STAT 515, Statistics I

JIANYU CHEN
DEPARTMENT OF MATHEMATICS AND STATISTICS
UNIVERSITY OF MASSACHUSETTS AMHERST
EMAIL: JCHEN@MATH.UMASS.EDU

Acknowledgement

Some contents of the lecture notes are based on Spring 2016 course notes by Prof. Yao Li and Dr. Zheng Wei.

Contents

1 Introduction

Statistics is a theory of information, with inference making as its objective.

$$\textit{Data} \Rightarrow \textit{Probability measurement} \Rightarrow \textit{Inference}/\textit{Prediction}$$

Example 1.1 *Based on the latest presidential election polls, who is going to be the next US president?*

Example 1.2 *Roll a six-sided dice and observe the number on the upper face. One can ask the following questions:*

(1) *What is the probability to get 1? 2? 5?*

(2) *What is the probability to get odd numbers?*

(3) *What is the probability to get 1, if the observed number is known to be odd?*

(4) *What is the mean value of the observed number?*

2 Probability

2.1 Set Notation

- Set: a collection of elements, denoted by capital letters, like A, B, etc.
- $A = \{a_1, a_2, \dots\}$: the elements of A are $a_1, a_2, \dots$.
- $\emptyset$: the empty set - a set that contains no elements;
- S: the universal/full set - a set of all elements under consideration;
- $A \subset B$: A is a subset of B, that is, every element in A is also in B.
- $A \cup B$: union of A and B, that is,

  $$A \cup B = \{x \in S : \ \ x \in A \text{ or } x \in B\}.$$

  We say that A and B are exhaustive, if $A \cup B = S$.
- $A \cap B$: intersection of A and B, that is,

  $$A \cap B = \{x \in S : \ \ x \in A \text{ and } x \in B\}.$$

  We say that A and B are mutually exclusive or mutually disjoint, if $A \cap B = \emptyset$.

- $\overline{A}$ or A^c or $S\backslash A$: the complement of A, that is,

$$\overline{A} = \{x \in S : \ x \notin A\}.$$

We also denote the complement of A in B by

$$B\backslash A = B \cap \overline{A} = \{x \in B : \ x \notin A\}.$$

We can visualize all these operations via Venn Diagram.

Exercise 2.1 $S = \{0, 1, \cdots, 9\}$. $A = \{1, 2, 3, 4, 5\}$. $B = \{3, 5\}$. $C = \{5, 7, 9\}$.

(1) *Find* $A \cup C$, $A \cap C$, $B \cup C$ *and* $B \cap C$.

(2) *Find* $\overline{A}$, $\overline{C}$, $A\backslash B$, $B\backslash C$.

Proposition 2.2

- *Distributive law:*

$$A \cap (B \cup C) = (A \cap B) \cup (A \cap C), \quad A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$$

- *DeMorgan's law:*

$$\overline{A \cap B} = \overline{A} \cup \overline{B}, \quad \overline{A \cup B} = \overline{A} \cap \overline{B}$$

Proof By definition. Exercise. □

2.2 Probabilistic Model - Discrete Case

Definition 2.3 *An experiment is the process by which an observation is made.*

Definition 2.4 *The events are the outcomes of an experiments.*

- *A simple event is an event that cannot be decomposed, and happens in only one way. Usually denoted by* E *with some subscript.*
- *A compound event is an event that can be decomposed into other events, and can happen in more than one distinctive way.*

Definition 2.5 *We refer a distinct outcome to a sample point. A sample space* S *associated with an experiment is the set consisting of all possible sample points.*

Mathematically, events can be represented by sets.

Example 2.6 *Below are some events associated with a single toss of a six-sided dice:*

- E_1*: Observe a 1.* $\Leftrightarrow$ $E_1 = \{1\}$.
- E_2 *: Observe a 2.* $\Leftrightarrow$ $E_2 = \{2\}$.
- E_3*: Observe a 3.* $\Leftrightarrow$ $E_3 = \{3\}$.
- E_4*: Observe a 4.* $\Leftrightarrow$ $E_4 = \{4\}$.
- E_5*: Observe a 5.* $\Leftrightarrow$ $E_5 = \{5\}$.
- E_6*: Observe a 6.* $\Leftrightarrow$ $E_6 = \{6\}$.
- A*: Observe an odd number.* $\Leftrightarrow$ $A = \{1,3,5\}$.
- B*: Observe a number less than 5.* $\Leftrightarrow$ $B = \{1,2,3,4\}$.
- C*: Observe a 2 or a 3.* $\Leftrightarrow$ $C = \{2,3\}$.
- S*: All possible observation.* $\Leftrightarrow$ $S = \{1,2,3,4,5,6\}$.

It is easy to see that

(1) A simple event corresponds to a sample point, that is, a simple event can be represented by a set consisting of exactly one sample point. Therefore, two distinct simple events are mutually exclusive.

(2) Any compound event can be written as unions of simple events, for instance, in the above example, $A = E_1 \cup E_3 \cup E_5$.

(3) Any event A is a subset of the sample space S.

Definition 2.7 *A discrete sample space is one that contains finite or countably many sample points.*

Definition 2.8 (Probability) *Let S be a sample space associated to some experiment. To every event A in S (A is a subset of S), we assign a number, $P(A)$, called the probability of A, so that the following axioms hold:*

Axiom 1: $P(A) \geq 0$;

Axiom 2: $P(S) = 1$;

Axiom 3: If $A_1, A_2, A_3, \ldots$ form a sequence of mutually exclusive events in S (that is, $A_i \cap A_j = \emptyset$ if $i \neq j$), then

$$P(A_1 \cup A_2 \cup A_3 \cup \ldots) = \sum_{i=1}^{\infty} P(A_i).$$

Example 2.9 *Let E_i be the event of observing i in a single toss of a fair dice, then we assign $P(E_i) = 1/6$. For arbitrary event A, we can decompose $A = E_{i_1} \cup E_{i_2} \cup \ldots$ and then assign $P(A)$ by Axiom 3. For instance, if A is the event to observe an odd number, then $A = E_1 \cup E_3 \cup E_5$, and hence*

$$P(A) = P(E_1) + P(E_3) + P(E_5) = 1/6 + 1/6 + 1/6 = 1/2.$$

In such assignment, Axiom 2, that is, $P(S) = 1$ is satisfied.

Exercise 2.10 *Every person's blood type is A, B, AB, O. In addition, each individual either has the Rhesus (Rh) factor (+) or does not (-). A medical technician records a person's blood type and Rh factor.*

(1) List the sample space for this experiment.

(2) Of the volunteers in a blood center, 1in 3 have O^+, 1 in 15 have O^-, 1 in 3 have A^+, and 1 in 16 have A^-. What is the probability that a randomly chosen volunteer has (a) type O blood? (b) neither type A nor type O blood?

Proposition 2.11 (Basic Properties of Probability)

(1) *Law of Complement: For any event A, $P(\overline{A}) = 1 - P(A)$. Consequently,*

   (a) $P(\emptyset) = 0$;

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

(2) *For any events A and B,*

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

   In particular, if $A \subset B$, then $P(A) \leq P(B)$.

(3) *Additive Law: For any events A and B,*

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

Proof Exercise. □

Exercise 2.12 *If A and B are exhaustive, $P(A) = 0.7$ and $P(B) = 0.9$, find $P(A \cap B)$, $P(A \backslash B)$, $P(B \backslash A)$ and $P(\overline{A} \cup \overline{B})$.*

Solution Since $A \cup B = S$, then $P(A \cup B) = P(S) = 1$.

$$P(A \cap B) = P(A) + P(B) - P(A \cup B) = 0.7 + 0.9 - 1 = 0.6;$$

$$P(A \backslash B) = P(A) - P(A \cap B) = 0.7 - 0.6 = 0.1;$$

$$P(B\backslash A) = P(B) - P(B \cap A) = 0.9 - 0.6 = 0.3;$$
$$P(\overline{A} \cup \overline{B}) = P(\overline{A \cap B}) = 1 - P(A \cap B) = 1 - 0.6 = 0.4.$$

□

Exercise 2.13 *For any two events A and B, prove that*

$$P(A \cap B) \geq 1 - P(\overline{A}) - P(\overline{B}).$$

2.3 Calculating the Probability of an Event - The Sample-Point Method

As shown in Example 2.9, once the probability is assigned to each simple event in a discrete sample space, we can determine the probability of any event. More precisely,

Method 2.14 (Sample-Point Method)

(1) *Define the experiment and determine the sample space S.*

(2) *Assign reasonable probabilities to the sample points in S such that $P(E_i) \geq 0$ and $\sum_i P(E_i) = 1$.*

(3) *Any event A is a union of simple events, that is, $A = E_{i_1} \cup E_{i_2} \cup \ldots$, then*

$$P(A) = \sum_j P(E_{i_j}).$$

Example 2.15 *A balanced coin is tossed three times. Calculate the probability that exactly two of the three tosses result in heads.*

Solution Denote the head by H, and tail by T. Then the three-time tossing of a balanced coin has 8 sample points, or equivalently, 8 simple events:

$$E_1 = \{HHH\},\ \ E_2 = \{HHT\},\ \ E_3 = \{HTH\},\ \ E_4 = \{HTT\},$$
$$E_5 = \{THH\},\ \ E_6 = \{THT\},\ \ E_7 = \{TTH\},\ \ E_8 = \{TTT\}.$$

Since the coin is balanced, then each simple event is equiprobable, that is, $P(E_i) = 1/8$. Let A denote the event that exactly two of the three tosses result in heads, that is,

$$A = \{HHT,\ HTH,\ THH\} = E_2 \cup E_3 \cup E_5,$$

and hence

$$P(A) = P(E_2) + P(E_3) + P(E_5) = 1/8 + 1/8 + 1/8 = 3/8.$$

□

Remark 2.16 *The sample-point method is very effective when every simple event has equal chance. In such case, $P(E_i) = 1/N$ where N is the sample size. For an event A that contains exactly N_A sample points, we must have $P(A) = \frac{N_A}{N}$.*

Exercise 2.17 *A vehicle arriving an intersection can turn right, turn left or go straight. An experiment consists of observing two vehicles moving through the intersection. Assuming equally likely sample points, what is the probability that at least one vehicle turns? What is the probability that at most one vehicle turns?*

2.4 Tools for Counting Sample Points

2.4.1 Multiplication Principle

Tool 2.18 (Multiplication Principle) *With m elements $a_1, \ldots, a_m$ and n elements $b_1, \ldots, b_n$, there are $mn = m \times n$ pairs of elements of the product-type (a_i, b_j).*

More generally, with m_i elements $a_1^i, \ldots, a_{m_i}^i$, $i = 1, 2, \ldots, k$, there are $m_1 m_2 \ldots m_k$ k-tuples of elements of the product-type $(a_{j_1}^1, a_{j_2}^2, \ldots, a_{j_k}^k)$.

Example 2.19 *An experiment invokes flipping a coin and rolling a 6-sided dice simultaneously. Then each sample point is of the form $(*, i)$, where $*$ is either H or T, and i ranges from 1 to 6. So there are $N = 2 \times 6 = 12$ sample points in total.*

If we further assume that both the coin and the dice is fair, then all sample points have equal chance. Let A be the event that we flip a head and roll an even number, that is,

$$A = \{(*, i) : \ * = H, \ i = 2, 4, 6\}.$$

Then $N_A = 1 \times 3 = 3$, and so $P(A) = \frac{N_A}{N} = \frac{3}{12} = 0.25$.

Example 2.20 *Flip a coin 10 times. Then there are totally $N = 2^{10}$ sample points.*

Exercise 2.21 *How many subsets are there for a set of 10 elements?*
(Hint: This is in some sense the same as flipping a coin 10 times. A subset A is formed by determining whether the i-th element belongs to A.)

2.4.2 Permutation

Recall that $n! = 1 \times 2 \times \cdots \times n$, and $0! = 1$ in convention.

Definition 2.22 (Permutation)

- *A permutation of n distinct objects is an ordered arrangement of them. Choosing objects one by one, we find that there are*

$$n \times (n-1) \times (n-2) \times \cdots \times 1 = n!$$

permutations of n distinct objects in total.

- *Let $1 \leq r \leq n$. The number of permutations of r objects from n distinct ones is given by*

$$n \times (n-1) \times \cdots \times (n-r+1) = \frac{n!}{(n-r)!} =: P_r^n.$$

Note that $P_0^n = 1$ and $P_n^n = n!$.

Example 2.23 *Randomly choose a president, a vice president and a secretary in a 50-person club. Since each job is distinct, there should be*

$$P_3^{50} = 50 \cdot 49 \cdot 48 = 117,600$$

possible choices.

Example 2.24 (Birthday Problem) *There are n people in a class. Ignoring leap years, twins, seasonal and weekday variation, and assume that the 365 possible birthdays are equally likely. What is the probability that at least two people in the classroom have the same birthday?*

On the one hand, each sample point is any n dates from 365 dates, then the sample size is 365^n by the multiplication principle. On the other hand, let A be the event that at least two people have the same birthday, then the complement $\overline{A}$ is the event that every people has distinct birthday. In other words, $\overline{A}$ contains all permutations of n dates from 365 dates, and thus $N_{\overline{A}} = P_n^{365}$. So

$$P(A) = 1 - P(\overline{A}) = 1 - \frac{N_{\overline{A}}}{N} = 1 - \frac{P_n^{365}}{365^n} = 1 - \frac{365!}{365^n(365-n)!}.$$

However, this formula is not easy to evaluate. Alternatively, we write

$$P(A) = 1 - \frac{365 \times 364 \times \cdots \times (365-n+1)}{365^n} = 1 - \prod_{k=0}^{n-1} \left(1 - \frac{k}{365}\right).$$

Using the Taylor expansion of e^x, we get

$$e^x = 1 + x + \frac{x^2}{2} + \frac{x^3}{3!} + \cdots \approx 1 + x, \quad \text{as } x \text{ close to } 0.$$

Thus, we could approximate $P(A)$ as:

$$P(A) \approx 1 - \prod_{k=0}^{n-1} e^{-\frac{k}{365}} = 1 - e^{-\frac{1}{365}\sum_{k=0}^{n-1} k} = 1 - e^{-\frac{n(n-1)}{730}} \approx 1 - e^{-\frac{n^2}{730}}.$$

Thus, we have

- *if there are $n = 23$ people, then $P(A) \approx 50.7\% > 50\%$;*
- *if there are $n = 40$ people, then $P(A) \approx 89.1\%$;*
- *if there are $n = 50$ people, then $P(A) \approx 97\%$.*

See more details in https://en.wikipedia.org/wiki/Birthday_problem.

2.4.3 Combination

Definition 2.25 (Combination)

- *The number of combinations of r objects from n distinct ones is given by*

  $$C_r^n = \binom{n}{r} := \frac{P_r^n}{r!} = \frac{n!}{r!(n-r)!}.$$

  The difference between permutation and combination is that permutation cares about the ordering of the r objects, while combination does not.

- *In general, the number of ways of partitioning n distinct objects into k distinct groups containing $n_1, n_2, \ldots, n_k$ objects, where $\sum_{i=1}^k n_i = n$, is given by*

  $$\binom{n}{n_1\ n_2\ \ldots\ n_k} = \frac{n!}{n_1!\, n_2!\ \ldots\ n_k!}.$$ [1]

Example 2.26 *There are 3 basketballs, 2 footballs and 5 volleyballs. Align all these balls in a row, then how many possible arrangements are there?*

Suppose that these 10 balls are already aligned and labelled in a row, then we just need to determine which group ball i belongs to. In other words, we partite 10 objects into 3 distinct groups containing 3, 2, 5 objects, so there are $\binom{10}{3\ 2\ 5} = 2,520$ arrangements.

[1] To understand this formula, take all $n!$ permutations of the n objects and divide them into k distinct groups containing $n_1, n_2, \ldots, n_k$ objects in order. This would cover all the partitions with permutation repetitions in each subgroup, that is, we count the same combination in the i-th group for $n_i!$ times.

Exercise 2.27 (Binomial and Mutli-nomial Expansion)

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

$$(x_1 + x_2 + \cdots + x_k)^n = \sum_{n_1+n_2+\cdots+n_k=n} \binom{n}{n_1\ n_2\ \ldots\ n_k} x_1^{n_1} x_2^{n_2} \ldots x_k^{n_k}.$$

Use the above expansion to show that

$$\sum_{i=0}^{n} \binom{n}{i} = 2^n, \quad \sum_{n_1+n_2+\cdots+n_k=n} \binom{n}{n_1\ n_2\ \ldots\ n_k} = k^n.$$

Example 2.28 *A balanced coin is tossed 100 times. What is the probability that exactly 50 of the 100 tosses result in heads?*

There are totally 2^{100} *sample points. The number of tossing 50 heads in the 100 tosses is* $\binom{100}{50}$*, so the probability is* $\dfrac{\binom{100}{50}}{2^{100}} \approx 0.0796$.

Example 2.29 *Lottery: From the numbers* $1, 2, \ldots, 44$*, a person may pick any six numbers for a ticket. The winning number is then randomly chosen from the 44 numbers. How many possible tickets are there in the following situations?*

(1) *Ordered and with replacement:* 44^6*;*

(2) *Ordered and without replacement:* P_6^{44}*;*

(3) *Unordered and without replacement:* $C_6^{44} = \binom{44}{6} = 7,059,052$.

2.5 Conditional Probability and Independence

2.5.1 Conditional Probability

The probability of an event will sometimes depend on whether we know that other events have occurred.

Example 2.30 *The (unconditional) probability of a 1 in the toss of a fair dice is* $1/6$*. If we know that an odd number is obtained, then the number on the upper face of the dice must be 1, 3, or 5, and thus the conditional probability of a 1 given that information will instead be* $1/3$.

Definition 2.31 *The conditional probability of an event A, given that an event B has occurred, is equal to*

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

provided $P(B) > 0$. The symbol $P(A|B)$ is read "probability of A given B". Note that the outcome of the event B allows us to update the probability of A.

Example 2.32 *Suppose a fair dice is tossed once. The probability of a 1, given that an odd number was obtained, is then*

$$\begin{aligned} P(E_1|A) = \frac{P(E_1 \cap A)}{P(A)} = \frac{P(E_1)}{P(E_1 \cup E_3 \cup E_5)} &= \frac{P(E_1)}{P(E_1) + P(E_3) + P(E_5)} \\ &= \frac{1/6}{1/6 + 1/6 + 1/6} = 1/3. \end{aligned}$$

The probability of a 2, given that an odd number was obtained, is then

$$P(E_2|A) = \frac{P(E_2 \cap A)}{P(A)} = \frac{P(\emptyset)}{P(A)} = 0.$$

Remark 2.33

- *if two events A and B are mutually exclusive, that is, $A \cap B = \emptyset$, then $P(A|B) = P(B|A) = 0$;*
- *if $A \subset B$, then $P(B|A) = 1$, that is, B definitely happens given that A happens.*

Example 2.34 *If two events A and B are such that $P(A) = 0.5$, $P(B) = 0.4$, and $P(A \cap B) = 0.3$, then*

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{0.3}{0.4} = 0.75, \quad P(A|A \cap B) = 1,$$

$$P(A|A \cup B) = \frac{P(A)}{P(A \cup B)} = \frac{P(A)}{P(A) + P(B) - P(A \cap B)} = \frac{0.5}{0.5 + 0.4 - 0.3} = \frac{5}{6}.$$

Example 2.35 *A certain population of employees are taking job competency exams. The following table shows the percentage passing or failing the exam, listed according to the sex.*

	Male	*Female*	*Total*
Pass	*24 %*	*40%*	*64%*
Fail	*16%*	*20%*	*36%*
Total	*40%*	*60%*	*100%*

Choose an employee randomly.

- *Given a chosen employee is female, the probability that she passes is*
$$P(Pass|Female) = \frac{P(Pass \cap Female)}{P(Female)} = \frac{40\%}{60\%} = \frac{2}{3}.$$
- *Given an employee who has passed the exam, the probability that the employee is male is*
$$P(Male|Pass) = \frac{P(Male \cap Pass)}{P(Pass)} = \frac{24\%}{64\%} = \frac{3}{8}.$$

2.5.2 Independence of Events

If the probability of an event A is unaffected by the occurrence or nonoccurence of another event B, then we would say that the events A and B are independent.

Definition 2.36 *Two events A are B are said to be independent if*
$$P(A \cap B) = P(A)P(B),$$
or equivalently, if $P(A|B) = P(A)$, or if $P(B|A) = P(B)$.

Otherwise, the events are said to be dependent.

Example 2.37 *Consider the following events in the toss of a dice:*

A: Observe an odd number —— $A = \{1,3,5\}$;

B: Observe an even number —— $B = \{2,4,6\}$;

C: Observe a 1 or 2 —— $C = \{1,2\}$.

Note that $P(A) = P(B) = \frac{1}{2}$, and $P(C) = \frac{1}{3}$.

- *A and B are dependent (in fact, mutually exclusive), since*
$$P(A \cap B) = P(\emptyset) = 0, \quad P(A)P(B) = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}.$$
- *A and C are independent since*
$$P(A \cap C) = P(\{1\}) = \frac{1}{6}, \quad P(A)P(C) = \frac{1}{2} \cdot \frac{1}{3} = \frac{1}{6}.$$

Proposition 2.38

(1) *If two events A and B are independent, then the following pairs of events are also independent: (a) A and $\overline{B}$; (b) $\overline{A}$ and B; (c) $\overline{A}$ and $\overline{B}$.*

(2) *The events A and $\overline{A}$ are dependent if $0 < P(A) < 1$.*

(3) *If an event A is self-independent, that is, A is independent with itself, then $P(A) = 0$ or 1.*

Proof

(1) $P(A \cap \overline{B}) = P(A) - P(A \cap B) = P(A) - P(A)P(B) = P(A)[1 - P(B)] = P(A)P(\overline{B})$. Similar arguments for others.

(2) Since $P(A \cap \overline{A}) = P(\emptyset) = 0$, but $P(A)P(\overline{A}) = P(A)[1 - P(A)] > 0$.

(3) If A is self-independent, then $P(A) = P(A \cap A) = P(A)^2$, so $P(A) = 0$ or 1.

□

Definition 2.39 *A collection of events $A_1, A_2, \ldots, A_n$ are said to be mutually independent if for any subindices $i_1, i_2, \ldots, i_k \in \{1, \ldots, n\}$, we have*

$$P(A_{i_1} \cap \cdots \cap A_{i_k}) = P(A_{i_1}) \cdot \cdots \cdot P(A_{i_k}).$$

There are $(2^n - n - 1)$ identities to check for the mutual independence of n events, which is usually too complicated when n is large. Nevertheless, in most applications, you can use common sense to determine that some events are mutually independent.

