Math 382 Lecture Notes
Probability and Statistics

Anwar Hossain and Oleg Makhnin

July 29, 2011

Contents

Chapter 1

Probability in the World Around Us

Probability theory is a tool to describe uncertainty. In science and engineering, the world around us is described by mathematical models. Most mathematical models are *deterministic*, that is, the model output is supposed to be known uniquely once all the inputs are specified. As an example of such model, consider the Newton's law $F = ma$ connecting the force F acting on an object of mass m resulting in the acceleration a. Once F and m are specified, we can determine exactly the object's acceleration.[1]

What is wrong with this model from practical point of view? Most obviously, the inputs in the model (F and m) are not precisely known. They may be measured, but there's usually a measurement error involved. Also, the model itself might be approximate or might not take into account all the factors influencing the model output. Finally, roundoff errors are sure to crop up during the calculations. Thus, our predictions of planetary motions, say, will be imperfect in the long run and will require further corrections as more recent observations become available.

At the other end of the spectrum, there are some phenomena that seem to completely escape any attempts at the rational description. These are random phenomena - ranging from lotteries to the heat-induced motion of the atoms. Upon closer consideration, there are still some laws governing these phenomena. However, they would not apply on case by case basis, but

[1] Now you are to stop and think: what are the factors that will make this model more uncertain?

rather to the results of many repetitions. For example, we cannot predict the result of one particular lottery drawing, but we can calculate probabilities of certain outcomes. We cannot describe the velocity of a single atom, but we can say something about the behavior of the velocities in the ensemble of all atoms.

This is the stuff that *probabilistic* models are made of. Another example of a field where probabilistic models are routinely used is actuarial science. It deals with lifetimes of humans and tries to predict how long any given person is expected to live, based on other variables describing the particulars of his/her life. Of course, this expected life span is a poor prediction when applied to any given person, but it works rather well when applied to many persons. It can help to decide the rates the insurance company should charge for covering any given person.

Today's science deals with enormously complex models, for example, the models of Earth's climate (there are many of them available, at different levels of complexity and resolution). The models should also take into account the uncertainties from many sources, including our imperfect knowledge of the current state of Earth, our imperfect understanding of all physical processes involved, and the uncertainty about future scenarios of human development.[2]

Understanding and communicating this uncertainty is greatly aided by the knowledge of the rules of probability.

[2]Not the least, our ability to calculate the output of such models is also limited by the current state of computational science.

Chapter 2

Probability

2.1 What is Probability

Probability theory is the branch of mathematics that studies the possible outcomes of given events together with the outcomes' relative likelihoods and distributions. In common usage, the word "probability" is used to mean the chance that a particular event (or set of events) will occur expressed on a linear scale from 0 (impossibility) to 1 (certainty), also expressed as a percentage between 0 and 100%. The analysis of data (possibly generated by probability models) is called **statistics**.

Probability is a way of summarizing the uncertainty of statements or events. It gives a numerical measure for the degree of certainty (or degree of uncertainty) of the occurrence of an event.

Another way to define probability is the ratio of the number of favorable outcomes to the total number of all possible outcomes. This is true if the outcomes are assumed to be equally likely. The collection of all possible outcomes is called the **sample space**.

If there are n total possible outcomes in a sample space $\mathcal{S}$, and m of those are favorable for an event A, then probability of event A is given as

$$P(A) = \frac{\text{number of favorable outcomes}}{\text{total number of possible outcomes}} = \frac{n(A)}{n(S)} = \frac{m}{n}$$

Example 2.1. *Find the probability of getting a 3 or 5 while throwing a die.*

Solution. Sample space $\mathcal{S} = \{1, 2, 3, 4, 5, 6\}$ and event $A = \{3, 5\}$.
We have $n(A) = 2$ and $n(S) = 6$.
So, $P(A) = n(A)/n(S) = 2/6 = 0.3333$ □

Axioms of Probability

All probability values are positive numbers not greater than 1, i.e. $0 \leq p \leq 1$. An event that is not likely to occur or impossible has probability zero, while an event that's certain to occur has probability one.

Examples:- P(A pregnant human being a female) = 1
P(A human male being pregnant) = 0.

Definition 2.1.

Random Experiment: A random experiment is the process of observing the outcome of a chance event.

Outcome: The elementary outcomes are all possible results of the random experiment.

Sample Space(SS): The sample space is the set or collection of all the outcomes of an experiment and is denoted by $\mathcal{S}$.

Example 2.2.
a) Flip a coin once, then the sample space is: $\mathcal{S} = \{H, T\}$
b) Flip a coin twice, then the sample space is: $\mathcal{S} = \{HH, HT, TH, TT\}$

We want to assign a numerical weight or **probability** to each outcome. We write the probability of A_i as $P(A_i)$. For example, in our coin toss experiment, we may assign $P(H) = P(T) = 0.5$. Each outcome comes up half the time.

