0.6)(0.4) = 3.6 \text{ (seeds)}^2.$$

The standard deviation of Y is $\sigma = \sqrt{3.6} \approx 1.9$ seeds. \square

BERNOULLI DISTRIBUTION: In the $b(n, p)$ family, when $n = 1$, the binomial pmf reduces to

$$p_Y(y) = \begin{cases} p^y(1-p)^{1-y}, & y = 0, 1 \\ 0, & \text{otherwise.} \end{cases}$$

This is called the **Bernoulli distribution**. Shorthand notation is $Y \sim b(1, p)$ or $Y \sim \text{Bern}(p)$.

3.7 Geometric distribution

TERMINOLOGY: Envision an experiment where Bernoulli trials are observed. If Y denotes the trial on which the first success occurs, then Y is said to follow a **geometric distribution** with parameter p , where p is the probability of success on any one trial.

GEOMETRIC PMF: The pmf for $Y \sim \text{geom}(p)$ is given by

$$p_Y(y) = \begin{cases} (1-p)^{y-1}p, & y = 1, 2, 3, \dots \\ 0, & \text{otherwise.} \end{cases}$$

RATIONALE: The form of this pmf makes intuitive sense; we first need $y - 1$ failures (each of which occurs with probability $1 - p$), and then a success on the y th trial (this occurs with probability p). By independence, we multiply

$$\underbrace{(1-p) \times (1-p) \times \cdots \times (1-p)}_{y-1 \text{ failures}} \times p = (1-p)^{y-1}p.$$

NOTE: Clearly $p_Y(y) > 0$ for all y . Does $p_Y(y)$ sum to one? Note that

$$\sum_{y=1}^{\infty} (1-p)^{y-1}p = p \sum_{x=0}^{\infty} (1-p)^x = \frac{p}{1-(1-p)} = 1.$$

In the last step, we realized that $\sum_{x=0}^{\infty} (1-p)^x$ is an infinite geometric sum with common ratio $1 - p$. \square

Example 3.16. Biology students are checking the eye color of fruit flies. For each fly, the probability of observing white eyes is $p = 0.25$. What is the probability the first white-eyed fly will be observed among the first five flies that are checked?

SOLUTION: Let Y denote the number of flies needed to observe the first white-eyed fly. We can envision each fly as a Bernoulli trial (each fly either has white eyes or not). If we assume that the flies are independent, then a geometric model is appropriate; i.e., $Y \sim \text{geom}(p = 0.25)$. We want to compute $P(Y \leq 5)$. We use the pmf to compute

$$\begin{aligned} P(Y = 1) &= p_Y(1) = (1 - 0.25)^{1-1}(0.25) = 0.25 \\ P(Y = 2) &= p_Y(2) = (1 - 0.25)^{2-1}(0.25) \approx 0.19 \\ P(Y = 3) &= p_Y(3) = (1 - 0.25)^{3-1}(0.25) \approx 0.14 \\ P(Y = 4) &= p_Y(4) = (1 - 0.25)^{4-1}(0.25) \approx 0.11 \\ P(Y = 5) &= p_Y(5) = (1 - 0.25)^{5-1}(0.25) \approx 0.08. \end{aligned}$$

Adding these probabilities, we get $P(Y \leq 5) \approx 0.77$. The pmf for the $\text{geom}(p = 0.25)$ model is depicted in Figure 3.3. \square

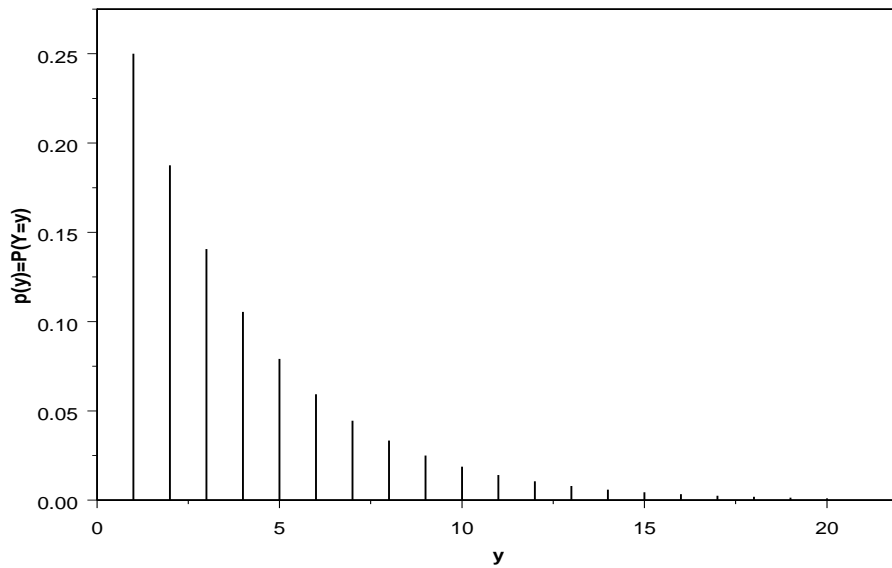


Figure 3.3: Probability histogram for the number of flies needed to find the first white-eyed fly. This represents the $\text{geom}(p = 0.25)$ model in Example 3.16.

GEOMETRIC MGF: Suppose that $Y \sim \text{geom}(p)$. The mgf of Y is given by

$$m_Y(t) = \frac{pe^t}{1 - qe^t},$$

where $q = 1 - p$, for $t < -\ln q$.

Proof. Exercise. \square

MEAN AND VARIANCE: Differentiating the mgf, we get

$$\frac{d}{dt}m_Y(t) = \frac{d}{dt} \left(\frac{pe^t}{1 - qe^t} \right) = \frac{pe^t(1 - qe^t) - pe^t(-qe^t)}{(1 - qe^t)^2}.$$

Thus,

$$E(Y) = \left. \frac{d}{dt}m_Y(t) \right|_{t=0} = \frac{pe^0(1 - qe^0) - pe^0(-qe^0)}{(1 - qe^0)^2} = \frac{p(1 - q) - p(-q)}{(1 - q)^2} = \frac{1}{p}.$$

Similar (but lengthier) calculations show

$$E(Y^2) = \left. \frac{d^2}{dt^2}m_Y(t) \right|_{t=0} = \frac{1 + q}{p^2}.$$

Finally,

$$V(Y) = E(Y^2) - [E(Y)]^2 = \frac{1+q}{p^2} - \left(\frac{1}{p}\right)^2 = \frac{q}{p^2}. \quad \square$$

NOTE: WMS derive the geometric mean and variance using a different approach (not using the mgf). See pp 116-117. \square

Example 3.17. At an orchard in Maine, “20-lb” bags of apples are weighed. Suppose that four percent of the bags are underweight and that each bag weighed is independent. If Y denotes the the number of bags observed to find the first underweight bag, then $Y \sim \text{geom}(p = 0.04)$. The mean of Y is

$$E(Y) = \frac{1}{p} = \frac{1}{0.04} = 25 \text{ bags.}$$

The variance of Y is

$$V(Y) = \frac{q}{p^2} = \frac{0.96}{(0.04)^2} = 600 \text{ (bags)}^2. \quad \square$$

3.8 Negative binomial distribution

NOTE: The negative binomial distribution can be motivated from two perspectives:

- as a generalization of the geometric
- as an “inverse” version of the binomial.

TERMINOLOGY: Imagine an experiment where Bernoulli trials are observed. If Y denotes the trial on which the r th success occurs, $r \geq 1$, then Y has a **negative binomial distribution** with waiting parameter r and probability of success p .

NEGATIVE BINOMIAL PMF: The pmf for $Y \sim \text{nib}(r, p)$ is given by

$$p_Y(y) = \begin{cases} \binom{y-1}{r-1} p^r (1-p)^{y-r}, & y = r, r+1, r+2, \dots \\ 0, & \text{otherwise.} \end{cases}$$

Of course, when $r = 1$, the $\text{nib}(r, p)$ pmf reduces to the $\text{geom}(p)$ pmf.

RATIONALE: The form of $p_Y(y)$ can be explained intuitively. If the r th success occurs on the y th trial, then $r - 1$ successes must have occurred during the 1st $y - 1$ trials. The total number of sample points (in the underlying sample space S) where this occurs is given by the binomial coefficient $\binom{y-1}{r-1}$, which counts the number of ways you can choose the locations of $r - 1$ successes in a string of the 1st $y - 1$ trials. The probability of any particular such ordering, by independence, is given by $p^{r-1}(1 - p)^{y-r}$. Thus, the probability of getting exactly $r - 1$ successes in the $y - 1$ trials is $\binom{y-1}{r-1}p^{r-1}(1 - p)^{y-r}$. On the y th trial, we observe the r th success (this occurs with probability p). Because the y th trial is independent of the previous $y - 1$ trials, we have

$$P(Y = y) = \underbrace{\binom{y-1}{r-1}p^{r-1}(1-p)^{y-r}}_{\text{pertains to 1st } y-1 \text{ trials}} \times p = \binom{y-1}{r-1}p^r(1-p)^{y-r}.$$

Example 3.18. A botanist is observing oak trees for the presence of a certain disease. From past experience, it is known that 30 percent of all trees are infected ($p = 0.30$). Treating each tree as a Bernoulli trial (i.e., each tree is infected/not), what is the probability that she will observe the 3rd infected tree ($r = 3$) on the 6th or 7th observed tree? *SOLUTION.* Let Y denote the tree on which she observes the 3rd infected tree. Then, $Y \sim \text{nib}(r = 3, p = 0.3)$. We want to compute $P(Y = 6 \text{ or } Y = 7)$. The $\text{nib}(3, 0.3)$ pmf gives

$$p_Y(6) = P(Y = 6) = \binom{6-1}{3-1}(0.3)^3(1-0.3)^{6-3} = 0.0926$$

$$p_Y(7) = P(Y = 7) = \binom{7-1}{3-1}(0.3)^3(1-0.3)^{7-3} = 0.0972.$$

Thus,

$$P(Y = 6 \text{ or } Y = 7) = P(Y = 6) + P(Y = 7) = 0.0926 + 0.0972 = 0.1898. \quad \square$$

RELATIONSHIP WITH THE BINOMIAL: Recall that in a binomial experiment, we fix the number of Bernoulli trials, n , and we observe the number of successes. In a negative binomial experiment, we fix the number of successes we are to observe, r , and we continue to observe Bernoulli trials until we reach that numbered success. In this sense, the negative binomial distribution is the “inverse” of the binomial distribution.

RECALL: Suppose that the real function $f(x)$ is infinitely differentiable at $x = a$. The **Taylor series expansion** of $f(x)$ about the point $x = a$ is given by

$$\begin{aligned} f(x) &= \sum_{n=0}^{\infty} \frac{f^{(n)}(a)}{n!} (x-a)^n \\ &= f(a) + \left[\frac{f'(a)}{1!} \right] (x-a)^1 + \left[\frac{f''(a)}{2!} \right] (x-a)^2 + \dots \end{aligned}$$

When $a = 0$, this is called the **McLaurin series expansion** of $f(x)$.

NEGATIVE BINOMIAL MGF: Suppose that $Y \sim \text{nib}(r, p)$. The mgf of Y is given by

$$m_Y(t) = \left(\frac{pe^t}{1-qe^t} \right)^r,$$

where $q = 1 - p$, for all $t < -\ln q$. Before we prove this, let's state and prove a lemma.

LEMMA. Suppose that r is a nonnegative integer. Then,

$$\sum_{y=r}^{\infty} \binom{y-1}{r-1} (qe^t)^{y-r} = (1-qe^t)^{-r}.$$

Proof of lemma. Consider the function $f(w) = (1-w)^{-r}$, where r is a nonnegative integer. It is easy to show that

$$\begin{aligned} f'(w) &= r(1-w)^{-(r+1)} \\ f''(w) &= r(r+1)(1-w)^{-(r+2)} \\ &\vdots \end{aligned}$$

In general, $f^{(z)}(w) = r(r+1)\cdots(r+z-1)(1-w)^{-(r+z)}$, where $f^{(z)}(w)$ denotes the z th derivative of f with respect to w . Note that

$$f^{(z)}(w) \Big|_{w=0} = r(r+1)\cdots(r+z-1).$$

Now, consider writing the McLaurin Series expansion of $f(w)$; i.e., a Taylor Series expansion of $f(w)$ about $w = 0$; this expansion is given by

$$f(w) = \sum_{z=0}^{\infty} \frac{f^{(z)}(0)w^z}{z!} = \sum_{z=0}^{\infty} \frac{r(r+1)\cdots(r+z-1)}{z!} w^z = \sum_{z=0}^{\infty} \binom{z+r-1}{r-1} w^z.$$

Letting $w = qe^t$ and $z = y - r$, the lemma is proven for $0 < q < 1$. \square

Now that we are finished with the lemma, let's find the mgf of $Y \sim \text{nib}(r, p)$. With $q = 1 - p$, we have

$$\begin{aligned} m_Y(t) = E(e^{tY}) &= \sum_{y=r}^{\infty} e^{ty} \binom{y-1}{r-1} p^r q^{y-r} \\ &= \sum_{y=r}^{\infty} e^{t(y-r)} e^{tr} \binom{y-1}{r-1} p^r q^{y-r} \\ &= (pe^t)^r \sum_{y=r}^{\infty} \binom{y-1}{r-1} (qe^t)^{y-r} = (pe^t)^r (1 - qe^t)^{-r}. \quad \square \end{aligned}$$

REMARK: Showing that the $\text{nib}(r, p)$ pmf sums to one can be done by using a similar series expansion as above. We omit it for brevity.

MEAN AND VARIANCE: For $Y \sim \text{nib}(r, p)$, with $q = 1 - p$,

$$E(Y) = \frac{r}{p} \quad \text{and} \quad V(Y) = \frac{rq}{p^2}.$$

3.9 Hypergeometric distribution

SETTING: Consider a collection of N objects (e.g., people, poker chips, plots of land, etc.) and suppose that we have two dichotomous classes, Class 1 and Class 2. For example, the objects and classes might be

Poker chips	red/blue
People	infected/not infected
Plots of land	respond to treatment/not.

From the collection of N objects, we sample n of them (without replacement), and record Y , the number of objects in Class 1.

REMARK: This sounds like a binomial setup! However, the difference here is that N , the **population size**, is finite (the population size, theoretically, is assumed to be infinite in the binomial model). Thus, if we sample from a population of objects **without replacement**, the “success” probability changes from trial to trial. This, violates the binomial

model assumptions! If N is large (i.e., in a very large population), the hypergeometric and binomial models will be similar, because the change in the probability of success from trial to trial will be small (maybe so small that it is not of practical concern).

HYPERGEOMETRIC DISTRIBUTION: Envision a collection of n objects sampled (at random and without replacement) from a population of size N , where r denotes the size of Class 1 and $N - r$ denotes the size of Class 2. Let Y denote the number of objects in the sample that belong to Class 1. Then, Y has a **hypergeometric distribution**, written $Y \sim \text{hyper}(N, n, r)$, where

$$\begin{aligned} N &= \text{total number of objects} \\ r &= \text{number of the 1st class (e.g., "success")} \\ N - r &= \text{number of the 2nd class (e.g., "failure")} \\ n &= \text{number of objects sampled.} \end{aligned}$$

HYPERGEOMETRIC PMF: The pmf for $Y \sim \text{hyper}(N, n, r)$ is given by

$$p_Y(y) = \begin{cases} \frac{\binom{r}{y} \binom{N-r}{n-y}}{\binom{N}{n}}, & y \in R \\ 0, & \text{otherwise,} \end{cases}$$

where the support set $R = \{y \in \mathcal{N} : \max(0, n - N + r) \leq y \leq \min(n, r)\}$.

BREAKDOWN: In the $\text{hyper}(N, n, r)$ pmf, we have three parts:

$$\begin{aligned} \binom{r}{y} &= \text{number of ways to choose } y \text{ Class 1 objects from } r \\ \binom{N-r}{n-y} &= \text{number of ways to choose } n - y \text{ Class 2 objects from } N - r \\ \binom{N}{n} &= \text{number of sample points.} \end{aligned}$$

REMARK: The hypergeometric pmf $p_Y(y)$ does sum to 1 over the support R , but we omit this proof for brevity (see Exercise 3.216, pp 156, WMS).

Example 3.19. In my fish tank at home, there are 50 fish. Ten have been tagged. If I catch 7 fish (and random, and without replacement), what is the probability that exactly two are tagged?

SOLUTION. Here, $N = 50$ (total number of fish), $n = 7$ (sample size), $r = 10$ (tagged

fish; Class 1), $N - r = 40$ (untagged fish; Class 2), and $y = 2$ (number of tagged fish caught). Thus,

$$P(Y = 2) = p_Y(2) = \frac{\binom{10}{2}\binom{40}{5}}{\binom{50}{7}} = 0.2964.$$

What about the probability that my catch contains at most two tagged fish?

SOLUTION. Here, we want

$$\begin{aligned} P(Y \leq 2) &= P(Y = 0) + P(Y = 1) + P(Y = 2) \\ &= \frac{\binom{10}{0}\binom{40}{7}}{\binom{50}{7}} + \frac{\binom{10}{1}\binom{40}{6}}{\binom{50}{7}} + \frac{\binom{10}{2}\binom{40}{5}}{\binom{50}{7}} \\ &= 0.1867 + 0.3843 + 0.2964 = 0.8674. \quad \square \end{aligned}$$

Example 3.20. A supplier ships parts to a company in lots of 25 parts. The company has an **acceptance sampling plan** which adopts the following acceptance rule:

“...sample 5 parts at random and without replacement. If there are no defectives in the sample, accept the entire lot; otherwise, reject the entire lot.”

Let Y denote the number of defectives in the sample. Then, $Y \sim \text{hyper}(25, 5, r)$, where r denotes the number defectives in the lot (in real life, r would be unknown). Define

$$OC(p) = P(Y = 0) = \frac{\binom{r}{0}\binom{25-r}{5}}{\binom{25}{5}},$$

where $p = r/25$ denotes the true proportion of defectives in the lot. The symbol $OC(p)$ denotes the probability of accepting the lot (which is a function of p). Consider the following table, whose entries are computed using the above probability expression:

r	p	$OC(p)$
0	0	1.00
1	0.04	0.80
2	0.08	0.63
3	0.12	0.50
4	0.16	0.38
5	0.20	0.29
10	0.40	0.06
15	0.60	0.01

REMARK: The graph of $OC(p)$ versus p is called an **operating characteristic curve**. For sensible sampling plans, $OC(p)$ is a decreasing function of p . Acceptance sampling is an important part of **statistical process control**, which is used in engineering and manufacturing settings. \square

MEAN AND VARIANCE: If $Y \sim \text{hyper}(N, n, r)$, then

$$\begin{aligned} E(Y) &= n \left(\frac{r}{N} \right) \\ V(Y) &= n \left(\frac{r}{N} \right) \left(\frac{N-r}{N} \right) \left(\frac{N-n}{N-1} \right). \end{aligned}$$

RELATIONSHIP WITH THE BINOMIAL: The binomial and hypergeometric models are similar. The key difference is that in a binomial experiment, p does not change from trial to trial, but it does in the hypergeometric setting. However, it can be shown that, for y fixed,

$$\lim_{N \rightarrow \infty} \frac{\binom{r}{y} \binom{N-r}{n-y}}{\binom{N}{n}} = \underbrace{\binom{n}{y} p^y (1-p)^{n-y}}_{b(n,p) \text{ pmf}},$$

as $r/N \rightarrow p$. The upshot is this: if N is large (i.e., the population size is large), a binomial probability calculation, with $p = r/N$, closely approximates the corresponding hypergeometric probability calculation.

Example 3.21. In a small town, there are 900 right-handed individuals and 100 left-handed individuals. We take a sample of size $n = 20$ individuals from this town (at random and without replacement). What is the probability that 4 or more people in the sample are left-handed?

SOLUTION. Let X denote the number of left-handed individuals in our sample. We compute the probability $P(X \geq 4)$ using both the binomial and hypergeometric models.

- **Hypergeometric:** Here, $N = 1000$, $r = 100$, $N - r = 900$, and $n = 20$. Thus,

$$P(X \geq 4) = 1 - P(X \leq 3) = 1 - \sum_{x=0}^3 \frac{\binom{100}{x} \binom{900}{20-x}}{\binom{1000}{20}} \approx 0.130947.$$

- **Binomial:** Here, $n = 20$ and $p = r/N = 0.10$. Thus,

$$P(X \geq 4) = 1 - P(X \leq 3) = 1 - \sum_{x=0}^3 \binom{20}{x} (0.1)^x (0.9)^{20-x} \approx 0.132953. \quad \square$$

REMARK: Of course, since the binomial and hypergeometric models are similar when N is large, their means and variances are similar too. Note the similarities; recall that the quantity $r/N \rightarrow p$, as $N \rightarrow \infty$:

$$E(Y) = n \left(\frac{r}{N} \right) \approx np$$

and

$$V(Y) = n \left(\frac{r}{N} \right) \left(\frac{N-r}{N} \right) \left(\frac{N-n}{N-1} \right) \approx np(1-p).$$

3.10 Poisson distribution

TERMINOLOGY: Let the number of occurrences in a given continuous interval of time or space be counted. A **Poisson process** enjoys the following properties:

- (1) the number of occurrences in non-overlapping intervals are **independent** random variables.
- (2) The probability of an occurrence in a sufficiently short interval is **proportional to the length** of the interval.
- (3) The probability of 2 or more occurrences in a sufficiently short interval is zero.

GOAL: Suppose that a process satisfies the above three conditions, and let Y denote the number of occurrences in an interval of length one. Our goal is to find an expression for $p_Y(y) = P(Y = y)$, the pmf of Y .

APPROACH: Envision partitioning the unit interval $[0, 1]$ into n subintervals, each of size $1/n$. Now, if n is sufficiently large (i.e., much larger than y), then we can approximate the probability that y events occur in this unit interval by finding the probability that exactly one event (occurrence) occurs in exactly y of the subintervals.

- By Property (2), we know that the probability of one event in any one subinterval is **proportional** to the subinterval's length, say λ/n , where λ is the proportionality constant.
- By Property (3), the probability of more than one occurrence in any subinterval is zero (for n large).
- Consider the occurrence/non-occurrence of an event in each subinterval as a **Bernoulli trial**. Then, by Property (1), we have a sequence of n Bernoulli trials, each with probability of "success" $p = \lambda/n$. Thus, a binomial (approximate) calculation gives

$$P(Y = y) \approx \binom{n}{y} \left(\frac{\lambda}{n}\right)^y \left(1 - \frac{\lambda}{n}\right)^{n-y}.$$

To improve the approximation for $P(Y = y)$, we let n get large without bound. Then,

$$\begin{aligned} \lim_{n \rightarrow \infty} P(Y = y) &= \lim_{n \rightarrow \infty} \binom{n}{y} \left(\frac{\lambda}{n}\right)^y \left(1 - \frac{\lambda}{n}\right)^{n-y} \\ &= \lim_{n \rightarrow \infty} \frac{n!}{y!(n-y)!} \lambda^y \left(\frac{1}{n}\right)^y \left(1 - \frac{\lambda}{n}\right)^n \left(\frac{1}{1 - \frac{\lambda}{n}}\right)^y \\ &= \lim_{n \rightarrow \infty} \underbrace{\frac{n(n-1)\cdots(n-y+1)}{n^y}}_{a_n} \underbrace{\frac{\lambda^y}{y!}}_{b_n} \underbrace{\left(1 - \frac{\lambda}{n}\right)^n}_{c_n} \underbrace{\left(\frac{1}{1 - \frac{\lambda}{n}}\right)^y}_{d_n}. \end{aligned}$$

Now, the limit of the product is the product of the limits:

$$\begin{aligned} \lim_{n \rightarrow \infty} a_n &= \lim_{n \rightarrow \infty} \frac{n(n-1)\cdots(n-y+1)}{n^y} = 1 \\ \lim_{n \rightarrow \infty} b_n &= \lim_{n \rightarrow \infty} \frac{\lambda^y}{y!} = \frac{\lambda^y}{y!} \\ \lim_{n \rightarrow \infty} c_n &= \lim_{n \rightarrow \infty} \left(1 - \frac{\lambda}{n}\right)^n = e^{-\lambda} \\ \lim_{n \rightarrow \infty} d_n &= \lim_{n \rightarrow \infty} \left(\frac{1}{1 - \frac{\lambda}{n}}\right)^y = 1. \end{aligned}$$

We have shown that

$$\lim_{n \rightarrow \infty} P(Y = y) = \frac{\lambda^y e^{-\lambda}}{y!}.$$

POISSON PMF: A discrete random variable Y is said to follow a **Poisson distribution** with rate λ if the pmf of Y is given by

$$p_Y(y) = \begin{cases} \frac{\lambda^y e^{-\lambda}}{y!}, & y = 0, 1, 2, \dots \\ 0, & \text{otherwise.} \end{cases}$$

We write $Y \sim \text{Poisson}(\lambda)$.

NOTE: Clearly $p_Y(y) > 0$ for all $y \in R$. That $p_Y(y)$ sums to one is easily seen as

$$\begin{aligned} \sum_{y \in R} p_Y(y) &= \sum_{y=0}^{\infty} \frac{\lambda^y e^{-\lambda}}{y!} \\ &= e^{-\lambda} \sum_{y=0}^{\infty} \frac{\lambda^y}{y!} = e^{-\lambda} e^{\lambda} = 1, \end{aligned}$$

since $\sum_{y=0}^{\infty} \lambda^y/y!$ is the McLaurin series expansion of e^{λ} . \square

EXAMPLES: Discrete random variables that might be modeled using a Poisson distribution include

- (1) the number of customers entering a post office in a given day.
- (2) the number of α -particles discharged from a radioactive substance in one second.
- (3) the number of machine breakdowns per month.
- (4) the number of blemishes on a piece of artificial turf.
- (5) the number of chocolate chips in a Chips-Ahoy cookie.