2.5.3 Multiplicative Law of Probability

Theorem 2.40 (Multiplicative Law of Probability) *The probability of the intersection of two events A and B is*

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

If A and B are independent, then

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

More generally, the probability of the intersection of n events $A_1, A_2, A_3, \ldots, A_n$ is

$$P(A_1 \cap A_2 \cap A_3 \cap \cdots \cap A_n) = P(A_1)P(A_2|A_1)P(A_3|A_1 \cap A_2) \ldots P(A_n|A_1 \cap A_2 \cap \cdots \cap A_{n-1}).$$

Example 2.41 *Suppose that A and B are two independent events with $P(A) = 0.5$ and $P(B) = 0.4$, then*

- $P(A \cap B) = P(A)P(B) = 0.5 \times 0.4 = 0.2$;

- $P(A \cup B) = P(A) + P(B) - P(A \cap B) = 0.5 + 0.4 - 0.2 = 0.7$;
- $P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{0.2}{0.4} = 0.5$.

Exercise 2.42 *A smoke detector system uses two devices, A and B. If the smoke is present, the probability that it will be detected by device A is 0.95; by device B is 0.90; by both is 0.88.*

(a) If the smoke is present, find the probability that it will be detected by at least one of the devices.

(b) Given device A detects the smoke, what is the probability that device B does not?

2.6 Calculating the Probability of an Event - The Event-Composition Method

Method 2.43 (Event-Composition Method)

(1) *Define the experiment and determine the sample space S.*

(2) *Express an event A as a composition of two or more events with known probability, using unions, intersections, and complements.*

(3) *Compute $P(A)$ by Additive Law, Multiplicative Law, Law of Complements, the Conditional Probability, Independences, etc.*

Example 2.44 *The odds are two to one that, when Alice and Bob play tennis, Alice wins. Suppose that Alice and Bob play two matches, and the result of the first match would not affect the second. What is the probability that Alice wins at least one match?*

Denote by A_i the event that Alice wins the i-th match, $i = 1, 2$, then $P(A_i) = 2/3$. Note that A_1 and A_2 are independent, and so the probability that Alice wins at least one match is

$$\begin{aligned} P(A_1 \cup A_2) &= P(A_1) + P(A_2) - P(A_1 \cap A_2) \\ &= P(A_1) + P(A_2) - P(A_1)P(A_2) \\ &= 2/3 + 2/3 - 2/3 \cdot 2/3 = 8/9. \end{aligned}$$

Example 2.45 *It is known that a patient with a disease will respond to treatment with probability equal to 0.6. If five patients with disease are treated and respond independently, find the probability that at least one will respond.*

Let A be the event that at least one patient will respond, and B_i be the event that the i-th patient will not respond, $i = 1, 2, \ldots 5$. Then

$$P(B_i) = 1 - P(\overline{B}_i) = 1 - 0.6 = 0.4.$$

Note that

$$\overline{A} = B_1 \cap B_2 \cap B_3 \cap B_4 \cap B_5.$$

Note that $B_1, B_2, \ldots, B_5$ are mutually independent, then

$$\begin{aligned} P(A) = 1 - P(\overline{A}) &= 1 - P(B_1 \cap B_2 \cap B_3 \cap B_4 \cap B_5) \\ &= 1 - P(B_1)P(B_2)\ldots P(B_5) \\ &= 1 - 0.4^5 \approx 0.9898. \end{aligned}$$

Example 2.46 *Observation of a waiting line at a local bank indicates that the probability that a new arrival will be a VIP member is $p = 1/5$. Find the probability that the r-th client is the first VIP. (Assume that conditions of arriving clients represent independent events)*

Let E_i be the event that the i-th client is VIP, and A_r be the event that the r-th client is the first VIP, then

$$A_r = \overline{E}_1 \cap \overline{E}_2 \cap \overline{E}_{r-1} \cap E_r.$$

Note that $P(E_i) = p$ and $P(\overline{E}_i) = 1 - p$, and $\overline{E}_1, \overline{E}_2, \ldots, \overline{E}_{r-1}, E_r$ are mutually independent. Thus,

$$P(A_r) = P(\overline{E}_1)\ldots P(\overline{E}_{r-1})P(E_r) = (1-p)^{r-1}p = 0.2 \cdot 0.8^{r-1}.$$

2.7 The Law of Total Probability and Bayes' Rule

The event-composition approach is sometimes facilitated by viewing the sample space S as a union of mutually exclusive subsets.

Definition 2.47 *If a collection of events $B_1, B_2, \ldots, B_k$ is such that*

- *$B_1, B_2, \ldots, B_k$ are mutually exclusive, that is, $B_i \cap B_j = \emptyset$ if $i \neq j$;*
- *$B_1, B_2, \ldots, B_k$ are exhaustive, that is, $B_1 \cup B_2 \cup \cdots \cup B_k = S$,*

then the collection of events $\{B_1, B_2, \ldots, B_k\}$ is called a partition of sample space S.

Given a partition $\{B_1, B_2, \ldots, B_k\}$ of S, any event A can be decomposed as

$$A = (A \cap B_1) \cup (A \cap B_2) \cup \cdots \cup (A \cap B_k), \tag{2–1}$$

and note that $A \cap B_1, \ldots, A \cap B_k$ are mutually exclusive.

Theorem 2.48 *Given a partition $\{B_1, B_2, \ldots, B_k\}$ of S such that $P(B_i) > 0$ for all $i = 1, \ldots, k$, any event A has probability*

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

Proof It directly follows from (2–1), Axiom 3 in the definition of Probability and the definition of conditional probability. □

As a direct consequence, we have the so-called Bayes' rule as a corollary.

Corollary 2.49 (Bayes' Rule) *Given a partition $\{B_1, B_2, \ldots, B_k\}$ of S such that $P(B_i) > 0$ for all $i = 1, \ldots, k$, and an event A with $P(A) > 0$, we have*

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

Example 2.50 *A population of voters contains 40% Republicans and 60% Democrats.It is reported that 30% of the Republicans and 70% of the Democrats favor an election issue.*

(1) *If a voter is randomly chosen, what is the probability that he or she will favor this election issue?*

(2) *If A randomly chosen voter is found to favor the issue, what is the probability that this person is a Democrat?*

Let R or D be the event that a voter is Republican or Democrat respectively, and let F be the event that the voter favors the issue. Note that $\{R, D\}$ is a partition of the sample space S.

(1) *By Law of Total Probability,*

$$P(F) = P(R)P(F|R)+P(D)P(F|D) = 40\%\times 30\%+60\%\times 70\% = 0.12+0.42 = 0.54.$$

(2) *By Bayes' Rule, the probability that a voter is Democrat given that he or she favors the issue is*

$$P(D|F) = \frac{P(D)P(F|D)}{P(R)P(F|R) + P(D)P(F|D)} = \frac{60\% \times 70\%}{40\% \times 30\% + 60\% \times 70\%} = \frac{0.42}{0.54} = 7/9.$$

Exercise 2.51 *A package, say P_1, of 24 balls, contains 8 green, 8 white, and 8 purple balls. Another Package, say P_2, of 24 balls, contains 6 green, 6 white and 12 purple balls. One of the two packages is selected at random.*

(1) *If 3 balls from this package (with replacement) were selected, all 3 are purple. Compute the conditional probability that package P_2 was selected.*

(2) *If 3 balls from this package (with replacement) were selected, they are 1 green, 1 white, and 1 purple. Compute the conditional probability that package P_2 was selected.*

3 Discrete Random Variables

3.1 Definition of Random Variables

We start with the following intuitive example.

Example 3.1 (Opinion Poll) *In an opinion poll, we decide to ask 100 people whether they agree or disagree with a certain bill.*

If we record a "1" as for agree and "0" for disagree, then the sample space S for this experiment has 2^{100} sample points, each an ordered string of 1's and 0's of length 100. It is tedious to list all sample points.

Suppose our interest is the number of people who agree out of 100. If we define variable Y to be the number of 1's recorded out of 100, then the new sample space for Y is the set $S_Y = \{0, 1, \ldots, 100\}$.

It frequently occurs that we are mainly interested in some functions of the outcomes as opposed to the outcome itself. In the example what we do is to define a new variable Y , the quantity of our interest.

Definition 3.2 *A random variable (RV) Y is a function from S into $\mathbb{R}$, where $\mathbb{R}$ is the set of real numbers. Notationally, we write $Y : S \to \mathbb{R}$, which assign each sample point $s \in S$ to a real number $y = Y(s)$.*

We usually use the late-alphabet capital letters, like X, Y, Z, W, to denote a random variable, and corresponding letters in lower case, like x, y, z, w, to denote a value that the random variable may assume. Also, we denote

$$(Y = y) = \{s \in S : \ Y(s) = y\},$$

that is, $(Y = y)$ is the set of all sample points assigned to the value y by the RV Y. Similarly, we denote

$$(Y > y) = \{s \in S : \ Y(s) > y\},$$

that is, $(Y > y)$ is the set of all sample points assigned to a value greater than y by the RV Y.

Example 3.3 (Opinion Poll) *Recall that*

$$S = \{s = a_1 a_2 \dots a_{100} : \ a_i = 0 \text{ or } 1\},$$

and Y is the random variable that represents the number of people who agree out of 100, that is,

$$Y(s) = Y(a_1 a_2 \dots a_{100}) = a_1 + a_2 + \cdots + a_{100} = \sum_{i=1}^{100} a_i.$$

Suppose that the minimum number to pass the bill is 60, then we need to consider the chance of the event $(Y \geq 60)$.

If $f : \mathbb{R} \to \mathbb{R}$ is a function from $\mathbb{R}$ to itself, and Y is a random variable over some sample space S, then we can define a new random variable $Z = f(Y)$ given by $Z(s) = f \circ Y(s) = f(Y(s))$.

Example 3.4 *In a 100-person club dinner, each person must order only one out of the two choices: beef or fish. Let Y be the number of people who choose beef.*

Assume that the price is \$15 for the beef dish, and \$18 for the fish. Then the total bill is a random variable Z given by

$$Z = 15Y + 18(100 - Y).$$

Definition 3.5 *A RV Y is said to be discrete if $R(Y)$, the range of Y - the set of all possible values for Y, is either finite or countably infinitely many.*

For instance, the random variables in all the above examples are finite.

3.2 Probability Distribution of a Discrete Random Variable

Definition 3.6 (Probability distribution function) *Given a discrete random variable Y, the function given by*

$$p_Y(y) := P(Y = y),$$

for all possible values y of Y, is called the probability distribution function (pdf) of Y.

If Y is the only random variable under consideration, we shall drop the subscript and simply denote the pdf by $p(y)$.

Example 3.7 *A supervisor in a manufacturing plant has three men and three women working for him. He wants to choose two workers randomly for a special job. Let Y denote the number of women in his selection. Find the probability distribution for Y and represent it by formula, a table, and a graph.*

The total number of choices is $\binom{6}{2}$, and the number of choices with y women workers, where $y = 0, 1, 2$, is given by $\binom{3}{y}\binom{3}{2-y}$. Thus,

$$p(y) = P(Y = y) = \frac{\binom{3}{y}\binom{3}{2-y}}{\binom{6}{2}} = \frac{\dfrac{3!}{y!(3-y)!}\dfrac{3!}{(2-y)!(1+y)!}}{\dfrac{6!}{2!4!}} = \frac{\dfrac{36}{y!(1+y)!(2-y)!(3-y)!}}{15},$$

so

$$p(0) = \frac{3}{15} = 0.2, \;\; p(1) = \frac{9}{15} = 0.6, \;\; p(2) = \frac{3}{15} = 0.2.$$

The probability distribution can then be represented by the following table:

y	$p(y)$
0	*0.2*
1	*0.6*
2	*0.2*

Proposition 3.8 *Given a discrete random variable Y with probability distribution function $p(y)$,*

(1) $0 \le p(y) \le 1$;

(2) $\sum_y p(y) = 1$, *where the summation ranges over all possible values y for Y.*

(3) *for any set $B \subset \mathbb{R}$, $P(Y \in B) = \sum_{y \in B} p(y)$.*

Proof This is a direct consequence of there axioms in the definition of probability, and the fact that the events $(Y = y_1)$, $(Y = y_2) \ldots$, $(Y = y_k)$ are mutually exclusive for all distinct values $y_1, y_2, \ldots, y_k$. □

Exercise 3.9 *Let $p(y) = cy^2$, $y = 1, 2, 3, 4$, be the pdf of some discrete random variable, find the value of c.*

Definition 3.10 (Cumulative distribution function) *Given a discrete random variable Y, the function given by*

$$F_Y(a) := P(Y \leq a) = \sum_{y \leq a} p_Y(y), \quad \text{for any } a \in \mathbb{R},$$

is called the cumulative distribution function (cdf) for Y.

If Y is the only random variable under consideration, we simply denote the cdf by $F(a)$.

Example 3.11 *Toss two fair coins simultaneously, and record head and tail by 0 and 1 respectively. Let Y denote the sum of the observed numbers. Find the pdf and cdf for Y.*

The sample points are 00, 01, 10, 11, each of which is of chance 1/4. Then the possible values for Y are $y = 0, 1, 2$, and the pdf of Y is given by

$$p(y) = P(Y = y) = \begin{cases} P(Y = 0) = P(\{00\} = 1/4, & y = 0, \\ P(Y = 1) = P(\{01, 10)) = 1/2, & y = 1, \\ P(Y = 2) = P(\{11\}) = 1/4, & y = 2 \end{cases}$$

The cdf of Y is then given by

$$F(a) = P(Y \leq a) = \begin{cases} 0, & a < 0, \\ p(0) = 1/4, & 0 \leq a < 1, \\ p(0) + p(1) = 3/4, & 1 \leq a < 2, \\ 1, & a \geq 2. \end{cases}$$

3.3 The Expected Value of a Random Variable, or a Function of a Random Variable

In this subsection, we always assume that Y is a discrete random variable with the probability distribution function $p(y)$.

Definition 3.12 (Expected value of a random variable) *The expected value (expectation) of Y is defined to be*

$$E(Y) = \sum_y yp(y).$$

Theorem 3.13 (Expected value of a function of a random variable) *Let $f : \mathbb{R} \to \mathbb{R}$ be a function. Then the expected value of $Z = f(Y)$ is given by*

$$E(Z) = E(f(Y)) = \sum_y f(y)p(y).$$

Definition 3.14 (Variance and standard deviation) *The variance of Y is defined to be*

$$V(Y) = E[(Y-\mu)^2] = \sum_y (y-\mu)^2 p(y),$$

where $\mu = E(Y)$. The standard deviation is then given by $\sqrt{V(Y)}$.

Example 3.15 *Roll a fair dice, and let X be the number observed. Find $E(X)$ and $V(X)$.*

Note that $p(x) = P(X = x) = 1/6$, $x = 1, 2, \ldots, 6$, then

$$\mu = E(X) = \sum_x xp(x) = \frac{1+2+\cdots+6}{6} = 3.5;$$

$$V(X) = \sum_x (x-\mu)^2 p(x) = \frac{1}{6}\sum_x (x-3.5)^2 = \frac{35}{12}.$$

Remark 3.16 *Recall that in a sample with outcomes $y_1, y_2, \ldots, y_n$, the sample mean is given by*

$$\bar{y} = \frac{y_1+y_2+\cdots+y_n}{n} = \frac{1}{n}\sum_{i=1}^n y_i = \sum_{i=1}^n y_i \frac{1}{n}.$$

The above formula can be theoretically viewed as the expected value of the sampling Y with outcomes $y_1, y_2, \ldots, y_n$ with equidisribution for large n. Similarly, one can see that $V(X)$ is also a good theoretical generalization of the sample variance.

Example 3.17 *In a gambling game a person draws a single card from an ordinary 52-card playing deck. A person is paid \$50 for drawing a jack, a queen or a king, and \$150 for drawing an ace. A person who draws any other card pays \$40. If a person plays this game, what is the expected gain?*

Let Y be the gain/loss of drawing cards, then it has pdf

$$p(50) = \frac{12}{52}, \;\; p(150) = \frac{4}{52}, \;\; p(-40) = \frac{36}{52}.$$

Then the expected gain is

$$E(Y) = \sum_y yp(y) = 50 \cdot \frac{12}{52} + 150 \cdot \frac{4}{52} + (-40) \cdot \frac{36}{52} = \frac{-240}{52} = -4.6153.$$

Theorem 3.18 (Properties of the expected value)

(1) $E(c) = c$ *for any constant* $c \in \mathbb{R}$ *(viewed as a constant random variable);*

(2) *Linearity: let* $Y_1, Y_2, \ldots, Y_k$ *be random variables over the same sample space* S*, and* $c_1, c_2, \ldots, c_k \in \mathbb{R}$*, then*

$$E(c_1Y_1 + c_2Y_2 + \cdots + c_kY_k) = c_1E(Y_1) + c_2E(Y_2) + \cdots + c_kE(Y_k).$$

(3) *The variance can be computed by*

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

Example 3.19 *Let* Y *be a random variable such that* $p(y) = \frac{y}{10}$*,* $y = 1, 2, 3, 4$*. Find* $E(Y)$*,* $E(Y^2)$*,* $E[Y(5-Y)]$ *and* $V(Y)$*.*

$$E(Y) = \sum_y yp(y) = 1 \cdot \frac{1}{10} + 2 \cdot \frac{2}{10} + 3 \cdot \frac{3}{10} + 4 \cdot \frac{4}{10} = 3;$$

$$E(Y^2) = \sum_y y^2p(y) = 1^2 \cdot \frac{1}{10} + 2^2 \cdot \frac{2}{10} + 3^2 \cdot \frac{3}{10} + 4^2 \cdot \frac{4}{10} = 10;$$

$$E[Y(5-Y)] = E(5Y - Y^2) = 5E(Y) - E(Y^2) = 5 \times 3 - 10 = 5;$$

$$V(Y) = E(Y^2) - [E(Y)]^2 = 10 - 3^2 = 1.$$

Remark 3.20 *If* Y *has expected value* μ *and variance* σ^2*, then we can renormalize* Y *to a new random variable* $Z = \dfrac{Y-\mu}{\sigma}$ *such that expected value 0 and variance 1. Indeed,*

$$E(Z) = E\left(\frac{Y-\mu}{\sigma}\right) = \frac{E(Y-\mu)}{\sigma} = \frac{E(Y)-\mu}{\sigma} = 0;$$

$$V(Z) = E[(Z - E(Z))^2] = E(Z^2) = E\left(\frac{(Y-\mu)^2}{\sigma^2}\right) = \frac{E[(Y-\mu)^2]}{\sigma^2} = 1.$$

Example 3.21 *The manager of a a stockroom in a factory has constructed the following probability distribution for the daily demand (number of times used) for a particular tool.*

y	0	1	2
$p(y)$	0.1	0.5	0.4

It costs the factory \$10 each time the tool is used. Find the mean and variance of the daily cost for use of the tool.

Let Y be the daily demand of the tool, and the daily cost is given by $10Y$. Then

$$E(10Y) = 10E(Y) = 10\sum_y yp(y) = 10(0 \cdot 0.1 + 1 \cdot 0.5 + 2 \cdot 0.4) = 13.$$

$$\begin{aligned} V(10Y) = E((10Y)^2) - [E(10Y)]^2 &= 100\sum_y y^2 p(y) - 13^2 \\ &= 100(0^2 \cdot 0.1 + 1^2 \cdot 0.5 + 2^2 \cdot 0.4) - 169 \\ &= 210 - 169 = 41. \end{aligned}$$

3.4 Well-known Discrete Probability Distributions

In practice, many experiments generate random variables with the same types of probability distribution. Note that a type of probability distribution may have some unknown constant(s) that determine its specific form, called parameters.

We are interested in the expected value, variance and standard deviation of these prototypes.

3.4.1 Uniform Distribution

Definition 3.22 *A random variable Y is said to have a discrete uniform distribution on $R(Y) = \{y_1, y_2, \ldots, y_n\}$ if $p(y_i) = \frac{1}{n}$ for each i.*

Note that the parameter here is the size n of $R(Y)$ - the set of outcomes for Y.

Example 3.23 *Let Y be the number observed in a fair dice toss, then Y has a discrete uniform distribution on $\{1,2,3,4,5,6\}$ with the pdf $p(y) = 1/6$ for any y.*

Then the expected value is

$$E(Y) = \sum_y yp(y) = \frac{1 + 2 + \cdots + 6}{6} = 3.5;$$

$$V(Y) = E(Y^2) - (E(Y))^2 = \frac{1^2 + 2^2 + \cdots + 6^2}{6} - 3.5^2 = \frac{35}{12}.$$

Example 3.24 *Let Y be with a discrete uniform distribution on*

$$\{1, 2, 3, \ldots, n\},$$

find $E(Y)$ and $V(Y)$.

Note that $p(y) = 1/n$, so

$$E(Y) = \frac{1 + 2 + \cdots + n}{n} = \frac{\frac{1}{2}n(n+1)}{n} = \frac{n+1}{2}.$$

$$E(Y^2) = \frac{1^2 + 2^2 + \cdots + n^2}{n} = \frac{\frac{1}{6}n(n+1)(2n+1)}{n} = \frac{(n+1)(2n+1)}{6},$$

and so

$$V(Y) = E(Y^2) - (E(Y))^2 = \frac{(n+1)(2n+1)}{6} - \left(\frac{n+1}{2}\right)^2 = \frac{n^2-1}{12}.$$

3.4.2 Binomial Distribution

A binomial experiment is described as follows:

(1) The experiment consists of n identical and independent trials.

(2) Each trial results in one of two outcomes: success (S) or failure (F);

(3) The probability of success in each single trial is equal to $p \in [0, 1]$, and that of failure is equal to $q = 1 - p$;

(4) The random variable under consideration is Y, the number of successes observed during the n trials.

The sample point in a binomial experiment with n trials is a string of length n with entries being either "S" or "F". And the random variable Y is the number of S's in a sample string.

Example 3.25 *Toss a coin 100 times. We call success if we toss head, and failure if tail. Assume that in a single toss, the chance to get a head is given by $2/3$ (and hence the chance for tail is $1 - 2/3 = 1/3$.)*

It is not hard to check this is a binomial experiment with 100 trials and $p = 2/3$.

Definition 3.26 (Binomial Distribution) *A random variable Y is said to have a binomial distribution based on n trials and success probability $p \in [0, 1]$ if and only if*

$$p(y) = \binom{n}{y} p^y (1-p)^{n-y}, \quad y = 0, 1, 2, \ldots, n.$$

We denote $Y \sim b(n, p)$.

Remark 3.27 *Note that $p(y)$ corresponds to the y-th term of the binomial expansion, that is,*

$$(p+q)^n = \sum_{y=0}^{n} \binom{n}{y} p^y q^{n-y},$$

where $q = 1-p$.