2.2 Review of set notation

Definition 2.2. Complement

The complement of event A is the set of all outcomes in a sample that are not included in the event A. The complement of event A is denoted by A'.

If the probability that an event occurs is p, then the probability that the event does not occur is $q = (1 - p)$. i.e. probability of the complement of an event $= 1-$ probability of the event.

$$\text{i.e. } P(A') = 1 - P(A)$$

Example 2.3. *Find the probability of not getting a 3 or 5 while throwing a die.*

Solution. Sample space $\mathcal{S} = \{1, 2, 3, 4, 5, 6\}$ and event $B = \{1, 2, 4, 6\}$.

$$n(B) = 4 \text{ and } n(S) = 6$$

$$\text{So, } P(B) = n(B)/n(S) = 4/6 = 0.6667$$

On the other hand, A (described in Example 2.1) and B are complementary events, i.e. $B = A'$.

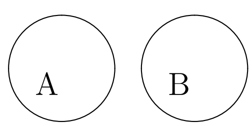
So, $P(B) = P(A') = 1 - P(A) = 1 - 0.3333 = 0.6667$ □

Definition 2.3. Intersections of Events

The event $A \cap B$ is the **intersection** of the events A and B and consists of outcomes that are contained within both events A and B. The probability of this event, is the probability that both events A and B occur [but not necessarily at the same time]. In the future, we will abbreviate intersection as AB.

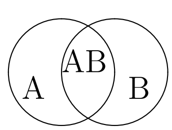
Definition 2.4. Mutually Exclusive Events

Two events are said to be mutually exclusive if $AB = \emptyset$ (i.e. they have empty intersection) so that they have no outcomes in common.



Definition 2.5. Unions of Events

The event $A \cup B$ is the **union** of events A and B and consists of the outcomes that are contained within at least one of the events A and B. The probability of this event $P(A \cup B)$, is the probability that at least one of the events A and B occurs.



Venn diagram

Venn diagram is often used to illustrate the relations between sets (events). The sets A and B are represented as circles; operations between them (intersections, unions and complements) can also be represented as parts of the diagram. The entire sample space $\mathcal{S}$ is the bounding box. See Figure 2.1

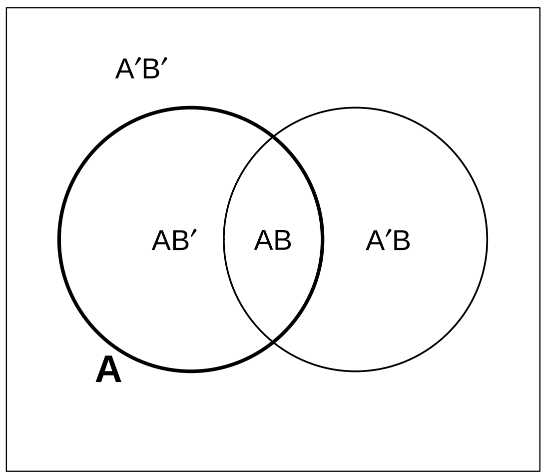


Figure 2.1: Venn diagram of events A (in bold) and B, represented as insides of circles, and various intersections

Example 2.4. Set notation

Suppose a set $\mathcal{S}$ consists of points labeled 1, 2, 3 and 4. We denote this by $\mathcal{S} = \{1, 2, 3, 4\}$.
If $A = \{1, 2\}$ and $B = \{2, 3, 4\}$, then A and B are subsets of $\mathcal{S}$, denoted by $A \subset \mathcal{S}$ and $B \subset \mathcal{S}$ (B is contained in $\mathcal{S}$). We denote the fact that 2 is an element of A by $2 \in A$.

The union of A and B, $A \cup B = \{1, 2, 3, 4\}$. If $C = \{4\}$, then $A \cup C = \{1, 2, 4\}$. The intersection $A \cap B = AB = \{2\}$. The complement $A' = \{3, 4\}$. $\square$

Distributive laws

$$A(B \cup C) = AB \cup AC$$

and

$$A \cup (BC) = (A \cup B)(A \cup C)$$

De Morgan's Law

$$(A \cup B)' = A'B'$$

$$(AB)' = A' \cup B'$$

Exercises

2.1.

Use the Venn diagrams to illustrate Distributive laws and De Morgan's law.

2.2.

Simplify the following (Draw the Venn diagrams to visualize)

a) $(A')'$

b) $(AB)' \cup A$

c) $(AB) \cup (AB')$

d) $(A \cup B \cup C)B$

2.3.