Example 3.22. The number of cars Y abandoned weekly on a highway is modeled using a Poisson distribution with $\lambda = 2.2$. In a given week, what is the probability that

- (a) no cars are abandoned?
- (b) exactly one car is abandoned?
- (c) at most one car is abandoned?
- (d) at least one car is abandoned?

SOLUTIONS. We have $Y \sim \text{Poisson}(\lambda = 2.2)$.

(a)

$$P(Y = 0) = p_Y(0) = \frac{(2.2)^0 e^{-2.2}}{0!} = e^{-2.2} = 0.1108$$

(b)

$$P(Y = 1) = p_Y(1) = \frac{(2.2)^1 e^{-2.2}}{1!} = 2.2e^{-2.2} = 0.2438$$

(c) $P(Y \leq 1) = P(Y = 0) + P(Y = 1) = p_Y(0) + p_Y(1) = 0.1108 + 0.2438 = 0.3456$

(d) $P(Y \geq 1) = 1 - P(Y = 0) = 1 - p_Y(0) = 1 - 0.1108 = 0.8892$. \square

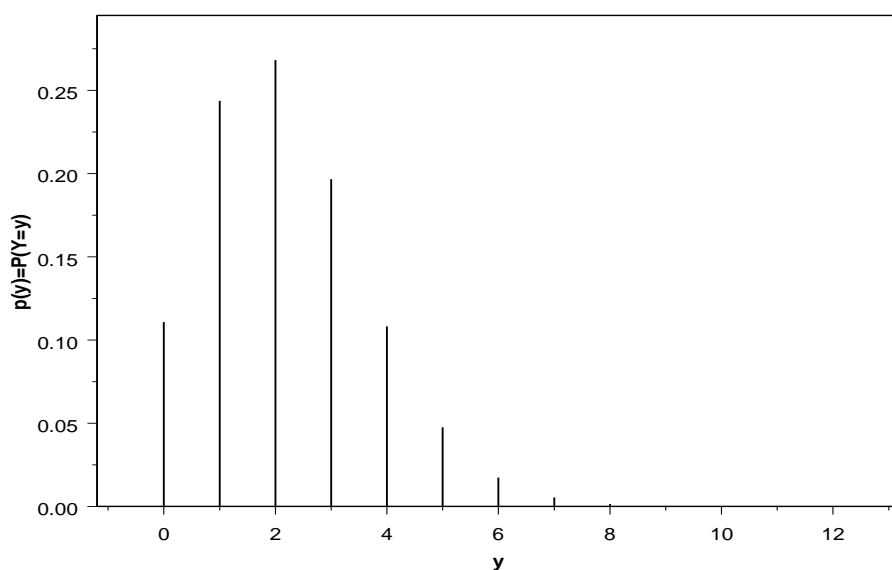


Figure 3.4: *Probability histogram for the number of abandoned cars. This represents the $\text{Poisson}(\lambda = 2.2)$ model in Example 3.22.*

REMARK: WMS's Appendix III, (Table 3, pp 843-847) includes an impressive table for Poisson probabilities of the form

$$F_Y(a) = P(Y \leq a) = \sum_{y=0}^a \frac{\lambda^y e^{-\lambda}}{y!}.$$

Recall that this function is called the **cumulative distribution function** of Y . This makes computing compound event probabilities much easier.

POISSON MGF: Suppose that $Y \sim \text{Poisson}(\lambda)$. The mgf of Y , for all t , is given by

$$\begin{aligned} m_Y(t) = E(e^{tY}) &= \sum_{y=0}^{\infty} e^{ty} \left(\frac{\lambda^y e^{-\lambda}}{y!} \right) \\ &= e^{-\lambda} \underbrace{\sum_{y=0}^{\infty} \frac{(\lambda e^t)^y}{y!}}_{= \exp(\lambda e^t)} \\ &= e^{-\lambda} e^{\lambda e^t} = \exp[\lambda(e^t - 1)]. \quad \square \end{aligned}$$

MEAN AND VARIANCE: With the mgf, we can derive the mean and variance. Differentiating the mgf, we get

$$m'_Y(t) = \frac{d}{dt} m_Y(t) = \frac{d}{dt} \exp[\lambda(e^t - 1)] = \lambda e^t \exp[\lambda(e^t - 1)].$$

Thus,

$$E(Y) = \left. \frac{d}{dt} m_Y(t) \right|_{t=0} = \lambda e^0 \exp[\lambda(e^0 - 1)] = \lambda.$$

Now, we need to find the second moment. Using the product rule, we have

$$\begin{aligned} \frac{d^2}{dt^2} m_Y(t) &= \frac{d}{dt} \underbrace{\lambda e^t \exp[\lambda(e^t - 1)]}_{m'_Y(t)} \\ &= \lambda e^t \exp[\lambda(e^t - 1)] + (\lambda e^t)^2 \exp[\lambda(e^t - 1)]. \end{aligned}$$

Thus,

$$E(Y^2) = \left. \frac{d^2}{dt^2} m_Y(t) \right|_{t=0} = \lambda e^0 \exp[\lambda(e^0 - 1)] + (\lambda e^0)^2 \exp[\lambda(e^0 - 1)] = \lambda + \lambda^2$$

so that

$$\begin{aligned} V(Y) &= E(Y^2) - [E(Y)]^2 \\ &= \lambda + \lambda^2 - \lambda^2 = \lambda. \quad \square \end{aligned}$$

REVELATION: The mean and variance of a Poisson random variable are always equal.

Example 3.23. Suppose that Y denotes the number of defects observed in one month at an automotive plant. From past experience, engineers believe that a Poisson model is appropriate and that $E(Y) = \lambda = 7$ defects/month.

QUESTION 1: What is the probability that, in a given month, we observe 11 or more defects?

SOLUTION. We want to compute

$$P(Y \geq 11) = 1 - \underbrace{P(Y \leq 10)}_{\text{Table 3}} = 1 - 0.901 = 0.099.$$

QUESTION 2: What is the probability that, in a given year, we have two or more months with 11 or more defects?

SOLUTION. First, we assume that the 12 months are independent (is this reasonable?), and call the event $A = \{11 \text{ or more defects in a month}\}$ a “success.” Thus, under our independence assumptions and viewing each month as a “trial,” we have a sequence of 12 Bernoulli trials with “success” probability $p = P(A) = 0.099$. Let X denote the number of months where we observe 11 or more defects. Then, $X \sim b(12, 0.099)$ and

$$\begin{aligned} P(X \geq 2) &= 1 - P(X = 0) - P(X = 1) \\ &= 1 - \binom{12}{0} (0.099)^0 (1 - 0.099)^{12} - \binom{12}{1} (0.099)^1 (1 - 0.099)^{11} \\ &= 1 - 0.2862 - 0.3774 = 0.3364. \quad \square \end{aligned}$$

POISSON PROCESSES OF ARBITRARY LENGTH: If events or occurrences in a Poisson process occur at a rate of λ per unit time or space, then the number of occurrences in an interval of length t follows a Poisson distribution with mean λt .

Example 3.24. Phone calls arrive at a call center according to a Poisson process, at a rate of $\lambda = 3$ per minute. If Y represents the number of calls received in 5 minutes, then $Y \sim \text{Poisson}(15)$. The probability that 8 or fewer calls come in during a 5-minute span is

$$P(Y \leq 8) = \sum_{y=0}^8 \frac{15^y e^{-15}}{y!} = 0.037,$$

using Table 3 (WMS). \square

POISSON-BINOMIAL LINK: We have seen that the hypergeometric and binomial models are related; as it turns out, so are the Poisson and binomial models. This should not be surprising because we derived the Poisson pmf by appealing to a binomial approximation.

RELATIONSHIP: Suppose that $Y \sim b(n, p)$. If n is large and p is small, then

$$p_Y(y) = \binom{n}{y} p^y (1-p)^{n-y} \approx \frac{\lambda^y e^{-\lambda}}{y!},$$

for $y \in R = \{0, 1, 2, \dots, n\}$, where $\lambda = np$.

Example 3.25. Hepatitis C (HCV) is a viral infection that causes cirrhosis and cancer of the liver. Since HCV is transmitted through contact with infectious blood, screening donors is important to prevent further transmission. The World Health Organization has projected that HCV will be a major burden on the US health care system before the year 2020. For public-health reasons, researchers take a sample of $n = 1875$ blood donors and screen each individual for HCV. If 3 percent of the entire population is infected, what is the probability that 50 or more are HCV-positive?

SOLUTION. Let Y denote the number of HCV-infected individuals in our sample. We compute the probability $P(Y \geq 50)$ using both the binomial and Poisson models.

- **Binomial:** Here, $n = 1875$ and $p = 0.03$. Thus,

$$P(Y \geq 50) = \sum_{y=50}^{1875} \binom{1875}{y} (0.03)^y (0.97)^{1875-y} \approx 0.818783.$$

- **Poisson:** Here, $\lambda = np = 1875(0.03) \approx 56.25$. Thus,

$$P(Y \geq 50) = \sum_{y=50}^{\infty} \frac{(56.25)^y e^{-56.25}}{y!} \approx 0.814932.$$

As we can see, the Poisson approximation is quite good. \square

RELATIONSHIP: One can see that the hypergeometric, binomial, and Poisson models are related in the following way:

$$\text{hyper}(N, n, r) \longleftrightarrow b(n, p) \longleftrightarrow \text{Poisson}(\lambda).$$

The first link results when N is large and $r/N \rightarrow p$. The second link results when n is large and p is small so that $\lambda/n \rightarrow p$. When these situations are combined, as you might suspect, one can approximate the hypergeometric model with a Poisson model!

4 Continuous Distributions

Complementary reading from WMS: Chapter 4.

4.1 Introduction

RECALL: In Chapter 3, we focused on discrete random variables. A discrete random variable Y can assume a finite or (at most) a countable number of values. We also learned about probability mass functions (pmfs). These functions tell us what probabilities to assign to each of the support points in R (a countable set).

PREVIEW: Continuous random variables have support sets that are not countable. In fact, most often, the support set for a continuous random variable Y is an interval of real numbers; e.g., $R = \{y : 0 \leq y \leq 1\}$, $R = \{y : 0 < y < \infty\}$, $R = \{y : -\infty < y < \infty\}$, etc. Thus, probabilities of events involving continuous random variables must be assigned in a different way.

4.2 Cumulative distribution functions

TERMINOLOGY: The **(cumulative) distribution function (cdf)** of a random variable Y , denoted by $F_Y(y)$, is given by the probability

$$F_Y(y) = P(Y \leq y),$$

for all $-\infty < y < \infty$. Note that the cdf is defined for all $y \in \mathcal{R}$ (the set of all real numbers), not just for those values of $y \in R$ (the support of Y). Every random variable, discrete or continuous, has a cdf.

Example 4.1. Suppose that the random variable Y has pmf

$$p_Y(y) = \begin{cases} \frac{1}{6}(3 - y), & y = 0, 1, 2 \\ 0, & \text{otherwise.} \end{cases}$$

We now compute probabilities of the form $P(Y \leq y)$:

- for $y < 0$, $F_Y(y) = P(Y \leq y) = 0$
- for $0 \leq y < 1$, $F_Y(y) = P(Y \leq y) = P(Y = 0) = \frac{3}{6}$
- for $1 \leq y < 2$, $F_Y(y) = P(Y \leq y) = P(Y = 0) + P(Y = 1) = \frac{3}{6} + \frac{2}{6} = \frac{5}{6}$
- for $y \geq 2$, $F_Y(y) = P(Y \leq y) = P(Y = 0) + P(Y = 1) + P(Y = 2) = \frac{3}{6} + \frac{2}{6} + \frac{1}{6} = 1$.

Putting this all together, we have the cdf for Y :

$$F_Y(y) = \begin{cases} 0, & y < 0 \\ \frac{3}{6}, & 0 \leq y < 1 \\ \frac{5}{6}, & 1 \leq y < 2 \\ 1, & y \geq 2. \end{cases}$$

It is instructive to plot the pmf of Y and the cdf of Y side by side.

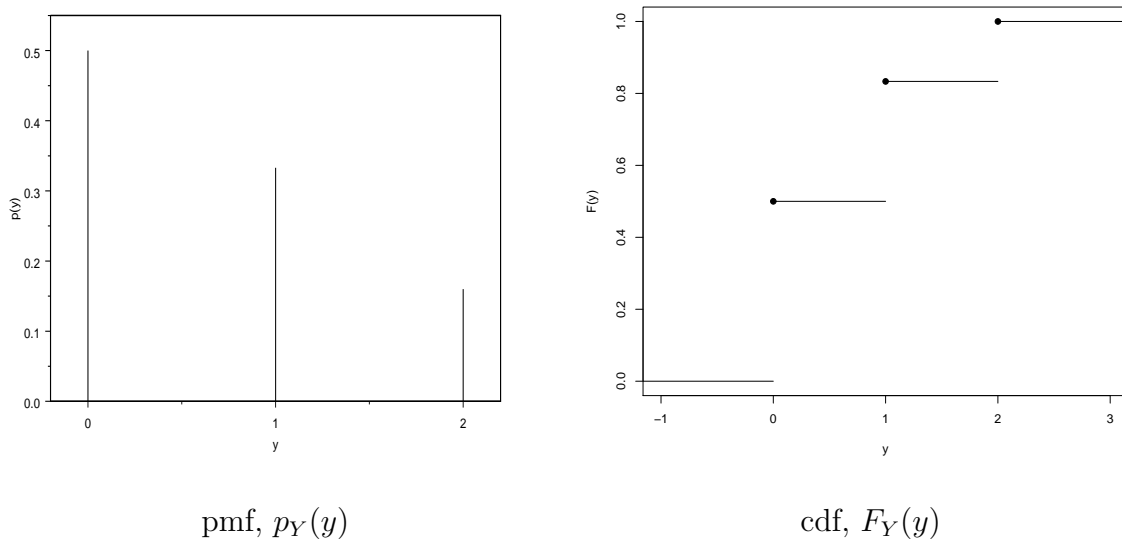


Figure 4.5: *Probability mass function $p_Y(y)$ and cumulative distribution function $F_Y(y)$ in Example 4.1.*

• **PMF**

- The height of the bar above y is the probability that Y assumes that value.
- For any y not equal to 0, 1, or 2, $p_Y(y) = 0$.

- **CDF**

- $F_Y(y)$ is a nondecreasing function.
- $0 \leq F_Y(y) \leq 1$; this makes sense since $F_Y(y) = P(Y \leq y)$ is a probability!
- The cdf $F_Y(y)$ in this example takes a “step” at the support points and stays constant otherwise. The height of the step at a particular point is equal to the probability associated with that point. \square

CDF PROPERTIES: Let Y be a random variable (discrete or continuous) and suppose that $F_Y(y)$ is the cdf for Y . Then

(i) $F_Y(y)$ satisfies the following:

$$\lim_{y \rightarrow -\infty} F_Y(y) = 0 \quad \text{and} \quad \lim_{y \rightarrow +\infty} F_Y(y) = 1.$$

(ii) $F_Y(y)$ is a right continuous function; that is, for any real a ,

$$\lim_{y \rightarrow a^+} F_Y(y) = F_Y(a).$$

(iii) $F_Y(y)$ is a non-decreasing function; that is,

$$y_1 \leq y_2 \implies F_Y(y_1) \leq F_Y(y_2).$$

EXERCISE: Graph the cdf for (a) $Y \sim b(5, 0.2)$ and (b) $Y \sim \text{Poisson}(2)$.

4.3 Continuous random variables

TERMINOLOGY: A random variable Y is said to be **continuous** if its cdf $F_Y(y)$ is a continuous function of y .

REMARK: The cdfs associated with discrete random variables are step functions (see Example 4.1). Such functions are not continuous; however, they are still right continuous.

OBSERVATION: We can immediately deduce that if Y is a continuous random variable, then

$$P(Y = y) = 0,$$

for all y . **That is, specific points are assigned zero probability in continuous probability models.** This must be true. If this was not true, and $P(Y = y) = p_0 > 0$, then $F_Y(y)$ would take a step of height p_0 at the point y . This would then imply that $F_Y(y)$ is not a continuous function.

TERMINOLOGY: Let Y be a continuous random variable with cdf $F_Y(y)$. The **probability density function (pdf)** for Y , denoted by $f_Y(y)$, is given by

$$f_Y(y) = \frac{d}{dy} F_Y(y),$$

provided that $\frac{d}{dy} F_Y(y) \equiv F'_Y(y)$ exists. Appealing to the Fundamental Theorem of Calculus, we know that

$$F_Y(y) = \int_{-\infty}^y f_Y(t) dt.$$

These are important facts that describe how the pdf and cdf of a continuous random variable are related. Because $F_Y(y) = P(Y \leq y)$, it should be clear that probabilities in continuous models are found by integration (compare this with how probabilities are obtained in discrete models).

PROPERTIES OF CONTINUOUS PDFs: Suppose that Y is a continuous random variable with pdf $f_Y(y)$ and support R . Then

- (1) $f_Y(y) > 0$, for all $y \in R$;
- (2) The function $f_Y(y)$ satisfies

$$\int_R f_Y(y) dy = 1.$$

CONTINUOUS MODELS: Probability density functions serve as **theoretical models** for continuous data (just as probability mass functions serve as models for discrete data). These models can be used to find probabilities associated with future (random) events.

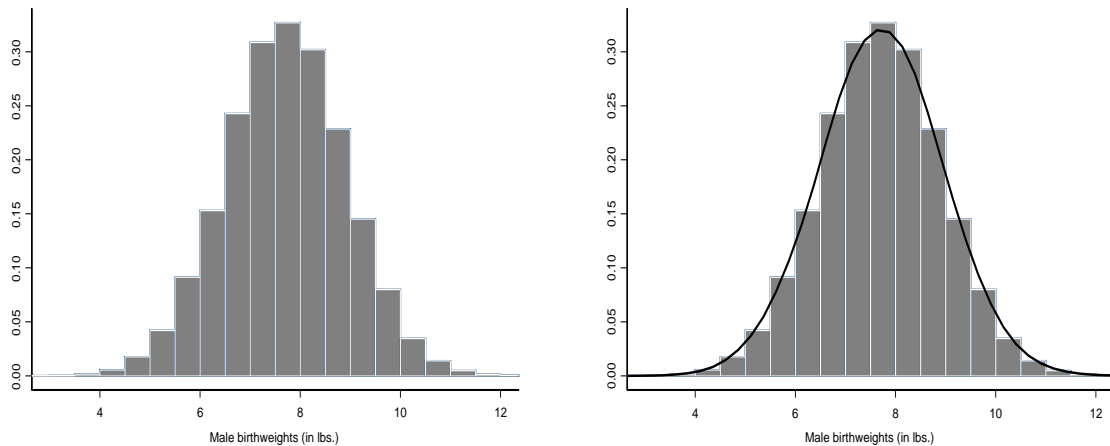


Figure 4.6: *Canadian male birth weight data. The histogram (left) is constructed from a sample of $n = 1250$ subjects. A normal probability density function has been fit to the empirical distribution (right).*

Example 4.2. A team of Montreal researchers who studied the birth weights of five million Canadian babies born between 1981 and 2003 say environmental contaminants may be to blame for a drop in the size of newborn baby boys. A subset ($n = 1250$ subjects) of the birth weights, measured in lbs, is given in Figure 4.6. \square

IMPORTANT: Suppose Y is a continuous random variable with pdf $f_Y(y)$ and cdf $F_Y(y)$. The probability of an event $\{Y \in B\}$ is computed by integrating $f_Y(y)$ over B , that is,

$$P(Y \in B) = \int_B f_Y(y) dy,$$

for any $B \subset \mathcal{R}$. If $B = \{y : a \leq y \leq b\}$; i.e., $B = [a, b]$, then

$$\begin{aligned} P(Y \in B) = P(a \leq Y \leq b) &= \int_a^b f_Y(y) dy \\ &= \int_{-\infty}^b f_Y(y) dy - \int_{-\infty}^a f_Y(y) dy \\ &= F_Y(b) - F_Y(a). \end{aligned}$$

Compare these to the analogous results for the discrete case (see page 29 in the notes). In the continuous case, $f_Y(y)$ replaces $p_Y(y)$ and integrals replace sums.

RECALL: We have already discovered that if Y is a continuous random variable, then $P(Y = a) = 0$ for any constant a . This can be also seen by writing

$$P(Y = a) = P(a \leq Y \leq a) = \int_a^a f_Y(y)dy = 0,$$

where $f_Y(y)$ is the pdf of Y . An immediate consequence of this is that if Y is continuous,

$$P(a \leq Y \leq b) = P(a \leq Y < b) = P(a < Y \leq b) = P(a < Y < b) = \int_a^b f_Y(y)dy.$$

Example 4.3. Suppose that Y has the pdf

$$f_Y(y) = \begin{cases} 2y, & 0 < y < 1 \\ 0, & \text{otherwise.} \end{cases}$$

Find the cdf of Y .

SOLUTION. We need to compute $F_Y(y) = P(Y \leq y)$ for all $y \in \mathcal{R}$. There are three cases to consider:

- when $y \leq 0$,

$$F_Y(y) = \int_{-\infty}^y f_Y(t)dt = \int_{-\infty}^y 0dt = 0;$$

- when $0 < y < 1$,

$$F_Y(y) = \int_{-\infty}^y f_Y(t)dt = \int_{-\infty}^0 0dt + \int_0^y 2t dt = 0 + t^2 \Big|_0^y = y^2;$$

- when $y \geq 1$,

$$F_Y(y) = \int_{-\infty}^y f_Y(t)dt = \int_{-\infty}^0 0dt + \int_0^1 2t dt + \int_1^y 0dt = 0 + 1 + 0 = 1.$$

Putting this all together, we have

$$F_Y(y) = \begin{cases} 0, & y < 0 \\ y^2, & 0 \leq y < 1 \\ 1, & y \geq 1. \end{cases}$$

The pdf $f_Y(y)$ and the cdf $F_Y(y)$ are plotted side by side in Figure 4.7.

EXERCISE: Find (a) $P(0.3 < Y < 0.7)$, (b) $P(Y = 0.3)$, and (c) $P(Y > 0.7)$. \square

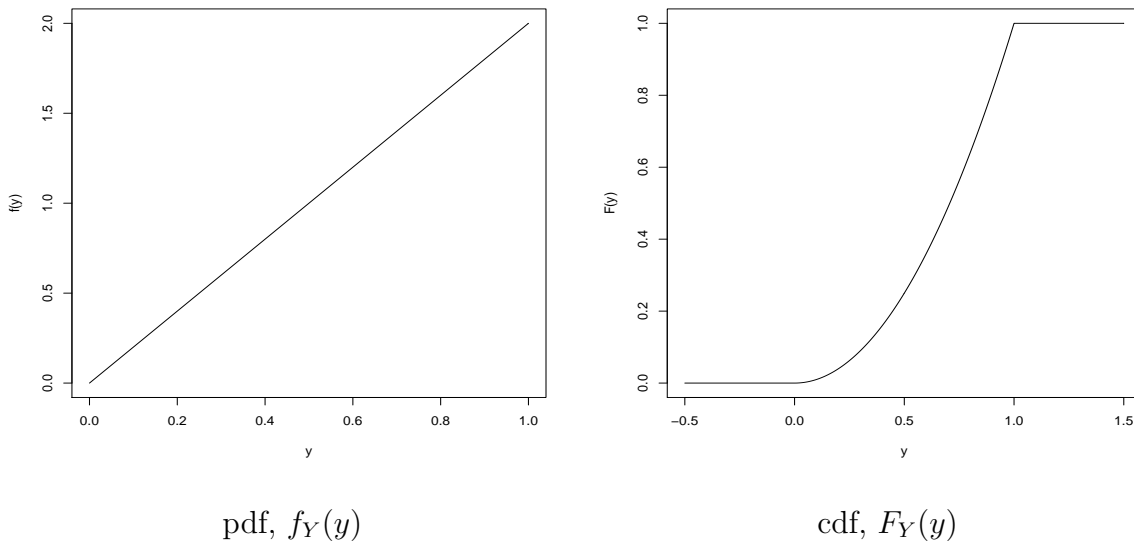


Figure 4.7: Probability density function $f_Y(y)$ and cumulative distribution function $F_Y(y)$ in Example 4.3.

Example 4.4. From the onset of infection, the survival time Y (measured in years) of patients with chronic active hepatitis receiving prednisolone is modeled with the pdf

$$f_Y(y) = \begin{cases} \frac{1}{10}e^{-y/10}, & y > 0 \\ 0, & \text{otherwise.} \end{cases}$$

Find the cdf of Y .