Theorem 3.28 *Let Y be the number of successes observed in a binomial experiment with n trials and single success probability p, then $Y \sim b(n,p)$.*

Proof $(Y = y)$ represents the set of sample points with y successes and $(n-y)$ failures, each of which is of probability $p^y(1-p)^{n-y}$. And there are $\binom{n}{y}$ sample points of this type, so

$$p(y) = P(Y = y) = \binom{n}{y} p^y (1-p)^{n-y}.$$

□

Example 3.29 *Suppose that a lot of 5000 electrical fuses contains 5% defectives. If a sample of 20 fuses is tested, find*

(a) the probability of observing at least one defective;

(b) the probability of observing exactly four defective;

(c) the probability of observing at least four defective;

(d) the probability of observing at least two but at most five defective.

We call a defective fuse to be a success, then we are dealing with a binomial experiment with 5 trials and success probability $p = 0.05$. Let Y the number of successes, that is, the number of defective fuses observed, then $Y \sim b(20, 0.05)$. Note that the outcomes of Y are $\{0,1,2,\ldots,20\}$.

(a) The event is to observe at least one defective, that is, $(Y \geq 1)$. We have that

$$P(Y \geq 1) = 1 - P(Y = 0) = 1 - p(0) = 1 - \binom{20}{0} 0.05^0 \cdot 0.95^{20} = 1 - 0.95^{20} = 0.6415.$$

(b) The event is to observe exactly 4 defective, that is, $(Y = 4)$. We have that

$$P(Y = 4) = p(4) = \binom{20}{4} 0.05^4 \cdot 0.95^{16} = 0.0133$$

Here I am using the software called R to compute binomial distribution. More precisely, you can input the command "dbinom(y,n,p)" to find $p(y)$. For instance, here I input dbinom$(4,20,.05)$.

(c) The event is to observe at least four defective, that is, $(Y \geq 4)$*. We have that*

$$P(Y \geq 4) = 1 - P(Y \leq 3) = 1 - F_Y(3) = 1 - .984 = 0.016.$$

Recall that here $F_Y(a) = P(Y \leq a)$ *is the cumulative distribution function (cdf) of* Y*. We use the R command "*$pbinom(a, n, p)$*" to find* $F_Y(a)$ *for a binomial random variable* Y*.*

(d) The event is to observe at least two but at most five defective, that is,

$$(2 \leq Y \leq 5) = (Y \leq 5) \backslash (Y \leq 1).$$

We have that

$$P(2 \leq Y \leq 5) = P(Y \leq 5) - P(Y \leq 1) = F_Y(5) - F_Y(1) = 0.9997 - 0.7358 = 0.2639.$$

Theorem 3.30 *Let* $Y \sim b(n, p)$*. Then*

$$\mu = E(Y) = np, \quad \sigma^2 = V(Y) = np(1-p).$$

Proof We postpone the proof till we introduce the moment-generating functions. One may prove this using elementary combinatoric arguments. □

Example 3.31 *An unfair coin has three-to-one odds for head versus tail. Toss this coin 10 times, find the expected value and variance of* Y*, the number of heads observed.*

Here $n = 10$ *and* $p = 3/4$*, then*

$$E(Y) = 10 \cdot 3/4 = 7.5, \quad V(Y) = 10 \cdot 3/4(1 - 3/4) = 1.875.$$

3.4.3 Geometric Distribution

A geometric experiment is described as follows:

(1) The experiment consists of identical and independent trials, but the number of trials is not given;

(2) Each trial results in one of two outcomes: success (S) or failure (F);

(3) The probability of success in each single trial is equal to $p \in [0, 1]$, and that of failure is equal to $q = 1 - p$;

(4) The random variable under consideration is Y, the number of trials on first the success occurs.

There are infinitely many sample points in this experiment:

$$S,\ FS,\ FFS,\ FFFS,\ FFFFS,\ \ldots$$

Example 3.32 *Toss a fair coin until a head is observed. Then this is a geometric experiment with success probability* $p = \frac{1}{2}$.

Definition 3.33 (Geometric Distribution) *A random variable* Y *is said to have a geometric distribution with single success probability* $p \in [0,1]$ *if and only if*

$$p(y) = p(1-p)^{y-1},\ \ y = 1,2,3,\ldots.$$

We denote $Y \sim Geo(p)$.

Clearly, Let Y the number of trials on which the first success occurs in a geometric experiment, we have $Y \sim Geo(p)$.

Theorem 3.34 *Let* $Y \sim Geo(p)$, *then*

$$\mu = E(Y) = \frac{1}{p}, \quad \sigma^2 = V(Y) = \frac{1-p}{p^2}.$$

Proof Postpone. □

Example 3.35 *Suppose that 30% of the applicants for a job have a master degree. Applicants are interviewed sequentially and are selected at random from the pool. Find*

(a) the probability that the first applicant with master is found on the fifth interview.

(b) the probability to meet the first applicant with master within 10 interviews.

(c) the expected number of interviews to meet the first applicant with master degree.

Let Y *be the number of interviews on which the first applicant with master degree is found. Then* $Y \sim Geo(0.3)$.

(a) The event is $(Y = 5)$, *then*

$$P(Y = 5) = p(5) = 0.3 \cdot (1-0.3)^{5-1} = 0.0720.$$

One can also use the R command "$dgeom(y-1, p)$*" to compute* $p(y)$. *For example, here we input* $dgeom(4, 0.3)$ *to calculate* $p(5)$.

(b) The event is $(Y \leq 10)$, *then*

$$P(Y \leq 10) = \sum_{y=1}^{10} p(y) = F_Y(10) = 0.9718.$$

One use the R command "pgeom$(y-1,p)$" to compute $F_Y(y)$. For example, here we input pgeom$(9,0.3)$ to calculate $F_Y(10)$.

(c) $E(Y)=\frac{1}{0.3}=3.3333$, so in average we are about to meet the first applicant with master in the 3rd or 4th interview.

Remark 3.36 *In fact, we can compute the cumulative distribution function as follows: If $Y\sim Geo(p)$, then for any positive integer a,*

$$P(Y>a)=\sum_{y>a}p(y)=p\sum_{y=a+1}(1-p)^{y-1}=p\frac{(1-p)^a}{1-(1-p)}=(1-p)^a,$$

and hence

$$F_Y(a)=P(Y\leq a)=1-P(Y>a)=1-(1-p)^a.$$

Also, we have the following memoryless property for a geometric distribution:

$$P(Y>a+b|Y>a)=\frac{P(Y>a+b)}{P(Y>a)}=\frac{(1-p)^{a+b}}{(1-p)^a}=(1-p)^b=P(Y>b).$$

It means that the failures in the first a trials would not increase or decrease the success rate in the next b trials.

3.4.4 Hypergeometric Distribution

A hypergeometric experiment is described as follows:

(1) In a population of N elements, there are two distinct types: r success (S) and $(N-r)$ failure (F);

(2) A sample of size n is randomly selected without replacement from this population;

(3) The random variable under consideration is Y, the number of successes in the sample.

Clearly Y in the hypergeometric experiment satisfies the following probability distribution:

Definition 3.37 (Hypergeometric Distribution) *A random variable Y is said to have a geometric distribution with population size N, success size r and sampling size n, or briefly, with parameters (N,r,n), if and only if*

$$p(y)=\frac{\binom{r}{y}\binom{N-r}{n-y}}{\binom{N}{n}},$$

for all integer y such that $0\leq y\leq r$ and $0\leq n-y\leq N-r$. We denote $Y\sim Hyp(N,r,n)$.

Example 3.38 *A bowl contains* 100 *chips, of which* 40 *are white, and the rest are green chips. Randomly select* 70 *chips from the bowl without replacement. Let* Y *be the number of white chips chosen.*

(a) Find the probability distribution function of Y*.*

(b) What is the probability to choose exactly 25 white chips.

(c) What is the probability to choose at most 30 white chips.

(a) $N = 100$*,* $r = 40$ *and* $n = 70$*. If* y *is an outcome of* Y*, we have* $0 \le y \le 40$ *and* $0 \le 70 - y \le 100 - 40$*, then* $10 \le y \le 40$*. Then*

$$p(y) = \frac{\binom{40}{y}\binom{60}{70-y}}{\binom{100}{70}}, \quad 10 \le y \le 40.$$

(b) The event is $(Y = 25)$*, so*

$$P(Y = 25) = p(25) = \frac{\binom{40}{25}\binom{60}{45}}{\binom{100}{70}} = 0.0728.$$

We can use the R command "$dhyper(y, r, N - r, n)$*" to compute* $p(y)$*, like here, we input* $dhyper(25, 40, 60, 70)$ *to compute* $p(25)$*.*

(c) The event is $(Y \le 30)$*, so*

$$P(Y \le 30) = \sum_{10 \le y \le 30} p(y) = F_Y(30) = 0.8676.$$

We use the R command "$phyper(a, r, N - r, n)$*" to compute* $F_Y(a)$*, like here, we input* $phyper(30, 40, 60, 70)$ *to compute* $F_Y(30)$*.*

Theorem 3.39 *Let* $Y \sim Hyp(N, r, n)$*, then*

$$E(Y) = \frac{rn}{N}, \quad V(Y) = n \cdot \frac{r}{N} \cdot \frac{N - r}{N} \cdot \frac{N - n}{N - 1}.$$

Proof Postpone. □

The hypergeometric distribution $Hyp(N, r, n)$ tends to a binomial distribution $b(n, p)$ with $p = r/N$ when $N \to \infty$, that is,

$$\lim_{N \to \infty} \frac{\binom{r}{y}\binom{N-r}{n-y}}{\binom{N}{n}} = \binom{n}{y} p^y (1 - p)^{n-y}.$$

A transparent explanation is as follows: When the population size N is very large compared to the sampling size n, the probability for picking a success sample would almost be $p = r/N$. Indeed, after m sample picking, the success rate may become

$$\frac{r-j}{N-j} = \frac{pN-j}{N-j} = \frac{p-j/N}{1-j/N} \approx p, \;\; j = 0, 1, \ldots, m, \;\; \text{when } N \to \infty.$$

In other words, when the population size is large, it does not matter to take replacement or not. One can also check that

$$E(Y) = \frac{rn}{N} = np, \quad V(Y) = np(1-p)\frac{1-n/N}{1-1/N} \to np(1-p).$$

Conversely, we might use a binomial distribution $b(n,p)$ to approximate a hypergeometric distribution $Hyp(N, pN, n)$ for large N and fixed n.

3.4.5 Poisson Distribution

The Poisson distribution is a good model for counting the number that events of a certain type occur in a given time period or in a fixed space, assuming that

- The probability of an event in an interval is proportional to the length of the interval.
- Two events cannot occur at the same time, and the occurrence of a first event would not affect the occurrence of a second event.

Example 3.40 *The number of clients arriving at a local bank on 9-11AM.*

Definition 3.41 (Poisson Distribution) *A random variable Y is said to have a Poisson distribution with a rate parameter $\lambda > 0$ if and only if*

$$p(y) = \frac{\lambda^y}{y!}e^{-\lambda}, \quad y = 0, 1, 2, 3, \ldots.$$

We denote $Y \sim Poisson(\lambda)$.

We first show that the definition of $p(y)$ is legit. Recall that the exponential function has the following Taylor expansion:

$$e^x = \sum_{k=0}^{\infty} \frac{x^k}{k!}, \quad \text{for all } x \in \mathbb{R},$$

and therefore,

$$\sum_y p(y) = \sum_{y=0}^{\infty} \frac{\lambda^y}{y!}e^{-\lambda} = e^{-\lambda}\sum_{y=0}^{\infty} \frac{\lambda^y}{y!} = e^{-\lambda}e^{\lambda} = 1.$$

Theorem 3.42 *Let $Y \sim Poisson(\lambda)$, then*

$$\mu = E(Y) = \lambda, \quad \sigma^2 = V(Y) = \lambda.$$

Proof Postpone. □

This result tells us the the parameter λ is in fact the expected value, or average of Y.

Example 3.43 *Suppose the average daily number of automobile accidents is 3, use Poisson model to find*

(a) the probability that there will be 5 accidents tomorrow.

(b) the probability that there will be less than 3 accidents tomorrow.

(a) Let Y be the number of accidents tomorrow, then Y can be modeled by a Poisson distribution with $\lambda = 3$, and hence

$$P(Y = 5) = p(5) = \frac{3^5}{5!}e^{-3} = 0.1008.$$

One can use the R command "dpois(y, λ)" to compute $p(y)$, like here, we input "dpois$(5, 3)$" to get $p(5)$.

(b) The event is $(Y \leq 2)$, then

$$P(Y \leq 2) = \sum_{y=0}^{2} p(y) = F_Y(2) = 0.4232.$$

One can use the R command "ppois(a, λ)" to compute $F_Y(a)$, like here, input "ppois$(2, 3)$" to get $F_Y(2)$.

Example 3.44 *The number of typing errorsmade by a typist has a poisson distribution with an average of four errors per page. If more than four errors appear on a given page, the typist must retype the whole page. What is the probability that a certain page does not have to be retyped?*

Let Y be the number of errors in a page, which can be modeled by a Poisson distribution with $\lambda = 4$. Then the probability that a certain page does not have to be retyped is

$$P(Y \leq 4) = \sum_{y=0}^{4} \frac{4^y}{y!}e^{-4} = F_Y(4) = 0.6288.$$

Here we use "ppois$(4, 4)$" to compute $F_Y(4)$.

Example 3.45 *If $Y \sim Poisson(\lambda)$, and $3P(Y=1) = P(Y=2)$, find $P(Y=3)$.*

Since $3p(1) = p(2)$, get

$$3\frac{\lambda^1}{1!}e^{-\lambda} = \frac{\lambda^2}{2!}e^{-\lambda} \implies \lambda = 6.$$

Then

$$P(Y=3) = \frac{6^3}{3!}e^{-6} = 36e^{-6}.$$

Example 3.46 *Given a Poisson random variable Y with standard deviation 2, and denote its mean (expected value) by μ, find $P(\mu - \sigma < Y < \mu + \sigma)$.*

Since $\lambda = \mu = \sigma^2 = 2^2 = 4$, we have $p(y) = \frac{4^y}{y!}e^{-4}$, $y = 1, 2, \ldots$. Then

$$P(\mu-\sigma < Y < \mu+\sigma) = P(2 < Y < 6) = p(3)+p(4)+p(5) = F_Y(5)-F_Y(2) = 0.5470.$$

Remark 3.47 *One can show that a binomial distribution $b(n,p)$ tends to a Poisson distribution $Poisson(\lambda)$ with $\lambda = np$, when $n \to \infty$, that is,*

$$\lim_{n\to\infty} \binom{n}{y} p^y (1-p)^{n-y} = \frac{\lambda^y}{y!}e^{-\lambda}.$$

It means that we can approximate a binomial distribution by a Poisson distribution when n is large, while $\lambda = np$ is fixed.

3.5 Moment-generating Functions

Let Y be a discrete random variable with probability distribution function $p(y)$.

Definition 3.48 *The k-th moment of Y, $k = 0, 1, 2, \ldots$, is defined to be*

$$\mu_k = E(Y^k) = \sum_y y^k p(y),$$

as long as the above summation converges.

Note that 0-th moment $\mu_0 = 1$. All well-known probability distributions that we discussed in Section 3.4.

Definition 3.49 *Assume that Y has finite k-th moments for all k. The moment-generating function is defined to be*

$$m(t) = E(e^{tY}), \quad t \in (-r, r),$$

for some $r > 0$ such that the above function $m(t)$ exists.

Recall that

$$e^x = \sum_{k=0}^{\infty} \frac{x^k}{k!}, \quad \text{for all } x \in \mathbb{R},$$

then the moment-generating function $m(t)$ has the following Taylor expansion:

$$m(t) = E(e^{tY}) = E\left(\sum_{k=0}^{\infty} \frac{t^k Y^k}{k!}\right) = \sum_{k=0}^{\infty} \frac{E(Y^k)}{k!} t^k = \sum_{k=0}^{\infty} \frac{\mu_k}{k!} t^k.$$

In particular, we have

$$\mu_k = E(Y^k) = m^{(k)}(0),$$

that is, the k-th moment of Y can be computed by taking the k-th derivative of $m(t)$ at $t = 0$. Now we are going to use this fact to compute the expectation and variance of previous well-known distributions.

Theorem 3.50

(1) *If $Y \sim b(n,p)$, then $E(Y) = np$ and $V(Y) = np(1-p)$;*

(2) *If $Y \sim Geo(p)$, then $E(Y) = \frac{1}{p}$ and $V(Y) = \dfrac{1-p}{p^2}$;*

(3) *IF $Y \sim Poisson(\lambda)$, then $E(Y) = \lambda$ and $V(Y) = \lambda$.*

Proof

(1) If $Y \sim b(n,p)$, then $p(y) = \binom{n}{y} p^y (1-p)^{n-y}$ for $y = 0, 1, 2, \ldots, n$, and the moment-generating function

$$m(t) = E(e^{tY}) = \sum_{y=0}^{n} e^{ty} p(y) = \sum_{y=0}^{n} \binom{n}{y} (e^t p)^y (1-p)^{n-y} = [e^t p + (1-p)]^n,$$

then

$$m'(t) = n[e^t p + (1-p)]^{n-1} e^t p, \implies E(Y) = m'(0) = np,$$

$$m''(t) = n(n-1)[e^t p + (1-p)]^{n-2} (e^t p)^2 + n[e^t p + (1-p)]^{n-1} e^t p.$$

$$\implies E(Y^2) = m''(0) = n(n-1)p^2 + np,$$

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

(2) If $Y \sim Geo(p)$, then $p(y) = p(1-p)^{y-1}$, $y = 1, 2, \ldots$, and the moment-generating function

$$\begin{aligned} m(t) = E(e^{tY}) = \sum_{y=1}^{\infty} e^{ty} p(y) &= \sum_{y=1}^{\infty} \frac{p}{1-p} [e^t(1-p)]^y \\ &= \frac{p}{1-p} \sum_{y=1}^{\infty} [e^t(1-p)]^y \\ &= \frac{p}{1-p} \frac{e^t(1-p)}{1-e^t(1-p)} \\ &= \frac{pe^t}{1-e^t(1-p)}. \end{aligned}$$

Then

$$m'(t) = \frac{pe^t[1-e^t(1-p)] - pe^t[-e^t(1-p)]}{[1-e^t(1-p)]^2} = \frac{pe^t}{[1-e^t(1-p)]^2},$$

$$m''(t) = \frac{pe^t[1-e^t(1-p)]^2 - pe^t 2[1-e^t(1-p)](-e^t)(1-p)}{[1-e^t(1-p)]^4} = \frac{pe^t + pe^{2t}(1-p)}{[1-e^t(1-p)]^3}.$$

Thus,

$$E(Y) = m'(0) = \frac{1}{p},\ E(Y^2) = m''(0) = \frac{2-p}{p^2},\ V(Y) = E(Y^2) - (E(Y))^2 = dfrac1 - pp^2.$$

(3) IF $Y \sim Poisson(\lambda)$, then $p(y) = \frac{\lambda^y}{y!} e^{-\lambda}$, $y = 0, 1, 2, \ldots$, then

$$m(t) = E(e^{tY}) = \sum_{y=0}^{\infty} e^{ty} p(y) = e^{-\lambda} \sum_{y=0}^{\infty} \frac{(e^t\lambda)^y}{y!} = e^{-\lambda} e^{e^t\lambda} = e^{\lambda(e^t-1)}.$$

Then

$$m'(t) = e^{\lambda(e^t-1)}\lambda e^t,\ \Longrightarrow\ E(Y) = m'(0) = \lambda,$$

$$m''(t) = e^{\lambda(e^t-1)}(\lambda e^t)^2 + e^{\lambda(e^t-1)}\lambda e^t = e^{\lambda(e^t-1)}[\lambda^2 e^{2t} + \lambda e^t],$$

$$\Longrightarrow E(Y^2) = m''(0) = \lambda^2 + \lambda,\ \ V(Y) = E(Y^2) - (E(Y))^2 = \lambda^2 + \lambda - \lambda^2 = \lambda.$$

□

Remark 3.51 *The moment-generating function $m(t)$, if exists, is uniquely correspondent to the probability distribution of Y. For example, if the moment-generating function of Y is given by $m(t) = e^{3(e^t-1)}$, then Y must be a Poisson random variable with $\lambda = 3$.*

3.6 Tchebysheff's Theorem

Given a random variable Y with mean/expectation μ and variance σ^2 (without knowing the probability distribution $p(y)$), we would like to estimate the probability that Y falls in the interval $(\mu - k\sigma, \mu + k\sigma)$ for any $k > 0$, that is,

$$P(\mu - k\sigma < Y < \mu + k\sigma) = P(|Y - \mu| < k\sigma) = ?$$

Fact: The Empirical Rule (68-95-99.7) tells us a Gaussian/Normal/Bell-curve distribution Y (we shall learn this later) satisfies that

- $k = 1$: $P(|Y - \mu| < \sigma) = 68\%$;
- $k = 2$: $P(|Y - \mu| < 2\sigma) = 95\%$;
- $k = 3$: $P(|Y - \mu| < 3\sigma) = 99.7\%$

In general, we can obtain a lower bound for $P(|Y - \mu| < k\sigma)$ by the following theorem.

Theorem 3.52 (Tchebysheff's Theorem) *Let Y be a random variable with mean μ and finite variance $\sigma^2 < \infty$. Then for any $k > 0$,*

$$P(|Y - \mu| < k\sigma) \geq 1 - \frac{1}{k^2}, \text{ or } P(|Y - \mu| \geq k\sigma) \leq \frac{1}{k^2}.$$

This theorem is true for both discrete and continuous random variables. Here we present the proof in the discrete case.

Proof Let $p(y)$ be the pdf of Y, then

$$\begin{aligned}
\sigma^2 = V(Y) = E(Y - \mu)^2 &= \sum_y (y - \mu)^2 p(y) \\
&\geq \sum_{y:|y-\mu|\geq k\sigma} (y - \mu)^2 p(y) \\
&\geq \sum_{y:|y-\mu|\geq k\sigma} (k\sigma)^2 p(y) \\
&= k^2\sigma^2 \sum_{y:|y-\mu|\geq k\sigma} p(y) \\
&= k^2\sigma^2 P(|Y - \mu| \geq k\sigma).
\end{aligned}$$

□

Example 3.53 *The number of customers per day at a sales counter, Y, has been observed for a long period of time and found to have mean 20 and standard deviation 2. The probability distribution of Y is not known. What can be said about the probability that, tomorrow, Y will be greater than 16 but less than 24?*

Note that $\mu = 20$ and $\sigma = 2$, then by Tchebysheff's theorem,

$$P(16 < Y < 24) = P(\mu - 2\sigma < Y < \mu + 2\sigma) \geq 1 - \frac{1}{2^2} = 0.75.$$

4 Continuous Random Variables

NOT all random variables of interest are discrete.