Represent by set notation the following events

a) both A and B occur

b) exactly one of A, B occurs

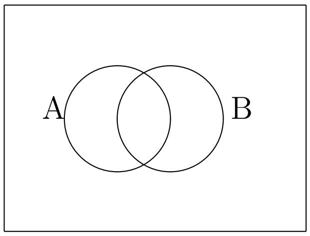
c) at least one of A, B, C occurs

d) at most one of A, B, C occurs

2.4.

Out of all items sent for refurbishing, 40% had mechanical defects, 50% had electrical defects, and 25% had both.
Denoting $A = \{\text{an item has a mechanical defect}\}$ and
$B = \{\text{an item has an electrical defect}\}$, fill the probabilities into the Venn

diagram and determine the quantities listed below.



a) $P(A)$

b) $P(AB)$

c) $P(A'B)$

d) $P(A'B')$

e) $P(A \cup B)$

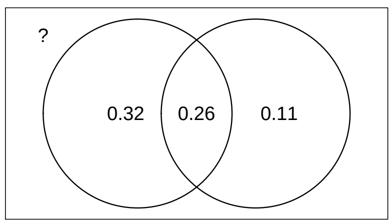
f) $P(A' \cup B')$

g) $P([A \cup B]')$

Ways to represent probabilities:

- **Venn diagram**

  We may write the probabilities inside the elementary pieces within a Venn diagram. For example, $P(AB') = 0.32$ and $P(A) = P(AB) + P(AB')$ [why?] $= 0.58$ The relative sizes of the pieces do not have to match the numbers.



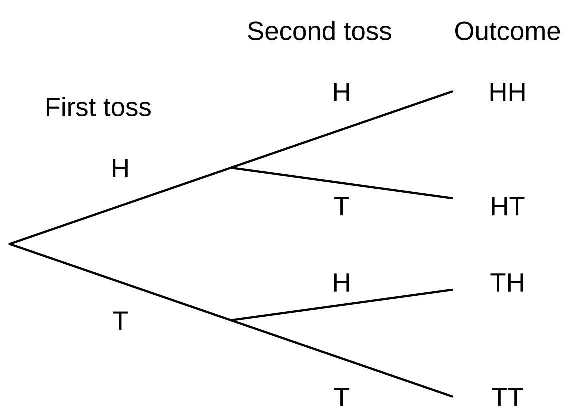
- **Two-way table**

  This is a popular way to represent statistical data. The cells of the table correspond to the intersections of row and column events. Note that the contents of the table add up accross rows and columns of the table. The bottom-right corner of the table contains $P(\mathcal{S}) = 1$

	B	B'	
A	0.26	0.32	0.58
A'	0.11	?	0.42
	0.37	0.63	1

- **Tree diagram**

  A tree diagram may be used to show the sequence of choices that lead to the complete description of outcomes. For example, when tossing two coins, we may represent this as follows



A tree diagram is also often useful for representing conditional probabilities (see below).

2.3 Types of Probability

There are three ways to define probability, namely classical, empirical and subjective probability.

Definition 2.6. Classical probability

A classical or theoretical probability is used when each outcome in a sample space is equally likely to occur. The classical probability for an event A is given by

$$P(A) = \frac{\text{Number of outcomes in A}}{\text{Total number of outcomes in } \mathcal{S}}$$

Example 2.5.

Roll a die and observe that $P(A) = P(\text{rolling a 3}) = 1/6$.

Definition 2.7. Empirical probability

Empirical (or Statistical) Probability is based on observed data. The empirical probability of an event A is the relative frequency of event A, that is

$$P(A) = \frac{\text{Frequency of event A}}{\text{Total frequency}}$$

Example 2.6.

Based on genetics, the proportion of male children among all children conceived should be around 0.5. However, based on the statistics from a large number of live births, the probability that a child being born is male is about 0.512. □

Example 2.7.

The following are the counts of fish of each type, that you have caught before.

Fish Types	Blue gill	Red gill	Crappy	Total
Number of times caught	13	17	10	40

Estimate the probability that the next fish you catch will be a Blue gill.

$$P(\text{Blue gill}) = \frac{13}{40} = 0.325 \quad \square$$

The empirical probability definition has a weakness that it depends on the results of a particular experiment. The next time this experiment is repeated, you are likely to get a somewhat different result.

However, as an experiment is repeated many times, the empirical probability of an event, based on the combined results, approaches the theoretical probability of the event.

Subjective Probability: Subjective probabilities result from intuition, educated guesses, and estimates. For example: given a patient's health and extent of injuries a doctor may feel that the patient has a 90% chance of a full recovery.