SOLUTION. We need to compute $F_Y(y) = P(Y \leq y)$ for all $y \in \mathcal{R}$. There are two cases to consider:

- when $y \leq 0$,

$$F_Y(y) = \int_{-\infty}^y f_Y(t)dt = \int_{-\infty}^y 0dt = 0;$$

- when $y > 0$,

$$\begin{aligned} F_Y(y) &= \int_{-\infty}^y f_Y(t)dt = \int_{-\infty}^0 0dt + \int_0^y \frac{1}{10}e^{-t/10}dt \\ &= 0 + \frac{1}{10} \left(-10e^{-t/10} \right) \Big|_0^y = 1 - e^{-y/10}. \end{aligned}$$

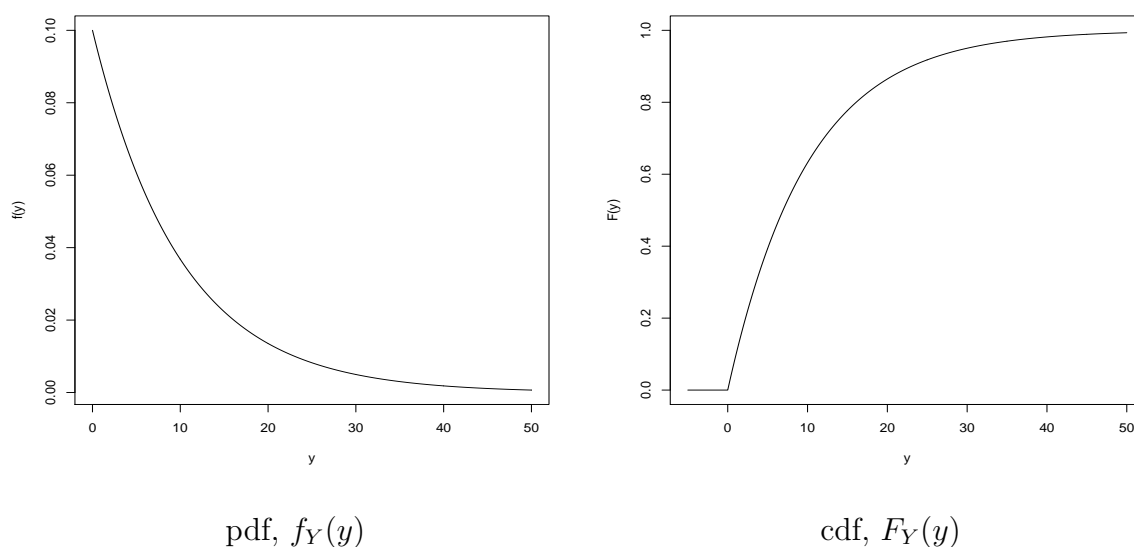


Figure 4.8: Probability density function $f_Y(y)$ and cumulative distribution function $F_Y(y)$ in Example 4.4.

Putting this all together, we have

$$F_Y(y) = \begin{cases} 0, & y \leq 0 \\ 1 - e^{-y/10}, & y > 0. \end{cases}$$

The pdf $f_Y(y)$ and the cdf $F_Y(y)$ are plotted side by side in Figure 4.8.

EXERCISE: What is the probability a patient survives 15 years after being diagnosed? less than 5 years? between 10 and 20 years? \square

Example 4.5. Suppose that Y has the pdf

$$f_Y(y) = \begin{cases} cye^{-y/2}, & y \geq 0 \\ 0, & \text{otherwise.} \end{cases}$$

Find the value of c that makes this a valid pdf.

SOLUTION. Because $f_Y(y)$ is a pdf, we know that

$$\int_0^{\infty} f_Y(y)dy = \int_0^{\infty} cye^{-y/2}dy = 1.$$

Using integration by parts with $u = cy$ and $dv = e^{-y/2}dy$, we have

$$\begin{aligned} 1 = \int_0^\infty cy e^{-y/2} dy &= \underbrace{-2cye^{-y/2}}_{=0} \Big|_0^\infty + \int_0^\infty 2ce^{-y/2} dy \\ &= 2c(-2)e^{-y/2} \Big|_0^\infty = 0 - (-4c) = 4c. \end{aligned}$$

Solving for c , we get $c = 1/4$. \square

QUANTILES: Suppose that Y is a continuous random variable with cdf $F_Y(y)$ and let $0 < p < 1$. The p th **quantile** of the distribution of Y , denoted by ϕ_p , solves

$$F_Y(\phi_p) = P(Y \leq \phi_p) = \int_{-\infty}^{\phi_p} f_Y(y) dy = p.$$

The **median** of the distribution of Y is the $p = 0.5$ quantile. That is, the median $\phi_{0.5}$ solves

$$F_Y(\phi_{0.5}) = P(Y \leq \phi_{0.5}) = \int_{-\infty}^{\phi_{0.5}} f_Y(y) dy = 0.5.$$

Another name for the p th quantile is the **100pth percentile**.

EXERCISE. Find the median of Y in Examples 4.3, 4.4, and 4.5.

REMARK: For Y discrete, there are some potential problems with the definition that ϕ_p solves $F_Y(\phi_p) = P(Y \leq \phi_p) = p$. The reason is that there may be many values of ϕ_p that satisfy this equation. For example, in Example 4.1, it is easy to see that the median $\phi_{0.5} = 0$ because $F_Y(0) = P(Y \leq 0) = 0.5$. However, $\phi_{0.5} = 0.5$ also satisfies $F_Y(\phi_{0.5}) = 0.5$. By convention, in discrete distributions, the p th quantile ϕ_p is taken to be the smallest value satisfying $F_Y(\phi_p) = P(Y \leq \phi_p) \geq p$.

4.4 Mathematical expectation

4.4.1 Expected value

TERMINOLOGY: Let Y be a continuous random variable with pdf $f_Y(y)$ and support R . The **expected value** of Y is given by

$$E(Y) = \int_R y f_Y(y) dy.$$

Mathematically, we require that

$$\int_{\mathcal{R}} |y| f_Y(y) dy < \infty.$$

If this is not true, we say that $E(Y)$ does not exist. If g is a real-valued function, then $g(Y)$ is a random variable and

$$E[g(Y)] = \int_{\mathcal{R}} g(y) f_Y(y) dy,$$

provided that this integral exists.

Example 4.6. Suppose that Y has pdf given by

$$f_Y(y) = \begin{cases} 2y, & 0 < y < 1 \\ 0, & \text{otherwise.} \end{cases}$$

Find $E(Y)$, $E(Y^2)$, and $E(\ln Y)$.

SOLUTION. The expected value of Y is given by

$$\begin{aligned} E(Y) &= \int_0^1 y f_Y(y) dy \\ &= \int_0^1 y(2y) dy \\ &= \int_0^1 2y^2 dy = 2 \left(\frac{y^3}{3} \Big|_0^1 \right) = 2 \left(\frac{1}{3} - 0 \right) = 2/3. \end{aligned}$$

The second moment is

$$\begin{aligned} E(Y^2) &= \int_0^1 y^2 f_Y(y) dy \\ &= \int_0^1 y^2(2y) dy \\ &= \int_0^1 2y^3 dy = 2 \left(\frac{y^4}{4} \Big|_0^1 \right) = 2 \left(\frac{1}{4} - 0 \right) = 1/2. \end{aligned}$$

Finally,

$$E(\ln Y) = \int_0^1 \ln y(2y) dy.$$

To solve this integral, use integration by parts with $u = \ln y$ and $dv = 2y dy$:

$$E(\ln Y) = \underbrace{y^2 \ln y}_0 \Big|_0^1 - \int_0^1 y dy = - \left(\frac{y^2}{2} \Big|_0^1 \right) = -\frac{1}{2}. \quad \square$$

PROPERTIES OF EXPECTATIONS: Let Y be a continuous random variable with pdf $f_Y(y)$ and support R , suppose that g, g_1, g_2, \dots, g_k are real-valued functions, and let c be any real constant. Then,

$$(a) \ E(c) = c$$

$$(b) \ E[cg(Y)] = cE[g(Y)]$$

$$(c) \ E[\sum_{j=1}^k g_j(Y)] = \sum_{j=1}^k E[g_j(Y)].$$

These properties are identical to those we discussed in the discrete case.

4.4.2 Variance

TERMINOLOGY: Let Y be a continuous random variable with pdf $f_Y(y)$, support R , and mean $E(Y) = \mu$. The **variance** of Y is given by

$$\sigma^2 \equiv V(Y) \equiv E[(Y - \mu)^2] = \int_R (y - \mu)^2 f_Y(y) dy.$$

The variance computing formula still applies in the continuous case, that is,

$$V(Y) = E(Y^2) - [E(Y)]^2.$$

Example 4.7. Suppose that Y has pdf given by

$$f_Y(y) = \begin{cases} 2y, & 0 < y < 1 \\ 0, & \text{otherwise.} \end{cases}$$

Find $\sigma^2 = V(Y)$.

SOLUTION. We computed $E(Y) = \mu = 2/3$ in Example 4.6. Using the definition above,

$$V(Y) = \int_0^1 \left(y - \frac{2}{3}\right)^2 (2y) dy.$$

Instead of doing this integral, it is easier to use the variance computing formula $V(Y) = E(Y^2) - [E(Y)]^2$. In Example 4.6, we computed the second moment $E(Y^2) = 1/2$. Thus,

$$V(Y) = E(Y^2) - [E(Y)]^2 = \frac{1}{2} - \left(\frac{2}{3}\right)^2 = 1/18. \quad \square$$

4.4.3 Moment generating functions

TERMINOLOGY: Let Y be a continuous random variable with pdf $f_Y(y)$ and support R . The **moment generating function (mgf)** for Y , denoted by $m_Y(t)$, is given by

$$m_Y(t) = E(e^{tY}) = \int_R e^{ty} f_Y(y) dy,$$

provided $E(e^{tY}) < \infty$ for all t in an open neighborhood about 0; i.e., there exists some $h > 0$ such that $E(e^{tY}) < \infty$ for all $t \in (-h, h)$. If $E(e^{tY})$ does not exist in an open neighborhood of 0, we say that the moment generating function does not exist.

Example 4.8. Suppose that the pdf of Y is given by

$$f_Y(y) = \begin{cases} e^{-y}, & y > 0 \\ 0, & \text{otherwise.} \end{cases}$$

Find the mgf of Y and use it to compute $E(Y)$ and $V(Y)$.

SOLUTION.

$$\begin{aligned} m_Y(t) = E(e^{tY}) &= \int_0^{\infty} e^{ty} f_Y(y) dy \\ &= \int_0^{\infty} e^{ty} e^{-y} dy \\ &= \int_0^{\infty} e^{ty-y} dy \\ &= \int_0^{\infty} e^{-y(1-t)} dy = - \left(\frac{1}{1-t} \right) e^{-y(1-t)} \Bigg|_{y=0}^{\infty}. \end{aligned}$$

In the last expression, note that

$$\lim_{y \rightarrow \infty} e^{-y(1-t)} < \infty$$

if and only if $1-t > 0$, i.e., $t < 1$. Thus, for $t < 1$, we have

$$m_Y(t) = - \left(\frac{1}{1-t} \right) e^{-y(1-t)} \Bigg|_{y=0}^{\infty} = 0 + \left(\frac{1}{1-t} \right) = \frac{1}{1-t}.$$

Note that $(-h, h)$ with $h = 1$ is an open neighborhood around zero for which $m_Y(t)$ exists. With the mgf, we can calculate the mean and variance. Differentiating the mgf,

we get

$$m'_Y(t) = \frac{d}{dt}m_Y(t) = \frac{d}{dt} \left(\frac{1}{1-t} \right) = \left(\frac{1}{1-t} \right)^2$$

so that

$$E(Y) = \frac{d}{dt}m_Y(t) \Big|_{t=0} = \left(\frac{1}{1-0} \right)^2 = 1.$$

To find the variance, we first find the second moment. The second derivative of $m_Y(t)$ is

$$\frac{d^2}{dt^2}m_Y(t) = \frac{d}{dt} \underbrace{\left(\frac{1}{1-t} \right)^2}_{m'_Y(t)} = 2 \left(\frac{1}{1-t} \right)^3.$$

The second moment is

$$E(Y^2) = \frac{d^2}{dt^2}m_Y(t) \Big|_{t=0} = 2 \left(\frac{1}{1-0} \right)^3 = 2.$$

The computing formula gives

$$V(Y) = E(Y^2) - [E(Y)]^2 = 2 - (1)^2 = 1.$$

EXERCISE: Find $E(Y)$ and $V(Y)$ without using the mgf. \square

4.5 Uniform distribution

TERMINOLOGY: A random variable Y is said to have a **uniform distribution** from θ_1 to θ_2 if its pdf is given by

$$f_Y(y) = \begin{cases} \frac{1}{\theta_2 - \theta_1}, & \theta_1 < y < \theta_2 \\ 0, & \text{otherwise.} \end{cases}$$

Shorthand notation is $Y \sim \mathcal{U}(\theta_1, \theta_2)$. Note that this is a valid density because $f_Y(y) > 0$ for all $y \in R = \{y : \theta_1 < y < \theta_2\}$ and

$$\int_{\theta_1}^{\theta_2} f_Y(y) dy = \int_{\theta_1}^{\theta_2} \left(\frac{1}{\theta_2 - \theta_1} \right) dy = \frac{y}{\theta_2 - \theta_1} \Big|_{\theta_1}^{\theta_2} = \frac{\theta_2 - \theta_1}{\theta_2 - \theta_1} = 1.$$

STANDARD UNIFORM: A popular member of the $\mathcal{U}(\theta_1, \theta_2)$ family is the $\mathcal{U}(0, 1)$ distribution; i.e., a uniform distribution with parameters $\theta_1 = 0$ and $\theta_2 = 1$. This model is used extensively in computer programs to simulate random numbers.

Example 4.9. Derive the cdf of $Y \sim \mathcal{U}(\theta_1, \theta_2)$.

SOLUTION. We need to compute $F_Y(y) = P(Y \leq y)$ for all $y \in \mathcal{R}$. There are three cases to consider:

- when $y \leq \theta_1$,

$$F_Y(y) = \int_{-\infty}^y f_Y(t) dt = \int_{-\infty}^y 0 dt = 0;$$

- when $\theta_1 < y < \theta_2$,

$$\begin{aligned} F_Y(y) &= \int_{-\infty}^y f_Y(t) dt = \int_{-\infty}^{\theta_1} 0 dt + \int_{\theta_1}^y \left(\frac{1}{\theta_2 - \theta_1} \right) dt \\ &= 0 + \frac{t}{\theta_2 - \theta_1} \Big|_{\theta_1}^y = \frac{y - \theta_1}{\theta_2 - \theta_1}; \end{aligned}$$

- when $y \geq \theta_2$,

$$F_Y(y) = \int_{-\infty}^y f_Y(t) dt = \int_{-\infty}^{\theta_1} 0 dt + \int_{\theta_1}^{\theta_2} \left(\frac{1}{\theta_2 - \theta_1} \right) dt + \int_{\theta_2}^y 0 dt = 0 + 1 + 0 = 1.$$

Putting this all together, we have

$$F_Y(y) = \begin{cases} 0, & y \leq \theta_1 \\ \frac{y - \theta_1}{\theta_2 - \theta_1}, & \theta_1 < y < \theta_2 \\ 1, & y \geq \theta_2. \end{cases}$$

The $\mathcal{U}(0, 1)$ pdf $f_Y(y)$ and cdf $F_Y(y)$ are plotted side by side in Figure 4.9.

EXERCISE: If $Y \sim \mathcal{U}(0, 1)$, find (a) $P(0.2 < Y < 0.4)$ and (b) $P(Y > 0.75)$. \square

MEAN AND VARIANCE: If $Y \sim \mathcal{U}(\theta_1, \theta_2)$, then

$$E(Y) = \frac{\theta_1 + \theta_2}{2} \quad \text{and} \quad V(Y) = \frac{(\theta_2 - \theta_1)^2}{12}.$$

UNIFORM MGF: If $Y \sim \mathcal{U}(\theta_1, \theta_2)$, then

$$m_Y(t) = \begin{cases} \frac{e^{\theta_2 t} - e^{\theta_1 t}}{t(\theta_2 - \theta_1)}, & t \neq 0 \\ 1, & t = 0 \end{cases}$$

EXERCISE: Derive the formulas for $E(Y)$ and $V(Y)$.

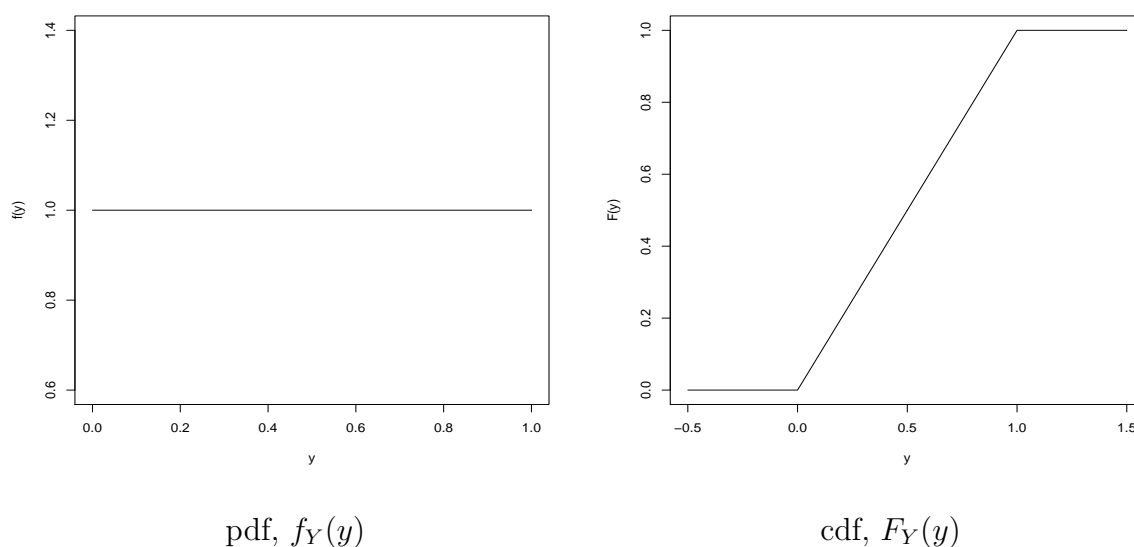


Figure 4.9: The $\mathcal{U}(0,1)$ probability density function and cumulative distribution function.

4.6 Normal distribution

TERMINOLOGY: A random variable Y is said to have a **normal distribution** if its pdf is given by

$$f_Y(y) = \begin{cases} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2}, & -\infty < y < \infty \\ 0, & \text{otherwise.} \end{cases}$$

Shorthand notation is $Y \sim \mathcal{N}(\mu, \sigma^2)$. There are two parameters in the normal distribution: the mean $E(Y) = \mu$ and the variance $V(Y) = \sigma^2$.

FACTS:

- (a) The $\mathcal{N}(\mu, \sigma^2)$ pdf is symmetric about μ ; that is, for any $a \in \mathcal{R}$,

$$f_Y(\mu - a) = f_Y(\mu + a).$$

- (b) The $\mathcal{N}(\mu, \sigma^2)$ pdf has points of inflection located at $y = \mu \pm \sigma$ (verify!).

- (c) $\lim_{y \rightarrow \pm\infty} f_Y(y) = 0$.

TERMINOLOGY: A normal distribution with mean $\mu = 0$ and variance $\sigma^2 = 1$ is called the **standard normal distribution**. It is conventional to let Z denote a random variable that follows a standard normal distribution; we write $Z \sim \mathcal{N}(0, 1)$.

IMPORTANT: Tabled values of the standard normal probabilities are given in Appendix III (Table 4, pp 848) of WMS. This table turns out to be helpful since the integral

$$F_Y(y) = P(Y \leq y) = \int_{-\infty}^y \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{t-\mu}{\sigma}\right)^2} dt$$

does not exist in closed form. Specifically, the table provides values of

$$1 - F_Z(z) = P(Z > z) = \int_z^{\infty} f_Z(u) du,$$

where $f_Z(u)$ denotes the nonzero part of the standard normal pdf; i.e.,

$$f_Z(u) = \frac{1}{\sqrt{2\pi}} e^{-u^2/2}.$$

To use the table, we need to first prove that any $\mathcal{N}(\mu, \sigma^2)$ distribution can be “transformed” to the (standard) $\mathcal{N}(0, 1)$ distribution (we’ll see how to do this later). Once we do this, we will see that there is only a need for one table of probabilities. Of course, probabilities like $F_Y(y) = P(Y \leq y)$ can be obtained using software too.

Example 4.10. Show that the $\mathcal{N}(\mu, \sigma^2)$ pdf integrates to 1.

Proof. Let $z = (y - \mu)/\sigma$ so that $dz = dy/\sigma$ and $dy = \sigma dz$. Define

$$\begin{aligned} I &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2} dy \\ &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz. \end{aligned}$$

We want to show that $I = 1$. Since $I > 0$, it suffices to show that $I^2 = 1$. Note that

$$\begin{aligned} I^2 &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-y^2/2} dy \\ &= \frac{1}{2\pi} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \exp\left[-\left(\frac{x^2 + y^2}{2}\right)\right] dx dy. \end{aligned}$$

Switching to polar coordinates; i.e., letting $x = r \cos \theta$ and $y = r \sin \theta$, we get $x^2 + y^2 = r^2(\cos^2 \theta + \sin^2 \theta) = r^2$, and $dx dy = r dr d\theta$; i.e., the Jacobian of the transformation from

(x, y) space to (r, θ) space. Thus, we write

$$\begin{aligned}
 I^2 &= \frac{1}{2\pi} \int_{\theta=0}^{2\pi} \int_{r=0}^{\infty} e^{-r^2/2} r dr d\theta \\
 &= \frac{1}{2\pi} \int_{\theta=0}^{2\pi} \left[\int_{r=0}^{\infty} r e^{-r^2/2} dr \right] d\theta \\
 &= \frac{1}{2\pi} \int_{\theta=0}^{2\pi} \left[-e^{-r^2/2} \Big|_{r=0}^{\infty} \right] d\theta \\
 &= \frac{1}{2\pi} \int_{\theta=0}^{2\pi} 1 d\theta = \frac{\theta}{2\pi} \Big|_{\theta=0}^{2\pi} = 1. \quad \square
 \end{aligned}$$

NORMAL MGF: Suppose that $Y \sim \mathcal{N}(\mu, \sigma^2)$. The mgf of Y is

$$m_Y(t) = \exp\left(\mu t + \frac{\sigma^2 t^2}{2}\right).$$

Proof. Using the definition of the mgf, we have

$$\begin{aligned}
 m_Y(t) = E(e^{tY}) &= \int_{-\infty}^{\infty} e^{ty} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2} dy \\
 &= \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} e^{ty - \frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2} dy.
 \end{aligned}$$

Define $b = ty - \frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2$, the exponent in the last integral. We are going to rewrite b in the following way:

$$\begin{aligned}
 b &= ty - \frac{1}{2} \left(\frac{y-\mu}{\sigma} \right)^2 \\
 &= ty - \frac{1}{2\sigma^2} (y^2 - 2\mu y + \mu^2) \\
 &= -\frac{1}{2\sigma^2} (y^2 - 2\mu y - 2\sigma^2 ty + \mu^2) \\
 &= -\frac{1}{2\sigma^2} \left[\underbrace{y^2 - 2(\mu + \sigma^2 t)y + \mu^2}_{\text{complete the square}} \right] \\
 &= -\frac{1}{2\sigma^2} \left[y^2 - 2(\mu + \sigma^2 t)y + \underbrace{(\mu + \sigma^2 t)^2 - (\mu + \sigma^2 t)^2}_{\text{add and subtract}} + \mu^2 \right] \\
 &= -\frac{1}{2\sigma^2} \{ [y - (\mu + \sigma^2 t)]^2 \} + \frac{1}{2\sigma^2} [(\mu + \sigma^2 t)^2 - \mu^2] \\
 &= -\frac{1}{2\sigma^2} (y - a)^2 + \frac{1}{2\sigma^2} (\mu^2 + 2\mu\sigma^2 t + \sigma^4 t^2 - \mu^2) \\
 &= -\frac{1}{2\sigma^2} (y - a)^2 + \underbrace{\mu t + \sigma^2 t^2 / 2}_{= c, \text{ say}}
 \end{aligned}$$

where $a = \mu + \sigma^2 t$. Noting that $c = \mu t + \sigma^2 t^2/2$ is free of y , we have

$$\begin{aligned} m_Y(t) &= \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} e^b dy \\ &= \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} e^{-\frac{1}{2\sigma^2}(y-a)^2+c} dy \\ &= e^c \left(\int_{-\infty}^{\infty} \underbrace{\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(y-a)^2}}_{\mathcal{N}(a,\sigma^2) \text{ density}} dy \right) = e^c, \end{aligned}$$

since the $\mathcal{N}(a, \sigma^2)$ pdf integrates to 1. Now, finally note

$$e^c \equiv \exp(c) = \exp\left(\mu t + \frac{\sigma^2 t^2}{2}\right). \quad \square$$

EXERCISE: Use the mgf to verify that $E(Y) = \mu$ and $V(Y) = \sigma^2$.