Example 4.1 *Let Y be the daily rainfall in London. The amount of rainfall could take on any value between 0 and 5 inches, that is, Y can take any value in the interval (0, 5).*

Example 4.2 *Let Y be the lifespan of a light bulb (measured in hours). Theoretically, Y can take any value in $(0, \infty)$.*

Definition 4.3 *A random variable is said to be continuous if it can take any value in an interval.*

Remark 4.4 *For a continuous random variable Y, its range $R(Y)$ - the set of all possible outcomes - is an interval and thus uncountably infinite. For most typical $y \in R(Y)$, the probability distribution at y given by $P(Y = y)$ is zero, which becomes useless.*

4.1 The Characteristics of a Continuous Random Variable

Definition 4.5 *Let Y be a random variable. The cumulative distribution function (cdf) of Y , denoted by $F_Y(y)$, is such that $F_Y(y) = P(Y \leq y)$ for $-\infty < y < \infty$.*

If Y is the only random variable under consideration, we simply denote $F(y)$.

Example 4.6 *Let $Y \sim b(2, 0.5)$, find $F(y)$.*

Recall that Y is discrete with pdf

$$p(y) = \binom{2}{y} 0.5^y 0.5^{2-y} = 0.25\binom{2}{y} = \begin{cases} 0.25, & y = 0, \\ 0.5, & y = 1, \\ 0.25, & y = 2. \end{cases}$$

Then

$$F(y) = \sum_{z \le y} p(z) = \begin{cases} 0, & y < 0, \\ 0,25, & 0 \le y < 1, \\ 0.75, & 1 \le y < 2, \\ 1, & y \ge 2. \end{cases}$$

In this example, $F(y)$ is a piecewise constant function.

Moreover, $F(y)$ is nondecreasing with $\lim_{y \to -\infty} F(y) = 0$ and $\lim_{y \to \infty} F(y) = 1$, which is a general fact.

Theorem 4.7 *Let $F(y)$ be the cdf of a random variable Y.*

(1) $F(-\infty) := \lim_{y \to -\infty} F(y) = 0$, and $F(\infty) := \lim_{y \to \infty} F(y) = 1$;

(2) $F(y)$ is a nondecreasing function of y, that is, if $y_1 < y_2$, then $F(y_1) \le F(y_2)$.

Remark 4.8 *Classical Calculus tells us that the discontinuity point of $F(y)$, if any, should be of jump type.*

We can now use the cumulative distribution $F(y)$ to describe a continuous random variable.

Definition 4.9 *A random variable Y is continuous if and only if its cdf $F(y)$ is continuous for all $-\infty < y < \infty$.*

Remark 4.10 *We can now characterize random variables as follows:*

- *Y is discrete $\Longleftrightarrow$ $F(y)$ is piecewise constant with jump discontinuities. Further, at each jump point y_0, $P(Y = y_0) > 0$;*
- *Y is continuous $\Longleftrightarrow$ $F(y)$ is continuous. Further, $P(Y = y) = 0$ for all y;*
- *Y is a mixture of discrete and continuous parts $\Longleftrightarrow$ $F(y)$ is not piecewise constant but has jump discontinuities.*

Definition 4.11 *Let $F(y)$ be the cdf of a continuous random variable Y. Then*

$$f(y) := \frac{dF(y)}{dy} = F'(y),$$

whenever the derivative exists, is called the probability density function (pdf) for Y.

Remark 4.12

- *Classical Calculus shows that $F'(y)$ actually exists for "almost every" y.*
- *By fundamental theorem of calculus and that $F(-\infty) = 0$,*

$$F(y) = \int_{-\infty}^{y} f(t)\, dt.$$

- *We use "pdf" for short to denote both the probability distribution function for a discrete RV and the probability density function for a continuous RV. We shall see that they kind of serves the same job in calculating probabilities.*

Example 4.13 *Let Y be a continuous RV with cdf*

$$F(y) = \begin{cases} 0, & y < 0, \\ y, & 0 \le y \le 1, \\ 1, & y > 1. \end{cases}$$

Find its pdf $f(y)$.

To compute $f(y) = F'(y)$, we need consider the subintervals separately:

- *if $y < 0$, then $f(y) = F'(y) = 0' = 0$;*
- *if $0 < y < 1$, then $f(y) = F'(y) = y' = 1$;*
- *if $y > 1$, then $f(y) = F'(y) = 1' = 0$;*
- *at $y = 0$, LHS derivative is 0 while RHS derivative is 1, so $f(0)$ is undefined; Similar reason at $y = 1$.*

To sum up,

$$f(y) = \begin{cases} 0, & y < 0 \text{ or } y > 1 \\ 1, & 0 < y < 1 \\ \text{undefined}, & y = 0, \text{ or } 1. \end{cases}$$

Example 4.14 *Let Y be a continuous RV with pdf*

$$f(y) = \begin{cases} 3y^2, & 0 \le y \le 1, \\ 0, & \text{otherwise}. \end{cases}$$

Find its cdf $F(y)$.

The cdf is given by the formula $F(y) = \int_{-\infty}^{y} f(z)dz$, but again, we need to consider different intervals for integrals.

- *if $y < 0$, then $F(y) = \int_{-\infty}^{y} 0dz = 0$;*

- *if* $0 \le y \le 1$, *then* $F(y) = \int_{-\infty}^{0} 0dz + \int_{0}^{y} 3z^2 dz = 0 + z^3|_0^y = y^3$;
- *if* $y > 1$, *then* $F(y) = \int_{-\infty}^{0} 0dz + \int_{0}^{1} 3z^2 dz + \int_{1}^{y} 0dz = 0 + z^3|_0^1 + 0 = 1$.

To sum up,

$$F(y) = \begin{cases} 0, & y < 0 \\ y^3, & 0 \le y \le 1 \\ 1, & y > 1. \end{cases}$$

Theorem 4.15 *Let* $f(y)$ *be a pdf of a continuous random variable, then*

(1) $f(y) \ge 0$ *for all* $-\infty < y < \infty$;

(2) *Normalization:* $\int_{-\infty}^{\infty} f(y)dy = 1$;

(3) $P(Y \in B) = \int_B f(y)dy$, *for any* $B \subset \mathbb{R}$. *In particular,*

$$P(a \le Y \le b) = \int_a^b f(y)dy.$$

Example 4.16 *Given*

$$f(y) = \begin{cases} cy^2, & 0 \le y \le 2, \\ 0, & otherwise. \end{cases}$$

(1) *Find the value of* c *such that* $f(y)$ *is probability density function of a continuous random variable.*

(2) *Find* $P(Y = 1)$ *and* $P(1 < Y < 2)$.

Use the normalization condition

$$1 = \int_{-\infty}^{\infty} f(y)dy = \int_0^2 cy^2 dy = \frac{cy^3}{3}|_0^2 = \frac{8c}{3},$$

so $c = \frac{3}{8}$.

$P(Y = 1) = 0$ *since* Y *is continuous.*

$$P(1 < Y < 2) = \int_1^2 f(y)dy = \int_1^2 \frac{3}{8}y^2 dy = \frac{y^3}{8}|_1^2 = \frac{2^3 - 1^3}{8} = \frac{7}{8}.$$

Definition 4.17 *Given a continuous random variable* Y *with pdf* $f(y)$, *its expected value is*

$$E(Y) = \int_{-\infty}^{\infty} yf(y)dy,$$

provided that the integral exists.

Similar to the discrete case, we have

Theorem 4.18

(1) *Expected value of a function of Y: let $g : \mathbb{R} \to \mathbb{R}$, then*

$$E(g(Y)) = \int_{-\infty}^{\infty} g(y)f(y)dy.$$

(2) *$E(c) = c$, where c is a constant viewed a constant random variable;*

(3) *Linearity: $E(c_1Y_1 + \cdots + c_kY_k) = c_1E(Y_1) + \cdots + c_kE(Y_k)$, where $Y_1, \ldots, Y_k$ are continuous random variables, and $c_1, \ldots, c_k$ are real numbers.*

(4) *Let $\mu = E(Y)$, then the variance of Y is*

$$V(Y) = E[(Y - E(Y)]^2 = \int_{-\infty}^{\infty} (y - \mu)^2 f(y)dy$$

or

$$V(Y) = E(Y^2) - [E(Y)]^2 = \int_{-\infty}^{\infty} y^2 f(y)dy - \mu^2.$$

Example 4.19 *Given a continuous random variable Y with pdf*

$$f(y) = \begin{cases} \frac{3}{8}y^2, & 0 \le y \le 2, \\ 0, & \text{otherwise}. \end{cases}$$

Find $E(Y)$ and $V(Y)$.

$$E(Y) = \int_{-\infty}^{\infty} yf(y)dy = \int_0^2 y\frac{3}{8}y^2 dy = \frac{3y^4}{32}|_0^2 = 1.5$$

$$V(Y) = E(Y^2) - E(Y)^2 = \int_0^2 y^2\frac{3}{8}y^2 dy - 1.5^2 = \frac{3y^5}{40}|_0^2 - 2.25 = 0.15$$

Definition 4.20 *Given a random variable Y and $0 < p < 1$, the smallest value ϕ such that*

$$F(\phi) = P(Y \le \phi) \ge p$$

is called the p-th quantile of Y, denoted by ϕ_p.

In the case when Y is continuous, ϕ_p is the smallest solution of the equation $F(\phi) = p$.

The $\frac{1}{2}$-th quantile $\phi_{.5}$ is also called the median of Y.

In old days without computers, the quantile table, for random variables of certain type, is a standard tool to estimate probabilities given known cdf or pdf.

The following lemma is about how to switch the cdf/pdf between random variables Y and Z with the relation that $Y = h(Z)$.

Lemma 4.21 (Change of Variables for Random Variables) *Let Y and Z be continuous random variables such that $Y = h(Z)$, where $h : \mathbb{R} \to \mathbb{R}$ be a strictly increasing differentiable function. Then*

$$F_Z(z) = F_Y(h(z)), \quad -\infty < z < \infty$$

and hence

$$f_Z(z) = f_Y(h(z))h'(z), \quad -\infty < z < \infty.$$

Proof Since h is one-to-one, let us denote its inverse by h^{-1}. We can write $Z = h^{-1}(Y)$.

$$F_Z(z) = P(Z \le z) = P(h^{-1}(Y) \le z) = P(Y \le h(z)) = F_Y(h(z)).$$

Hence

$$f_Z(z) = F_Z'(z) = [F_Y(h(z))]' = F_Y'(h(z))h'(z) = f_Y(h(z))h'(z).$$

□

4.2 Well-known Continuous Probability Distributions

4.2.1 Uniform Distribution on an Interval

Definition 4.22 *Given $-\infty < \theta_1 < \theta_2 < \infty$, a random variable Y is said to have a continuous uniform probability distribution on the interval $[\theta_1, \theta_2]$ if its probability density function is*

$$f(y) = \begin{cases} \dfrac{1}{\theta_2 - \theta_1}, & \theta_1 \le y \le \theta_2, \\ \\ 0, & \textit{elsewhere}. \end{cases}$$

We denote $Y \sim U(\theta_1, \theta_2)$.

Example 4.23 *Suppose that a bus always arrives at a particular stop between 8:00 and 8:10 am, and that the probability that the bus will arrive in any given subinterval of time is proportional only to the length of the subinterval. For instance, the bus is as likely to arrive between 8:00 and 8:02 as it is to arrive between 8:06 and 8:08.*

We denote the 8:00 by $\theta_1 = 0$ *and 8:10 by* $\theta_2 = 10$*, and let* Y *be the time moment that the bus arrives. Then* $Y \sim U(0, 10)$*.*

Theorem 4.24 *Let* $Y \sim U(\theta_1, \theta_2)$*, then*

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

Proof

$$\mu = E(Y) = \int_{-\infty}^{\infty} yf(y)dy = \int_{\theta_1}^{\theta_2} \frac{ydy}{\theta_2 - \theta_1} = \frac{y^2}{2(\theta_2 - \theta_1)}|_{\theta_1}^{\theta_2} = \frac{\theta_2^2 - \theta_1^2}{2(\theta_2 - \theta_1)} = \frac{\theta_1 + \theta_2}{2}$$

$$\begin{aligned} V(Y) = E(Y^2) - (E(Y))^2 = \int_{\theta_1}^{\theta_2} \frac{y^2 dy}{\theta_2 - \theta_1} - \left(\frac{\theta_1 + \theta_2}{2}\right)^2 &= \frac{\theta_2^3 - \theta_1^3}{3(\theta_2 - \theta_1)} - \frac{\theta_1^2 + 2\theta_1\theta_2 + \theta_2^2}{4} \\ &= \frac{(\theta_2 - \theta_1)^2}{12} \end{aligned}$$

□

In the above bus waiting problem, the expected time moment that the bus arrives is $E(Y) = \frac{0+10}{2} = 5$, that is, 8:05am. And the variance is $\frac{(10-0)^2}{12} = \frac{25}{3}$.

Example 4.25 *The cycle time for trucks hauling concrete to a highway construction site is uniformly distributed over the interval 50 to 70 minutes. What is the probability that the cycle time exceeds 65 minutes if it is known that the cycle time exceeds 55 minutes?*

Let Y *be the cycle time, then* $Y \sim U(50, 70)$*, and so* $f(y) = \frac{1}{20}$ *for* $50 \leq y \leq 70$*, and* $f(y) = 0$ *elsewhere.*

$$P(Y > 65 | Y > 55) = \frac{P(Y > 65)}{P(Y > 55)} = \frac{\int_{65}^{70} \frac{1}{20} dy}{\int_{55}^{70} \frac{1}{20} dy} = \frac{70 - 65}{70 - 55} = \frac{1}{3}.$$

Example 4.26 *The failure of a circuit board interrupts work that utilizes a computing system until a new board is delivered. The delivery time, Y , is uniformly distributed on the interval one to five days. The total cost of a board failure and interruption is $C = 10 + 3Y^2$. Find the expected cost.*

$Y \sim U(1,5)$, *then* $E(Y) = \frac{1+5}{2} = 3$ *and* $V(Y) = \frac{(5-1)^2}{12} = \frac{4}{3}$,

$$E(C) = E(10+3Y^2) = 10+3E(Y^2) = 10+3[V(Y)+(E(Y))^2] = 10+3(\frac{4}{3}+3^2) = 41.$$

4.2.2 Gaussian/Normal Distribution

The most widely used continuous probability distribution is the Gaussian distribution or normal distribution, whose density function has the familiar "bell" shape.

Definition 4.27 *A random variable Y is said to have a normal probability distribution if for $\sigma > 0$ and $-\infty < \mu < \infty$, the probability density function of Y is*

$$f(y) = f_Y(y) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(y-\mu)^2}{2\sigma^2}}, \quad -\infty < y < \infty. \tag{4–1}$$

Denote $Y \sim N(\mu, \sigma^2)$. We shall see that $E(Y) = \mu$ and $V(Y) = \sigma^2$.

In particular, a random variable $Z \sim N(0,1)$ is called standard normal distribution, and its pdf is given by

$$f(z) = f_Z(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}, \quad -\infty < z < \infty. \tag{4–2}$$

Example 4.28 *The final exam scores in a class are expected to have mean $\mu = 75$ and standard variance $\sigma = 10$. That is, let Y be the score of a randomly chosen student, then $Y \sim N(75, 10^2)$.*

Remark 4.29 *On the on hand, despite the usefulness, it is very unfortunate that we cannot compute the cdf*

$$F(y) = P(Y \le y) = \int_{-\infty}^{y} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(t-\mu)^2}{2\sigma^2}} dt$$

by any accurate formula. A traditional method before computer was invented is to use the quantile table for normal distributions. For our course, we use the R command "pnorm(y, μ, σ)" to obtain $F(y)$.

On the other hand, if $Y \sim N(\mu, \sigma^2)$, then the density function $f(y)$ is symmetrically distributed around μ, which allows us to do simplification. For instance, for any $a \geq 0$,

$$F(\mu - a) = P(Y \leq \mu - a) = P(Y \geq \mu + a) = 1 - F(\mu + a),$$

in particular, $F(\mu) = P(Y \geq \mu) = P(Y \leq \mu) = 0.5$. Also,

$$F(\mu + a) - F(\mu - a) = P(\mu - a \leq Y \leq \mu + a) = 2P(\mu \leq Y \leq \mu + a).$$

Example 4.30 *Let $Z \sim N(0, 1)$, find*

(1) $P(Z > 2) = 1 - P(Z \leq 2) = 1 - F(2) = 1 - pnorm(2, 0, 1) = 1 - 0.9772 = 0.0228$

(2) $P(-2 \leq Z \leq 2) = P(Y \leq 2) - P(Y < -2) = 2P(Y \leq 2) - 1 = 2F(2) - 1 = 0.9545$

(3) $P(0 \leq Z \leq 1.73) = P(Y \leq 1.73) - P(Y < 0) = F(1.73) - 0.5 = pnorm(1.73, 0, 1) - 0.5 = 0.4582$

Example 4.31 *The GPAs of a large population of college students are approximately normally distributed with mean 2.4 and standard deviation 0.8.*

(1) *What fraction of the students will possesses a GPA in excess of 3.0?*

(2) *If students with GPA less than 1.5 will drop college, what percentage of students will drop?*

(3) *Suppose that three students are randomly selected. What is the probability that all three will possess GPA in excess of 3.0?*

Let Y be the GPA of a randomly chosen student, then $Y \sim N(2.4, 0.8^2)$.

(1)

$$P(Y > 3.0) = 1 - P(Y \leq 3.0) = 1 - F(3.0) = 1 - pnorm(3.0, 2.4, 0.8) = 0.2266.$$

(2)

$$P(Y < 1.5) = F(1.5) = pnorm(1.5, 2.4, 0.8) = 0.1303.$$

(3) *Let Y_1, Y_2, Y_3 be the GPAs of these three randomly chosen students. Then the events $(Y_1 > 3.0)$, $(Y_2 > 3.0)$, $(Y > 3.0)$ are mutually independent, and by (1), they are of the same probability 0.2266. Thus,*

$$P((Y_1 > 3.0) \cap (Y_2 > 3.0) \cap (Y_3 > 3.0)) = P(Y_1 > 3.0)P(Y_2 > 3.0)P(Y_3 > 3.0) = 0.2266^3.$$

The following theorem asserts that we can renormalize a normal distribution $Y \sim N(\mu, \sigma^2)$ to a standard distribution $Z = \dfrac{Y-\mu}{\sigma} \sim N(0,1)$, which might greatly simplify our computation.

Theorem 4.32 $Y \sim N(\mu, \sigma^2) \iff \dfrac{Y-\mu}{\sigma} \sim N(0,1)$.

Proof Let $Z = \dfrac{Y-\mu}{\sigma}$, or equivalently, $Y = \mu + \sigma Z = h(Z)$, where $h : \mathbb{R} \to \mathbb{R}$ is the affine function $h(z) = \mu + \sigma z$. By Lemma 4.21, we have

$$f_Z(z) = f_Y(h(z))h'(z) = \sigma f_Y(\mu + \sigma z).$$

Comparing (4–1) and (4–2), get $Y \sim N(\mu, \sigma^2) \iff \dfrac{Y-\mu}{\sigma} \sim N(0,1)$.

□

Example 4.33 (68-95-99.7 Rule) *Given $Y \sim N(\mu, \sigma^2)$, we would like to estimate*

$$P(\mu - k\sigma < Y < \mu + k\sigma), \quad k = 1, 2, 3, \ldots$$

To do so, set $Z = \dfrac{Y-\mu}{\sigma} \sim N(0,1)$, and then

$$P_k = P(\mu - k\sigma < Y < \mu + k\sigma) = P(-k < Z < k).$$

- $k = 1$: $P_1 = P(-1 < Z < 1) = pnorm(1,0,1) - pnorm(-1,0,1) \cong 68\%$
- $k = 2$: $P_2 = P(-2 < Z < 2) = pnorm(2,0,1) - pnorm(-2,0,1) \cong 95\%$
- $k = 3$: $P_3 = P(-3 < Z < 3) = pnorm(3,0,1) - pnorm(-3,0,1) \cong 99.7\%$

Theorem 4.34 *Let $Y \sim N(\mu, \sigma^2)$, then $E(Y) = \mu$ and $V(Y) = \sigma^2$.*

Proof We first show the special case: if $Z \sim N(0,1)$, then $E(Z) = 0$ and $V(Z) = 1$. Using the fact that $\int_{-\infty}^{\infty} e^{-\frac{x^2}{2}} dx = \sqrt{2\pi}$, we can show that the moment-generating function $m(t) = E(e^{tZ}) = e^{\frac{t^2}{2}}$. Indeed,

$$\begin{aligned} m(t) = E(e^{tZ}) = \int_{-\infty}^{\infty} e^{tz} f_Z(z) dz &= \int_{-\infty}^{\infty} e^{tz} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{\frac{t^2}{2} - \frac{1}{2}(z-t)^2} dz \\ &\overset{x=z-t}{=} e^{\frac{t^2}{2}} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{x^2}{2}} dx = e^{\frac{t^2}{2}} \end{aligned}$$

Therefore, $m'(t) = te^{\frac{t^2}{2}}$ and $m''(t) = (1+t^2)e^{\frac{t^2}{2}}$, so

$$E(Z) = m'(0) = 0, \quad V(Z) = E(Z^2) - (E(Z))^2 = m''(0) - (m'(0))^2 = 1 - 0^2 = 1.$$

In the general case when $Y \sim N(\mu, \sigma^2)$, by Theorem 4.32, we have that $Z = \frac{Y-\mu}{\sigma} \sim N(0,1)$. Therefore,

$$E(Y) = E(\mu + \sigma Z) = \mu + \sigma E(Z) = \mu$$

and

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

□

4.2.3 Gamma Distribution

The Gamma probability distribution is widely used in engineering, science, and business, to model continuous variables that are always positive and have skewed (non-symmetric) distributions, for instance, the lengths of time between malfunctions for aircraft engines, the lengths of time between arrivals at a supermarket checkout queue.

Definition 4.35 *A random variable Y is said to have a gamma distribution with parameters $\alpha > 0$ and $\beta > 0$ if the density function of Y is*

$$f(y) = \begin{cases} \dfrac{y^{\alpha-1}e^{-y/\beta}}{\beta^\alpha \Gamma(\alpha)}, & 0 \le y < \infty \\ 0, & \textit{elsewhere}, \end{cases}$$

where $\Gamma(\alpha) = \int_0^\infty y^{\alpha-1}e^{-y}dy$ is called the gamma function. We denote $Y \sim$ gamma(α, β).

α is called the shape parameter, determining the shape or concavity of $f(y)$, while β is called the scaling parameter.

Here is some properties of the gamma function:

- $\Gamma(n) = (n-1)!$, in particular, $\Gamma(1) = 0! = 1$.
- $\Gamma(\alpha) = (\alpha - 1)\Gamma(\alpha - 1)$ for all $\alpha > 1$.

Theorem 4.36 *Let $Y \sim$ gamma(α, β), then*

$$E(Y) = \alpha\beta, \quad V(Y) = \alpha\beta^2.$$

Proof Show that the moment-generating function of $Y \sim gamma(\alpha, \beta)$ is given by

$$m(t) = E(e^{tY}) = (1 - \beta t)^{-\alpha}, \;\; t < \frac{1}{\beta},$$

then

$$m'(t) = \alpha\beta(1 - \beta t)^{-\alpha-1}, \;\; m''(t) = \alpha(\alpha + 1)\beta^2(1 - \beta t)^{-\alpha-2},$$

so

$$E(Y) = m'(0) = \alpha\beta, \;\; V(Y) = E(Y^2) - (E(Y))^2 = m''(0) - m'(0)^2 = \alpha\beta^2.$$

□

Again, one can use software R to compute $F(y) = P(Y \leq y)$ for a gamma random variable Y by the command "$pgamma(y, \alpha, 1/\beta)$".

Example 4.37 *Four-week summer rainfall totals in a section of the Midwest United States have approximately a gamma distribution with $\alpha = 1.6$ and $\beta = 2.0$.*

(1) *Find the mean and variance of the four-week rainfall totals.*

(2) *What is the probability that the four-week rainfall total exceeds 4 inches?*

$Y \sim gamma(1.6, 2.0)$, then

$$E(Y) = 1.6 \times 2.0 = 3.2, \;\; V(Y) = 1.6 \times 2.0^2 = 6.4.$$

and

$$P(Y > 4) = 1 - P(Y \leq 4) = 1 - F(4) = 1 - pgamma(4, 1.6, 0.5) = 0.2896$$

There are two special cases of gamma random variables: One is the chi-square distribution, the other is the exponential distribution.

Definition 4.38 *A random variable Y is said to have a chi-square distribution with ν degrees of freedom if Y is a gamma-distributed random variable with parameters $\alpha = \nu/2$ and $\beta = 2$. Consequently, the pdf of a chi-square distribution is*

$$f(y) = \begin{cases} \dfrac{y^{\nu/2-1}e^{-y/2}}{2^{\nu/2}\Gamma(\nu/2)}, & 0 \leq y < \infty \\ 0, & elsewhere. \end{cases}$$

We denote $Y \sim \chi^2(\nu)$. Note that

$$E(Y) = \nu, \;\; V(Y) = 2\nu.$$

Use R command "$pchisq(y, \nu)$" to compute $F(y)$ for a χ^2 distribution.

Definition 4.39 *A random variable Y is said to have an exponential distribution with parameter $\beta > 0$ if Y is a gamma-distributed random variable with parameters $\alpha = 1$ and $\beta > 0$. Consequently, the pdf of an exponential distribution is*

$$f(y) = \begin{cases} \frac{1}{\beta} e^{-y/\beta}, & 0 \le y < \infty \\ 0, & \text{elsewhere.} \end{cases}$$

We denote $Y \sim exp(\beta)$. Note that the cumulative distribution is

$$F(y) = \int_{-\infty}^{y} f(y)dy = \begin{cases} 1 - e^{-y/\beta}, & 0 \le y < \infty, \\ 0, & \text{elsewhere.} \end{cases}$$

and

$$E(Y) = \beta, \;\; V(Y) = \beta^2.$$

We can use R command "$pexp(y, 1/\beta)$" to compute $F(y)$ for an exponential distribution.

Remark 4.40 *Note that if $Y \sim exp(\beta)$, then for any $y \ge 0$,*

$$P(Y > y) = 1 - F(y) = e^{-y/\beta}.$$

Consequently, the exponential distribution Y satisfies the memoryless property, that is,

$$P(Y > a + b | Y > a) = P(Y > b), \;\; \text{for any } a, b > 0,$$

which is useful for modeling the lifetime of electronic components.

Example 4.41 *Let Y be an exponential distribution with mean 40. Find*

- *the probability density function: mean=$E(Y) = \beta = 40$, so*

  $$f(y) = \frac{1}{40} e^{-y/40}, \;\; y \ge 0,$$

  and $f(y) = 0$ for $y < 0$.
- $P(Y < 36) = F(36) = 1 - e^{-36/40} = 0.5934$
- $P(Y > 40 | Y > 4) = P(Y > 36) = 1 - P(Y < 36) = 0.4066$.

Example 4.42 *A light bulb that will burn out after t hours according to an exponential distribution. The average lifetime of a light bulb is 100 hours. What is the probability that the light bulb will not burn out before 20 hours.*

Let Y be the lifetime of the bulb, in other words, Y is the time moment that it burns out. Then $\beta = E(Y) = 100$, and so Y has density

$$f(t) = \frac{1}{100} e^{-t/100}, \quad t > 0.$$

Then

$$P(Y > 20) = \int_{20}^{\infty} \frac{1}{100} e^{-t/100} dt = e^{-0.2} = 0.8187.$$

Example 4.43 *The length of time Y necessary to complete a key operation in the construction of houses has an exponential distribution with mean 10 hours. The formula $C = 100 + 4Y + 3Y^2$ relates the cost C of completing this operation. Find the expected cost.*

Since $Y \sim exp(10)$, $E(Y) = 10$ and $V(Y) = 10^2 = 100$, and

$$\begin{aligned} E(C) = 100 + 4E(Y) + 3E(Y^2) &= 100 + 4 \times 10 + 3[V(Y) + (E(Y))^2] \\ &= 140 + 3[100 + 10^2] = 740 \end{aligned}$$

4.3 Tchebysheff's Theorem (Revisited)

Theorem 4.44 (Tchebysheff's Theorem) *Let Y be a random variable with mean μ and finite variance $\sigma^2 < \infty$. Then for any $k > 0$,*

$$P(|Y - \mu| < k\sigma) \geq 1 - \frac{1}{k^2}, \quad \text{or} \quad P(|Y - \mu| \geq k\sigma) \leq \frac{1}{k^2}.$$

The proof in continuous case is almost the same as in discrete case, just replacing $\sum$ by $\int$, and probability distribution $p(y)$ by density $f(y)$.

Example 4.45 *A machine used to fill cereal boxes disperse, on the average, μ ounces per box. The manufacturer wants the actual ounces dispensed Y to be within 1 ounce of μ at least 75% of the time. What is the largest value of σ, the standard deviation of Y, that can be toilerated if the manufacturer's objectives are to be met?*

By Tchebysheff's theorem,

$$P(|Y - \mu| < 1) = P(|Y - \mu| < k\sigma) \geq 1 - \frac{1}{k^2} \geq 75\%$$

so $k \geq 2$, and so $\sigma = \frac{1}{k} \leq \frac{1}{2}$.

5 Multivariate Probability Distributions

Suppose that $Y_1, Y_2, \ldots, Y_n$ denote the outcomes of n successive trials of an experiment. A specific set of outcomes, or sample measurements, may be expressed in terms of the intersection of n events

$$(Y_1 = y_1),\ (Y_2 = y_2),\ \ldots\ (Y_n = y_n)$$

which we will denote as

$$(Y_1 = y_1, Y_2 = y_2, \ldots, Y_n = y_n)$$

or more compactly, as (y_1, y_2, y_n). The random variables $Y_1, Y_2, \ldots, Y_n$ might be related in some way, or might be totally irrelevant.

Calculation of the probability of this intersection is essential in making inferences about the population from which the sample was drawn and is a major reason for studying multivariate probability distributions.

Example 5.1 *In a physical exam of a sing person, let Y_1 be the age, Y_2 the height, Y_3 the weight, Y_4 the blood type and Y_5 the pulse rate. The doctor will be interested in the personal data $(Y_1, Y_2, Y_3, Y_4, Y_5)$. A sample measurement would be something like*

$$(y_1, y_2, y_3, y_4, y_5) = (30yrs, 6ft, 140lbs, type\ O, 65/min)$$

5.1 Bivariate Probability Distributions

Many random variables can be defined over the same sample space.

Example 5.2 *Tossing a pair of dice. The sample space contains 36 sample points:*

$$(i,j):\ \ i = 1, 2, \ldots, 6,\ j = 1, 2, \ldots, 6.$$

Let Y_1 be the number on dice 1, and Y_2 the number on dice 2. The probability of $(Y_1 = y_1, Y_2 = y_2)$ for any possible values of y_1 and y_2 is then given by

$$p(y_1, y_2) = P(Y_1 = y_1, Y_2 = y_2) = P(Y_1 = y_1)P(Y_2 = y_2) = \frac{1}{6} \times \frac{1}{6} = \frac{1}{36}.$$

Definition 5.3 *For any two random variables Y_1 and Y_2, the joint/bivariate cumulative distribution function (cdf) for Y_1 and Y_2 is given by*

$$F(y_1, y_2) = P(Y_1 \leq y_1, Y_2 \leq y_2),\ \ -\infty < y_1, y_2 < \infty.$$

Here are some basic properties of the bivariate cdf $F(y_1, y_2)$:

- $F(-\infty, -\infty) = F(y_1, -\infty) = F(-\infty, y_2) = 0$, and $F(\infty, \infty) = 1$.
- Given $-\infty \leq a_i < b_i \leq \infty$, $i = 1, 2$,

  $P(a_1 < Y_1 \leq b_1, a_2 < Y_2 \leq b_2) = F(b_1, b_2) + F(a_1, a_2) - F(a_1, b_2) - F(a_2, b_1) \geq 0.$

  To memorize this formula, one can consider the rectangle $R = (a_1, b_1] \times (a_2, b_2]$, and then the above identity can be rephrased as

  $$P((Y_2, Y_2) \in R) = \sum F(\textit{antidiagonal}) - \sum F(\textit{diagonal}).$$

Example 5.4 *Let Y_1 be a discrete uniform distribution on $\{0, 1\}$, and Y_2 be a continuous uniform distribution on $[0, 1]$. Assume that Y_1 an Y_2 are irrelevant.*

Note that the cdf for Y_1 is

$$F_{Y_1}(y_1) = P(Y_1 \leq y_1) = \begin{cases} 0, & y_1 \leq 0, \\ \frac{1}{2}, & 0 < y_1 < 1, \\ 1, & y_1 \geq 1. \end{cases}$$

and the cdf for Y_2 is

$$F_{Y_2}(y_2) = P(Y_2 \leq y_2) = \begin{cases} 0, & y_2 \leq 0, \\ y_2, & 0 < y_2 < 1, \\ 1, & y_2 \geq 1. \end{cases}$$

Thus, the bivariant cdf of (Y_1, Y_2) is given by

$$F(y_1, y_2) = P(Y_1 \leq y_1)P(Y_2 \leq y_2) = F_{Y_1}(y_1)F_{Y_2}(y_2) = \begin{cases} 0, & y_1 \leq 0 \textit{ or } y_2 \leq 0 \\ \frac{1}{2}y_2, & 0 < y_1, y_2 < 1, \\ \frac{1}{2}, & 0 < y_1 < 1, y_2 \geq 1, \\ y_2, & y_1 \geq 1, 0 < y_2 < 1, \\ 1, & y_1, y_2 \geq 1. \end{cases}$$

5.1.1 Discrete Case

Definition 5.5 *For any two discrete random variables Y_1 and Y_2, the joint/bivariate probability distribution function (pdf) for Y_1 and Y_2 is given by*

$$p(y_1, y_2) = P(Y_1 = y_1, Y_2 = y_2), \quad -\infty < y_1, y_2 < \infty.$$

Theorem 5.6 *Let Y_1 and Y_2 be discrete random variables with joint probability distribution function $p(y_1, y_2)$, then*

(1) $p(y_1, y_2) \geq 0$ *for all* y_1, y_2;

(2) $\sum_{y_1, y_2} p(y_1, y_2) = 1$;

(3) *For any $A \subset \mathbb{R} \times \mathbb{R}$, the probability of the event $((Y_1, Y_2) \in A)$ is given by*

$$P((Y_1, Y_2) \in A) = \sum_{(y_1, y_2) \in A} p(y_1, y_2).$$

In particular,

$$F(b_1, b_2) = P(Y_1 \leq b_1, Y_2 \leq b_2) = \sum_{y_1 \leq b_1, y_2 \leq b_2} p(y_1, y_2).$$

Example 5.7 *Tossing a pair of fair dices, and let Y_1 be the number on dice 1, Y_2 the number on dice 2. Each sample point (y_1, y_2) are of equal chance, and hence Y_1 and Y_2 have a bivariate pdf given by*

$$p(y_1, y_2) = \frac{1}{36}, \;\; y_1, y_2 = 1, 2, 3, 4, 5, 6.$$

The event that the first dice is tossed no more than 2 and the second tossed no less than 4 is given by

$$\begin{aligned} P(Y_1 \leq 2, Y_2 \geq 4) &= P(Y_1 = 1, 2, \; Y_2 = 4, 5, 6) \\ &= p(1,4) + p(1,5) + p(1,6) + p(2,4) + p(2,5) + p(2,6) \\ &= 6 \times \frac{1}{36} = \frac{1}{6}. \end{aligned}$$

Example 5.8 *A local supermarket has three checkout counters. Two customers arrive at the counters at different times when the counters are serving no other customers. Each customer chooses a counter at random, independently of the other. Let Y_1 denote the number of customers who choose counter A and Y_2, the number who select counter B. Find the joint probability function of Y_1 and Y_2.*

We first list all possible status of this two customers:

$$AA, \; AB, \; AC, \; BA, \; BB, \; BC, \; CA, \; CB, \; CC.$$

and each status is of chance $\frac{1}{9}$*. Then the joint pdf* $p(y_1, y_2)$ *is given by*

$$\begin{aligned}
p(0,0) &= P(CC) = \frac{1}{9}, \\
p(0,1) &= P(BC,\ CB) = \frac{2}{9}, \\
p(0,2) &= P(BB) = \frac{1}{9}, \\
p(1,0) &= P(AC,\ CA) = \frac{2}{9}, \\
p(1,1) &= P(AB,\ BA) = \frac{2}{9}, \\
p(2,0) &= P(AA) = \frac{1}{9}.
\end{aligned}$$

5.1.2 Continuous Case

Definition 5.9 *Two random variables* Y_1 *and* Y_2 *are said to be jointly continuous if there joint cumulative distribution* $F(y_1, y_2)$ *is continuous in both arguments.*

A function $f(y_1, y_2)$ *is called the joint probability density function (pdf) of the jointly continuous random variables* Y_1 *and* Y_2 *if*

$$F(y_1, y_2) = \int_{-\infty}^{y_2} \int_{-\infty}^{y_1} f(z_1, z_2) dz_1 dz_2, \quad -\infty < y_1, y_2 < \infty.$$

Theorem 5.10 *Let* Y_1 *and* Y_2 *be jointly continuous random variables with joint probability density function* $f(y_1, y_2)$*, then*

(1) $f(y_1, y_2) \geq 0$ *for all* y_1, y_2*;*

(2) $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(y_1, y_2) dy_1 dy_2 = 1$*;*

(3) *For any* $A \subset \mathbb{R} \times \mathbb{R}$*, the probability of the event* $((Y_1, Y_2) \in A)$ *is given by*

$$P((Y_1, Y_2) \in A) = \iint_A f(y_1, y_2) dy_1 dy_2.$$

In particular,

$$F(b_1, b_2) = P(Y_1 \leq b_1, Y_2 \leq b_2) = \int_{\infty}^{b_2} \int_{-\infty}^{b_1} f(y_1, y_2) dy_1 dy_2;$$

$$P(a_1 \leq Y_1 \leq b_1, a_2 \leq Y_2 \leq b_2) = \int_{a_2}^{b_2} \int_{a_1}^{b_1} f(y_1, y_2) dy_1 dy_2.$$

Example 5.11 *Suppose that a radioactive particle is randomly located in a square with sides of unit length. That is, if two regions within the unit square and of equal area are considered, the particle is equally likely to be in either region. Let Y_1 and Y_2 denote the coordinates of the particle?s location. A reasonable model for the relative frequency histogram for Y_1 and Y_2 is the bivariate analogue of the univariate uniform density function:*

$$f(y_1, y_2) = \begin{cases} 1, & 0 \le y_1 \le 1,\ 0 \le y_2 \le 1, \\ 0, & \textit{elsewhere}. \end{cases}$$

(a) Find $F(0.2, 0.4)$.

$$\begin{aligned} F(0.2, 0.4) = \int_{-\infty}^{0.4} \int_{-\infty}^{0.2} f(y_1, y_2) dy_1 dy_2 &= \int_{0}^{0.4} \int_{0}^{0.2} 1 dy_1 dy_2 \\ &= \int_{0}^{0.4} dy_2 \cdot \int_{0}^{0.2} dy_1 \\ &= 0.4 \times 0.2 = 0.08. \end{aligned}$$

(b) Find $P(0.1 \le Y_1 \le 0.3, 0 \le Y_2 \le 0.5)$.

$$\begin{aligned} P(0.1 \le Y_1 \le 0.3, 0 \le Y_2 \le 0.5) &= \int_{0}^{0.5} \int_{0.1}^{0.3} f(y_1, y_2) dy_1 dy_2 \\ &= \int_{0}^{0.5} \int_{0.1}^{0.3} 1 dy_1 dy_2 \\ &= \int_{0}^{0.5} dy_2 \cdot \int_{0.1}^{0.3} dy_1 \\ &= 0.5 \times (0.3 - 0.1) = 0.1. \end{aligned}$$

Example 5.12 *Gasoline is to be stocked in a bulk tank once at the beginning of each week and then sold to individual customers. Let Y_1 denote the proportion of the capacity of the bulk tank that is available after the tank is stocked at the beginning of the week. Because of the limited supplies, Y_1 varies from week to week. Let Y_2 denote the proportion of the capacity of the bulk tank that is sold during the week. Because Y_1 and Y_2 are both proportions, both variables take on values between 0 and 1. Further, the amount sold, y_2, cannot exceed the amount available, y_1. Suppose that the joint density function for Y_1 and Y_2 is given by*

$$f(y_1, y_2) = \begin{cases} 3y_1, & 0 \le y_2 \le y_1 \le 1, \\ 0, & \textit{elsewhere}. \end{cases}$$

Find the probability that less than one-half of the tank will be stocked and more than one-quarter of the tank will be sold.

We have

$$\begin{aligned}
P(0 \le Y_1 \le 0.5, Y_2 > 0.25) &= \iint\limits_{\substack{0\le y_2\le y_1\le 1,\\ 0\le y_1\le 0.5,\\ y_2>0.25}} f(y_1,y_2)dy_1dy_2 \\
&= \iint\limits_{0.25<y_2\le y_1\le 0.5} 3y_1dy_1dy_2 \\
&= \int_{0.25}^{0.5} dy_2 \left(\int_{y_2}^{0.5} 3y_1dy_1 \right) \\
&= \int_{0.25}^{0.5} \left(\frac{3}{2}y_1^2 \Big|_{y_2}^{0.5} \right) dy_2 \\
&= \int_{0.25}^{0.5} \frac{3}{2}\left(\frac{1}{4} - y_2^2 \right) dy_2 \\
&= \int_{0.25}^{0.5} \left(\frac{3}{8} - \frac{3}{2}y_2^2 \right) dy_2 \\
&= \left(\frac{3}{8}y_2 - \frac{1}{2}y_2^3 \right) \Big|_{0.25}^{0.5} \\
&= \frac{3}{8}(\frac{1}{2} - \frac{1}{4}) - \frac{1}{2}((\frac{1}{2})^3 - (\frac{1}{4})^3) = \frac{5}{128}.
\end{aligned}$$

Example 5.13 *Let (Y_1, Y_2) denote the coordinates of a point chosen at random inside a unit circle whose center is at the origin. That is, Y_1 and Y_2 have a joint density function given by*

$$f(y_1,y_2) = \begin{cases} \frac{1}{\pi}, & y_1^2 + y_2^2 \le 1, \\ 0, & \textit{elsewhere}. \end{cases}$$

Find $P(Y_1 \le Y_2)$, $P(Y_1 \le 3Y_2)$, and $P(Y_1 \le tY_2)$ for any $t \in \mathbb{R}$.

Let $\mathbb{D} = \{(y_1, y_2) : y_1^2 + y_2^2 \leq 1\}$ *be the unit disk, then*

$$\begin{aligned} P(Y_1 \leq Y_2) &= \iint_{y_1 \leq y_2} f(y_1, y_2) dy_1 dy_2 \\ &= \iint_{\mathbb{D} \cap \{y_1 \leq y_2\}} \frac{1}{\pi} dy_1 dy_2 \\ &= \frac{1}{\pi} Area(\mathbb{D} \cap \{y_1 \leq y_2\}) \\ &= \frac{1}{\pi} Area \text{ (} half \text{ } unit \text{ } disk \text{)} \\ &= \frac{1}{\pi} \frac{1}{2} \pi 1^2 = \frac{1}{2}. \end{aligned}$$

For the same reason, $P(Y_1 \leq tY_2) = \frac{1}{2}$ *for any* t*.*

Example 5.14 *Let* Y_1 *and* Y_2 *have the joint probability density function given by*

$$f(y_1, y_2) = \begin{cases} k(1 - y_2), & 0 \leq y_1 \leq y_2 \leq 1, \\ 0, & elsewhere. \end{cases}$$

Find the value of k *that makes this a probability density function.*

We have the normalization property, that is,

$$\begin{aligned} 1 = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(y_1, y_2) &= \iint_{0 \leq y_1 \leq y_2 \leq 1} k(1 - y_2) dy_1 dy_2 \\ &= \int_0^1 dy_1 \int_{y_1}^1 k(1 - y_2) dy_2 \\ &= k \int_0^1 \left(y_2 - \frac{1}{2} y_2^2 \right) \Big|_{y_1}^1 dy_1 \\ &= k \int_0^1 \left(\frac{1}{2} - y_1 + \frac{1}{2} y_1^2 \right) dy_1 \\ &= k \left(\frac{1}{2} y_1 - \frac{1}{2} y_1^2 + \frac{1}{6} y_1^3 \right) \Big|_0^1 \\ &= k \frac{1}{6}, \end{aligned}$$

so $k = 6$*.*

We can use the same method to deal with multivariate probability distributions.

5.2 Marginal and Conditional Probability Distributions

Given the joint distribution of two random variables Y_1 and Y_2, how can we determine the individual distribution of Y_1 or Y_2?

5.2.1 Discrete Case

In this subsection, we always let Y_1 and Y_2 be jointly discrete random variables with probability distribution function $p(y_1, y_2)$.