IMPORTANT: Suppose that $Y \sim \mathcal{N}(\mu, \sigma^2)$. The random variable

$$Z = \frac{Y - \mu}{\sigma} \sim \mathcal{N}(0, 1).$$

Proof. Let $Z = (Y - \mu)/\sigma$. The mgf of Z is given by

$$\begin{aligned} m_Z(t) = E(e^{tZ}) &= E\left\{\exp\left[t\left(\frac{Y - \mu}{\sigma}\right)\right]\right\} \\ &= E(e^{tY/\sigma - \mu t/\sigma}) \\ &= e^{-\mu t/\sigma} E(e^{tY/\sigma}) \\ &= e^{-\mu t/\sigma} m_Y(t/\sigma) \\ &= e^{-\mu t/\sigma} \exp\left[\mu(t/\sigma) + \frac{\sigma^2(t/\sigma)^2}{2}\right] = e^{t^2/2}, \end{aligned}$$

which is the mgf of a $\mathcal{N}(0, 1)$ random variable. Thus, by the **uniqueness** of moment generating functions, we know that $Z \sim \mathcal{N}(0, 1)$. \square

USEFULNESS: From the last result, we know that if $Y \sim \mathcal{N}(\mu, \sigma^2)$, then the event

$$\{y_1 < Y < y_2\} = \left\{\frac{y_1 - \mu}{\sigma} < \frac{Y - \mu}{\sigma} < \frac{y_2 - \mu}{\sigma}\right\} = \left\{\frac{y_1 - \mu}{\sigma} < Z < \frac{y_2 - \mu}{\sigma}\right\}.$$

As a result,

$$\begin{aligned} P(y_1 < Y < y_2) &= P\left(\frac{y_1 - \mu}{\sigma} < Z < \frac{y_2 - \mu}{\sigma}\right) \\ &= F_Z\left(\frac{y_2 - \mu}{\sigma}\right) - F_Z\left(\frac{y_1 - \mu}{\sigma}\right), \end{aligned}$$

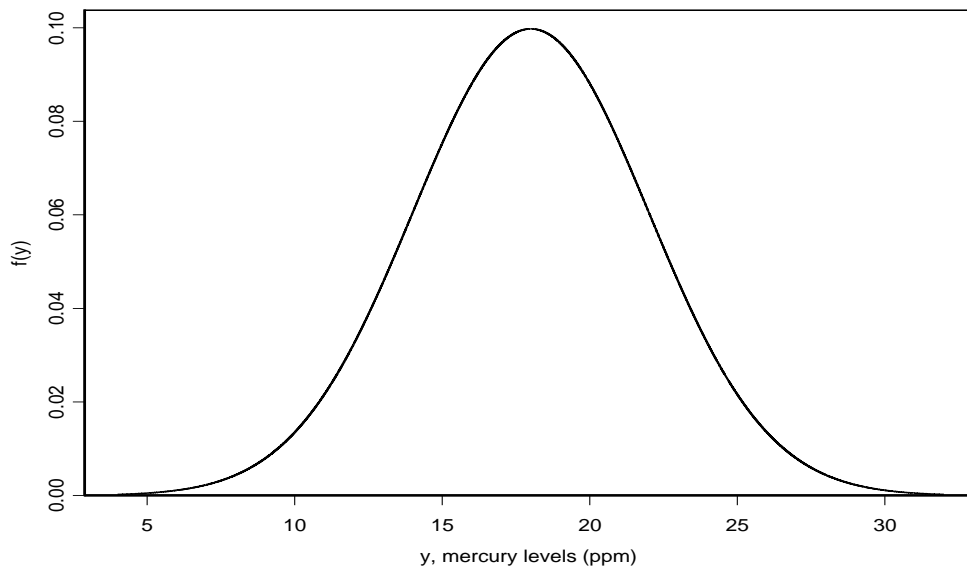


Figure 4.10: Probability density function, $f_Y(y)$, in Example 4.11. A model for mercury contamination in large-mouth bass.

where $F_Z(\cdot)$ is the cdf of the $\mathcal{N}(0, 1)$ distribution. Note also that $F_Z(-z) = 1 - F_Z(z)$, for $z > 0$ (verify!). The standard normal table (Table 4, pp 848) gives values of $1 - F_Z(z)$, for $z > 0$.

Example 4.11. Young large-mouth bass were studied to examine the level of mercury contamination, Y (measured in parts per million), which varies according to a normal distribution with mean $\mu = 18$ and variance $\sigma^2 = 16$, depicted in Figure 4.10.

(a) What proportion of contamination levels are between 11 and 21 parts per million?

SOLUTION. We want $P(11 < Y < 21)$. By standardizing, we see that

$$\begin{aligned}
 P(11 < Y < 21) &= P\left(\frac{11 - 18}{4} < \frac{Y - 18}{4} < \frac{21 - 18}{4}\right) \\
 &= P\left(\frac{11 - 18}{4} < Z < \frac{21 - 18}{4}\right) \\
 &= P(-1.75 < Z < 0.75) \\
 &= F_Z(0.75) - F_Z(-1.75) = 0.7734 - 0.0401 = 0.7333.
 \end{aligned}$$

(b) For this model, ninety percent of all contamination levels are above what mercury level?

SOLUTION. We want to find $\phi_{0.10}^Y$, the 10th percentile of $Y \sim \mathcal{N}(18, 16)$; i.e., $\phi_{0.10}^Y$ solves

$$F_Y(\phi_{0.10}^Y) = P(Y \leq \phi_{0.10}^Y) = 0.10.$$

We'll start by finding $\phi_{0.10}^Z$, the 10th percentile of $Z \sim \mathcal{N}(0, 1)$; i.e., $\phi_{0.10}^Z$ solves

$$F_Z(\phi_{0.10}^Z) = P(Z \leq \phi_{0.10}^Z) = 0.10.$$

From the standard normal table (Table 4), we see that

$$\phi_{0.10}^Z \approx -1.28.$$

We are left to solve the equation:

$$\frac{\phi_{0.10}^Y - 18}{4} = \phi_{0.10}^Z \approx -1.28 \implies \phi_{0.10}^Y \approx -1.28(4) + 18 = 12.88.$$

Thus, 90 percent of all contamination levels are greater than 12.88 parts per million. \square

4.7 The gamma family of distributions

INTRODUCTION: In this section, we examine an important family of probability distributions; namely, those in the **gamma family**. There are three well-known “named distributions” in this family:

- the exponential distribution
- the gamma distribution
- the χ^2 distribution.

NOTE: The exponential and gamma distributions are popular models for **lifetime** random variables; i.e., random variables that record “time to event” measurements, such as the lifetimes of an electrical component, death times for human subjects, waiting times in Poisson processes, etc. Other lifetime distributions include the lognormal, Weibull, loggamma, among others.

4.7.1 Exponential distribution

TERMINOLOGY: A random variable Y is said to have an **exponential distribution** with parameter $\beta > 0$ if its pdf is given by

$$f_Y(y) = \begin{cases} \frac{1}{\beta}e^{-y/\beta}, & y > 0 \\ 0, & \text{otherwise.} \end{cases}$$

Shorthand notation is $Y \sim \text{exponential}(\beta)$. The value of β determines the scale of the distribution, so it is called a **scale parameter**.

EXERCISE: Show that the exponential pdf integrates to 1.

EXPONENTIAL MGF: Suppose that $Y \sim \text{exponential}(\beta)$. The mgf of Y is given by

$$m_Y(t) = \frac{1}{1 - \beta t},$$

for $t < 1/\beta$.

Proof. From the definition of the mgf, we have

$$\begin{aligned} m_Y(t) = E(e^{tY}) &= \int_0^\infty e^{ty} \left(\frac{1}{\beta} e^{-y/\beta} \right) dy = \frac{1}{\beta} \int_0^\infty e^{ty - y/\beta} dy \\ &= \frac{1}{\beta} \int_0^\infty e^{-y[(1/\beta) - t]} dy \\ &= \frac{1}{\beta} \left\{ - \left(\frac{1}{\frac{1}{\beta} - t} \right) e^{-y[(1/\beta) - t]} \right\} \Bigg|_{y=0}^\infty \\ &= \left(\frac{1}{1 - \beta t} \right) \left\{ e^{-y[(1/\beta) - t]} \Big|_{y=\infty}^0 \right\}. \end{aligned}$$

In the last expression, note that

$$\lim_{y \rightarrow \infty} e^{-y[(1/\beta) - t]} < \infty$$

if and only if $(1/\beta) - t > 0$, i.e., $t < 1/\beta$. Thus, for $t < 1/\beta$, we have

$$m_Y(t) = \left(\frac{1}{1 - \beta t} \right) e^{-y[(1/\beta) - t]} \Big|_{y=\infty}^0 = \left(\frac{1}{1 - \beta t} \right) - 0 = \frac{1}{1 - \beta t}.$$

Note that $(-h, h)$ with $h = 1/\beta$ is an open neighborhood around 0 for which $m_Y(t)$ exists. \square

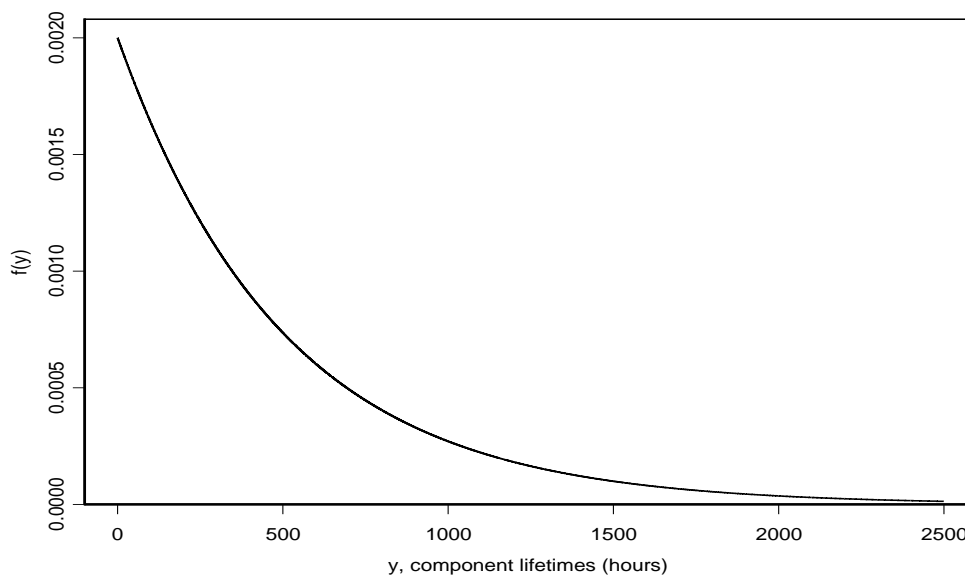


Figure 4.11: The probability density function, $f_Y(y)$, in Example 4.12. A model for electrical component lifetimes.

MEAN AND VARIANCE: Suppose that $Y \sim \text{exponential}(\beta)$. The mean and variance of Y are given by

$$E(Y) = \beta \quad \text{and} \quad V(Y) = \beta^2.$$

Proof: Exercise. \square

Example 4.12. The lifetime of an electrical component has an exponential distribution with mean $\beta = 500$ hours. What is the probability that a randomly selected component fails before 100 hours? lasts between 250 and 750 hours?

SOLUTION. With $\beta = 500$, the pdf for Y is given by

$$f_Y(y) = \begin{cases} \frac{1}{500}e^{-y/500}, & y > 0 \\ 0, & \text{otherwise.} \end{cases}$$

This pdf is depicted in Figure 4.11. Thus, the probability of failing before 100 hours is

$$P(Y < 100) = \int_0^{100} \frac{1}{500}e^{-y/500} dy \approx 0.181.$$

Similarly, the probability of failing between 250 and 750 hours is

$$P(250 < Y < 750) = \int_{250}^{750} \frac{1}{500} e^{-y/500} dy \approx 0.383. \quad \square$$

EXPONENTIAL CDF: Suppose that $Y \sim \text{exponential}(\beta)$. Then, the cdf of Y exists in closed form and is given by

$$F_Y(y) = \begin{cases} 0, & y \leq 0 \\ 1 - e^{-y/\beta}, & y > 0. \end{cases}$$

Proof. Exercise. \square

THE MEMORYLESS PROPERTY: Suppose that $Y \sim \text{exponential}(\beta)$, and let r and s be positive constants. Then

$$P(Y > r + s | Y > r) = P(Y > s).$$

That is, given that the lifetime Y has exceeded r , the probability that Y exceeds $r+s$ (i.e., an additional s units) is the same as if we were to look at Y unconditionally lasting until time s . Put another way, that Y has actually “made it” to time r has been forgotten. The exponential random variable is the only continuous random variable that possesses the memoryless property.

RELATIONSHIP WITH A POISSON PROCESS: Suppose that we are observing events according to a Poisson process with rate $\lambda = 1/\beta$, and let the random variable W denote the time until the first occurrence. Then, $W \sim \text{exponential}(\beta)$.

Proof: Clearly, W is a continuous random variable with nonnegative support. Thus, for $w \geq 0$, we have

$$\begin{aligned} F_W(w) = P(W \leq w) &= 1 - P(W > w) \\ &= 1 - P(\{\text{no events in } [0, w]\}) \\ &= 1 - \frac{e^{-\lambda w} (\lambda w)^0}{0!} \\ &= 1 - e^{-\lambda w}. \end{aligned}$$

Substituting $\lambda = 1/\beta$, we have $F_W(w) = 1 - e^{-w/\beta}$, the cdf of an exponential random variable with mean β . Thus, the result follows. \square

Example 4.13. Suppose that customers arrive at a check-out according to a Poisson process with mean $\lambda = 12$ per hour. What is the probability that we will have to wait longer than 10 minutes to see the first customer? NOTE: 10 minutes is 1/6th of an hour. SOLUTION. The time until the first arrival, say W , follows an exponential distribution with mean $\beta = 1/\lambda = 1/12$, so that the cdf of W , for $w > 0$, is $F_W(w) = 1 - e^{-12w}$. Thus, the desired probability is

$$P(W > 1/6) = 1 - P(W \leq 1/6) = 1 - F_W(1/6) = 1 - [1 - e^{-12(1/6)}] = e^{-2} \approx 0.135. \quad \square$$

4.7.2 Gamma distribution

TERMINOLOGY: The **gamma function** is a real function of t , defined by

$$\Gamma(t) = \int_0^{\infty} y^{t-1} e^{-y} dy,$$

for all $t > 0$. The gamma function satisfies the recursive relationship

$$\Gamma(\alpha) = (\alpha - 1)\Gamma(\alpha - 1),$$

for $\alpha > 1$. From this fact, we can deduce that if α is an integer, then

$$\Gamma(\alpha) = (\alpha - 1)!$$

For example, $\Gamma(5) = 4! = 24$.

TERMINOLOGY: A random variable Y is said to have a **gamma distribution** with parameters $\alpha > 0$ and $\beta > 0$ if its pdf is given by

$$f_Y(y) = \begin{cases} \frac{1}{\Gamma(\alpha)\beta^\alpha} y^{\alpha-1} e^{-y/\beta}, & y > 0 \\ 0, & \text{otherwise.} \end{cases}$$

Shorthand notation is $Y \sim \text{gamma}(\alpha, \beta)$. The gamma distribution is indexed by two parameters:

α = the **shape** parameter

β = the **scale** parameter.

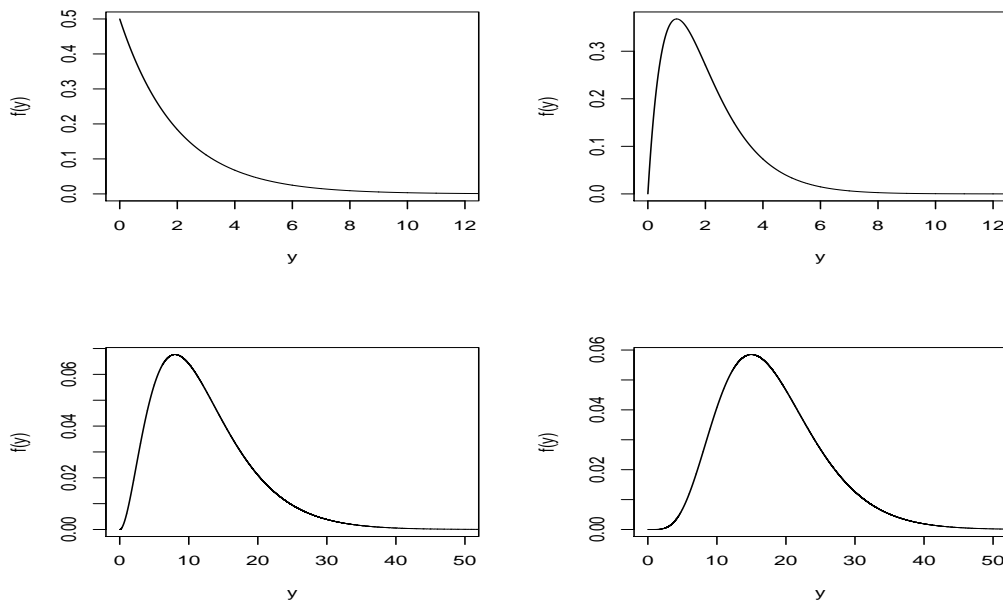


Figure 4.12: Four gamma pdfs. Upper left: $\alpha = 1$, $\beta = 2$. Upper right: $\alpha = 2$, $\beta = 1$. Lower left: $\alpha = 3$, $\beta = 4$. Lower right: $\alpha = 6$, $\beta = 3$.

REMARK: By changing the values of α and β , the gamma pdf can assume many shapes. This makes the gamma distribution popular for modeling lifetime data. Note that when $\alpha = 1$, the gamma pdf reduces to the exponential(β) pdf. That is, the exponential pdf is a “special” gamma pdf.

Example 4.14. Show that the gamma(α, β) pdf integrates to 1.

SOLUTION. Change the variable of integration to $u = y/\beta$ so that $du = dy/\beta$ and $dy = \beta du$. We have

$$\begin{aligned}
 \int_0^{\infty} f_Y(y) dy &= \int_0^{\infty} \frac{1}{\Gamma(\alpha)\beta^\alpha} y^{\alpha-1} e^{-y/\beta} dy \\
 &= \frac{1}{\Gamma(\alpha)} \int_0^{\infty} \frac{1}{\beta^\alpha} (\beta u)^{\alpha-1} e^{-\beta u/\beta} \beta du \\
 &= \frac{1}{\Gamma(\alpha)} \int_0^{\infty} u^{\alpha-1} e^{-u} du \\
 &= \frac{\Gamma(\alpha)}{\Gamma(\alpha)} = 1. \quad \square
 \end{aligned}$$

GAMMA MGF: Suppose that $Y \sim \text{gamma}(\alpha, \beta)$. The mgf of Y is

$$m_Y(t) = \left(\frac{1}{1 - \beta t} \right)^\alpha,$$

for $t < 1/\beta$.

Proof. From the definition of the mgf, we have

$$\begin{aligned} m_Y(t) = E(e^{tY}) &= \int_0^\infty e^{ty} \left[\frac{1}{\Gamma(\alpha)\beta^\alpha} y^{\alpha-1} e^{-y/\beta} \right] dy \\ &= \int_0^\infty \frac{1}{\Gamma(\alpha)\beta^\alpha} y^{\alpha-1} e^{ty-y/\beta} dy \\ &= \int_0^\infty \frac{1}{\Gamma(\alpha)\beta^\alpha} y^{\alpha-1} e^{-y[(1/\beta)-t]} dy \\ &= \int_0^\infty \frac{1}{\Gamma(\alpha)\beta^\alpha} y^{\alpha-1} e^{-y/[(1/\beta)-t]^{-1}} dy \\ &= \frac{\eta^\alpha}{\beta^\alpha} \int_0^\infty \frac{1}{\Gamma(\alpha)\eta^\alpha} y^{\alpha-1} e^{-y/\eta} dy, \end{aligned}$$

where $\eta = [(1/\beta) - t]^{-1}$. If $\eta > 0 \iff t < 1/\beta$, then the last integral equals 1, because the integrand is the $\text{gamma}(\alpha, \eta)$ pdf and integration is over $R = \{y : 0 < y < \infty\}$.

Thus,

$$m_Y(t) = \left(\frac{\eta}{\beta} \right)^\alpha = \left\{ \frac{1}{\beta[(1/\beta) - t]} \right\}^\alpha = \left(\frac{1}{1 - \beta t} \right)^\alpha.$$

Note that $(-h, h)$ with $h = 1/\beta$ is an open neighborhood around 0 for which $m_Y(t)$ exists. \square

MEAN AND VARIANCE: If $Y \sim \text{gamma}(\alpha, \beta)$, then

$$E(Y) = \alpha\beta \quad \text{and} \quad V(Y) = \alpha\beta^2.$$

NOTE: Upon closer inspection, we see that the nonzero part of the $\text{gamma}(\alpha, \beta)$ pdf

$$f_Y(y) = \frac{1}{\Gamma(\alpha)\beta^\alpha} y^{\alpha-1} e^{-y/\beta}$$

consists of two parts:

- the **kernel** of the pdf: $y^{\alpha-1} e^{-y/\beta}$
- a **constant** out front: $1/\Gamma(\alpha)\beta^\alpha$.

The kernel is the “guts” of the formula, while the constant out front is simply the “right quantity” that makes $f_Y(y)$ a valid pdf; i.e., the constant which makes $f_Y(y)$ integrate to 1. Note that because

$$\int_0^{\infty} \frac{1}{\Gamma(\alpha)\beta^\alpha} y^{\alpha-1} e^{-y/\beta} dy = 1,$$

it follows immediately that

$$\int_0^{\infty} y^{\alpha-1} e^{-y/\beta} dy = \Gamma(\alpha)\beta^\alpha.$$

This fact is extremely fascinating in its own right, and it is very helpful too; we will use it repeatedly.

Example 4.15. Suppose that Y has pdf given by

$$f_Y(y) = \begin{cases} cy^2e^{-y/4}, & y > 0 \\ 0, & \text{otherwise.} \end{cases}$$

(a) What is the value of c that makes this a valid pdf?

(b) What is the mgf of Y ?

(c) What are the mean and variance of Y ?

SOLUTIONS. Note that $y^2e^{-y/4}$ is a gamma kernel with $\alpha = 3$ and $\beta = 4$. Thus, the constant out front is

$$c = \frac{1}{\Gamma(\alpha)\beta^\alpha} = \frac{1}{\Gamma(3)4^3} = \frac{1}{2(64)} = \frac{1}{128}.$$

The mgf of Y is

$$m_Y(t) = \left(\frac{1}{1 - \beta t} \right)^\alpha = \left(\frac{1}{1 - 4t} \right)^3,$$

for $t < 1/4$. Finally,

$$E(Y) = \alpha\beta = 3(4) = 12$$

$$V(Y) = \alpha\beta^2 = 3(4^2) = 48.$$

RELATIONSHIP WITH A POISSON PROCESS: Suppose that we are observing events according to a Poisson process with rate $\lambda = 1/\beta$, and let the random variable W denote the time until the α th occurrence. Then, $W \sim \text{gamma}(\alpha, \beta)$.

Proof: Clearly, W is a continuous random variable with nonnegative support. Thus, for $w \geq 0$, we have

$$\begin{aligned} F_W(w) = P(W \leq w) &= 1 - P(W > w) \\ &= 1 - P(\{\text{fewer than } \alpha \text{ events in } [0, w]\}) \\ &= 1 - \sum_{j=0}^{\alpha-1} \frac{e^{-\lambda w} (\lambda w)^j}{j!}. \end{aligned}$$

The pdf of W , $f_W(w)$, is equal to $F'_W(w)$, provided that this derivative exists. For $w > 0$,

$$\begin{aligned} f_W(w) = F'_W(w) &= \lambda e^{-\lambda w} - e^{-\lambda w} \underbrace{\sum_{j=1}^{\alpha-1} \left[\frac{j(\lambda w)^{j-1} \lambda}{j!} - \frac{(\lambda w)^j \lambda}{j!} \right]}_{\text{telescoping sum}} \\ &= \lambda e^{-\lambda w} - e^{-\lambda w} \left[\lambda - \frac{\lambda(\lambda w)^{\alpha-1}}{(\alpha-1)!} \right] \\ &= \frac{\lambda(\lambda w)^{\alpha-1} e^{-\lambda w}}{(\alpha-1)!} = \frac{\lambda^\alpha}{\Gamma(\alpha)} w^{\alpha-1} e^{-\lambda w}. \end{aligned}$$

Substituting $\lambda = 1/\beta$,

$$f_W(w) = \frac{1}{\Gamma(\alpha)\beta^\alpha} w^{\alpha-1} e^{-w/\beta},$$

for $w > 0$, which is the pdf for the gamma(α, β) distribution. \square

Example 4.16. Suppose that customers arrive at a check-out according to a Poisson process with mean $\lambda = 12$ per hour. What is the probability that we will have to wait longer than 10 minutes to see the third customer? NOTE: 10 minutes is 1/6th of an hour.

SOLUTION. The time until the third arrival, say W , follows a gamma distribution with parameters $\alpha = 3$ and $\beta = 1/\lambda = 1/12$, so that the pdf of W , for $w > 0$,

$$f_W(w) = 864w^2 e^{-12w}.$$

Thus, the desired probability is

$$\begin{aligned} P(W > 1/6) &= 1 - P(W \leq 1/6) \\ &= 1 - \int_0^{1/6} 864w^2 e^{-12w} dw \approx 0.677. \quad \square \end{aligned}$$

4.7.3 χ^2 distribution

TERMINOLOGY: Let ν be a positive integer. In the gamma(α, β) family, when

$$\alpha = \nu/2$$

$$\beta = 2,$$

we call the resulting distribution a χ^2 **distribution** with ν degrees of freedom. We write $Y \sim \chi^2(\nu)$.

NOTE: At this point, it suffices to accept the fact that the χ^2 distribution is simply a “special” gamma distribution. However, it should be noted that the χ^2 distribution is used extensively in applied statistics. In fact, many statistical procedures used in practice are valid because of this model.

χ^2 *PDF:* If $Y \sim \chi^2(\nu)$, then the pdf of Y is

$$f_Y(y) = \begin{cases} \frac{1}{\Gamma(\frac{\nu}{2})2^{\nu/2}} y^{(\nu/2)-1} e^{-y/2}, & y > 0 \\ 0, & \text{otherwise.} \end{cases}$$

χ^2 *MGF:* Suppose that $Y \sim \chi^2(\nu)$. The mgf of Y is

$$m_Y(t) = \left(\frac{1}{1-2t} \right)^{\nu/2},$$

for $t < 1/2$.

MEAN AND VARIANCE: If $Y \sim \chi^2(\nu)$, then

$$E(Y) = \nu \quad \text{and} \quad V(Y) = 2\nu.$$

TABLED VALUES FOR CDF: Because the χ^2 distribution is so pervasive in applied statistics, tables of probabilities are common. Appendix III, Table 6 (WMS, pp 850-851) provides the upper α quantiles χ_α^2 which satisfy

$$\alpha = P(Y > \chi_\alpha^2) = \int_{\chi_\alpha^2}^{\infty} \frac{1}{\Gamma(\frac{\nu}{2})2^{\nu/2}} y^{(\nu/2)-1} e^{-y/2} dy$$

for different values of α and degrees of freedom ν .