Definition 5.15 *The marginal probability functions of Y_1 and Y_2, respectively, are given by*

$$p_1(y_1) = \sum_{all\ y_2} p(y_1, y_2), \quad p_2(y_2) = \sum_{all\ y_1} p(y_1, y_2).$$

Example 5.16 *From a group of three Republicans, two Democrats, and one independent, a committee of two people is to be randomly selected. Let Y_1 denote the number of Republicans and Y_2 denote the number of Democrats on the committee. Find the joint probability function of Y_1 and Y_2 and then find the marginal probability function of Y_1 and Y_2.*

The joint pdf is

$$p(y_1, y_2) = P(Y_1 = y_1, Y_2 = y_2) = \frac{\binom{3}{y_1}\binom{2}{y_2}\binom{1}{2-y_1-y_2}}{\binom{6}{2}}$$

To make sense of the above formula, we must have

$$0 \le y_1 \le 3,\ 0 \le y_2 \le 2,\ 0 \le 2 - y_1 - y_2 \le 1,$$

and hence the possible outcomes for (y_1, y_2) are

$$(0,1),\ (1,0),\ (1,1),\ (0,2),\ (2,0).$$

By direct computation, we have

$$p(0,1) = \frac{2}{15},\ p(1,0) = \frac{3}{15},\ p(1,1) = \frac{6}{15},\ p(0,2) = \frac{1}{15},\ p(2,0) = \frac{3}{15}.$$

The marginal probability function of Y_1 is given by

$$\begin{aligned} p_1(0) &= p(0,1) + p(0,2) = \frac{2}{15} + \frac{1}{15} = \frac{1}{5}, \\ p_1(1) &= p(1,0) + p(1,1) = \frac{3}{15} + \frac{6}{15} = \frac{3}{5}, \\ p_1(2) &= p(2,0) = \frac{3}{15} = \frac{1}{5}. \end{aligned}$$

The marginal probability function of Y_2 is given by

$$\begin{aligned} p_2(0) &= p(1,0)+p(2,0) = \frac{3}{15}+\frac{3}{15} = \frac{2}{5}, \\ p_2(1) &= p(0,1)+p(1,1) = \frac{2}{15}+\frac{6}{15} = \frac{8}{15}, \\ p_2(2) &= p(0,2) = \frac{1}{15}. \end{aligned}$$

Definition 5.17 *If Y_1 and Y_2 have marginal probability functions $p_1(y_1)$ and $p_2(y_2)$, respectively, then the conditional discrete probability function of Y_1 given Y_2 is*

$$p(y_1|y_2) := P(Y_1 = y_1|Y_2 = y_2) = \frac{P(Y_1 = y_1, Y_2 = y_2)}{P(Y_2 = y_2)} = \frac{p(y_1,y_2)}{p_2(y_2)}$$

provided that $p_2(y_2) > 0$.

Similarly, we can define the conditional probability function of Y_2 given Y_1.

In Example 5.16, the conditional probability function of Y_2 given Y_1 is given by

$$p(y_2|y_1) = \frac{p(y_1,y_2)}{p_1(y_1)},$$

then

$$\begin{aligned} p(0|0) &= 0, \quad p(0|1) = \tfrac{3/15}{3/5} = \frac{1}{3}, \quad p(0|2) = \frac{3/15}{1/5} = 1 \\ p(1|0) &= \frac{2/15}{1/5} = \frac{2}{3}, \quad p(1|1) = \tfrac{6/15}{3/5} = \frac{2}{3}, \quad p(1|2) = \frac{0}{1/5} = 0 \\ p(2|0) &= \frac{1/15}{1/5} = \frac{1}{3}, \quad p(2|1) = \tfrac{0}{3/5} = 0, \quad p(2|2) = \frac{0}{1/5} = 0 \end{aligned}$$

5.2.2 Continuous Case

In this subsection, we always let Y_1 and Y_2 be jointly continuous random variables with probability density function $f(y_1,y_2)$.

Definition 5.18 *The marginal density functions of Y_1 and Y_2, respectively, are given by*

$$f_1(y_1) = \int_{-\infty}^{\infty} f(y_1,y_2)dy_2, \quad f_2(y_2) = \int_{-\infty}^{\infty} f(y_1,y_2)dy_1.$$

Example 5.19 *Let*

$$f(y_1, y_2) = \begin{cases} 2y_1, & 0 \le y_1 \le 1,\ 0 \le y_2 \le 1, \\ 0, & \textit{elsewhere}. \end{cases}$$

Find the marginal density functions for Y_1 *and* Y_2.

$$f_1(y_1) = \int_{-\infty}^{\infty} f(y_1, y_2)dy_2 = \begin{cases} \int_0^1 2y_1 dy_2 = 2y_1, & 0 \le y_1 \le 1, \\ 0, & \textit{elsewhere}. \end{cases}$$

$$f_2(y_2) = \int_{-\infty}^{\infty} f(y_1, y_2)dy_1 = \begin{cases} \int_0^1 2y_1 dy_1 = y_1^2|_0^1 = 1, & 0 \le y_2 \le 1, \\ 0, & \textit{elsewhere}. \end{cases}$$

Definition 5.20

- *The conditional density of* Y_1 *given* $Y_2 = y_2$ *is given by*

$$f(y_1|y_2) = \frac{f(y_1, y_2)}{f_2(y_2)}$$

  provided $f_2(y_2) > 0$. *Note that if* $f_2(y_2) = 0$, *then* $f(y_1|y_2)$ *is undefined.*

- *The conditional cumulative distribution function of* Y_1 *given* $Y_2 = y_2$ *is*

$$F(y_1|y_2) = P(Y_1 \le y_1|Y_2 = y_2) := \int_{-\infty}^{y_1} f(z|y_2)dz.$$

Similarly, we can define the conditional density and conditional cumulative distribution of Y_2 *given* $Y_1 = y_1$ *provided* $f_1(y_1) > 0$.

Example 5.21 *A soft-drink machine has a random amount* Y_2 *in supply at the beginning of a given day and dispenses a random amount* Y_1 *during the day (with measurements in gallons). It is not resupplied during the day, and hence* $Y_1 \le Y_2$. *It has been observed that* Y_1 *and* Y_2 *have a joint density given by*

$$f(y_1, y_2) = \begin{cases} 1/2, & 0 \le y_1 \le y_2 \le 2, \\ 0, & \textit{elsewhere}. \end{cases}$$

That is, the points (y_1, y_2) *are uniformly distributed over the triangle with the given boundaries.*

(a) Find the conditional density of Y_1 *given* $Y_2 = y_2$.

$$f_2(y_2) = \int_{-\infty}^{\infty} f(y_1, y_2) dy_1 = \begin{cases} \int_0^{y_2} \frac{1}{2} dy_1 = \frac{1}{2} y_2, & 0 \le y_2 \le 2, \\ \\ 0, \textit{elsewhere}. \end{cases}$$

Then

$$f(y_1|y_2) = \frac{f(y_1, y_2)}{f_2(y_2)} = \begin{cases} \frac{\frac{1}{2}}{\frac{1}{2} y_2} = \frac{1}{y_2}, & 0 < y_2 \le 2, \textit{ and } 0 \le y_1 \le y_2 \\ 0, & 0 < y_2 \le 2, \textit{ and } y_1 \le 0, \textit{ or } y_1 > y_2, \\ \textit{undefined}, & y_2 \le 0 \textit{ or } y_2 > 2. \end{cases}$$

(b) Evaluate the probability that less than 1/2 gallon will be sold, given that the machine contains 1.5 gallons at the start of the day.

$$P(Y_1 \le 0.5 | Y_2 = 1.5) = F(0.5|1.5) = \int_{-\infty}^{0.5} f(y_1|1.5) dy_1 = \int_0^{0.5} \frac{1}{1.5} dy_1 = \frac{1}{3}.$$

(c) Evaluate the probability that greater than 1 gallon will be sold, given that the machine contains 2 gallons at the start of the day.

$$\begin{aligned} P(Y_1 \ge 1 | Y_2 = 2) &= 1 - P(Y_1 < 1 | Y_2 = 2) \\ &= 1 - F(1|2) \\ &= 1 - \int_{-\infty}^{1} f(y_1|2) dy_1 \\ &= 1 - \int_0^1 \frac{1}{2} dy_1 = \frac{1}{2}. \end{aligned}$$

5.3 Independence of Random Variables

Recall that two events A and B are independent if $P(A \cap B) = P(A)P(B)$.

We can define independence of random variables via the cumulative distribution function (cdf).

Definition 5.22 *Suppose that Y_1 and Y_2 have cdfs $F_1(y_1)$ and $F_2(y_2)$ respectively, and (Y_1, Y_2) has the bivariate cdf $F(y_1, y_2)$. Y_1 and Y_2 are said to be independent if*

(5–1) $$F(y_1, y_2) = F_1(y_1) F(y_2),$$

for any $-\infty < y_1, y_2 < \infty$. Otherwise, we say that Y_1 and Y_2 are dependent.

- Condition (5–1) is equivalent to that

$$P(Y_1 \le y_1,\ Y_2 \le y_2) = P(Y_1 \le y_1)P(Y_2 \le y_2),$$

that is, the events $(Y_1 \le y_1)$ and $(Y_2 \le y_2)$ are independent. More generally, we can show that Condition (5–1) implies that the events $(Y_1 \in A_1)$ and $(Y_2 \in A_2)$ are independent for any $A_1, A_2 \subset \mathbb{R}$.

- The random variables $Y_1, Y_2, \ldots, Y_n$ are independent if

$$F(y_1, y_2, \ldots, y_n) = F_1(y_1)F(y_2)\ldots F_n(y_n).$$

We can check the independence by pdf: probability distribution function in the discrete case, and probability density function in the continuous case.

Theorem 5.23 *Y_1 and Y_2 are independent if and only if*

- *Discrete Case: $p(y_1, y_2) = p_1(y_1)p_2(y_2)$ for all y_1, y_2;*
- *Continuous Case: $f(y_1, y_2) = f_1(y_1)f_2(y_2)$ for all y_1, y_2.*

Example 5.24 *Let Y_1, Y_2 be the random variables in Example 5.16. Then Y_1 and Y_2 are dependent since*

$$0 = p(0,0) \ne p_1(0)p_2(0) = \frac{1}{5}\cdot\frac{2}{5} = \frac{2}{25}.$$

Example 5.25 *Suppose that Y_1 and Y_2 have bivariate pdf given by*

$$f(y_1, y_2) = \begin{cases} 2, & 0 \le y_1 \le y_2 \le 1, \\ 0, & \textit{elsewhere}. \end{cases}$$

Then Y_1 and Y_2 are dependent.

We first compute the density functions for Y_1 and Y_2:

$$f_1(y_1) = \int_{-\infty}^{\infty} f(y_1, y_2)dy_2 = \begin{cases} \int_{y_1}^{1} 2dy_2 = 2(1-y_1), & 0 \le y_1 \le 1, \\ 0, & \textit{elsewhere}. \end{cases}$$

$$f_2(y_2) = \int_{-\infty}^{\infty} f(y_1, y_2)dy_1 = \begin{cases} \int_{0}^{y_2} 2dy_1 = 2y_2, & 0 \le y_2 \le 1, \\ 0, & \textit{elsewhere}. \end{cases}$$

Then clearly $f(y_1, y_2) \ne f_1(y_1)f(y_2)$, which means that Y_1 and Y_2 are dependent.

Here is an easy way to check two continuous random variables of some particular form are independent.

Theorem 5.26 *Suppose Y_1 and Y_2 have bivariate pdf*

$$f(y_1,y_2)\begin{cases}>0, & a_1\le y_1\le b_1, a_2\le y_2\le b_2,\\ =0, & \textit{elsewhere}.\end{cases}$$

Then Y_1 and Y_2 are independent if and only if $f(y_1,y_2)$ is separable, that is,

$$f(y_1,y_2)=g(y_1)h(y_2),$$

where g and h are non-negative functions of one single variable.

In fact, if $f(y_1,y_2)$ is separable, then there is $c>0$ such that $f_1(y_1)=cg(y_1)$ and $f_2(y_2)=\frac{1}{c}h(y_2)$. The constant c is subject to the normalization condition that $\int_{-\infty}^{\infty}f_1(y_1)dy_1=\int_{-\infty}^{\infty}f_2(y_2)dy_2=1$.

Example 5.27 *Suppose that Y_1 and Y_2 have bivariate pdf given by*

$$f(y_1,y_2)=\begin{cases}6y_1y_2^2, & 0\le y_1\le 1, 0\le y_2\le 1,\\ 0, & \textit{elsewhere}.\end{cases}$$

Show that Y_1 and Y_2 are independent, and find the pdf's of Y_1 and Y_2.

By Theorem 5.26, we have $f(y_1,y_2)=g(y_1)h(y_2)$, where

$$g(y_1)=\begin{cases}6y_1, & 0\le y_1\le 1,\\ 0, & \textit{elsewhere}\end{cases},\quad h(y_2)=\begin{cases}y_2^2, & 0\le y_2\le 1,\\ 0, & \textit{elsewhere}\end{cases}$$

Set $f_(y_1)=cg(y_1)$, then

$$1=\int_{-\infty}^{\infty}f_1(y_1)dy_1=\int_0^1 c\cdot 6y_1dy_1=6c\,\frac{1}{2}y_1^2|_0^1=3c,$$

so $c=\frac{1}{3}$, and hence

$$f_1(y_1)=\frac{1}{3}g(y_1)=\begin{cases}2y_1, & 0\le y_1\le 1,\\ 0, & \textit{elsewhere}\end{cases},\quad f_2(y_2)=3h(y_2)=\begin{cases}3y_2^2, & 0\le y_2\le 1,\\ 0, & \textit{elsewhere}\end{cases}$$

Example 5.28 *Let Y_1 and Y_2 denote the lifetime, in the unit of hours, for two components of different type in an electronic system. The joint density of Y_1 and Y_2 is*

$$f(y_1,y_2)=\begin{cases}\frac{1}{8}y_1e^{-(y_1+y_2)/2}, & y_1>0, y_2>0,\\ 0, & \textit{elsewhere}.\end{cases}$$

Are Y_1 and Y_2 independent?

Yes, since

$$\frac{1}{8}y_1e^{-(y_1+y_2)/2}=\frac{1}{8}y_1e^{-y_1/2}\cdot e^{-y_2/2}.$$

5.4 Expected Value of a Function of Random Variables

Given a random variable Y, recall that the expected value of $g(Y)$ is given by

- Discrete: $E[g(Y)] = \sum_y g(y)\, p(y)$;
- Continuous: $E[g(Y)] = \int_{-\infty}^{\infty} g(y)\, f(y) dy$.

Definition 5.29 *Given random variables $Y_1, Y_2, \ldots, Y_n$ with joint pdf, the expected value of $g(Y_1, Y_2, \ldots, Y_n)$ is given by*

- *Discrete:* $E[g(Y_1, Y_2, \ldots, Y_n)] = \sum_{(y_1, y_2, \ldots, y_n)} g(y_1, y_2, \ldots, y_n)\, p(y_1, y_2, \ldots, y_n)$;
- *Continuous:*

$$E[g(Y_1, Y_2, \ldots, Y_n)] = \int_{y_n} \cdots \int_{y_2} \int_{y_1} g(y_1, y_2, \ldots, y_n) f(y_1, y_2, \ldots, y_n) dy_1 dy_2 \ldots dy_n.$$

We shall mostly focus on the case that $n = 2$. This definition actually generalizes previous ones. In particular, if Y_1 and Y_2 have bivariate pdf $f(y_1, y_2)$, we can compute $E(Y_1)$ by

$$E(Y_1) = \iint y_1 f(y_1, y_2) dy_1 dy_2$$

directly, without knowing the marginal pdf for Y_1.

Example 5.30 *Let Y_1 and Y_2 have joint density given by*

$$f(y_1, y_2) = \begin{cases} 2y_1, & 0 \le y_1 \le 1, 0 \le y_2 \le 1, \\ 0, & \textit{elsewhere}. \end{cases}$$

(a) Find $E(Y_1 Y_2)$:

$$\begin{aligned} E(Y_1 Y_2) = \iint y_1 y_2 f(y_1, y_2) dy_1 dy_2 &= \int_0^1 \int_0^1 y_1 y_2 2y_1 dy_1 dy_2 \\ &= \int_0^1 2y_2 \left(\int_0^1 y_1^2 dy_1 \right) dy_2 \\ &= \int_0^1 2y_2 \frac{1}{3} y_1^3 |_0^1 dy_2 \\ &= \frac{2}{3} \int_0^1 y_2 dy_2 \\ &= \frac{2}{3} \frac{1}{2} y_2^2 |_0^1 = \frac{1}{3}. \end{aligned}$$

(b) Find $E(Y_1)$:

$$\begin{aligned} E(Y_1) = \iint y_1 f(y_1, y_2) dy_1 dy_2 &= \int_0^1 \int_0^1 y_1 2y_1 dy_1 dy_2 \\ &= \int_0^1 2\left(\int_0^1 y_1^2 dy_1\right) dy_2 \\ &= \int_0^1 2\frac{1}{3} y_1^3 |_0^1 dy_2 \\ &= \frac{2}{3}\int_0^1 dy_2 = \frac{2}{3}. \end{aligned}$$

(c) Find $V(Y_1)$:

$$\begin{aligned} E(Y_1^2) = \iint y_1^2 f(y_1, y_2) dy_1 dy_2 &= \int_0^1 \int_0^1 y_1^2 2y_1 dy_1 dy_2 \\ &= \int_0^1 2\left(\int_0^1 y_1^3 dy_1\right) dy_2 \\ &= \int_0^1 2\frac{1}{4} y_1^4 |_0^1 dy_2 \\ &= \frac{1}{2}\int_0^1 dy_2 = \frac{1}{2}. \end{aligned}$$

Then $V(Y_1) = E(Y_1^2) - [E(Y_1)]^2 = \frac{1}{2} - (\frac{2}{3})^2 = \frac{1}{18}$.

To simplify computations of the expected value, we recall some basic properties:

Theorem 5.31

(1) $E(c) = c$;

(2) $E(c_1Y_1 + c_2Y_2 + \cdots + c_kY_k) = c_1E(Y_1) + c_2E(Y_2) + c_kE(Y_k)$;

(3) *if Y_1 and Y_2 are independent, then $E(Y_1Y_2) = E(Y_1)E(Y_2)$. More generally, $E[g(Y_1)h(Y_2)] = E[g(Y_1)]E[h(Y_2)]$ for any real-valued functions $g, h : \mathbb{R} \to \mathbb{R}$.*

Proof The only new property here is (3): since Y_1 and Y_2 are independent, then $f(y_1, y_2) = f_1(y_1)f_2(y_2)$, and hence

$$\begin{aligned} E[g(Y_1)h(Y_2)] = \iint g(y_1)h(y_2)f(y_1, y_2) dy_1 dy_2 &= \iint g(y_1)h(y_2)f_1(y_1)f_2(y_2) dy_1 dy_2 \\ &= \int g(y_1)f_1(y_1) dy_1 \int h(y_2)f_2(y_2) dy_2 \\ &= E[g(Y_1)]E[h(Y_2)]. \end{aligned}$$

□

Example 5.32 *Suppose that a radioactive particle is randomly located with coordinates* (Y_1, Y_2) *in a unit disk. A reasonable model for the joint density function for* Y_1 *and* Y_2 *is*

$$f(y_1, y_2) = \begin{cases} 1, & 0 \le y_1 \le 1, 0 \le y_2 \le 1, \\ 0, & \textit{elsewhere}. \end{cases}$$

(a). $E(Y_1) = \iint y_1 f(y_1, y_2) dy_1 dy_2 = \int_0^1 \int_0^1 y_1 dy_1 dy_2 = \int_0^1 y_1 dy_1 \int_0^1 dy_2 = \frac{1}{2} \cdot 1 = \frac{1}{2}$.

(b). $E(Y_1 - Y_2)$*: by the similar formula as (a) and symmetry between* Y_1 *and* Y_2*, we have* $E(Y_2) = \frac{1}{2}$ *as well. So* $E(Y_1 - Y_2) = E(Y_1) - E(Y_2) = 0$.

(c). $E(Y_1 Y_2)$*: By Theorem 5.26,* Y_1 *and* Y_2 *are independent, so* $E(Y_1 Y_2) = E(Y_1)E(Y_2) = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$.

(d). $E(Y_1^2 + Y_2^2)$*:*

$$E(Y_1^2) = \iint y_1^2 f(y_1, y_2) dy_1 dy_2 = \int_0^1 \int_0^1 y_1^2 dy_1 dy_2 = \int_0^1 y_1^2 dy_1 \int_0^1 dy_2 = \frac{1}{3}.$$

By symmetry, we can get $E(Y_2^2) = \frac{1}{3}$ *as well. So* $E(Y_1^2 + Y_2^2) = E(Y_1^2) + E(Y_2^2) = \frac{2}{3}$.

(e). $V(Y_1 Y_2)$*:*

$$V(Y_1 Y_2) = E(Y_1^2 Y_2^2) - [E(Y_1 Y_2)]^2 = E(Y_1^2)E(Y_2^2) - [E(Y_1 Y_2)]^2 = \frac{1}{3}\frac{1}{3} - (\frac{1}{4})^2 = \frac{7}{144}.$$

5.5 Covariance of Random Variables

In real lift, two random variables Y_1 and Y_2 of interest are often dependent. To measure the dependency between Y_1 and Y_2, we need the definition of covariance.

Definition 5.33 *Let* $\mu_1 = E(Y_1)$ *and* $\mu_2 = E(Y_2)$*, then the covariance of* Y_1 *and* Y_2 *is given by*

$$\text{Cov}(Y_1, Y_2) = E[(Y_1 - \mu_1)(Y_2 - \mu_2)].$$

The quantity $(Y_1 - \mu_1)(Y_2 - \mu_2)$ has the following meaning: the center of the distribution of (Y_1, Y_2) is $C(\mu_1, \mu_2)$. Given a sample point $P(y_1, y_2)$ in the 2-dimensional coordinate plane, we know that P deviates from C by $(y_1 - \mu_1)$ units in Y_1-direction, and $(y_2 - \mu_2)$ units in Y_2-direction, that being said, P deviates from C by a rectangle with signed area $(y_1 - \mu_1)(y_2 - \mu_2)$. Then the covariance $\text{Cov}(Y_1, Y_2)$ is the average value of such area deviation.

Proposition 5.34

(1) $\mathrm{Cov}(Y_1, Y_2)$ *can be positive, negative or zero.*

(2) $V(Y) = \mathrm{Cov}(Y, Y) \geq 0$*;*

(3) $\mathrm{Cov}(Y_1, Y_2) = E(Y_1Y_2) - E(Y_1)E(Y_2)$*.*

(4) *If* Y_1 *and* Y_2 *are independent, then* $\mathrm{Cov}(Y_1, Y_2) = 0$*. Indeed,*

$$\mathrm{Cov}(Y_1, Y_2) = E(Y_1Y_2) - E(Y_1)E(Y_2) = E(Y_1)E(Y_2) - E(Y_1)E(Y_2) = 0.$$

However, the converse might not be true. $\mathrm{Cov}(Y_1, Y_2) = 0$ *only means that* Y_1 *and* Y_2 *have no linear dependence.*

(5) *By Cauchy-Schwartz inequality, we have*

$$|\mathrm{Cov}(Y_1, Y_2)| \leq \sqrt{V(Y_1)}\sqrt{V(Y_2)} = \sigma_1\sigma_2.$$

Moreover, the equality only holds when Y_1 *and* Y_2 *are linearly dependent, i.e.,* $Y_2 - \mu_2 = k(Y_1 - \mu_1)$ *or* $Y_1 = \mu_1$*. The proof is sketched as follows: assume that* Y_1 *and* Y_2 *are continuous and have bivariate density function* $f(y_1, y_2)$*, then*

$$\begin{aligned}
|\mathrm{Cov}(Y_1, Y_2)| &= \left|\iint (y_1 - \mu_1)(y_2 - \mu_2) f(y_1, y_2) dy_1 dy_2\right| \\
&\leq \iint [|y_1 - \mu_1|\sqrt{f(y_1, y_2)}][|y_2 - \mu_2|\sqrt{f(y_1, y_2)}] dy_1 dy_2 \\
&\leq \sqrt{\iint [|y_1 - \mu_1|\sqrt{f(y_1, y_2)}]^2 dy_1 dy_2}\sqrt{\iint [|y_2 - \mu_2|\sqrt{f(y_1, y_2)}]^2 dy_1 dy_2} \\
&= \sqrt{\iint (y_1 - \mu_1)^2 f(y_1, y_2) dy_1 dy_2}\sqrt{\iint (y_2 - \mu_2)^2 f(y_1, y_2) dy_1 dy_2} \\
&= \sqrt{V(Y_1)}\sqrt{V(Y_2)}.
\end{aligned}$$

Remark 5.35

(1) *The covariance* $\mathrm{Cov}(Y_1, Y_2)$ *only captures the linear dependence between* Y_1 *and* Y_2*. In other words, we regard* Y_2 *as a function of* Y_1*. The larger* $|\mathrm{Cov}(Y_1, Y_2)|$ *is, the most likely that the distribution of* (Y_1, Y_2) *can be approximated by the line* $y_2 - \mu_2 = k(y_1 - \mu_1)$ *for some* k*. In particular, if* $\mathrm{Cov}(Y_1, Y_2) > 0$*, then* $k > 0$*, which means that* Y_2 *increases as* Y_1 *increases; otherwise if* $\mathrm{Cov}(Y_1, Y_2) < 0$*, then* $k < 0$*, which means that* Y_2 *decreases as* Y_1 *increases.*

(2) *The magnitude* $|\mathrm{Cov}(Y_1, Y_2)|$ *sometimes may not be a good measure since there is no a priori information on its maximum. We instead introduce the correlation coefficient*

$$\rho = \rho(Y_1, Y_2) := \frac{\mathrm{Cov}(Y_1, Y_2)}{\sqrt{V(Y_1)}\sqrt{V(Y_2)}}.$$

provided both $V(Y_1)$ *and* $V(Y_2)$ *are positive. Then*

- $|\rho| \leq 1$ *(by Cauchy-Schwartz);*
- $\rho = 0$ *means that* Y_1 *and* Y_2 *has no linear relation;*
- $\rho = \pm 1$ *means that* Y_1 *and* Y_2 *has perfect linear relation, that is,* $Y_2 - \mu_2 = k(Y_1 - \mu_1)$ *almost surely.*

Example 5.36 *Let* Y_1 *and* Y_2 *have joint density given by*

$$f(y_1, y_2) = \begin{cases} 2y_1, & 0 \leq y_1 \leq 1, 0 \leq y_2 \leq 1, \\ 0, & \textit{elsewhere}. \end{cases}$$

Find $\mathrm{Cov}(Y_1, Y_2)$ *and* $\rho(Y_1, Y_2)$.

By Theorem 5.26, Y_1 *and* Y_2 *must be independent, and hence* $\mathrm{Cov}(Y_1, Y_2) = 0$, $\rho(Y_1, Y_2) = 0$.

Example 5.37 *Let* Y_1 *and* Y_2 *have joint density given by*

$$f(y_1, y_2) = \begin{cases} 2, & 0 \leq y_1 \leq y_2 \leq 1, \\ 0, & \textit{elsewhere}. \end{cases}$$

Find $\mathrm{Cov}(Y_1, Y_2)$.

$$\begin{aligned} \mathrm{Cov}(Y_1, Y_2) &= E(Y_1 Y_2) - E(Y_1)E(Y_2) \\ &= \iint y_1 y_2 f(y_1, y_2) dy_1 dy_2 - \iint y_1 f(y_1, y_2) dy_1 dy_2 \iint y_2 f(y_1, y_2) dy_1 dy_2 \\ &= \int_0^1 \int_0^{y_2} y_1 y_2 2 dy_1 dy_2 - \int_0^1 \int_0^{y_2} y_1 2 dy_1 dy_2 \int_0^1 \int_0^{y_2} y_2 2 dy_1 dy_2 \\ &= \frac{1}{4} - \frac{1}{3}\frac{2}{3} = \frac{1}{36}. \end{aligned}$$

By straightforward computation, we deduce a formula of the variance of linear combinations of random variables.

Theorem 5.38 *Consider the random variables*

$$U = \sum_{i=1}^{n} a_i Y_i = a_1 Y_1 + \cdots + a_n Y_n,$$

$$V = \sum_{j=1}^{m} b_j Z_j = b_1 Z_1 + \cdots + b_m Z_m,$$

then

$$\begin{aligned}
\mathrm{Cov}(U,V) &= \sum_{i=1}^{n}\sum_{j=1}^{m} a_i b_j \mathrm{Cov}(Y_i, Z_j) \\
&= \begin{pmatrix} a_1 & a_2 & \dots & a_n \end{pmatrix} \begin{pmatrix} \mathrm{Cov}(Y_1,Z_1) & \mathrm{Cov}(Y_1,Z_2) & \cdots & \mathrm{Cov}(Y_1,Z_m) \\ \mathrm{Cov}(Y_2,Z_1) & \mathrm{Cov}(Y_2,Z_2) & \cdots & \mathrm{Cov}(Y_2,Z_m) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}(Y_n,Z_1) & \mathrm{Cov}(Y_n,Z_2) & \cdots & \mathrm{Cov}(Y_n,Z_m) \end{pmatrix} \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{pmatrix}
\end{aligned}$$

In particular, we have the variance

$$V(U) = \sum_{i=1}^{n} a_i^2 V(Y_i) + 2 \sum_{1 \le i < j \le n} a_i a_j \mathrm{Cov}(Y_i, Y_j).$$

An even simpler example is that

$$V(aY + b) = a^2 V(Y)$$

Since $V(b) = 0$ *and* $\mathrm{Cov}(Y, b) = 0$.

Example 5.39 *Let* $Y_1, Y_2, \ldots, Y_n$ *be independent random variables with* $E(Y_i) = \mu$ *and* $V(Y_i) = \sigma^2$. *(These variables may denote the outcomes of* n *independent trials of an experiment.) Define the sample mean variable*

$$\overline{Y} = \frac{Y_1 + Y_2 + \cdots + Y_n}{n}.$$

Then $E(\overline{Y}) = \mu$, $V(\overline{Y}) = \frac{\sigma^2}{n}$.

Indeed,

$$E(\overline{Y}) = \frac{E(Y_1) + E(Y_2) + \cdots + E(Y_n)}{n} = \mu,$$

and since $\mathrm{Cov}(Y_i, Y_j) = 0$ *if* $i \neq j$, *since they are independent, thus*

$$V(\overline{Y}) = \sum_{i=1}^{n} \left(\frac{1}{n}\right)^2 V(Y_i) = \frac{1}{n^2} \sum_{i=1}^{n} \sigma^2 = \frac{\sigma^2}{n}.$$

Consequently,

$$V(\sqrt{n}\left(\overline{Y}-\mu\right)) = \sigma^2.$$

This example will be helpful when we introduce the law of large numbers (LLN) and the central limit theorem (CLT).

5.6 Conditional Expectations

A random variable Y_1 often relies on outcomes of another random variable Y_2.

Example 5.40 *A quality control plan for an assembly line involves sampling $n = 10$ finished items per day and counting Y, the number of defectives. If P denotes the probability of observing a defective, then Y has a binomial distribution $b(n,P)$, assuming that a large number of items are produced by the line. But P varies from day to day and is assumed to have a uniform distribution on the interval from 0 to 1/4, that is, $P \sim U(0,\frac{1}{4})$.*

Here we get the number Y of defectives depends on the outcome of the defective variable P. The conditional probability of Y given $P = 1/8$ is then given by the binomial distribution $b(n,1/8)$, that is,

$$P(Y = y|P = 1/8) = \binom{10}{y}(1/8)^y(7/8)^{10-y}, \;\; y = 0,1,\ldots,10.$$

We recall the conditional distribution of Y_1 given that $Y_2 = y_2$:

- Discrete: the conditional probability distribution function of Y_1 given that $Y_2 = y_2$ is
$$p(y_1|y_2) = \frac{p(y_1,y_2)}{p_2(y_2)},$$
provided $p_2(y_2) = P(Y_2 = y_2) > 0$;
- Continuous: the conditional probability density function of Y_1 given that $Y_2 = y_2$ is
$$f(y_1|y_2) = \frac{f(y_1,y_2)}{f_2(y_2)},$$
provided $f_2(y_2) > 0$.

Definition 5.41 *The conditional expectation of $g(Y_1)$ given that $Y_2 = y_2$ is*

Discrete:

$$E(g(Y_1)|Y_2 = y_2) = \sum_{y_1} g(y_1)p(y_1|y_2),$$

Continuous:

$$E(g(Y_1)|Y_2 = y_2) = \int_{-\infty}^{\infty} g(y_1)f(y_1|y_2)dy_1.$$

Example 5.42 *The random variables Y_1 and Y_2 have the joint density function given by*

$$f(y_1, y_2) = \begin{cases} 1/2, & 0 \le y_1 \le y_2 \le 2, \\ 0, & \textit{elsewhere}. \end{cases}$$

Find the conditional expectation of Y_1 given that $Y_2 = 1.5$.

$$f_2(y_2) = \int f(y_1, y_2)dy_1 = \begin{cases} \int_0^{y_2} 1/2 dy_1 = \frac{y_2}{2}, & 0 \le y_2 \le 2, \\ 0, & \textit{elsewhere}. \end{cases}$$

Then

$$f(y_1|y_2) = \frac{f(y_1, y_2)}{f_2(y_2)} = \begin{cases} \frac{1}{y_2}, & 0 \le y_1 \le y_2 \le 2, \\ 0, & 0 \le y_2 \le 2, y_1 < 0 or y_1 > y_2, \\ \textit{undefined}, & \textit{elsewhere}. \end{cases}$$

Given that $Y_2 = 1.5$, we have

$$E(Y_1|Y_2 = 1.5) = \int y_1 f(y_1|1.5)dy_1 = \int_0^{1.5} \frac{y_1}{1.5} dy_1 = 0.75.$$

If we allow the outcome of $Y_2 = y_2$ to vary, then actually $y_2 \mapsto E(Y_1|Y_2 = y_2)$ can be viewed as a function, or random variable, denoted by $E(Y_1|Y_2)$. As in the Example 5.42, $E(Y_1|Y_2) = Y_2/2$ with outcome in $[0, 2]$.

Theorem 5.43

$$E(E(Y_1|Y_2)) = E(Y_1).$$

The proof of this theorem is routine. The idea of this formula is the following: Take average of the conditional average of Y_1 given Y_2 is exactly the average of Y_1.

In Example 5.42, The expected value of Y is then

$$E(Y) = E(E(Y|P)) = E(nP) = E(10P) = 10E(P) = 10\frac{0 + 1/4}{2} = 1.25.$$

6 Functions of Random Variables

Given random variables $Y_1, Y_2, \ldots, Y_n$, and a real-valued function $g : \mathbb{R}^n \to \mathbb{R}$, we obtain a new random variable

$$U = g(Y_1, Y_2, \ldots, Y_n).$$

A very important random variable of this type would be the sample mean variable

$$U = \frac{Y_1 + Y_2 + \cdots + Y_n}{n} = \frac{1}{n}\sum_{i=1}^{n} Y_i.$$

Sometimes, we are really interested in the actual distribution of U instead of only knowing the expected value $E(U)$ and the variance $V(U)$. There are typically three standard methods.

6.1 The Method of Distribution Functions

We compute the cumulative distribution function of U by

$$F_U(u) = P(U \leq u) = P(g(Y_1, Y_2, \ldots, Y_n) \leq u) =: P(A),$$

where A is the event given by

$$A = \{(y_1, y_2, \ldots, y_n) : g(y_1, y_2, \ldots, y_n) \leq u\} \subset \mathbb{R}^n.$$

If $Y_1, Y_2, \ldots, Y_n$ have joint density function $f(y_1, y_2, \ldots, y_n)$, then

$$F_U(u) = P(A) = \int \cdots \int_A f(y_1, \ldots, y_n) dy_1 \ldots dy_n,$$

and the density function of U is given by $f_U(u) = F'_U(u)$.

Example 6.1 *A process for refining sugar yields up to 1 ton of pure sugar per day, but the actual amount produced, Y, is a random variable because of machine breakdowns and other slowdowns. Suppose that Y has density function given by*

$$f(y) = \begin{cases} 2y, & 0 \leq y \leq 1, \\ 0, & \textit{elsewhere}. \end{cases}$$

The company is paid at the rate of \$300 per ton for the refined sugar, but it also has a fixed overhead cost of \$100 per day. Thus the daily profit, in hundreds of dollars, is $U = 3Y - 1$. Find the probability density function for U.

$$\begin{aligned} F_U(u) = P(U \le u) &= P(3Y - 1 \le u) \\ &= P(Y \le \frac{u+1}{3}) \\ &= \int_{-\infty}^{\frac{u+1}{3}} f(y)dy \end{aligned}$$

Then

$$\begin{aligned} f_U(u) = F'_U(u) &= \frac{d}{du}\int_{-\infty}^{\frac{u+1}{3}} f(y)dy \\ &= f\left(\frac{u+1}{3}\right)\frac{1}{3} \\ &= \begin{cases} \frac{1}{3}\cdot 2^{\frac{u+1}{3}}, & 0 \le \frac{u+1}{3} \le 1, \\ 0, & \textit{elsewhere}. \end{cases} \\ &= \begin{cases} \frac{2}{9}(u+1), & -1 \le u \le 2, \\ 0, & \textit{elsewhere}. \end{cases} \end{aligned}$$

Example 6.2 *The joint density function of X and Y is given by*

$$f(x,y) = \begin{cases} 3x, & 0 \le y \le x \le 1, \\ 0, & \textit{elsewhere}. \end{cases}$$

Find the pdf of $U = X - Y$.

We denote the triangle $\Delta = \{0 \le y \le x \le 1\}$, outside which the joint pdf $f(x,y) = 0$. For any $u \in \mathbb{R}$, note that the intersection of domains

$$\{y < x - u\} \cap \Delta = \begin{cases} \Delta, & u < 0, \\ \{u < x \le 1,\ 0 \le y < x - u\}, & 0 \le u \le 1, \\ \emptyset, & u > 1. \end{cases}$$

Hence,

$$\begin{aligned}
F_U(u) = P(U \le u) &= 1 - P(X - Y > u) \\
&= 1 - \iint_{\{y<x-u\}} f(x,y)dxdy \\
&= \begin{cases} 1 - \int\int_{\Delta} f(x,y)dxdy, & u < 0, \\ 1 - \int_u^1 \int_0^{x-u} 3x dy dx, & 0 \le u \le 1, \\ 1 - \int\int_{\emptyset} f(x,y)dxdy, & u > 1. \end{cases} \\
&= \begin{cases} 1 - 1, & u < 0, \\ 1 - \int_u^1 3x(x-u)dx, & 0 \le u \le 1, \\ 1 - 0, & u > 1. \end{cases} \\
&= \begin{cases} 0, & u < 0, \\ 1 - \left(x^3 - \frac{3}{2}ux^2\right)\Big|_u^1, & 0 \le u \le 1, \\ 1, & u > 1. \end{cases} \\
&= \begin{cases} 0, & u < 0, \\ (3u - u^3)/2, & 0 \le u \le 1, \\ 1, & u > 1. \end{cases}
\end{aligned}$$

Then

$$f_U(u) = F_U'(u) = \begin{cases} \frac{3}{2}(1-u^2), & 0 \le u \le 1, \\ 0, & \textit{elsewhere}. \end{cases}$$

6.2 The Method of Transformations

Let Y be a random variable, and $U = g(Y)$. Assume that $g : \mathbb{R} \to \mathbb{R}$ is invertible, that is, g is strictly increasing or strictly decreasing. In such case, the inverse $h = g^{-1}$ exists, and set $Y = h(U)$.

Theorem 6.3

$$F_U(u) = F_Y(h(u)), \text{ and } f_U(u) = f_Y(h(u))|h'(u)|.$$

In Example 6.1, since $U = 3Y - 1$, then $Y = \frac{U+1}{3}$, that is, $h(u) = (u+1)/3$, and hence

$$f_U(u) = f_Y((u+1)/3) \cdot 1/3 = \begin{cases} \frac{2}{9}(u+1), & -1 \le u \le 2, \\ 0, & \textit{elsewhere}. \end{cases}$$

Suppose now that X and Y has joint pdf $f_{(X,Y)}(x,y)$, and $U = g(X,Y)$. Assume that for any fixed $X = x$ (which is viewed as a parameter), the function $u = g(x,y)$, regarded as a function u of the variable y, is invertible, and the inverse is given by $y = h(x,u)$. By Theorem 6.3, we have

$$f_U(u|X=x) = f_Y(h(x,u)|X=x)\left|\frac{\partial}{\partial u}h(x,u)\right|.$$

Multiplying the pdf $f_X(x)$ on both sides, we get the joint pdf of (X,U) by

Theorem 6.4

$$f_{(X,U)}(x,u) = f_{(X,Y)}(x,h(x,u))\left|\frac{\partial}{\partial u}h(x,u)\right|.$$

Thus, we can further obtain the pdf of U by

$$f_U(u) = \int_{-\infty}^{\infty} f_{(X,U)}(x,u)dx.$$

In Example 6.2, $U = X - Y$, that is, $u = g(x,y) = x - y$. For any fixed $X = x$, this function is invertible, and has inverse $y = h(x,u) = x - u$. Thus, $\left|\frac{\partial}{\partial u}h(x,u)\right| = |-1| = 1$. Hence the joint pdf of (X,U) is

$$\begin{aligned} f_{(X,U}(x,u) = f_{(X,Y)}(x,y)\left|\frac{\partial}{\partial u}h(x,u)\right| &= \begin{cases} 3x, & 0 \le x-u \le x \le 1, \\ 0, & elsewhere. \end{cases} \cdot 1 \\ &= \begin{cases} 3x, & 0 \le u \le x \le 1, \\ 0, & elsewhere. \end{cases} \end{aligned}$$

and hence

$$\begin{aligned} f_U(u) = \int_{-\infty}^{\infty} f_{(X,U)}(x,u)dx &= \begin{cases} \int_u^1 3x dx, & 0 \le u \le 1, \\ 0, & elsewhere. \end{cases} \\ &= \begin{cases} \frac{3}{2}(1-u^2), & 0 \le u \le 1, \\ 0, & elsewhere. \end{cases} \end{aligned}$$

6.3 The Method of Moment-generating Functions

Given a random variable Y, the moment-generating function (MGF) is given by

$$m_Y(t) = E\left(e^{tY}\right).$$

In general, $m_Y(t)$ may not be defined for all t, but it is defined near $t = 0$ for all the well known distributions in our course. In earlier lectures, we have shown that MGF can be used to calculate the moments of Y, that is,

Theorem 6.5 $E(Y^k) = m_Y^{(k)}(0)$.

Example 6.6 *A random variable Y has pdf (probability density function)*

$$f(y) = \begin{cases} e^y, & y < 0 \\ 0, & \textit{elsewhere}. \end{cases}$$

Compute (1) MGF $m_Y(t)$; (2) $E(e^{2Y})$; (3) $V(Y)$.

Solutions: (1) near $t = 0$, we could assume $t > -1$, then

$$\begin{aligned} m_Y(t) = E(e^{tY}) = \int_{-\infty}^{\infty} e^{ty} f(y)dy = \int_{-\infty}^{0} e^{ty} e^y dy &= \int_{-\infty}^{0} e^{(t+1)y} dy \\ &= \frac{1}{t+1} e^{(t+1)y}\Big|_{-\infty}^{0} = (t+1)^{-1}. \end{aligned}$$

(2) $E(e^{2Y}) = m_Y(2) = \frac{1}{3}$;
(3) Take derivatives of $m_Y(t)$:

$$m_Y'(t) = -(t+1)^{-2}, \text{ and } m_Y''(t) = 2(t+1)^{-3}.$$

So $E(Y) = m_Y'(0) = -1$, $E(Y^2) = m_Y''(0) = 2$, and thus, $V(Y) = E(Y^2) - [E(Y)]^2 = 2 - (-1)^2 = 1$.

Theorem 6.7 (Uniqueness Theorem)

$$X \text{ and } Y \text{ have the same distribution} \iff m_X(t) = m_Y(t).$$

Summary of moment-generating functions (MGF) for well known distributions:

(1) Continuous distributions:
   (a) Uniform $U(\theta_1, \theta_2)$: $m(t) = \dfrac{e^{t\theta_2} - e^{t\theta_1}}{t(\theta_2 - \theta_1)}$;
   (b) Normal $N(\mu, \sigma^2)$: $m(t) = \exp\left(\mu t + \frac{1}{2}\sigma^2 t^2\right)$;
   (c) Gamma $Gamma(\alpha, \beta)$: $m(t) = (1 - \beta t)^{-\alpha}$. In particular,
      (i) exponential $\exp(\beta) = Gamma(1, \beta)$ has $m(t) = (1 - \beta t)^{-1}$;
      (ii) Chi-square $\chi^2(\nu) = Gamma(\nu/2, 2)$ has $m(t) = (1 - 2t)^{-\nu/2}$.

(2) Discrete distributions:
   (a) Binomial $B(n, p)$: $m(t) = [pe^t + (1 - p)]^n$;
   (b) Geometric $Geo(p)$: $m(t) = \dfrac{pe^t}{1 - (1 - p)e^t}$;

(c) Poisson *Poisson*(λ): $m(t) = \exp[\lambda(e^t - 1)]$.