4.8 Beta distribution

TERMINOLOGY: A random variable Y is said to have a **beta distribution** with parameters $\alpha > 0$ and $\beta > 0$ if its pdf is given by

$$f_Y(y) = \begin{cases} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} y^{\alpha-1} (1-y)^{\beta-1}, & 0 < y < 1 \\ 0, & \text{otherwise.} \end{cases}$$

Since the support of Y is $R = \{y : 0 < y < 1\}$, the beta distribution is a popular probability model for proportions. Shorthand notation is $Y \sim \text{beta}(\alpha, \beta)$.

NOTE: Upon closer inspection, we see that the nonzero part of the $\text{beta}(\alpha, \beta)$ pdf

$$f_Y(y) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} y^{\alpha-1} (1-y)^{\beta-1}$$

consists of two parts:

- the **kernel** of the pdf: $y^{\alpha-1}(1-y)^{\beta-1}$
- a **constant** out front: $\Gamma(\alpha + \beta)/\Gamma(\alpha)\Gamma(\beta)$.

Again, the kernel is the “guts” of the formula, while the constant out front is simply the “right quantity” that makes $f_Y(y)$ a valid pdf; i.e., the constant which makes $f_Y(y)$ integrate to 1. Note that because

$$\int_0^1 \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} y^{\alpha-1} (1-y)^{\beta-1} dy = 1,$$

it follows immediately that

$$\int_0^1 y^{\alpha-1} (1-y)^{\beta-1} dy = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha + \beta)}.$$

BETA PDF SHAPES: The beta pdf is very flexible. That is, by changing the values of α and β , we can come up with many different pdf shapes. See Figure 4.13 for examples.

- When $\alpha = \beta$, the pdf is **symmetric** about the line $y = \frac{1}{2}$.
- When $\alpha < \beta$, the pdf is **skewed right** (i.e., smaller values of y are more likely).

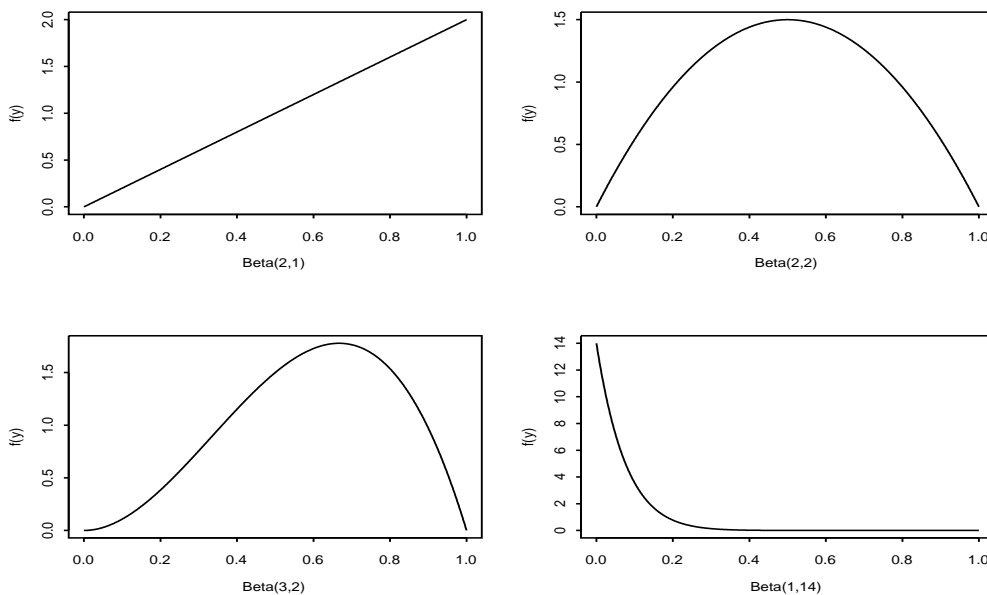


Figure 4.13: Four beta pdfs. Upper left: $\alpha = 2, \beta = 1$. Upper right: $\alpha = 2, \beta = 2$. Lower left: $\alpha = 3, \beta = 2$. Lower right: $\alpha = 1, \beta = 14$.

- When $\alpha > \beta$, the pdf is **skewed left** (i.e., larger values of y are more likely).
- When $\alpha = \beta = 1$, the beta pdf reduces to the $\mathcal{U}(0, 1)$ pdf!

BETA MGF: The beta(α, β) mgf exists, but not in closed form. Hence, we'll compute moments directly.

MEAN AND VARIANCE: If $Y \sim \text{beta}(\alpha, \beta)$, then

$$E(Y) = \frac{\alpha}{\alpha + \beta} \quad \text{and} \quad V(Y) = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}.$$

Proof. We will derive $E(Y)$ only. From the definition of expected value, we have

$$\begin{aligned} E(Y) &= \int_0^1 y f_Y(y) dy = \int_0^1 y \left[\frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} y^{\alpha-1} (1 - y)^{\beta-1} \right] dy \\ &= \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \int_0^1 \underbrace{y^{(\alpha+1)-1} (1 - y)^{\beta-1}}_{\text{beta}(\alpha+1, \beta) \text{ kernel}} dy. \end{aligned}$$

Note that the last integrand is a beta kernel with parameters $\alpha + 1$ and β . Because integration is over $R = \{y : 0 < y < 1\}$, we have

$$\int_0^1 y^{(\alpha+1)-1}(1-y)^{\beta-1} = \frac{\Gamma(\alpha+1)\Gamma(\beta)}{\Gamma(\alpha+1+\beta)}$$

and thus

$$\begin{aligned} E(Y) &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(\alpha+1)\Gamma(\beta)}{\Gamma(\alpha+1+\beta)} \\ &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)} \frac{\Gamma(\alpha+1)}{\Gamma(\alpha+1+\beta)} \\ &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)} \frac{\alpha\Gamma(\alpha)}{(\alpha+\beta)\Gamma(\alpha+\beta)} = \frac{\alpha}{\alpha+\beta}. \end{aligned}$$

To derive $V(Y)$, first find $E(Y^2)$ using similar calculations. Use the variance computing formula $V(Y) = E(Y^2) - [E(Y)]^2$ and simplify. \square

Example 4.17. At a health clinic, suppose that Y , the proportion of individuals infected with a new flu virus (e.g., H1N1, etc.), varies daily according to a beta distribution with pdf

$$f_Y(y) = \begin{cases} 20(1-y)^{19}, & 0 < y < 1 \\ 0, & \text{otherwise.} \end{cases}$$

This distribution is displayed in Figure 4.14.

QUESTIONS.

- What are the parameters in this distribution; i.e., what are α and β ?
- What is the mean proportion of individuals infected?
- Find $\phi_{0.95}$, the 95th percentile of this distribution.
- Treating daily infection counts as independent (from day to day), what is the probability that during any given 5-day span, there is at least 2 days where the infection proportion is above 10 percent?

SOLUTIONS.

- $\alpha = 1$ and $\beta = 20$.
- $E(Y) = 1/(1+20) \approx 0.048$.
- The 95th percentile $\phi_{0.95}$ solves

$$P(Y \leq \phi_{0.95}) = \int_0^{\phi_{0.95}} 20(1-y)^{19} dy = 0.95.$$

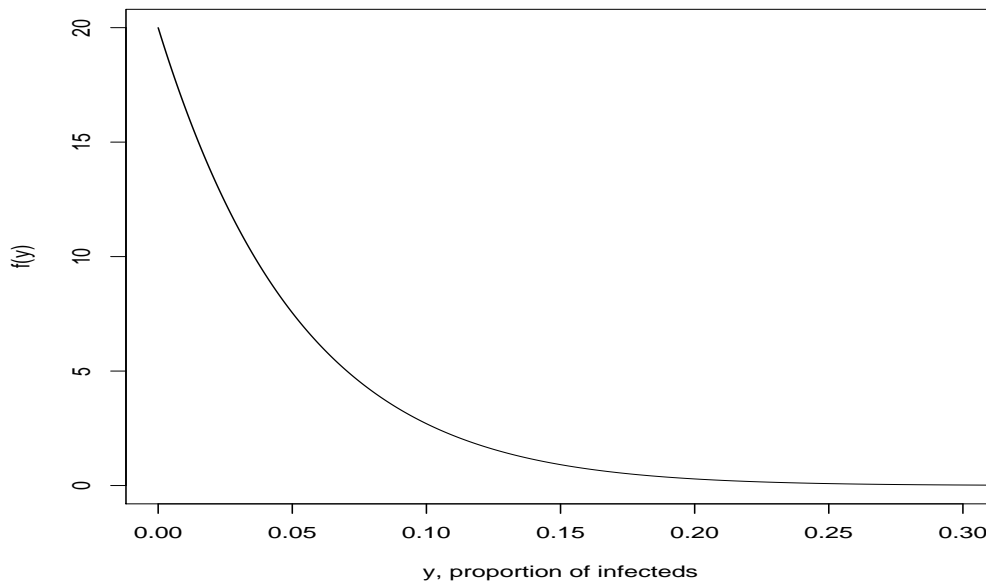


Figure 4.14: The probability density function, $f_Y(y)$, in Example 4.17. A model for the proportion of infected individuals.

Let $u = 1 - y$ so that $du = -dy$. The limits on the integral must change:

$$y : 0 \longrightarrow \phi_{0.95}$$

$$u : 1 \longrightarrow 1 - \phi_{0.95}$$

Thus, we are left to solve

$$0.95 = - \int_1^{1-\phi_{0.95}} 20u^{19} du = u^{20} \Big|_{1-\phi_{0.95}}^1 = 1 - (1 - \phi_{0.95})^{20}$$

for $\phi_{0.95}$. We get

$$\phi_{0.95} = 1 - (0.05)^{1/20} \approx 0.139.$$

(d) First, we compute

$$P(Y > 0.1) = \int_{0.1}^1 20(1-y)^{19} dy = \int_0^{0.9} 20u^{19} du = u^{20} \Big|_0^{0.9} = (0.9)^{20} \approx 0.122.$$

This is the probability that the infection proportion exceeds 0.10 on any given day. Now, we treat each day as a “trial,” and let X denote the number of days where “the

infection proportion is above 10 percent” (i.e., a “success”). Because days are assumed independent, $X \sim b(5, 0.122)$ and

$$\begin{aligned} P(X \geq 2) &= 1 - P(X = 0) - P(X = 1) \\ &= 1 - \binom{5}{0}(0.122)^0(1 - 0.122)^5 - \binom{5}{1}(0.122)^1(1 - 0.122)^4 \approx 0.116. \quad \square \end{aligned}$$

4.9 Chebyshev’s Inequality

MARKOV’S INEQUALITY: Suppose that X is a nonnegative random variable with pdf (pmf) $f_X(x)$ and let c be a positive constant. Markov’s Inequality puts a bound on the upper tail probability $P(X > c)$; that is,

$$P(X > c) \leq \frac{E(X)}{c}.$$

Proof. First, define the event $B = \{x : x > c\}$. We know that

$$\begin{aligned} E(X) &= \int_0^{\infty} x f_X(x) dx = \int_B x f_X(x) dx + \int_{\overline{B}} x f_X(x) dx \\ &\geq \int_B x f_X(x) dx \\ &\geq \int_B c f_X(x) dx = cP(X > c). \quad \square \end{aligned}$$

CHEBYSHEV’S INEQUALITY: Let Y be any random variable, discrete or continuous, with mean μ and variance $\sigma^2 < \infty$. For $k > 0$,

$$P(|Y - \mu| > k\sigma) \leq \frac{1}{k^2}.$$

Proof. Applying Markov’s Inequality with $X = (Y - \mu)^2$ and $c = k^2\sigma^2$, we have

$$P(|Y - \mu| > k\sigma) = P[(Y - \mu)^2 > k^2\sigma^2] \leq \frac{E[(Y - \mu)^2]}{k^2\sigma^2} = \frac{\sigma^2}{k^2\sigma^2} = \frac{1}{k^2}. \quad \square$$

REMARK: The beauty of Chebyshev’s result is that it applies to any random variable Y . In words, $P(|Y - \mu| > k\sigma)$ is the probability that the random variable Y will differ from the mean μ by more than k standard deviations. If we do not know how Y is distributed,

we can not compute $P(|Y - \mu| > k\sigma)$ exactly, but, at least we can put an upper bound on this probability; this is what Chebyshev's result allows us to do. Note that

$$P(|Y - \mu| > k\sigma) = 1 - P(|Y - \mu| \leq k\sigma) = 1 - P(\mu - k\sigma \leq Y \leq \mu + k\sigma).$$

Thus, it must be the case that

$$P(|Y - \mu| \leq k\sigma) = P(\mu - k\sigma \leq Y \leq \mu + k\sigma) \geq 1 - \frac{1}{k^2}.$$

Example 4.18. Suppose that Y represents the amount of precipitation (in inches) observed annually in Barrow, AK. The exact probability distribution for Y is unknown, but, from historical information, it is posited that $\mu = 4.5$ and $\sigma = 1$. What is a lower bound on the probability that there will be between 2.5 and 6.5 inches of precipitation during the next year?

SOLUTION: We want to compute a lower bound for $P(2.5 \leq Y \leq 6.5)$. Note that

$$P(2.5 \leq Y \leq 6.5) = P(|Y - \mu| \leq 2\sigma) \geq 1 - \frac{1}{2^2} = 0.75.$$

Thus, we know that $P(2.5 \leq Y \leq 6.5) \geq 0.75$. The chances are good that the annual precipitation will be between 2.5 and 6.5 inches.

4.10 Expectations of piecewise functions and mixed distributions

4.10.1 Expectations of piecewise functions

RECALL: Suppose that Y is a continuous random variable with pdf $f_Y(y)$ and support R . Let $g(Y)$ be a function of Y . The **expected value** of $g(Y)$ is given by

$$E[g(Y)] = \int_R g(y)f_Y(y)dy,$$

provided that this integral exists.

REMARK: In mathematical expectation examples up until now, we have always considered functions g which were continuous and differentiable everywhere; e.g., $g(y) = y^2$,

$g(y) = e^{ty}$, $g(y) = \ln y$, etc. We now extend the notion of mathematical expectation to handle piecewise functions (which may not even be continuous).

EXTENSION: Suppose that Y is a continuous random variable with pdf $f_Y(y)$ and support R , where R can be expressed as the union of k disjoint sets; i.e.,

$$R = B_1 \cup B_2 \cup \cdots \cup B_k,$$

where $B_i \subset \mathcal{R}$ and $B_i \cap B_j = \emptyset$, for $i \neq j$. Let $g : R \rightarrow \mathcal{R}$ be a function which can be written as

$$g(y) = \sum_{i=1}^k g_i(y) \mathcal{I}_{B_i}(y),$$

where $g_i : B_i \rightarrow \mathcal{R}$ is a continuous function and $\mathcal{I}_{B_i} : B_i \rightarrow \{0, 1\}$ is the **indicator function** that $y \in B_i$; i.e.,

$$\mathcal{I}_{B_i}(y) = \begin{cases} 1, & y \in B_i \\ 0, & y \notin B_i. \end{cases}$$

The expected value of the function

$$g(Y) = \sum_{i=1}^k g_i(Y) \mathcal{I}_{B_i}(Y),$$

is equal to

$$\begin{aligned} E[g(Y)] &= \int_R g(y) f_Y(y) dy = \int_R \sum_{i=1}^k g_i(y) \mathcal{I}_{B_i}(y) f_Y(y) dy \\ &= \sum_{i=1}^k \int_R g_i(y) \mathcal{I}_{B_i}(y) f_Y(y) dy \\ &= \sum_{i=1}^k \int_{B_i} g_i(y) f_Y(y) dy. \end{aligned}$$

That is, to compute the expectation for a piecewise function $g(Y)$, we simply compute the expectation of each $g_i(Y)$ over each set B_i and add up the results. Note that if Y is discrete, the same formula applies except integration is replaced by summation and the pdf $f_Y(y)$ is replaced by a pmf $p_Y(y)$.

Example 4.19. An insurance policy reimburses a policy holder up to a limit of 10,000 dollars. If the loss exceeds 10,000 dollars, the insurance company will pay 10,000 dollars

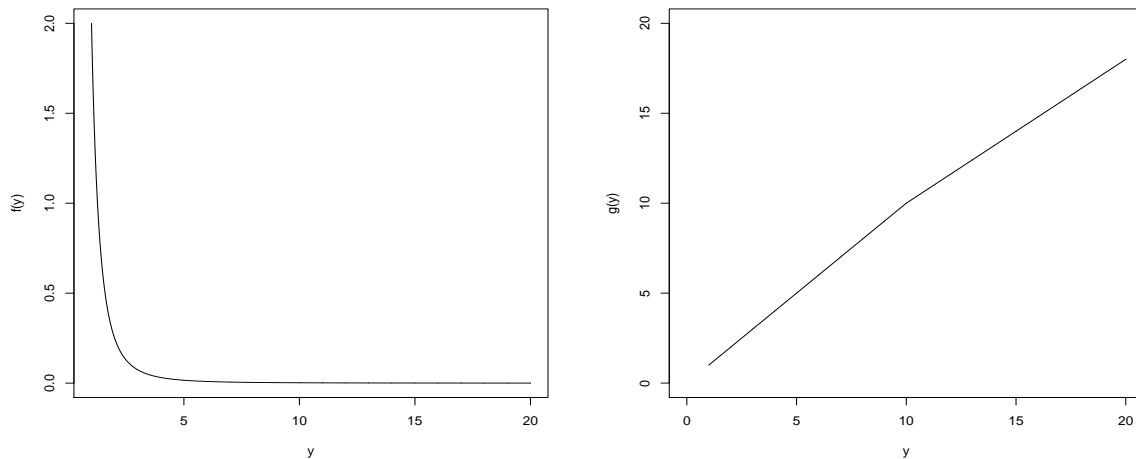


Figure 4.15: *Left: The probability density function for the loss incurred in Example 4.19. Right: The function which describes the amount of benefit paid.*

plus 80 percent of the loss that exceeds 10,000 dollars. Suppose that the policy holder's loss Y (measured in \$1,000s) is a random variable with pdf

$$f_Y(y) = \begin{cases} 2/y^3, & y > 1 \\ 0, & \text{otherwise.} \end{cases}$$

This pdf is plotted in Figure 4.15 (left). What is the expected value of the benefit paid to the policy holder?

SOLUTION. Let $g(Y)$ denote the benefit paid to the policy holder; i.e.,

$$g(Y) = \begin{cases} Y, & 1 < Y < 10 \\ 10 + 0.8(Y - 10), & Y \geq 10. \end{cases}$$

This function is plotted in Figure 4.15 (right). We have

$$\begin{aligned} E[g(Y)] &= \int_1^{\infty} g(y) f_Y(y) dy \\ &= \int_1^{10} y \left(\frac{2}{y^3} \right) dy + \int_{10}^{\infty} [10 + 0.8(y - 10)] \left(\frac{2}{y^3} \right) dy = 1.98. \end{aligned}$$

Thus, the expected benefit paid to the policy holder is \$1,980.

EXERCISE: Find $V[g(Y)]$, the variance of the benefit paid to the policy holder.

4.10.2 Mixed distributions

TERMINOLOGY: We define a **mixed distribution** as one with discrete and continuous parts (more general definitions are available). In particular, suppose that Y_1 is a discrete random variable with cdf $F_{Y_1}(y)$ and that Y_2 is a continuous random variable with cdf $F_{Y_2}(y)$. A mixed random variable Y has cdf

$$F_Y(y) = c_1 F_{Y_1}(y) + c_2 F_{Y_2}(y),$$

for all $y \in \mathcal{R}$, where the constants c_1 and c_2 satisfy $c_1 + c_2 = 1$. These constants are called **mixing constants**. It is straightforward to show that the function $F_Y(y)$ satisfies the cdf requirements (see pp 64, notes).

RESULT: Let Y have the mixed distribution

$$F_Y(y) = c_1 F_{Y_1}(y) + c_2 F_{Y_2}(y),$$

where Y_1 is a discrete random variable with cdf $F_{Y_1}(y)$ and Y_2 is a continuous random variable with cdf $F_{Y_2}(y)$. Let $g(Y)$ be a function of Y . Then,

$$E[g(Y)] = c_1 E[g(Y_1)] + c_2 E[g(Y_2)],$$

where each expectation is taken with respect to the appropriate distribution.

Example 4.20. A standard experiment in the investigation of carcinogenic substances is one in which laboratory animals (e.g., rats, etc.) are exposed to a toxic substance. Suppose that the time from exposure until death follows an exponential distribution with mean $\beta = 10$ hours. Suppose additionally that the animal is sacrificed after 24 hours if death has not been observed. Let Y denote the death time for an animal in this experiment. Find the cdf for Y and compute $E(Y)$.

SOLUTION. Let Y_2 denote the time until death for animals who die before 24 hours. We are given that $Y_2 \sim \text{exponential}(10)$; the cdf of Y_2 is

$$F_{Y_2}(y) = \begin{cases} 0, & y \leq 0 \\ 1 - e^{-y/10}, & y > 0. \end{cases}$$

The probability that an animal has not died before 24 hours is

$$P(Y_2 < 24) = F_{Y_2}(24) = 1 - e^{-24/10} \approx 0.909.$$

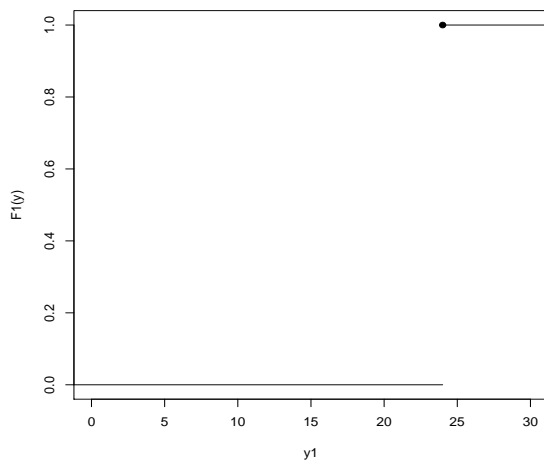
There is one discrete point in the distribution for Y , namely, at the value $y = 24$ which occurs with probability

$$1 - P(Y_1 < 24) \approx 1 - 0.909 = 0.091.$$

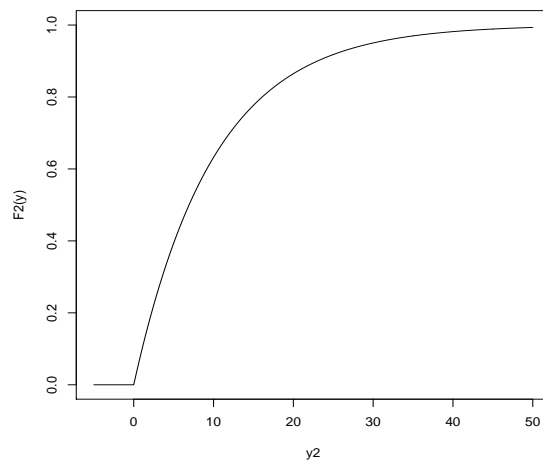
Define Y_1 to be a random variable with cdf

$$F_{Y_1}(y) = \begin{cases} 0, & y < 24 \\ 1, & y \geq 24, \end{cases}$$

that is, Y_1 has a **degenerate distribution** at the value $y = 24$. Here are the cdfs of Y_1 and Y_2 , plotted side by side.



cdf, $F_{Y_1}(y)$



cdf, $F_{Y_2}(y)$

The cdf of Y is

$$F_Y(y) = c_1 F_{Y_1}(y) + c_2 F_{Y_2}(y),$$

where $c_1 = 0.091$ and $c_2 = 0.909$. The mean of Y is

$$E(Y) = 0.091E(Y_1) + 0.909E(Y_2) = 0.091(24) + 0.909(10) = 11.274 \text{ hours. } \square$$

EXERCISE: Find $V(Y)$.

5 Multivariate Distributions

Complementary reading from WMS: Chapter 5.

5.1 Introduction

REMARK: Up until now, we have discussed **univariate** random variables (and their associated probability distributions, moment generating functions, means, variances, etc.). In practice, however, one is often interested in multiple random variables. Consider the following examples:

- In an educational assessment program, we want to predict a student's posttest score (Y_2) from her pretest score (Y_1).
- In a clinical trial, physicians want to characterize the concentration of a drug (Y) in one's body as a function of the time (X) from injection.
- An insurance company wants to estimate the amount of loss related to collisions Y_1 and liability Y_2 (both measured in 1000s of dollars).
- Agronomists want to understand the relationship between yield (Y , measured in bushels/acre) and the nitrogen content of the soil (X).
- In a marketing study, the goal is to forecast next month's sales, say Y_n , based on sales figures from the previous $n - 1$ periods, say Y_1, Y_2, \dots, Y_{n-1} .

NOTE: In each of these examples, it is natural to posit a relationship between (or among) the random variables that are involved. This relationship can be described mathematically through a probabilistic model. This model, in turn, allows us to make probability statements involving the random variables (just as univariate models allow us to do this with a single random variable).

TERMINOLOGY: If Y_1 and Y_2 are random variables, then

$$\mathbf{Y} = (Y_1, Y_2)$$

is called a **bivariate random vector**. If Y_1, Y_2, \dots, Y_n denote n random variables, then

$$\mathbf{Y} = (Y_1, Y_2, \dots, Y_n)$$

is called an n -**variate random vector**.

5.2 Discrete random vectors

TERMINOLOGY: Let Y_1 and Y_2 be discrete random variables. Then, (Y_1, Y_2) is called a **discrete random vector**, and the **joint probability mass function (pmf)** of Y_1 and Y_2 is given by

$$p_{Y_1, Y_2}(y_1, y_2) = P(Y_1 = y_1, Y_2 = y_2),$$

for all $(y_1, y_2) \in R$. The set $R \subseteq \mathcal{R}^2$ is the two dimensional support of (Y_1, Y_2) . The function $p_{Y_1, Y_2}(y_1, y_2)$ has the following properties:

- (1) $p_{Y_1, Y_2}(y_1, y_2) > 0$, for all $(y_1, y_2) \in R$
- (2) $\sum_R p_{Y_1, Y_2}(y_1, y_2) = 1$.