Example 6.8 *Find the corresponding distributions for the following MGF:*

(1) $m(t) = \left(\dfrac{e^t + 1}{2}\right)^{10}$;

(2) $m(t) = \dfrac{e^t}{3 - 2e^t}$;

(3) $m(t) = e^{-1+(1-t)^2}$;

(4) $m(t) = \dfrac{1}{(1-3t)^2}$;

(5) $m(t) = e^{e^t - 1}$;

(6) $m(t) = \dfrac{e^{2t} - e^t}{t}$.

Solutions: Rewrite these functions to match the MGF of well known distributions.

(1) $m(t) = \left(\frac{1}{2}e^t + (1 - \frac{1}{2})\right)^{10}$ — $B(10, \frac{1}{2})$;

(2) $m(t) = \dfrac{\frac{1}{3}e^t}{1 - (1 - \frac{1}{3})e^t}$ — $Geo(\frac{1}{3})$;

(3) $m(t) = \exp\left(-2t + \frac{1}{2}(\sqrt{2})^2 t^2\right)$ — $N(-2, (\sqrt{2})^2)$;

(4) $m(t) = (1-3t)^{-2}$ — $Gamma(2, 3)$;

(5) $m(t) = \exp\left(1 \cdot (e^t - 1)\right)$ — $Poisson(1)$;

(6) $m(t) = \dfrac{e^{2t} - e^{1t}}{t(2-1)}$ — $U(1, 2)$.

Using the uniqueness theorem, we can show certain distributions are stable under scaling and shifting.

Theorem 6.9 *If $Y = aX + b$, then $m_Y(t) = e^{bt} m_X(at)$.*

Proof $m_Y(t) = E(e^{tY}) = E(e^{(t(aX+b)}) = E(e^{bt}e^{(at)X}) = e^{bt}E(e^{(at)X}) = e^{bt}m_X(at)$. □

Example 6.10

(1) *If X follows a normal distribution, i.e., $X \sim N(\mu, \sigma^2)$, then $Y = aX + b$ also follows normal distribution, in fact, $Y \sim N(a\mu + b, (a\sigma)^2)$, since*

$$m_Y(t) = e^{bt} m_X(at) = e^{bt} e^{\mu at + \frac{1}{2}\sigma^2 (at)^2} = e^{(a\mu+b)t + \frac{1}{2}(a\sigma)^2 t^2}.$$

(2) *If X follows the uniform distribution over the interval (θ_1, θ_2), i.e., $X \sim U(\theta_1, \theta_2)$, we know that $m_X(t) = \dfrac{e^{t\theta_2} - e^{t\theta_1}}{t(\theta_2 - \theta_1)}$, then $Y = aX + b$ also follows uniform distribution, in fact, $Y = aX + b \sim U(a\theta_1 + b,\ a\theta_2 + b)$ since*

$$m_Y(t) = e^{bt} m_X(at) = e^{bt} \frac{e^{at\theta_2} - e^{at\theta_1}}{t(\theta_2 - \theta_1)} = \frac{e^{t(a\theta_2 + b)} - e^{t(a\theta_1 + b)}}{t[(a\theta_2 + b) - (a\theta_1 + b)]}.$$

The method of moment-generating functions are also effective in dealing with a random variable generating by many independent random variables.

Theorem 6.11 *If $Y_1, Y_2, \ldots, Y_n$ are independent random variables, and $U = Y_1 + Y_2 + \cdots + Y_n$, then*

$$m_U(t) = m_{Y_1}(t) \times \cdots \times m_{Y_n}(t).$$

Proof By independence,

$$m_U(t) = E\left(e^{tU}\right) = E\left(e^{t(Y_1 + \cdots + Y_n)}\right) = E\left(e^{tY_1}\right) \ldots E\left(e^{tY_n}\right) = m_{Y_1}(t) \times \cdots \times m_{Y_n}(t).$$

□

Example 6.12 *Let $Y_1, \ldots, Y_n$ be independent random variables such that*

- *each $Y_i \sim B(1, p)$, then $U = Y_1 + \cdots + Y_n \sim B(n, p)$, since*

$$m_U(t) = m_{Y_1}(t) \times \cdots \times m_{Y_n}(t) = \left[pe^t + (1 - p)\right]^n.$$

- *each $Y_i \sim \chi^2(\nu_i)$, then $U = Y_1 + \cdots + Y_n \sim \chi^2(\nu)$ with $\nu = \sum_{i=1}^n \nu_i$, since*

$$m_U(t) = m_{Y_1}(t) \times \cdots \times m_{Y_n}(t) = \prod_{i=1}^{n} (1 - 2t)^{-\nu_i/2} = (1 - 2t)^{-\nu/2}.$$

Example 6.13 *Let Y_1 and Y_2 are two independent random variables such that $Y_i \sim$ Possion(λ_i), $i = 1, 2$, find (1) the pdf of $Y_1 + Y_2$; (2) the conditional distribution of Y_1 given that $Y_1 + Y_2 = n$.*

Solution. (1) Note that

$$m_{Y_1 + Y_2}(t) = m_{Y_1}(t) m_{Y_2}(t) = e^{\lambda_1(e^t - 1)} e^{\lambda_2(e^t - 1)} = e^{(\lambda_1 + \lambda_2)(e^t - 1)},$$

which implies that $Y_1 + Y_2 \sim$ Possion(λ) with $\lambda = \lambda_1 + \lambda_2$, so

$$p_{Y_1 + Y_2}(n) = P(Y_1 + Y_2 = n) = \frac{\lambda^n}{n!} e^{-\lambda}.$$

(2) the conditional pdf of Y_1 given that $Y_1 + Y_2 = n$ is

$$\begin{aligned} p(k|Y_1+Y_2=n) = P(Y_1=k|Y_1+Y_2=n) &= \frac{P(Y_1=k,\ Y_1+Y_2=n)}{P(Y_1+Y_2=n)} \\ &= \frac{P(Y_1=k,\ Y_2=n-k)}{P(Y_1+Y_2=n)} \\ &= \frac{P(Y_1=k)P(Y_2=n-k)}{P(Y_1+Y_2=n)} \\ &= \frac{\frac{\lambda_1^k}{k!}e^{-\lambda_1}\frac{\lambda_2^{n-k}}{(n-k)!}e^{-\lambda_2}}{\frac{\lambda^n}{n!}e^{-\lambda}} \\ &= \frac{n!}{k!(n-k)!}\left(\frac{\lambda_1}{\lambda}\right)^k\left(\frac{\lambda_2}{\lambda}\right)^{n-k} \\ &= \binom{n}{k}p^k(1-p)^{n-k} \end{aligned}$$

if we set $p = \frac{\lambda_1}{\lambda}$. In other words, $Y_1|Y_1+Y_2=n \sim B(n,p)$.

A typical situation in statistics is that $Y_1, Y_2, \ldots Y_n$ are i.i.d. (independent and identical distributed), and we want to consider the sample mean variable

$$\overline{Y}_n = \frac{Y_1 + \cdots + Y_n}{n},$$

whose moment generating function is given by

$$m_{\overline{Y}_n}(t) = E\left(e^{t\frac{Y_1+\cdots+Y_n}{n}}\right) = E\left(e^{\frac{t}{n}Y_1}\right)\ldots E\left(e^{\frac{t}{n}Y_n}\right) = \left[m_{Y_1}\left(\frac{t}{n}\right)\right]^n. \tag{6–1}$$

Example 6.14 *If $Y_1, Y_2, \ldots, Y_n$ are i.i.d. and $Y_1 \sim N(\mu, \sigma^2)$, then*

$$m_{\overline{Y}_n}(t) = \left[m_{Y_1}\left(\frac{t}{n}\right)\right]^n = \left[\exp\left(\mu\frac{t}{n} + \frac{1}{2}\sigma^2\left(\frac{t}{n}\right)^2\right)\right]^n = \exp\left(\mu t + \frac{1}{2}\frac{\sigma^2}{n}t^2\right)$$

which, by uniqueness theorem, implies that $\overline{Y}_n \sim N(\mu, \overline{\sigma}_n^2)$, whose standard deviation is $\overline{\sigma}_n = \frac{\sigma}{\sqrt{n}}$. Alternatively, we have the normalized sample mean variable

$$U_n := \frac{\overline{Y}_n - \mu}{\overline{\sigma}_n} = \frac{\overline{Y}_n - \mu}{\sigma/\sqrt{n}}$$

follows the standard normal distribution, i.e., $U_n \sim N(0,1)$.

In general, even if we do not have any information on the distributions of Y_i's but only the expected value μ and variance σ^2, we still have

Theorem 6.15 (The Central Limit Theorem (CLT)) *Let $Y_1, \ldots, Y_n$ be i.i.d. random variables such that $\mu = E(Y_1)$ and $V(Y_1) = \sigma^2$, then we have*

$$U_n := \frac{\overline{Y}_n - \mu}{\sigma/\sqrt{n}} \Rightarrow N(0, 1^2), \quad as \quad n \to \infty.$$

Here $\Rightarrow$ means the distribution of U_n converges to the distribution of standard normal distribution. One can verify it by showing the convergence of the mgf of U_n to that of $N(0, 1^2)$. Also, in applications, we can take $U_n \approx N(0, 1)$ as n is sufficiently large, say, $n \geq 100$.

Proof We first do the simplest case when $\mu = 0$ and $\sigma = 1$. By Taylor expansion and Theorem 6.5, as t close to 0,

$$\begin{aligned} m_{Y_1}(t) = m_{Y_1}(0) + m'_{Y_1}(0)t + \frac{1}{2}m''_{Y_1}(0)t^2 + O(t^3) &= 1 + \mu t + \frac{1}{2}\sigma^2 t^2 + O(t^3) \\ &\approx 1 + \frac{1}{2}t^2. \end{aligned}$$

Note that $U_n = \sqrt{n}\overline{Y}_n$, and by (6–1), its mgf is given by

$$\begin{aligned} m_{U_n}(t) = E\left(e^{t\sqrt{n}\overline{Y}_n}\right) = m_{\overline{Y}_n}\left(t\sqrt{n}\right) &= \left[m_{Y_1}\left(\frac{t}{\sqrt{n}}\right)\right]^n \\ &\approx \left[1 + \frac{t^2}{2n}\right]^n \\ &= \left\{\left[1 + \frac{t^2}{2n}\right]^{\frac{2n}{t^2}}\right\}^{\frac{t^2}{2}} \\ &\to e^{\frac{t^2}{2}}, \ as \ n \to \infty. \end{aligned}$$

In other words, the moment generating function of U_n converges to that of $N(0, 1)$, which implies that $U_n \approx N(0, 1^2)$ for large n.

In the general case with arbitrary values for μ and σ^2, we first normalize each Y_i, that is, $Z_i = \frac{Y_i - \mu}{\sigma}$ for $i = 1, \ldots, n$, then it is easy to see that

$$\overline{Z}_n = \frac{Z_1 + \cdots + Z_n}{n} = \frac{\overline{Y}_n - \mu}{\sigma}.$$

From the above simple situation,

$$\frac{\overline{Z}_n}{1/\sqrt{n}} \approx N(0, 1^2), \quad \text{as} \quad n \to \infty,$$

which is exactly

$$U_n := \frac{\overline{Y}_n - \mu}{\sigma/\sqrt{n}} \approx N(0,1^2), \quad \text{as} \quad n \to \infty.$$

□

7 Sampling Distributions

In real-life, we are often interested in the sampling process.

Mathematically, let us fix a sample space S with Probability $P(\cdot)$.

Definition 7.1 *A random sample is a collection of random variables $Y_1, Y_2, \ldots, Y_n$ (defined on the sample space S) that are independent and identically distributed (iid).*

Example 7.2 *Flip a coin for n times, and let*

$$Y_i = \begin{cases} 1, & \text{if the } i\text{-th flip is head,} \\ 0, & \text{otherwise.} \end{cases}$$

Then clearly $Y_1, Y_2, \ldots, Y_n$ is a random sample.

Definition 7.3 *A statistic is a function of a random sample. For instance,*

- *Sample mean:*
$$\overline{Y}_n = \frac{Y_1 + Y_2 + \cdots + Y_n}{n}$$
- *Sample variance:*
$$S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \mu)^2,$$
where $\mu = E(Y_i)$.
- *Maximum and Minimum:*
$$Y_{max} = \max\{Y_1, Y_2, \ldots, Y_n\}, \quad Y_{min} = \min\{Y_1, Y_2, \ldots, Y_n\}.$$

Given a random sample $Y_1, Y_2, \ldots, Y_n$ with mean $\mu = E(Y_i)$ and variance $\sigma^2 = V(Y_i)$, we shall consider the sample mean:

$$\overline{Y}_n = \frac{Y_1 + Y_2 + \cdots + Y_n}{n}.$$

As we know,

$$E(\overline{Y}_n) = \mu, \quad V(\overline{Y}_n) = \frac{\sigma^2}{n}.$$

7.1 Normally Distributed Sample Mean

Let us consider a normally distributed random sample $Y_1, Y_2, \ldots, Y_n$ with mean μ and variance σ^2, that is, each $Y_i \sim N(\mu, \sigma^2)$, and $Y_1, \ldots, Y_n$ are independent.

By Example 6.14, the sample mean of size n

$$\overline{Y}_n = \frac{Y_1 + Y_2 + \cdots + Y_n}{n} \sim N\left(\mu, \frac{\sigma^2}{n}\right),$$

or after normalization,

$$U_n := \frac{\overline{Y}_n - \mu}{\sigma/\sqrt{n}} \sim N(0, 1^2).$$

Example 7.4 *Independent random samples of $n = 100$ observations each are drawn from normal populations. The parameters of these populations are:*

- *Population 1: $\mu_1 = 300$ and $\sigma_1 = 60$;*
- *Population 2: $\mu_2 = 290$ and $\sigma_2 = 80$.*

Find the probability that the mean of sample 1 is greater than the mean of sample 2 by more than 20.

Solution. Let $\{X_i\}_{1\le i\le n}$ and $\{Y_i\}_{1\le i\le n}$ be the sample variables of population 1 and 2 respectively. Then

$$\overline{X}_n \sim N(\mu_1, \sigma_1^2/n), \quad \overline{Y}_n \sim N(\mu_2, \sigma_2^2/n),$$

and hence

$$\overline{X}_n - \overline{Y}_n \sim N(\mu_1 - \mu_2, (\sigma_1^2 + \sigma_2^2)/n) = N(300 - 290, (60^2 + 80^2)/100) = N(10, 10^2).$$

Therefore,

$$P(\overline{X}_n - \overline{Y}_n \ge 20) = \int_{20}^{\infty} \frac{1}{10\sqrt{2\pi}} e^{-\frac{(z-10)^2}{200}} dz \approx 16\%.$$

Example 7.5 *A bottling machine can be regulated so that it discharges an average of μ ounces per bottle. It has been observed that the amount of fill dispensed by the machine is normally distributed with $\sigma = 1.0$ ounce. A sample of $n = 9$ filled bottles is randomly selected from the output of the machine on a given day (all bottled with the same machine setting), and the ounces of fill are measured for each.*

(1) *Find the probability that the sample mean will be within 1/3 ounce of the true mean μ for the chosen machine setting.*

(2) *How many observations should be included in the sample if we want the sample mean be within 1/3 ounce of μ with probability 95%?*

Solution. In this example, we have a normally distributed random sample $Y_1, \ldots, Y_9$ of size $n = 9$, with unknown mean μ, and variance $\sigma^2 = 1^2$. Then

$$U_9 = \frac{\overline{Y}_9 - \mu}{1/\sqrt{9}} = 3(\overline{Y}_9 - \mu) \sim N(0, 1^2).$$

Then our target probability is

$$P(|\overline{Y}_9 - \mu| \leq 1/3) = P(|\overline{Y}_9 - \mu| \leq 1/3\sigma) = P(|U_9 - 0| \leq \sigma) = 68\%$$

by the 68-95-99.7 rule of the standard normal distribution.

Similarly to (1), since $\sigma = 1$,

$$U_n = \frac{\overline{Y}_n - \mu}{\sigma/\sqrt{n}} = \sqrt{n}(\overline{Y}_n - \mu) \sim N(0, 1^2).$$

To require

$$P(|\overline{Y}_n - \mu| \leq 1/3) = P(|U_n - 0| \leq \frac{1}{3}\sqrt{n}) = 95\%,$$

we need $\frac{1}{3}\sqrt{n} = 2 \times 1$, that is, $n = 36$ at least.

7.2 The Central Limit Theorem

What can we say about the sample mean

$$\overline{Y}_n = \frac{Y_1 + \cdots + Y_n}{n},$$

if we are dealing with a random sample $Y_1, Y_2, \ldots, Y_n$ with mean μ and variance σ^2, but with arbitrary probability distribution (could be either discrete or continuous)?

Theorem 7.6 (The Central Limit Theorem (CLT)) *For sufficiently large n,*

$$U_n := \frac{\overline{Y}_n - \mu}{\sigma/\sqrt{n}} \approx N(0, 1^2).$$

Example 7.7 *Achievement test scores of all high school seniors in a state have mean 60 and variance 64. A random sample of $n = 100$ students from one large high school had a mean score of 57.6. Is there evidence to suggest that this high school is inferior?*

Suppose that this high school is not inferior, then a random chosen student should follow the distribution given by the state statistics, that is, if we let Y_i to be the score of the i-th student, then $\mu = E(Y_i) = 60$ and $\sigma^2 = V(Y_i) = 64$. In experience, $n = 100$ is large enough to apply CLT. Then

$$U_{100} = \frac{\overline{Y}_{100} - 60}{\sqrt{64}/\sqrt{100}} = \frac{5}{4}(\overline{Y}_{100} - 60) \approx N(0, 1^2),$$

and hence

$$P(\overline{Y}_{100} > 57.6) \geq P(|\overline{Y}_{100} - 60| < 2.4) = P(|U_{100}| < 3) = 99.7\%,$$

and hence

$$P(\overline{Y}_{100} \leq 57.6) = 1 - P(\overline{Y}_{100} > 57.6) \leq 0.15\%,$$

which means that $\overline{Y}_{100} \leq 57.6$ is quite unlikely, and hence this high school is inferior.

Example 7.8 *The service times for customers coming through a checkout counter in a retail store are independent random variables with mean 1.5 minutes and variance 1.0. Approximate the probability that 100 customers can be served in less than 2 hours of total service time.*

Let Y_i be the service time of the i-th customer, then $\mu = E(Y_i) = 1.5$ and $\sigma^2 = V(Y_i) = 1.0$. By CLT,

$$U_{100} = \frac{\overline{Y}_{100} - 1.5}{1/\sqrt{100}} = 10\overline{Y}_{100} - 15 \sim N(0, 1^2),$$

and then the probability is

$$\begin{aligned} P(Y_1 + \cdots + Y_{100} \leq 120) = P(100\overline{Y}_{100} \leq 120) &= P(10\overline{Y}_{100} - 15 \leq -3) \\ &= P(U_{100} \leq -3) \\ &= \frac{1}{2}[1 - P(|U_{100} - 0| \leq 3)] \\ &= \frac{1}{2}(1 - 99.7\%) = 0.15\%. \end{aligned}$$

7.3 The Normal Approximation to the Binomial Distribution

The Normal Approximation to the Binomial Distribution is an important consequence of the Central Limit Theorem, which turns out to be very useful in real life.

In a binomial experiment with trial number n and success rate p, we let X_i be the outcome of the i-th trial, then

$$P(X_i = 1) = p, \quad \text{and} \quad P(X_i = 0) = 1 - p.$$

Therefore, $\mu = E(X_i) = p$ and $\sigma^2 = V(X_i) = p(1-p)$.

The number of successes is denoted by Y, which obeys the binomial distribution $Y \sim b(n, p)$. Clearly,

$$Y = X_1 + X_2 + \cdots + X_n = n\overline{X}_n,$$

where $\overline{X}_n$ is the sample mean of $X_1, \ldots, X_n$. By the central limit theorem,

$$U_n = \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} = \frac{Y/n - p}{\sqrt{p(1-p)}/\sqrt{n}} = \frac{Y - np}{\sqrt{np(1-p)}} \approx N(0, 1^2).$$

The above formula suggests that we can actually approximate the binomial distribution Y (which is discrete) by the normal distribution U_n (which is continuous), when n is large.

Example 7.9 *An airline company is considering a new policy of booking as many as 400 persons on an airplane that can seat only 378. Past studies have revealed that only 90% of the booked passengers actually arrive for the flight. Find the probability that not enough seats are available if 400 persons are booked.*

Let Y be the number of booked passengers who actually arrive, then $Y \sim b(400, 0.9)$. Then the probability for not enough seats is given by

$$P(Y > 378) = \sum_{k=379}^{400} P(Y = k) = \sum_{k=379}^{400} \binom{400}{k} 0.9^k 0.1^{400-k}.$$

To calculate the above formula by hand is impossible. Instead, we use the normal approximation,

$$U_{400} = \frac{Y - 400 \cdot 0.9}{\sqrt{400 \cdot 0.9(1-0.9)}} = \frac{Y - 360}{6} \approx N(0, 1^2),$$

then

$$\begin{aligned} P(Y > 378) &= P\left(U_{400} > \frac{378 - 360}{6}\right) \\ &= P(U_{400} > 3) \approx \frac{1}{2}[1 - P(U_{400} \le 3)] = \frac{1}{2}(1 - 99.7\%) = 0.15\%. \end{aligned}$$

Example 7.10 *Given a coin, let p denote the probability to get head. Assume that we know a priori that $p = 0.4$. How many times do we need to flip the coin so that we can convince other people that this coin is unfair, in fact, biased for tail, with 97.5% confidence?*

Let Y be the number of heads out of n flips, then $Y \sim B(n,p)$ with $p = 0.4$, and hence

$$U_n = \frac{Y - np}{\sqrt{np(1-p)}} \approx N(0, 1^2),$$

To assure people that the coin is unfair, we just need to make sure

$$P(Y/n < 0.5) = P\left(U_n < \frac{\sqrt{n}(0.5-p)}{\sqrt{p(1-p)}}\right) \geq 97.5\%,$$

that is,

$$\frac{\sqrt{n}(0.5-p)}{\sqrt{p(1-p)}} = 2, \quad \text{and hence } n = \frac{4p(1-p)}{(0.5-p)^2}.$$

Plug in $p = 0.4$, we get $n = 96$.

1. Roll a fair dice 1,000 times and add up the values. Use CLT to estimate the probability that the sum is at least 3,600.

2. Shear strength measurements for spot welds have been found to have standard deviation 10 psi (pounds per square inch). If 100 test welds are to measured, what is the probability that the sample mean will be within 1 psi of the true mean?

3. Given a coin, let p denote the probability to get head. Assume that we know a priori that $p = 0.4$. How many times do we need to flip the coin so that we can convince other people that this coin is unfair, in fact, biased for tail, with 97.5% confidence?

4. The time to process orders at the service counter of a pharmacy store are exponentially distributed with mean 10 minutes. If 100 customers visit the counter in a day, what is the probability that at least half of them need to wait more than 10 minutes?