RESULT: Suppose that (Y_1, Y_2) is a discrete random vector with pmf $p_{Y_1, Y_2}(y_1, y_2)$. Then,

$$P[(Y_1, Y_2) \in B] = \sum_B p_{Y_1, Y_2}(y_1, y_2),$$

for any set $B \subset \mathcal{R}^2$. That is, the probability of the event $\{(Y_1, Y_2) \in B\}$ is obtained by adding up the probability (mass) associated with each support point in B . If $B = (-\infty, y_1] \times (-\infty, y_2]$, then

$$P[(Y_1, Y_2) \in B] = P(Y_1 \leq y_1, Y_2 \leq y_2) \equiv F_{Y_1, Y_2}(y_1, y_2) = \sum_{t_1 \leq y_1} \sum_{t_2 \leq y_2} p_{Y_1, Y_2}(t_1, t_2)$$

is called the **joint cumulative distribution function (cdf)** of (Y_1, Y_2) .

Example 5.1. Tornadoes are natural disasters that cause millions of dollars in damage each year. An actuary determines that the annual numbers of tornadoes in two Iowa counties (Lee and Van Buren) are jointly distributed as indicated in the table below. Let Y_1 and Y_2 denote the number of tornadoes seen each year in Lee and Van Buren counties, respectively.

$p_{Y_1, Y_2}(y_1, y_2)$	$y_2 = 0$	$y_2 = 1$	$y_2 = 2$	$y_2 = 3$
$y_1 = 0$	0.12	0.06	0.05	0.02
$y_1 = 1$	0.13	0.15	0.12	0.03
$y_1 = 2$	0.05	0.15	0.10	0.02

(a) What is the probability that there is no more than one tornado seen in the two counties combined?

SOLUTION. We want to compute $P(Y_1 + Y_2 \leq 1)$. Note that the support points which correspond to the event $\{Y_1 + Y_2 \leq 1\}$ are $(0, 0)$, $(0, 1)$ and $(1, 0)$. Thus,

$$\begin{aligned} P(Y_1 + Y_2 \leq 1) &= p_{Y_1, Y_2}(0, 0) + p_{Y_1, Y_2}(1, 0) + p_{Y_1, Y_2}(0, 1) \\ &= 0.12 + 0.13 + 0.06 = 0.31. \end{aligned}$$

(b) What is the probability that there are two tornadoes in Lee County?

SOLUTION. We want to compute $P(Y_1 = 2)$. Note that the support points which correspond to the event $\{Y_1 = 2\}$ are $(2, 0)$, $(2, 1)$, $(2, 2)$ and $(2, 3)$. Thus,

$$\begin{aligned} P(Y_1 = 2) &= p_{Y_1, Y_2}(2, 0) + p_{Y_1, Y_2}(2, 1) + p_{Y_1, Y_2}(2, 2) + p_{Y_1, Y_2}(2, 3) \\ &= 0.05 + 0.15 + 0.10 + 0.02 = 0.32. \quad \square \end{aligned}$$

5.3 Continuous random vectors

TERMINOLOGY: Let Y_1 and Y_2 be continuous random variables. Then, (Y_1, Y_2) is called a **continuous random vector**, and the **joint probability density function (pdf)** of Y_1 and Y_2 is denoted by $f_{Y_1, Y_2}(y_1, y_2)$. The joint pdf $f_{Y_1, Y_2}(y_1, y_2)$ is a three-dimensional function whose domain is R , the two-dimensional support of (Y_1, Y_2) .

PROPERTIES: The function $f_{Y_1, Y_2}(y_1, y_2)$ has the following properties:

- (1) $f_{Y_1, Y_2}(y_1, y_2) > 0$, for all $(y_1, y_2) \in R$
- (2) The function $f_{Y_1, Y_2}(y_1, y_2)$ integrates to 1 over its support R ; i.e.,

$$\int_R f_{Y_1, Y_2}(y_1, y_2) dy_1 dy_2 = 1.$$

We realize this is a double integral since R is a two-dimensional set.

RESULT: Suppose that (Y_1, Y_2) is a continuous random vector with joint pdf $f_{Y_1, Y_2}(y_1, y_2)$.

Then,

$$P[(Y_1, Y_2) \in B] = \int_B f_{Y_1, Y_2}(y_1, y_2) dy_1 dy_2,$$

for any set $B \subset \mathcal{R}^2$. We realize that this is a double integral since B is a two-dimensional set in the (y_1, y_2) plane. Therefore, geometrically, $P[(Y_1, Y_2) \in B]$ is the volume under the three-dimensional function $f_{Y_1, Y_2}(y_1, y_2)$ over the two-dimensional set B .

TERMINOLOGY: Suppose that (Y_1, Y_2) is a continuous random vector with joint pdf $f_{Y_1, Y_2}(y_1, y_2)$. The **joint cumulative distribution function (cdf)** for (Y_1, Y_2) is given by

$$F_{Y_1, Y_2}(y_1, y_2) \equiv P(Y_1 \leq y_1, Y_2 \leq y_2) = \int_{-\infty}^{y_2} \int_{-\infty}^{y_1} f_{Y_1, Y_2}(t_1, t_2) dt_1 dt_2,$$

for all $(y_1, y_2) \in \mathcal{R}^2$. It follows upon differentiation that the joint pdf is given by

$$f_{Y_1, Y_2}(y_1, y_2) = \frac{\partial^2}{\partial y_1 \partial y_2} F_{Y_1, Y_2}(y_1, y_2),$$

wherever these mixed partial derivatives are defined.

Example 5.2. A bank operates with a drive-up facility and a walk-up window. On a randomly selected day, let

Y_1 = proportion of time the drive-up facility is in use

Y_2 = proportion of time the walk-up facility is in use.

Suppose that the joint pdf of (Y_1, Y_2) is given by

$$f_{Y_1, Y_2}(y_1, y_2) = \begin{cases} \frac{6}{5}(y_1 + y_2^2), & 0 < y_1 < 1, 0 < y_2 < 1 \\ 0, & \text{otherwise.} \end{cases}$$

Note that the support in this example is

$$R = \{(y_1, y_2) : 0 < y_1 < 1, 0 < y_2 < 1\}.$$

It is very helpful to plot the support of (Y_1, Y_2) in the (y_1, y_2) plane.

(a) What is the probability that neither facility is busy more than $1/4$ of the day? That is, what is $P(Y_1 \leq 1/4, Y_2 \leq 1/4)$?

SOLUTION. Here, we want to integrate the joint pdf $f_{Y_1, Y_2}(y_1, y_2)$ over the set

$$B = \{(y_1, y_2) : 0 < y_1 < 1/4, 0 < y_2 < 1/4\}.$$

The desired probability is

$$\begin{aligned} P(Y_1 \leq 1/4, Y_2 \leq 1/4) &= \int_{y_1=0}^{1/4} \int_{y_2=0}^{1/4} \frac{6}{5}(y_1 + y_2^2) dy_2 dy_1 \\ &= \frac{6}{5} \int_{y_1=0}^{1/4} \left[\left(y_1 y_2 + \frac{y_2^3}{3} \right) \Big|_{y_2=0}^{1/4} \right] dy_1 \\ &= \frac{6}{5} \int_{y_1=0}^{1/4} \left(\frac{y_1}{4} + \frac{1}{192} \right) dy_1 \\ &= \frac{6}{5} \left[\left(\frac{y_1^2}{8} + \frac{y_1}{192} \right) \Big|_{y_1=0}^{1/4} \right] = \frac{6}{5} \left(\frac{1}{128} + \frac{1}{768} \right) \approx 0.0109. \end{aligned}$$

(b) Find the probability that the proportion of time the drive-up facility is in use is less than the proportion of time the walk-up facility is in use; i.e., compute $P(Y_1 < Y_2)$.

SOLUTION. Here, we want to integrate the joint pdf $f_{Y_1, Y_2}(y_1, y_2)$ over the set

$$B = \{(y_1, y_2) : 0 < y_1 < y_2 < 1\}.$$

The desired probability is

$$\begin{aligned} P(Y_1 < Y_2) &= \int_{y_2=0}^1 \int_{y_1=0}^{y_2} \frac{6}{5}(y_1 + y_2^2) dy_1 dy_2 \\ &= \frac{6}{5} \int_{y_2=0}^1 \left[\left(\frac{y_1^2}{2} + y_1 y_2^2 \right) \Big|_{y_1=0}^{y_2} \right] dy_2 \\ &= \frac{6}{5} \int_{y_2=0}^1 \left(\frac{y_2^2}{2} + y_2^3 \right) dy_2 \\ &= \frac{6}{5} \left[\left(\frac{y_2^3}{6} + \frac{y_2^4}{4} \right) \Big|_{y_2=0}^1 \right] = \frac{6}{5} \left(\frac{1}{6} + \frac{1}{4} \right) = 0.5. \quad \square \end{aligned}$$

5.4 Marginal distributions

DISCRETE CASE: The joint pmf of (Y_1, Y_2) in Example 5.1 is depicted below (in the inner rectangular part of the table). The marginal distributions of Y_1 and Y_2 are catalogued in the margins of the table.

$p_{Y_1, Y_2}(y_1, y_2)$	$y_2 = 0$	$y_2 = 1$	$y_2 = 2$	$y_2 = 3$	$p_{Y_1}(y_1)$
$y_1 = 0$	0.12	0.06	0.05	0.02	0.25
$y_1 = 1$	0.13	0.15	0.12	0.03	0.43
$y_1 = 2$	0.05	0.15	0.10	0.02	0.32
$p_{Y_2}(y_2)$	0.30	0.36	0.27	0.07	1

TERMINOLOGY: Let (Y_1, Y_2) be a **discrete** random vector with pmf $p_{Y_1, Y_2}(y_1, y_2)$. The **marginal pmf** of Y_1 is

$$p_{Y_1}(y_1) = \sum_{y_2} p_{Y_1, Y_2}(y_1, y_2)$$

and the **marginal pmf** of Y_2 is

$$p_{Y_2}(y_2) = \sum_{y_1} p_{Y_1, Y_2}(y_1, y_2).$$

MAIN POINT: In the two-dimensional discrete case, marginal pmfs are obtained by “summing over” the other variable.

TERMINOLOGY: Let (Y_1, Y_2) be a **continuous** random vector with pdf $f_{Y_1, Y_2}(y_1, y_2)$. Then the **marginal pdf** of Y_1 is

$$f_{Y_1}(y_1) = \int_{-\infty}^{\infty} f_{Y_1, Y_2}(y_1, y_2) dy_2$$

and the **marginal pdf** of Y_2 is

$$f_{Y_2}(y_2) = \int_{-\infty}^{\infty} f_{Y_1, Y_2}(y_1, y_2) dy_1.$$

MAIN POINT: In the two-dimensional continuous case, marginal pdfs are obtained by “integrating over” the other variable.

Example 5.3. In a simple genetics model, the proportion, say Y_1 , of a population with trait 1 is always less than the proportion, say Y_2 , of a population with trait 2. Suppose that the random vector (Y_1, Y_2) has joint pdf

$$f_{Y_1, Y_2}(y_1, y_2) = \begin{cases} 6y_1, & 0 < y_1 < y_2 < 1 \\ 0, & \text{otherwise.} \end{cases}$$

(a) Find the marginal distributions $f_{Y_1}(y_1)$ and $f_{Y_2}(y_2)$.

SOLUTION. To find $f_{Y_1}(y_1)$, we integrate $f_{Y_1, Y_2}(y_1, y_2)$ over y_2 . For $0 < y_1 < 1$,

$$f_{Y_1}(y_1) = \int_{y_2=y_1}^1 6y_1 dy_2 = 6y_1(1 - y_1).$$

Thus, the marginal distribution of Y_1 is given by

$$f_{Y_1}(y_1) = \begin{cases} 6y_1(1 - y_1), & 0 < y_1 < 1 \\ 0, & \text{otherwise.} \end{cases}$$

That is, $Y_1 \sim \text{beta}(2, 2)$. To find $f_{Y_2}(y_2)$, we integrate $f_{Y_1, Y_2}(y_1, y_2)$ over y_1 . For values of $0 < y_2 < 1$,

$$f_{Y_2}(y_2) = \int_{y_1=0}^{y_2} 6y_1 dy_1 = 3y_1^2 \Big|_0^{y_2} = 3y_2^2.$$

Thus, the marginal distribution of Y_2 is given by

$$f_{Y_2}(y_2) = \begin{cases} 3y_2^2, & 0 < y_2 < 1 \\ 0, & \text{otherwise.} \end{cases}$$

That is, $Y_2 \sim \text{beta}(3, 1)$.

(b) Find the probability that the proportion of individuals with trait 2 exceeds 1/2.

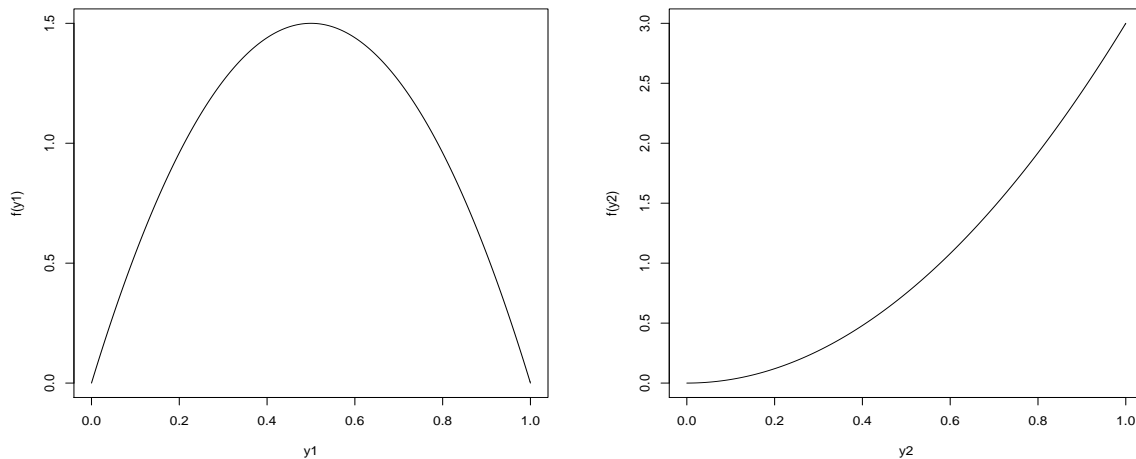
SOLUTION. Here, we want to find $P(B)$, where the set

$$B = \{(y_1, y_2) : 0 < y_1 < y_2, y_2 > 1/2\}.$$

This probability can be computed two different ways:

(i) using the **joint** distribution $f_{Y_1, Y_2}(y_1, y_2)$ and computing

$$P[(Y_1, Y_2) \in B] = \int_{y_2=1/2}^1 \int_{y_1=0}^{y_2} 6y_1 dy_1 dy_2.$$



$$Y_1 \sim \text{beta}(2, 2)$$

$$Y_2 \sim \text{beta}(3, 1)$$

Figure 5.16: Marginal distributions in Example 5.3.

(ii) using the **marginal** distribution $f_{Y_2}(y_2)$ and computing

$$P(Y_2 > 1/2) = \int_{y_2=1/2}^1 3y_2^2 dy_2.$$

Either way, you will get the same answer! Notice that in (i), you are computing the volume under $f_{Y_1, Y_2}(y_1, y_2)$ over the set B . In (ii), you are finding the area under $f_{Y_2}(y_2)$ over the set $\{y_2 : y_2 > 1/2\}$.

(c) Find the probability that the proportion of individuals with trait 2 is at least twice that of the proportion of individuals with trait 1.

SOLUTION. Here, we want to compute $P(Y_2 \geq 2Y_1)$; i.e., we want to compute $P(D)$, where the set

$$D = \{(y_1, y_2) : y_2 \geq 2y_1\}.$$

This equals

$$P[(Y_1, Y_2) \in D] = \int_{y_2=0}^1 \int_{y_1=0}^{y_2/2} 6y_1 dy_1 dy_2 = 0.25.$$

This is the volume under $f_{Y_1, Y_2}(y_1, y_2)$ over the set D . \square

5.5 Conditional distributions

RECALL: For events A and B in a non-empty sample space S , we defined

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

for $P(B) > 0$. Now, suppose that (Y_1, Y_2) is a discrete random vector. If we let $B = \{Y_2 = y_2\}$ and $A = \{Y_1 = y_1\}$, we obtain

$$P(A|B) = \frac{P(Y_1 = y_1, Y_2 = y_2)}{P(Y_2 = y_2)} = \frac{p_{Y_1, Y_2}(y_1, y_2)}{p_{Y_2}(y_2)}.$$

This leads to the definition of a discrete conditional distribution.

TERMINOLOGY: Suppose that (Y_1, Y_2) is a discrete random vector with joint pmf $p_{Y_1, Y_2}(y_1, y_2)$. We define the **conditional probability mass function (pmf)** of Y_1 , given $Y_2 = y_2$, as

$$p_{Y_1|Y_2}(y_1|y_2) = \frac{p_{Y_1, Y_2}(y_1, y_2)}{p_{Y_2}(y_2)},$$

whenever $p_{Y_2}(y_2) > 0$. Similarly, the conditional probability mass function of Y_2 , given $Y_1 = y_1$, is

$$p_{Y_2|Y_1}(y_2|y_1) = \frac{p_{Y_1, Y_2}(y_1, y_2)}{p_{Y_1}(y_1)},$$

whenever $p_{Y_1}(y_1) > 0$.

Example 5.4. The joint pmf of (Y_1, Y_2) in Example 5.1 is depicted below (in the inner rectangular part of the table). The marginal distributions of Y_1 and Y_2 are catalogued in the margins of the table.

$p_{Y_1, Y_2}(y_1, y_2)$	$y_2 = 0$	$y_2 = 1$	$y_2 = 2$	$y_2 = 3$	$p_{Y_1}(y_1)$
$y_1 = 0$	0.12	0.06	0.05	0.02	0.25
$y_1 = 1$	0.13	0.15	0.12	0.03	0.43
$y_1 = 2$	0.05	0.15	0.10	0.02	0.32
$p_{Y_2}(y_2)$	0.30	0.36	0.27	0.07	1

QUESTION: What is the conditional pmf of Y_1 , given $Y_2 = 1$?

SOLUTION. Straightforward calculations show that

$$\begin{aligned} p_{Y_1|Y_2}(y_1 = 0|y_2 = 1) &= \frac{p_{Y_1,Y_2}(y_1 = 0, y_2 = 1)}{p_{Y_2}(y_2 = 1)} = \frac{0.06}{0.36} = 2/12 \\ p_{Y_1|Y_2}(y_1 = 1|y_2 = 1) &= \frac{p_{Y_1,Y_2}(y_1 = 1, y_2 = 1)}{p_{Y_2}(y_2 = 1)} = \frac{0.15}{0.36} = 5/12 \\ p_{Y_1|Y_2}(y_1 = 2|y_2 = 1) &= \frac{p_{Y_1,Y_2}(y_1 = 2, y_2 = 1)}{p_{Y_2}(y_2 = 1)} = \frac{0.15}{0.36} = 5/12. \end{aligned}$$

Thus, the conditional pmf of Y_1 , given $Y_2 = 1$, is given by

y_1	0	1	2
$p_{Y_1 Y_2}(y_1 y_2 = 1)$	2/12	5/12	5/12

This conditional pmf tells us how Y_1 is distributed if we are given that $Y_2 = 1$.

EXERCISE. Find the conditional pmf of Y_2 , given $Y_1 = 0$. \square

TERMINOLOGY: Suppose that (Y_1, Y_2) is a continuous random vector with joint pdf $f_{Y_1,Y_2}(y_1, y_2)$. We define the **conditional probability density function (pdf)** of Y_1 , given $Y_2 = y_2$, as

$$f_{Y_1|Y_2}(y_1|y_2) = \frac{f_{Y_1,Y_2}(y_1, y_2)}{f_{Y_2}(y_2)}.$$

Similarly, the conditional probability density function of Y_2 , given $Y_1 = y_1$, is

$$f_{Y_2|Y_1}(y_2|y_1) = \frac{f_{Y_1,Y_2}(y_1, y_2)}{f_{Y_1}(y_1)}.$$

Example 5.5. Consider the bivariate pdf in Example 5.3,

$$f_{Y_1,Y_2}(y_1, y_2) = \begin{cases} 6y_1, & 0 < y_1 < y_2 < 1 \\ 0, & \text{otherwise.} \end{cases}$$

This model describes the distribution of the random vector (Y_1, Y_2) , where Y_1 , the proportion of a population with trait 1, is always less than Y_2 , the proportion of a population with trait 2. Derive the conditional distributions $f_{Y_1|Y_2}(y_1|y_2)$ and $f_{Y_2|Y_1}(y_2|y_1)$.

SOLUTION. In Example 5.3, we derived the marginal pdfs to be

$$f_{Y_1}(y_1) = \begin{cases} 6y_1(1 - y_1), & 0 < y_1 < 1 \\ 0, & \text{otherwise} \end{cases}$$

and

$$f_{Y_2}(y_2) = \begin{cases} 3y_2^2, & 0 < y_2 < 1 \\ 0, & \text{otherwise.} \end{cases}$$

First, we derive $f_{Y_1|Y_2}(y_1|y_2)$, so fix $Y_2 = y_2$. Remember, once we condition on $Y_2 = y_2$ (i.e., once we fix $Y_2 = y_2$), we then regard y_2 as simply a constant. **This is an important point to understand!** For values of $0 < y_1 < y_2$, it follows that

$$f_{Y_1|Y_2}(y_1|y_2) = \frac{f_{Y_1, Y_2}(y_1, y_2)}{f_{Y_2}(y_2)} = \frac{6y_1}{3y_2^2} = \frac{2y_1}{y_2^2},$$

and, thus, this is the value of $f_{Y_1|Y_2}(y_1|y_2)$ when $0 < y_1 < y_2$. For values of $y_1 \notin (0, y_2)$, the conditional density $f_{Y_1|Y_2}(y_1|y_2) = 0$. Summarizing,

$$f_{Y_1|Y_2}(y_1|y_2) = \begin{cases} 2y_1/y_2^2, & 0 < y_1 < y_2 \\ 0, & \text{otherwise.} \end{cases}$$

To reiterate, in this (conditional) pdf, the value of y_2 is fixed and known. It is Y_1 that is varying. This function describes how Y_1 is distributed for y_2 fixed.

Now, to derive the conditional pdf of Y_2 given Y_1 , we fix $Y_1 = y_1$; then, for all values of $y_1 < y_2 < 1$, we have

$$f_{Y_2|Y_1}(y_2|y_1) = \frac{f_{Y_1, Y_2}(y_1, y_2)}{f_{Y_1}(y_1)} = \frac{6y_1}{6y_1(1-y_1)} = \frac{1}{1-y_1}.$$

This is the value of $f_{Y_2|Y_1}(y_2|y_1)$ when $y_1 < y_2 < 1$. When $y_2 \notin (y_1, 1)$, the conditional pdf is $f_{Y_2|Y_1}(y_2|y_1) = 0$. Remember, once we condition on $Y_1 = y_1$, then we regard y_1 simply as a constant. Summarizing,

$$f_{Y_2|Y_1}(y_2|y_1) = \begin{cases} \frac{1}{1-y_1}, & y_1 < y_2 < 1 \\ 0, & \text{otherwise.} \end{cases}$$

That is, conditional on $Y_1 = y_1$, $Y_2 \sim \mathcal{U}(y_1, 1)$. **Again, in this (conditional) pdf, the value of y_1 is fixed and known. It is Y_2 that is varying. This function describes how Y_2 is distributed for y_1 fixed.** \square

RESULT: The use of conditional distributions allows us to define conditional probabilities of events associated with one random variable when we know the value of another random

variable. If Y_1 and Y_2 are jointly **discrete**, then for any set $B \subset \mathcal{R}$,

$$P(Y_1 \in B | Y_2 = y_2) = \sum_B p_{Y_1|Y_2}(y_1|y_2)$$

$$P(Y_2 \in B | Y_1 = y_1) = \sum_B p_{Y_2|Y_1}(y_2|y_1).$$

If Y_1 and Y_2 are jointly **continuous**, then for any set $B \subset \mathcal{R}$,

$$P(Y_1 \in B | Y_2 = y_2) = \int_B f_{Y_1|Y_2}(y_1|y_2) dy_1$$

$$P(Y_2 \in B | Y_1 = y_1) = \int_B f_{Y_2|Y_1}(y_2|y_1) dy_2.$$

Example 5.6. A health-food store stocks two different brands of grain. Let Y_1 denote the amount of brand 1 in stock and let Y_2 denote the amount of brand 2 in stock (both Y_1 and Y_2 are measured in 100s of lbs). The joint distribution of Y_1 and Y_2 is given by

$$f_{Y_1, Y_2}(y_1, y_2) = \begin{cases} 24y_1y_2, & y_1 > 0, y_2 > 0, 0 < y_1 + y_2 < 1 \\ 0, & \text{otherwise.} \end{cases}$$

(a) Find the conditional pdf $f_{Y_1|Y_2}(y_1|y_2)$.

(b) Compute $P(Y_1 > 0.5 | Y_2 = 0.3)$.

(c) Find $P(Y_1 > 0.5)$.

SOLUTIONS. (a) To find the conditional pdf $f_{Y_1|Y_2}(y_1|y_2)$, we first need to find the marginal pdf of Y_2 . The marginal pdf of Y_2 , for $0 < y_2 < 1$, is

$$f_{Y_2}(y_2) = \int_{y_1=0}^{1-y_2} 24y_1y_2 dy_1 = 24y_2 \left(\frac{y_1^2}{2} \Big|_0^{1-y_2} \right) = 12y_2(1-y_2)^2,$$

and 0, otherwise. We recognize this as a beta(2, 3) pdf; i.e., $Y_2 \sim \text{beta}(2, 3)$. The conditional pdf of Y_1 , given $Y_2 = y_2$, is

$$f_{Y_1|Y_2}(y_1|y_2) = \frac{f_{Y_1, Y_2}(y_1, y_2)}{f_{Y_2}(y_2)} = \frac{24y_1y_2}{12y_2(1-y_2)^2}$$

$$= \frac{2y_1}{(1-y_2)^2},$$

for $0 < y_1 < 1 - y_2$, and 0, otherwise. Summarizing,

$$f_{Y_1|Y_2}(y_1|y_2) = \begin{cases} \frac{2y_1}{(1-y_2)^2}, & 0 < y_1 < 1 - y_2 \\ 0, & \text{otherwise.} \end{cases}$$

(b) To compute $P(Y_1 > 0.5|Y_2 = 0.3)$, we work with the conditional pdf $f_{Y_1|Y_2}(y_1|y_2)$, which for $y_2 = 0.3$, is given by

$$f_{Y_1|Y_2}(y_1|y_2) = \begin{cases} \left(\frac{200}{49}\right) y_1, & 0 < y_1 < 0.7 \\ 0, & \text{otherwise.} \end{cases}$$

Thus,

$$P(Y_1 > 0.5|Y_2 = 0.3) = \int_{0.5}^{0.7} \left(\frac{200}{49}\right) y_1 dy_1 \approx 0.489.$$

(c) To compute $P(Y_1 > 0.5)$, we can either use the marginal pdf $f_{Y_1}(y_1)$ or the joint pdf $f_{Y_1, Y_2}(y_1, y_2)$. Marginally, it turns out that $Y_1 \sim \text{beta}(2, 3)$ as well (verify!). Thus,

$$P(Y_1 > 0.5) = \int_{0.5}^1 12y_1(1 - y_1)^2 dy_1 \approx 0.313.$$

REMARK: Notice how $P(Y_1 > 0.5|Y_2 = 0.3) \neq P(Y_1 > 0.5)$; that is, knowledge of the value of Y_2 has affected the way that we assign probability to events involving Y_1 . Of course, one might expect this because of the support in the joint pdf $f_{Y_1, Y_2}(y_1, y_2)$. \square

5.6 Independent random variables

TERMINOLOGY: Suppose (Y_1, Y_2) is a random vector (discrete or continuous) with joint cdf $F_{Y_1, Y_2}(y_1, y_2)$, and denote the marginal cdfs of Y_1 and Y_2 by $F_{Y_1}(y_1)$ and $F_{Y_2}(y_2)$, respectively. We say the random variables Y_1 and Y_2 are **independent** if and only if

$$F_{Y_1, Y_2}(y_1, y_2) = F_{Y_1}(y_1)F_{Y_2}(y_2)$$

for all values of y_1 and y_2 . Otherwise, we say that Y_1 and Y_2 are **dependent**.

RESULT: Suppose that (Y_1, Y_2) is a random vector (discrete or continuous) with joint pdf (pmf) $f_{Y_1, Y_2}(y_1, y_2)$, and denote the marginal pdfs (pmfs) of Y_1 and Y_2 by $f_{Y_1}(y_1)$ and $f_{Y_2}(y_2)$, respectively. Then, Y_1 and Y_2 are independent if and only if

$$f_{Y_1, Y_2}(y_1, y_2) = f_{Y_1}(y_1)f_{Y_2}(y_2)$$

for all values of y_1 and y_2 . Otherwise, Y_1 and Y_2 are dependent.

Proof. Exercise. \square

Example 5.7. Suppose that the pmf for the discrete random vector (Y_1, Y_2) is given by

$$p_{Y_1, Y_2}(y_1, y_2) = \begin{cases} \frac{1}{18}(y_1 + 2y_2), & y_1 = 1, 2, y_2 = 1, 2 \\ 0, & \text{otherwise.} \end{cases}$$

The marginal distribution of Y_1 , for values of $y_1 = 1, 2$, is given by

$$p_{Y_1}(y_1) = \sum_{y_2=1}^2 p_{Y_1, Y_2}(y_1, y_2) = \sum_{y_2=1}^2 \frac{1}{18}(y_1 + 2y_2) = \frac{1}{18}(2y_1 + 6),$$

and $p_{Y_1}(y_1) = 0$, otherwise. Similarly, the marginal distribution of Y_2 , for values of $y_2 = 1, 2$, is given by

$$p_{Y_2}(y_2) = \sum_{y_1=1}^2 p_{Y_1, Y_2}(y_1, y_2) = \sum_{y_1=1}^2 \frac{1}{18}(y_1 + 2y_2) = \frac{1}{18}(3 + 4y_2),$$

and $p_{Y_2}(y_2) = 0$, otherwise. Note that, for example,

$$\frac{3}{18} = p_{Y_1, Y_2}(1, 1) \neq p_{Y_1}(1)p_{Y_2}(1) = \frac{8}{18} \times \frac{7}{18} = \frac{14}{81};$$

thus, the random variables Y_1 and Y_2 are dependent. \square

Example 5.8. Let Y_1 and Y_2 denote the proportions of time (out of one workday) during which employees I and II, respectively, perform their assigned tasks. Suppose that the random vector (Y_1, Y_2) has joint pdf

$$f_{Y_1, Y_2}(y_1, y_2) = \begin{cases} y_1 + y_2, & 0 < y_1 < 1, 0 < y_2 < 1 \\ 0, & \text{otherwise.} \end{cases}$$

It is straightforward to show (verify!) that

$$f_{Y_1}(y_1) = \begin{cases} y_1 + \frac{1}{2}, & 0 < y_1 < 1 \\ 0, & \text{otherwise} \end{cases} \quad \text{and} \quad f_{Y_2}(y_2) = \begin{cases} y_2 + \frac{1}{2}, & 0 < y_2 < 1 \\ 0, & \text{otherwise.} \end{cases}$$

Thus, since

$$f_{Y_1, Y_2}(y_1, y_2) = y_1 + y_2 \neq \left(y_1 + \frac{1}{2}\right) \left(y_2 + \frac{1}{2}\right) = f_{Y_1}(y_1)f_{Y_2}(y_2),$$

for $0 < y_1 < 1$ and $0 < y_2 < 1$, Y_1 and Y_2 are dependent. \square

A *CONVENIENT RESULT*: Let (Y_1, Y_2) be a random vector (discrete or continuous) with pdf (pmf) $f_{Y_1, Y_2}(y_1, y_2)$. If the support set R does not constrain y_1 by y_2 (or y_2 by y_1), and additionally, we can factor the joint pdf (pmf) $f_{Y_1, Y_2}(y_1, y_2)$ into two nonnegative expressions

$$f_{Y_1, Y_2}(y_1, y_2) = g(y_1)h(y_2),$$

then Y_1 and Y_2 are independent. Note that $g(y_1)$ and $h(y_2)$ are simply functions; **they need not be pdfs (pmfs)**, although they sometimes are. The only requirement is that $g(y_1)$ is a function of y_1 only, $h(y_2)$ is a function of y_2 only, and that both are nonnegative. *If the support involves a constraint, the random variables are automatically dependent.*

Example 5.9. In Example 5.6, Y_1 denoted the amount of brand 1 grain in stock and Y_2 denoted the amount of brand 2 grain in stock. Recall that the joint pdf of (Y_1, Y_2) was given by

$$f_{Y_1, Y_2}(y_1, y_2) = \begin{cases} 24y_1y_2, & y_1 > 0, y_2 > 0, 0 < y_1 + y_2 < 1 \\ 0, & \text{otherwise.} \end{cases}$$

Here, the support is $R = \{(y_1, y_2) : y_1 > 0, y_2 > 0, 0 < y_1 + y_2 < 1\}$. Since knowledge of y_1 affects the value of y_2 , and vice versa, the support involves a constraint, and Y_1 and Y_2 are dependent. \square

Example 5.10. Suppose that the random vector (X, Y) has joint pdf

$$f_{X, Y}(x, y) = \begin{cases} [\Gamma(\alpha)\Gamma(\beta)]^{-1}\lambda e^{-\lambda x}(\lambda x)^{\alpha+\beta-1}y^{\alpha-1}(1-y)^{\beta-1}, & x > 0, 0 < y < 1 \\ 0, & \text{otherwise.} \end{cases}$$

for $\lambda > 0$, $\alpha > 0$, and $\beta > 0$. Since $R = \{(x, y) : x > 0, 0 < y < 1\}$ does not involve a constraint, it follows immediately that X and Y are independent, since we can write

$$f_{X, Y}(x, y) = \underbrace{\lambda e^{-\lambda x}(\lambda x)^{\alpha+\beta-1}}_{g(x)} \times \underbrace{\frac{y^{\alpha-1}(1-y)^{\beta-1}}{\Gamma(\alpha)\Gamma(\beta)}}_{h(y)},$$

where $g(x)$ and $h(y)$ are nonnegative functions. Note that we are not saying that $g(x)$ and $h(y)$ are marginal distributions of X and Y , respectively (in fact, they are not the marginal distributions, although they are proportional to the marginals). \square

EXTENSION: We generalize the notion of **independence** to n -variate random vectors. We use the conventional notation

$$\mathbf{Y} = (Y_1, Y_2, \dots, Y_n)$$

and

$$\mathbf{y} = (y_1, y_2, \dots, y_n).$$

We denote the joint cdf of \mathbf{Y} by $F_{\mathbf{Y}}(\mathbf{y})$ and the joint pdf (pmf) of \mathbf{Y} by $f_{\mathbf{Y}}(\mathbf{y})$.

TERMINOLOGY: Suppose that the random vector $\mathbf{Y} = (Y_1, Y_2, \dots, Y_n)$ has joint cdf $F_{\mathbf{Y}}(\mathbf{y})$, and suppose that the random variable Y_i has cdf $F_{Y_i}(y_i)$, for $i = 1, 2, \dots, n$. Then, Y_1, Y_2, \dots, Y_n are independent random variables if and only if

$$F_{\mathbf{Y}}(\mathbf{y}) = \prod_{i=1}^n F_{Y_i}(y_i);$$

that is, the joint cdf can be factored into the product of the marginal cdfs. Alternatively, Y_1, Y_2, \dots, Y_n are independent random variables if and only if

$$f_{\mathbf{Y}}(\mathbf{y}) = \prod_{i=1}^n f_{Y_i}(y_i);$$

that is, the joint pdf (pmf) can be factored into the product of the marginals.

Example 5.11. In a small clinical trial, $n = 20$ patients are treated with a new drug. Suppose that the response from each patient is a measurement $Y \sim \mathcal{N}(\mu, \sigma^2)$. Denoting the 20 responses by $\mathbf{Y} = (Y_1, Y_2, \dots, Y_{20})$, then, assuming independence, the joint distribution of the 20 responses is, for $\mathbf{y} \in \mathcal{R}^{20}$,

$$f_{\mathbf{Y}}(\mathbf{y}) = \prod_{i=1}^{20} \underbrace{\frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{1}{2}\left(\frac{y_i-\mu}{\sigma}\right)^2}}_{f_{Y_i}(y_i)} = \left(\frac{1}{\sqrt{2\pi\sigma}}\right)^{20} e^{-\frac{1}{2}\sum_{i=1}^{20}\left(\frac{y_i-\mu}{\sigma}\right)^2}.$$

What is the probability that at least one patient's response is greater than $\mu + 2\sigma$?

SOLUTION. Define the event

$$B = \{\text{at least one patient's response exceeds } \mu + 2\sigma\}.$$

We want to compute $P(B)$. Note that

$$\bar{B} = \{\text{all 20 responses are less than } \mu + 2\sigma\}$$

and recall that $P(B) = 1 - P(\bar{B})$. We will compute $P(\bar{B})$ because it is easier. The probability that the first patient's response Y_1 is less than $\mu + 2\sigma$ is given by

$$F_{Y_1}(\mu + 2\sigma) = P(Y_1 < \mu + 2\sigma) = P(Z < 2) = F_Z(2) = 0.9772,$$

where $Z \sim \mathcal{N}(0, 1)$ and $F_Z(\cdot)$ denotes the standard normal cdf. This probability is same for each patient, because each patient's response follows the same $\mathcal{N}(\mu, \sigma^2)$ distribution. Because the patients' responses are independent random variables,

$$\begin{aligned} P(\bar{B}) &= P(Y_1 < \mu + 2\sigma, Y_2 < \mu + 2\sigma, \dots, Y_{20} < \mu + 2\sigma) \\ &= \prod_{i=1}^{20} F_{Y_i}(\mu + 2\sigma) \\ &= [F_Z(2)]^{20} \approx 0.63. \end{aligned}$$

Finally, $P(B) = 1 - P(\bar{B}) \approx 1 - 0.63 = 0.37$. \square

5.7 Expectations of functions of random variables

RESULT: Suppose that $\mathbf{Y} = (Y_1, Y_2, \dots, Y_n)$ has joint pdf $f_{\mathbf{Y}}(\mathbf{y})$, or joint pmf $p_{\mathbf{Y}}(\mathbf{y})$, and suppose that $g(\mathbf{Y}) = g(Y_1, Y_2, \dots, Y_n)$ is a real vector valued function of Y_1, Y_2, \dots, Y_n ; i.e., $g : \mathcal{R}^n \rightarrow \mathcal{R}$. Then,

- if \mathbf{Y} is discrete,

$$E[g(\mathbf{Y})] = \sum_{y_1} \sum_{y_2} \cdots \sum_{y_n} g(\mathbf{y}) p_{\mathbf{Y}}(\mathbf{y}),$$

- and if \mathbf{Y} is continuous,

$$E[g(\mathbf{Y})] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} g(\mathbf{y}) f_{\mathbf{Y}}(\mathbf{y}) d\mathbf{y}.$$

If these quantities are not finite, then we say that $E[g(\mathbf{Y})]$ does not exist.

PROPERTIES OF EXPECTATIONS: Let $\mathbf{Y} = (Y_1, Y_2, \dots, Y_n)$ be a discrete or continuous random vector, suppose that g, g_1, g_2, \dots, g_k are real vector valued functions from $\mathcal{R}^n \rightarrow \mathcal{R}$, and let c be any real constant. Then,

- (a) $E(c) = c$
- (b) $E[cg(\mathbf{Y})] = cE[g(\mathbf{Y})]$
- (c) $E[\sum_{j=1}^k g_j(\mathbf{Y})] = \sum_{j=1}^k E[g_j(\mathbf{Y})]$.

Example 5.12. In Example 5.6, Y_1 denotes the amount of grain 1 in stock and Y_2 denotes the amount of grain 2 in stock. Both Y_1 and Y_2 are measured in 100s of lbs. The joint distribution of Y_1 and Y_2 is

$$f_{Y_1, Y_2}(y_1, y_2) = \begin{cases} 24y_1y_2, & y_1 > 0, y_2 > 0, 0 < y_1 + y_2 < 1 \\ 0, & \text{otherwise.} \end{cases}$$

What is the expected total amount of grain ($Y_1 + Y_2$) in stock?

SOLUTION. Let the function $g : \mathcal{R}^2 \rightarrow \mathcal{R}$ be defined by $g(y_1, y_2) = y_1 + y_2$. We would like to compute $E[g(Y_1, Y_2)] = E(Y_1 + Y_2)$. From the last result, we know that

$$\begin{aligned} E(Y_1 + Y_2) &= \int_{y_1=0}^1 \int_{y_2=0}^{1-y_1} (y_1 + y_2) \times 24y_1y_2 \, dy_2 dy_1 \\ &= \int_{y_1=0}^1 \int_{y_2=0}^{1-y_1} (24y_1^2y_2 + 24y_1y_2^2) \, dy_2 dy_1 \\ &= \int_{y_1=0}^1 \left[\left(24y_1^2 \frac{y_2^2}{2} \Big|_0^{1-y_1} \right) + \left(24y_1 \frac{y_2^3}{3} \Big|_0^{1-y_1} \right) \right] dy_1 \\ &= \int_{y_1=0}^1 12y_1^2(1-y_1)^2 dy_1 + \int_{y_1=0}^1 8y_1(1-y_1)^3 dy_1 \\ &= 12 \int_{y_1=0}^1 y_1^2(1-y_1)^2 dy_1 + 8 \int_{y_1=0}^1 y_1(1-y_1)^3 dy_1 \\ &= 12 \left[\frac{\Gamma(3)\Gamma(3)}{\Gamma(6)} \right] + 8 \left[\frac{\Gamma(2)\Gamma(4)}{\Gamma(6)} \right] = 4/5. \end{aligned}$$

The expected total amount of grain in stock is 80 lbs.

REMARK: In the calculation above, we twice used the fact that

$$\int_0^1 y^{\alpha-1}(1-y)^{\beta-1} dy = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}.$$

ANOTHER SOLUTION: To compute $E(Y_1 + Y_2)$, we could have taken a different route. In Example 5.6, we discovered that the marginal distributions were

$$Y_1 \sim \text{beta}(2, 3)$$

$$Y_2 \sim \text{beta}(2, 3)$$

so that

$$E(Y_1) = E(Y_2) = \frac{2}{2+3} = \frac{2}{5}.$$

Because expectations are linear, we have

$$E(Y_1 + Y_2) = \frac{2}{5} + \frac{2}{5} = \frac{4}{5}. \quad \square$$

RESULT: Suppose that Y_1 and Y_2 are **independent** random variables. Let $g(Y_1)$ be a function of Y_1 only, and let $h(Y_2)$ be a function of Y_2 only. Then,

$$E[g(Y_1)h(Y_2)] = E[g(Y_1)]E[h(Y_2)],$$

provided that all expectations exist.

Proof. Without loss, assume that (Y_1, Y_2) is a continuous random vector (the discrete case is analogous). Suppose that (Y_1, Y_2) has joint pdf $f_{Y_1, Y_2}(y_1, y_2)$ with support $R \subset \mathcal{R}^2$.

Note that

$$\begin{aligned} E[g(Y_1)h(Y_2)] &= \int_{\mathcal{R}^2} g(y_1)h(y_2)f_{Y_1, Y_2}(y_1, y_2)dy_2dy_1 \\ &= \int_{\mathcal{R}} \int_{\mathcal{R}} g(y_1)h(y_2)f_{Y_1}(y_1)f_{Y_2}(y_2)dy_2dy_1 \\ &= \int_{\mathcal{R}} g(y_1)f_{Y_1}(y_1)dy_1 \int_{\mathcal{R}} h(y_2)f_{Y_2}(y_2)dy_2 \\ &= E[g(Y_1)]E[h(Y_2)]. \quad \square \end{aligned}$$

COROLLARY: If Y_1 and Y_2 are **independent** random variables, then

$$E(Y_1Y_2) = E(Y_1)E(Y_2).$$

This is a special case of the previous result obtained by taking $g(Y_1) = Y_1$ and $h(Y_2) = Y_2$.

5.8 Covariance and correlation

5.8.1 Covariance

TERMINOLOGY: Suppose that Y_1 and Y_2 are random variables (discrete or continuous) with means $E(Y_1) = \mu_1$ and $E(Y_2) = \mu_2$, respectively. The **covariance** between Y_1 and Y_2 is given by

$$\begin{aligned}\text{Cov}(Y_1, Y_2) &\equiv E[(Y_1 - \mu_1)(Y_2 - \mu_2)] \\ &= E(Y_1 Y_2) - E(Y_1)E(Y_2).\end{aligned}$$

The latter expression is often easier to work with and is called the **covariance computing formula**. The covariance is a numerical measure that describes how two variables are linearly related.

- If $\text{Cov}(Y_1, Y_2) > 0$, then Y_1 and Y_2 are positively linearly related.
- If $\text{Cov}(Y_1, Y_2) < 0$, then Y_1 and Y_2 are negatively linearly related.
- If $\text{Cov}(Y_1, Y_2) = 0$, then Y_1 and Y_2 are not linearly related.

RESULT: If Y_1 and Y_2 are independent, then $\text{Cov}(Y_1, Y_2) = 0$.

Proof. Suppose that Y_1 and Y_2 are independent. Using the covariance computing formula,

$$\begin{aligned}\text{Cov}(Y_1, Y_2) &= E(Y_1 Y_2) - E(Y_1)E(Y_2) \\ &= E(Y_1)E(Y_2) - E(Y_1)E(Y_2) = 0. \quad \square\end{aligned}$$

IMPORTANT: If two random variables are independent, then they have zero covariance. However, zero covariance does not necessarily imply independence, as we see now.

Example 5.13. *An example of two dependent variables with zero covariance.* Suppose that $Y_1 \sim \mathcal{U}(-1, 1)$, and let $Y_2 = Y_1^2$. It is straightforward to show that

$$\begin{aligned}E(Y_1) &= 0 \\ E(Y_1 Y_2) &= E(Y_1^3) = 0 \\ E(Y_2) &= E(Y_1^2) = V(Y_1) = 1/3.\end{aligned}$$

Thus,

$$\text{Cov}(Y_1, Y_2) = E(Y_1 Y_2) - E(Y_1)E(Y_2) = 0 - 0(1/3) = 0.$$

However, clearly Y_1 and Y_2 are not independent; in fact, they are perfectly related! It is just that the relationship is not linear (it is quadratic). The covariance only measures linear relationships. \square

Example 5.14. Gasoline is stocked in a tank once at the beginning of each week and then sold to customers. Let Y_1 denote the proportion of the capacity of the tank that is available after it is stocked. Let Y_2 denote the proportion of the capacity of the bulk tank that is sold during the week. Suppose that the random vector (Y_1, Y_2) has joint pdf

$$f_{Y_1, Y_2}(y_1, y_2) = \begin{cases} 3y_1, & 0 < y_2 < y_1 < 1 \\ 0, & \text{otherwise.} \end{cases}$$

Compute $\text{Cov}(Y_1, Y_2)$.

SOLUTION. It is perhaps easiest to use the covariance computing formula

$$\text{Cov}(Y_1, Y_2) = E(Y_1 Y_2) - E(Y_1)E(Y_2).$$

The marginal distribution of Y_1 is beta(3, 1). The marginal distribution of Y_2 is

$$f_{Y_2}(y_2) = \begin{cases} \frac{3}{2}(1 - y_2^2), & 0 < y_2 < 1 \\ 0, & \text{otherwise.} \end{cases}$$

Thus, the marginal first moments are

$$\begin{aligned} E(Y_1) &= \frac{3}{3+1} = 0.75 \\ E(Y_2) &= \int_0^1 y_2 \times \frac{3}{2}(1 - y_2^2) dy_2 = 0.375. \end{aligned}$$

Now, we need to compute $E(Y_1 Y_2)$. This is given by

$$E(Y_1 Y_2) = \int_{y_1=0}^1 \int_{y_2=0}^{y_1} y_1 y_2 \times 3y_1 dy_2 dy_1 = 0.30.$$

Thus, the covariance is

$$\text{Cov}(Y_1, Y_2) = E(Y_1 Y_2) - E(Y_1)E(Y_2) = 0.30 - (0.75)(0.375) = 0.01875. \quad \square$$

IMPORTANT: Suppose that Y_1 and Y_2 are random variables (discrete or continuous).

$$V(Y_1 + Y_2) = V(Y_1) + V(Y_2) + 2\text{Cov}(Y_1, Y_2)$$

$$V(Y_1 - Y_2) = V(Y_1) + V(Y_2) - 2\text{Cov}(Y_1, Y_2).$$

Proof. Suppose that Y_1 and Y_2 are random variables with means $E(Y_1) = \mu_1$ and $E(Y_2) = \mu_2$, respectively. Let $Z = Y_1 + Y_2$. From the definition of variance, we have

$$\begin{aligned} V(Z) &= E[(Z - \mu_Z)^2] \\ &= E\{[(Y_1 + Y_2) - E(Y_1 + Y_2)]^2\} \\ &= E[(Y_1 + Y_2 - \mu_1 - \mu_2)^2] \\ &= E\{[(Y_1 - \mu_1) + (Y_2 - \mu_2)]^2\} \\ &= E[(Y_1 - \mu_1)^2 + (Y_2 - \mu_2)^2 + \underbrace{2(Y_1 - \mu_1)(Y_2 - \mu_2)}_{\text{cross product}}] \\ &= E[(Y_1 - \mu_1)^2] + E[(Y_2 - \mu_2)^2] + 2E[(Y_1 - \mu_1)(Y_2 - \mu_2)] \\ &= V(Y_1) + V(Y_2) + 2\text{Cov}(Y_1, Y_2). \end{aligned}$$

That $V(Y_1 - Y_2) = V(Y_1) + V(Y_2) - 2\text{Cov}(Y_1, Y_2)$ is shown similarly. \square

RESULT: Suppose that Y_1 and Y_2 are **independent** random variables (discrete or continuous).

$$V(Y_1 + Y_2) = V(Y_1) + V(Y_2)$$

$$V(Y_1 - Y_2) = V(Y_1) + V(Y_2).$$

Proof. In the light of the last result, this is obvious. \square

Example 5.15. A small health-food store stocks two different brands of grain. Let Y_1 denote the amount of brand 1 in stock and let Y_2 denote the amount of brand 2 in stock (both Y_1 and Y_2 are measured in 100s of lbs). The joint distribution of Y_1 and Y_2 is

$$f_{Y_1, Y_2}(y_1, y_2) = \begin{cases} 24y_1y_2, & y_1 > 0, y_2 > 0, 0 < y_1 + y_2 < 1 \\ 0, & \text{otherwise.} \end{cases}$$

What is the variance for the total amount of grain in stock? That is, find $V(Y_1 + Y_2)$.

SOLUTION: We know that

$$V(Y_1 + Y_2) = V(Y_1) + V(Y_2) + 2\text{Cov}(Y_1, Y_2).$$

Marginally, Y_1 and Y_2 are both $\text{beta}(2, 3)$; see Example 5.6. Thus,

$$E(Y_1) = E(Y_2) = \frac{2}{2+3} = \frac{2}{5}$$

and

$$V(Y_1) = V(Y_2) = \frac{2(3)}{(2+3+1)(2+3)^2} = \frac{1}{25}.$$

We need to compute $\text{Cov}(Y_1, Y_2)$. Note that

$$E(Y_1 Y_2) = \int_{y_1=0}^1 \int_{y_2=0}^{1-y_1} y_1 y_2 \times 24 y_1 y_2 dy_2 dy_1 = \frac{2}{15}.$$

Thus,

$$\begin{aligned} \text{Cov}(Y_1, Y_2) &= E(Y_1 Y_2) - E(Y_1)E(Y_2) \\ &= \frac{2}{15} - \left(\frac{2}{5}\right)\left(\frac{2}{5}\right) \approx -0.027. \end{aligned}$$

Finally,

$$\begin{aligned} V(Y_1 + Y_2) &= V(Y_1) + V(Y_2) + 2\text{Cov}(Y_1, Y_2) \\ &= \frac{1}{25} + \frac{1}{25} + 2(-0.027) \approx 0.027. \quad \square \end{aligned}$$

RESULTS: Suppose that Y_1 and Y_2 are random variables (discrete or continuous). The covariance function satisfies the following:

- (a) $\text{Cov}(Y_1, Y_2) = \text{Cov}(Y_2, Y_1)$
- (b) $\text{Cov}(Y_1, Y_1) = V(Y_1)$.
- (c) $\text{Cov}(a + bY_1, c + dY_2) = bd\text{Cov}(Y_1, Y_2)$, for any constants a, b, c , and d .

Proof. Exercise. \square

5.8.2 Correlation

GENERAL PROBLEM: Suppose that X and Y are random variables and that we want to predict Y as a linear function of X . That is, we want to consider functions of the form $Y = \beta_0 + \beta_1 X$, for fixed constants β_0 and β_1 . In this situation, the “error in prediction” is given by

$$Y - (\beta_0 + \beta_1 X).$$

This error can be positive or negative, so in developing a measure of prediction error, we want one that maintains the magnitude of error but ignores the sign. Thus, we define the **mean squared error of prediction**, given by

$$Q(\beta_0, \beta_1) \equiv E\{[Y - (\beta_0 + \beta_1 X)]^2\}.$$

A two-variable calculus argument shows that the mean squared error of prediction $Q(\beta_0, \beta_1)$ is minimized when

$$\beta_1 = \frac{\text{Cov}(X, Y)}{V(X)}$$

and

$$\beta_0 = E(Y) - \left[\frac{\text{Cov}(X, Y)}{V(X)} \right] E(X) = E(Y) - \beta_1 E(X).$$

Note that the value of β_1 , algebraically, is equal to

$$\begin{aligned} \beta_1 &= \frac{\text{Cov}(X, Y)}{V(X)} \\ &= \left[\frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y} \right] \frac{\sigma_Y}{\sigma_X} \\ &= \rho \left(\frac{\sigma_Y}{\sigma_X} \right), \end{aligned}$$

where

$$\rho = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}.$$

The quantity ρ is called the **correlation coefficient** between X and Y .

SUMMARY: The **best linear predictor** of Y , given X , is $Y = \beta_0 + \beta_1 X$, where

$$\begin{aligned} \beta_1 &= \rho \left(\frac{\sigma_Y}{\sigma_X} \right) \\ \beta_0 &= E(Y) - \beta_1 E(X). \end{aligned}$$

NOTES ON THE CORRELATION COEFFICIENT:

- (1) $-1 \leq \rho \leq 1$ (this can be proven using the Cauchy-Schwartz Inequality, from calculus).
- (2) If $\rho = 1$, then $Y = \beta_0 + \beta_1 X$, where $\beta_1 > 0$. That is, X and Y are perfectly positively linearly related; i.e., the bivariate probability distribution of (X, Y) lies entirely on a straight line with positive slope.
- (3) If $\rho = -1$, then $Y = \beta_0 + \beta_1 X$, where $\beta_1 < 0$. That is, X and Y are perfectly negatively linearly related; i.e., the bivariate probability distribution of (X, Y) lies entirely on a straight line with negative slope.
- (4) If $\rho = 0$, then X and Y are not linearly related.

NOTE: If X and Y are independent random variables, then $\rho = 0$. However, again, the implication does not go the other way; that is, if $\rho = 0$, this does not necessarily mean that X and Y are independent.

NOTE: In assessing the strength of the linear relationship between X and Y , the correlation coefficient is often preferred over the covariance since ρ is measured on a bounded, unitless scale. On the other hand, $\text{Cov}(X, Y)$ can be any real number and its units may not even make practical sense.

Example 5.16. In Example 5.14, we considered the bivariate model

$$f_{Y_1, Y_2}(y_1, y_2) = \begin{cases} 3y_1, & 0 < y_2 < y_1 < 1 \\ 0, & \text{otherwise.} \end{cases}$$

for Y_1 , the proportion of the capacity of the tank after being stocked, and Y_2 , the proportion of the capacity of the tank that is sold. Compute the correlation ρ between Y_1 and Y_2 .

SOLUTION: In Example 5.14, we computed $\text{Cov}(Y_1, Y_2) = 0.01875$, so all we need is σ_{Y_1} and σ_{Y_2} , the marginal standard deviations. In Example 5.14, we also found that

$Y_1 \sim \text{beta}(3, 1)$ and

$$f_{Y_2}(y_2) = \begin{cases} \frac{3}{2}(1 - y_2^2), & 0 < y_2 < 1 \\ 0, & \text{otherwise.} \end{cases}$$

The variance of Y_1 is

$$V(Y_1) = \frac{3(1)}{(3 + 1 + 1)(3 + 1)^2} = \frac{3}{80} \implies \sigma_{Y_1} = \sqrt{\frac{3}{80}} \approx 0.194.$$

Simple calculations using $f_{Y_2}(y_2)$ show that $E(Y_2^2) = 1/5$ and $E(Y_2) = 3/8$ so that

$$V(Y_2) = \frac{1}{5} - \left(\frac{3}{8}\right)^2 = 0.059 \implies \sigma_{Y_2} = \sqrt{0.059} \approx 0.244.$$

Finally, the correlation is

$$\rho = \frac{\text{Cov}(Y_1, Y_2)}{\sigma_{Y_1}\sigma_{Y_2}} \approx \frac{0.01875}{(0.194)(0.244)} \approx 0.40. \quad \square$$

5.9 Expectations and variances of linear functions of random variables

TERMINOLOGY: Suppose that Y_1, Y_2, \dots, Y_n are random variables and that a_1, a_2, \dots, a_n are constants. The function

$$U = \sum_{i=1}^n a_i Y_i = a_1 Y_1 + a_2 Y_2 + \dots + a_n Y_n$$

is called a **linear combination** of the random variables Y_1, Y_2, \dots, Y_n .

EXPECTED VALUE OF A LINEAR COMBINATION:

$$E(U) = E\left(\sum_{i=1}^n a_i Y_i\right) = \sum_{i=1}^n a_i E(Y_i)$$

VARIANCE OF A LINEAR COMBINATION:

$$\begin{aligned} V(U) &= V\left(\sum_{i=1}^n a_i Y_i\right) = \sum_{i=1}^n a_i^2 V(Y_i) + 2 \sum_{i < j} a_i a_j \text{Cov}(Y_i, Y_j) \\ &= \sum_{i=1}^n a_i^2 V(Y_i) + \sum_{i \neq j} a_i a_j \text{Cov}(Y_i, Y_j) \end{aligned}$$

Example 5.17. Achievement tests are commonly seen in educational or employment settings. For a large population of test-takers, let Y_1 , Y_2 , and Y_3 represent scores for different parts of an exam. Suppose that $Y_1 \sim \mathcal{N}(12, 4)$, $Y_2 \sim \mathcal{N}(16, 9)$, and $Y_3 \sim \mathcal{N}(20, 16)$. Suppose additionally that Y_1 and Y_2 are independent, $\text{Cov}(Y_1, Y_3) = 0.8$, and $\text{Cov}(Y_2, Y_3) = -6.7$. Two different summary measures are computed to assess a subject's performance:

$$U_1 = 0.5Y_1 - 2Y_2 + Y_3 \quad \text{and} \quad U_2 = 3Y_1 - 2Y_2 - Y_3.$$

Find $E(U_1)$ and $V(U_1)$.

SOLUTION: The expected value of U_1 is

$$\begin{aligned} E(U_1) &= E(0.5Y_1 - 2Y_2 + Y_3) = 0.5E(Y_1) - 2E(Y_2) + E(Y_3) \\ &= 0.5(12) - 2(16) + 20 = -6. \end{aligned}$$

The variance of U_1 is

$$\begin{aligned} V(U_1) &= V(0.5Y_1 - 2Y_2 + Y_3) \\ &= (0.5)^2V(Y_1) + (-2)^2V(Y_2) + (1)^2V(Y_3) \\ &\quad + 2(0.5)(-2)\text{Cov}(Y_1, Y_2) + 2(0.5)(1)\text{Cov}(Y_1, Y_3) + 2(-2)(1)\text{Cov}(Y_2, Y_3) \\ &= (0.25)(4) + 4(9) + 16 + 2(0.5)(-2)(0) + 2(0.5)(0.8) + 2(-2)(-6.7) = 80.6. \end{aligned}$$

EXERCISE: Find $E(U_2)$ and $V(U_2)$. \square

COVARIANCE BETWEEN TWO LINEAR COMBINATIONS: Suppose that

$$\begin{aligned} U_1 &= \sum_{i=1}^n a_i Y_i = a_1 Y_1 + a_2 Y_2 + \cdots + a_n Y_n \\ U_2 &= \sum_{j=1}^m b_j X_j = b_1 X_1 + b_2 X_2 + \cdots + b_m X_m. \end{aligned}$$

Then,

$$\text{Cov}(U_1, U_2) = \sum_{i=1}^n \sum_{j=1}^m a_i b_j \text{Cov}(Y_i, X_j).$$

EXERCISE: In Example 5.17, compute $\text{Cov}(U_1, U_2)$.

5.10 The multinomial model

RECALL: When we discussed the binomial model in Chapter 3, each (Bernoulli) trial resulted in either a “success” or a “failure;” that is, on each trial, there were only two outcomes possible (e.g., infected/not, germinated/not, defective/not, etc.).

TERMINOLOGY: A **multinomial experiment** is simply a generalization of a binomial experiment. In particular, consider an experiment where

- the experiment consists of n trials (n is fixed),
- the outcome for any trial belongs to exactly one of $k \geq 2$ categories,
- the probability that an outcome for a single trial falls into category i is p_i , for $i = 1, 2, \dots, k$, where each p_i remains constant from trial to trial, and
- the trials are independent.

DEFINITION: In a multinomial experiment, define

$$\begin{aligned} Y_1 &= \text{number of outcomes in category 1} \\ Y_2 &= \text{number of outcomes in category 2} \\ &\vdots \\ Y_k &= \text{number of outcomes in category } k \end{aligned}$$

so that $Y_1 + Y_2 + \dots + Y_k = n$, and denote $\mathbf{Y} = (Y_1, Y_2, \dots, Y_k)$. We call \mathbf{Y} a **multinomial** random vector and write $\mathbf{Y} \sim \text{mult}(n, p_1, p_2, \dots, p_k)$.

NOTE: When there are $k = 2$ categories (e.g., success/failure), the multinomial model reduces to a binomial model! When $k = 3$, \mathbf{Y} is said to have a **trinomial** distribution.

JOINT PMF: In general, If $\mathbf{Y} \sim \text{mult}(n, p_1, p_2, \dots, p_k)$, the pmf for \mathbf{Y} is given by

$$p_{\mathbf{Y}}(\mathbf{y}) = \begin{cases} \frac{n!}{y_1! y_2! \dots y_k!} p_1^{y_1} p_2^{y_2} \dots p_k^{y_k}, & y_i = 0, 1, \dots, n; \sum_i y_i = n \\ 0, & \text{otherwise.} \end{cases}$$

Example 5.18. At a number of clinic sites throughout Nebraska, chlamydia and gonorrhea testing is performed on individuals using urine or swab specimens. Define the following categories:

Category 1 : subjects with neither chlamydia nor gonorrhea

Category 2 : subjects with chlamydia but not gonorrhea

Category 3 : subjects with gonorrhea but not chlamydia

Category 4 : subjects with both chlamydia and gonorrhea.

For these $k = 4$ categories, empirical evidence suggests that $p_1 = 0.90$, $p_2 = 0.06$, $p_3 = 0.01$, and $p_4 = 0.03$. At one site, suppose that $n = 20$ individuals are tested on a given day. What is the probability exactly 16 are disease free, 2 are chlamydia positive but gonorrhea negative, and the remaining 2 are positive for both infections?

SOLUTION. Define $\mathbf{Y} = (Y_1, Y_2, Y_3, Y_4)$, where Y_i counts the number of subjects in category i . Assuming that subjects are independent,

$$\mathbf{Y} \sim \text{mult}(n = 20, p_1 = 0.90, p_2 = 0.06, p_3 = 0.01, p_4 = 0.03).$$

We want to compute

$$\begin{aligned} P(Y_1 = 16, Y_2 = 2, Y_3 = 0, Y_4 = 2) &= \frac{20!}{16! 2! 0! 2!} (0.90)^{16} (0.06)^2 (0.01)^0 (0.03)^2 \\ &\approx 0.017. \end{aligned}$$

FACTS: If $\mathbf{Y} = (Y_1, Y_2, \dots, Y_k) \sim \text{mult}(n, p_1, p_2, \dots, p_k)$, then

- the marginal distribution of Y_i is $b(n, p_i)$, for $i = 1, 2, \dots, k$.
- $E(Y_i) = np_i$, for $i = 1, 2, \dots, k$.
- $V(Y_i) = np_i(1 - p_i)$, for $i = 1, 2, \dots, k$.
- the joint distribution of (Y_i, Y_j) is trinomial($n, p_i, p_j, 1 - p_i - p_j$).
- $\text{Cov}(Y_i, Y_j) = -np_i p_j$, for $i \neq j$.

5.11 The bivariate normal distribution

TERMINOLOGY: The random vector (Y_1, Y_2) has a **bivariate normal distribution** if its joint pdf is given by

$$f_{Y_1, Y_2}(y_1, y_2) = \begin{cases} \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} e^{-Q/2}, & (y_1, y_2) \in \mathcal{R}^2 \\ 0, & \text{otherwise,} \end{cases}$$

where

$$Q = \frac{1}{1-\rho^2} \left[\left(\frac{y_1 - \mu_1}{\sigma_1} \right)^2 - 2\rho \left(\frac{y_1 - \mu_1}{\sigma_1} \right) \left(\frac{y_2 - \mu_2}{\sigma_2} \right) + \left(\frac{y_2 - \mu_2}{\sigma_2} \right)^2 \right].$$

We write $(Y_1, Y_2) \sim \mathcal{N}_2(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$. There are 5 parameters associated with this bivariate distribution: the marginal means (μ_1 and μ_2), the marginal variances (σ_1^2 and σ_2^2), and the correlation ρ .

FACTS ABOUT THE BIVARIATE NORMAL DISTRIBUTION:

1. Marginally, $Y_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$ and $Y_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$.
2. Y_1 and Y_2 are independent $\iff \rho = 0$. This is only true for the bivariate normal distribution (remember, this does not hold in general).
3. The conditional distribution

$$Y_1 | \{Y_2 = y_2\} \sim \mathcal{N} \left[\mu_1 + \rho \left(\frac{\sigma_1}{\sigma_2} \right) (y_2 - \mu_2), \sigma_1^2(1 - \rho^2) \right].$$

4. The conditional distribution

$$Y_2 | \{Y_1 = y_1\} \sim \mathcal{N} \left[\mu_2 + \rho \left(\frac{\sigma_2}{\sigma_1} \right) (y_1 - \mu_1), \sigma_2^2(1 - \rho^2) \right].$$

EXERCISE: Suppose that $(Y_1, Y_2) \sim \mathcal{N}_2(0, 0, 1, 1, 0.5)$. What is $P(Y_2 > 0.5 | Y_1 = 0.2)$?

ANSWER: Conditional on $Y_1 = y_1 = 0.2$, $Y_2 \sim \mathcal{N}(0.1, 0.75)$. Thus,

$$P(Y_2 > 0.5 | Y_1 = 0.2) = P(Z > 0.46) = 0.3228.$$

5.12 Conditional expectation

5.12.1 Conditional means and curves of regression

TERMINOLOGY: Suppose that X and Y are continuous random variables and that $g(X)$ and $h(Y)$ are functions of X and Y , respectively. The **conditional expectation** of $g(X)$, given $Y = y$, is

$$E[g(X)|Y = y] = \int_{\mathcal{R}} g(x)f_{X|Y}(x|y)dx.$$

Similarly, the conditional expectation of $h(Y)$, given $X = x$, is

$$E[h(Y)|X = x] = \int_{\mathcal{R}} h(y)f_{Y|X}(y|x)dy.$$

If X and Y are discrete, then sums replace integrals.

IMPORTANT: It is important to see that, in general,

- $E[g(X)|Y = y]$ is a function of y , and
- $E[h(Y)|X = x]$ is a function of x .

CONDITIONAL MEANS: In the definition above, if $g(X) = X$ and $h(Y) = Y$, we get (in the continuous case),

$$\begin{aligned} E(X|Y = y) &= \int_{\mathcal{R}} xf_{X|Y}(x|y)dx \\ E(Y|X = x) &= \int_{\mathcal{R}} yf_{Y|X}(y|x)dy. \end{aligned}$$

$E(X|Y = y)$ is called the **conditional mean** of X , given $Y = y$. $E(Y|X = x)$ is the conditional mean of Y , given $X = x$.

Example 5.19. In a simple genetics model, the proportion, say X , of a population with Trait 1 is always less than the proportion, say Y , of a population with trait 2. In Example 5.3, we saw that the random vector (X, Y) has joint pdf

$$f_{X,Y}(x, y) = \begin{cases} 6x, & 0 < x < y < 1 \\ 0, & \text{otherwise.} \end{cases}$$

In Example 5.5, we derived the conditional distributions

$$f_{X|Y}(x|y) = \begin{cases} 2x/y^2, & 0 < x < y \\ 0, & \text{otherwise} \end{cases} \quad \text{and} \quad f_{Y|X}(y|x) = \begin{cases} \frac{1}{1-x}, & x < y < 1 \\ 0, & \text{otherwise.} \end{cases}$$

Thus, the conditional mean of X , given $Y = y$ is

$$\begin{aligned} E(X|Y = y) &= \int_0^y x f_{X|Y}(x|y) dx \\ &= \int_0^y x \left(\frac{2x}{y^2} \right) dx = \frac{2}{y^2} \left(\frac{x^3}{3} \Big|_0^y \right) = \frac{2y}{3}. \end{aligned}$$

Similarly, the conditional mean of Y , given $X = x$ is

$$\begin{aligned} E(Y|X = x) &= \int_x^1 y f_{Y|X}(y|x) dy \\ &= \int_x^1 y \left(\frac{1}{1-x} \right) dy = \frac{1}{1-x} \left(\frac{y^2}{2} \Big|_x^1 \right) = \frac{1}{2}(x+1). \end{aligned}$$

That $E(Y|X = x) = \frac{1}{2}(x+1)$ is not surprising because $Y|\{X = x\} \sim \mathcal{U}(x, 1)$. \square

TERMINOLOGY: Suppose that (X, Y) is a bivariate random vector.

- The graph of $E(X|Y = y)$ versus y is called the **curve of regression** of X on Y .
- The graph of $E(Y|X = x)$ versus x is the curve of regression of Y on X .

The curve of regression of Y on X , from Example 5.19, is depicted in Figure 5.17.

5.12.2 Iterated means and variances

REMARK: In general, $E(X|Y = y)$ is a function of y , and y is fixed (not random). Thus, $E(X|Y = y)$ is a **fixed number**. However, $E(X|Y)$ is a function of Y ; thus, $E(X|Y)$ is a **random variable**! Furthermore, as with any random variable, it has a mean and variance associated with it!!

ITERATED LAWS: Suppose that X and Y are random variables. Then the **laws of iterated expectation and variance**, respectively, are given by

$$E(X) = E[E(X|Y)]$$

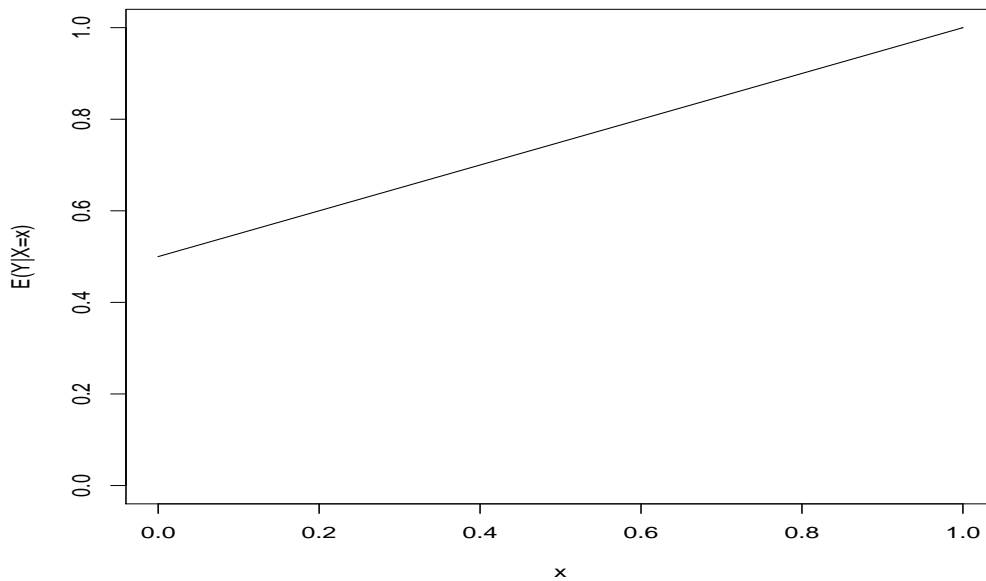


Figure 5.17: The curve of regression $E(Y|X = x)$ versus x in Example 5.19.

and

$$V(X) = E[V(X|Y)] + V[E(X|Y)].$$

NOTE: When considering the quantity $E[E(X|Y)]$, the inner expectation is taken with respect to the conditional distribution $f_{X|Y}(x|y)$. However, since $E(X|Y)$ is a function of Y , the outer expectation is taken with respect to the marginal distribution $f_Y(y)$.

Proof. We will prove that $E(X) = E[E(X|Y)]$ for the continuous case. Note that

$$\begin{aligned} E(X) &= \int_{\mathcal{R}} \int_{\mathcal{R}} x f_{X,Y}(x, y) dx dy \\ &= \int_{\mathcal{R}} \int_{\mathcal{R}} x f_{X|Y}(x|y) f_Y(y) dx dy \\ &= \int_{\mathcal{R}} \underbrace{\left[\int_{\mathcal{R}} x f_{X|Y}(x|y) dx \right]}_{E(X|Y=y)} f_Y(y) dy = E[E(X|Y)]. \quad \square \end{aligned}$$

Example 5.20. Suppose that in a field experiment, we observe Y , the number of plots, out of n , that respond to a treatment. However, we don't know the value of p , the probability of response, and furthermore, we think that it may be a function of location,

temperature, precipitation, etc. In this situation, it might be appropriate to regard p as a random variable. Specifically, suppose that the random variable P varies according to a beta(α, β) distribution. That is, we assume a **hierarchical structure**:

$$\begin{aligned} Y|P = p &\sim \text{binomial}(n, p) \\ P &\sim \text{beta}(\alpha, \beta). \end{aligned}$$

The (unconditional) mean of Y can be computed using the iterated expectation rule:

$$E(Y) = E[E(Y|P)] = E[nP] = nE(P) = n\left(\frac{\alpha}{\alpha + \beta}\right).$$

The (unconditional) variance of Y is given by

$$\begin{aligned} V(Y) &= E[V(Y|P)] + V[E(Y|P)] \\ &= E[nP(1 - P)] + V[nP] \\ &= nE(P - P^2) + n^2V(P) \\ &= nE(P) - n\{V(P) + [E(P)]^2\} + n^2V(P) \\ &= n\left(\frac{\alpha}{\alpha + \beta}\right) - n\left[\frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)} + \left(\frac{\alpha}{\alpha + \beta}\right)^2\right] + \frac{n^2\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)} \\ &= n\left(\frac{\alpha}{\alpha + \beta}\right)\left[1 - \left(\frac{\alpha}{\alpha + \beta}\right)\right] + \underbrace{\frac{n(n - 1)\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}}_{\text{extra variation}}. \end{aligned}$$

Unconditionally, the random variable Y follows a **beta-binomial** distribution. This is a popular probability model for situations wherein one observes binomial type responses but where the variance is suspected to be larger than the usual binomial variance. \square

BETA-BINOMIAL PMF: The probability mass function for a **beta-binomial** random variable Y is given by

$$\begin{aligned} p_Y(y) &= \int_0^1 f_{Y,P}(y, p) dp = \int_0^1 f_{Y|P}(y|p) f_P(p) dp \\ &= \int_0^1 \binom{n}{y} p^y (1 - p)^{n-y} \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1} (1 - p)^{\beta-1} dp \\ &= \binom{n}{y} \frac{\Gamma(\alpha + \beta)\Gamma(y + \alpha)\Gamma(n + \beta - y)}{\Gamma(\alpha)\Gamma(\beta)\Gamma(n + \alpha + \beta)}, \end{aligned}$$

for $y = 0, 1, \dots, n$, and $p_Y(y) = 0$, otherwise.