Regardless of the way probabilities are defined, they always follow the same laws, which we will explore starting with the following Section.

2.4 Laws of Probability

As we have seen in the previous section, the probabilities are not always based on the assumption of equal outcomes.

Definition 2.8. Axioms of Probability

For an experiment with a sample space $\mathcal{S} = \{e_1, e_2, \ldots, e_n\}$ we can assign probabilities $P(e_1), P(e_2), \ldots, P(e_n)$ provided that

a) $0 \leq P(e_i) \leq 1$

b) $P(\mathcal{S}) = \sum_{i=1}^{n} P(e_i) = 1$.

If a set (event) A consists of outcomes $e_1, e_2, \ldots, e_k\}$, then

$$P(A) = \sum_{i=1}^{k} P(e_i)$$

This definition just tells us which probability assignments are legal, but not necessarily which ones would work in practice. However, once we have assigned the probability to each outcome, they are subject to further rules which we will describe below.

Theorem 2.1. Complement Rule

For any event A,
$$P(A') = 1 - P(A) \tag{2.1}$$

Theorem 2.2. Addition Law

If A and B are two different events then
$$P(A \cup B) = P(A) + P(B) - P(A \cap B) \tag{2.2}$$

Proof. Consider the Venn diagram. $P(A \cup B)$ is the probability of the sum of all sample points in $A \cup B$. Now $P(A) + P(B)$ is the sum of probabilities of sample points in A and in B. Since we added up the sample points in $(A \cap B)$ twice, we need to subtract once to obtain the sum of probabilities in $(A \cup B)$, which is $P(A \cup B)$. □

Example 2.8. *Probability that John passes a Math exam is 4/5 and that he passes a Chemistry exam is 5/6. If the probability that he passes both exams is 3/4, find the probability that he will pass at least one exam.*

Solution. Let M = John passes Math exam, and C = John passes Chemistry exam.
$$P(\text{John passes at least one exam}) = P(M \cup C) =$$
$$= P(M) + P(C) - P(M \cap C) = 4/5 + 5/6 - 3/4 = 53/60$$
□

Corollary. If two events A and B are mutually exclusive, then $P(A \cup B) = P(A) + P(B)$.
This follows immediately from (2.2). Since A and B are mutually exclusive, $P(A \cap B) = 0$.

Example 2.9. *What is the probability of getting a total of 7 or 11, when two dice are rolled?*

Solution. Let A be the event that the total is 7 and B be the event that it is 11. The sample space for this experiment is

$$\mathcal{S} = \{(1,1),(1,2),......,(2,1),(2,2),.........,(6,6)\}, \quad n(\mathcal{S}) = 36$$

$$A = \{(1,6),(2,5),(3,4),(4,3),(5,2),(6,1)\} \text{ and } n(A) = 6.$$

So, $P(A) = 6/36 = 1/6$.

$$B = \{(5,6),(6,5)\} \text{ and } n(B) = 2$$

So, $P(B) = 2/36 = 1/18$.
Since we cannot have a total equal to both 7 and 11, A and B are mutually exclusive, i.e. $P(A \cap B) = 0$.
So, we have $P(A \cup B) = P(A) + P(B) = 1/6 + 1/18 = 2/9$. □

Exercises

2.5.

Two cards are drawn from a pack, without replacement. What is the probability that both are greater than 2 and less than 8?

2.6.

A permutation of the word "white" is chosen at random. Find the probability that it begins with a vowel. Also find the probability that it ends with a consonant.

2.7.

As a foreign language, 40% of the students took Spanish and 30% took French, while 60% took at least one of these languages. What percent of students took both Spanish and French?

2.8.

Find the probability that a leap year will have 53 Sundays.

2.5 Counting Rules useful in Probability

In some experiments it is helpful to list the elements of the sample space systematically by means of a tree diagram, see page 15.

In many cases, we shall be able to solve a probability problem by counting the number of points in the sample space without actually listing each element.

Theorem 2.3. Multiplication principle

If one operation can be performed in n_1 ways, and if for each of these a second operation can be performed in n_2 ways, then the two operations can be performed together in $n_1 n_2$ ways.

Example 2.10. *How large is the sample space when a pair of dice is thrown?*

Solution. The first die can be thrown in $n_1 = 6$ ways and the second in $n_2 = 6$ ways. Therefore, the pair of dice can land in $n_1 n_2 = 36$ possible ways. □

Theorem 2.3 can naturally be extended to more than two operations: if we have n_1, n_2,...,n_k consequent choices, then the total number of ways is $n_1 n_2 \cdots n_k$.

The term *permutations* refers to an arrangement of objects when the order matters (for example, letters in a word).

Theorem 2.4. Permutations

The number of permutations of n distinct objects taken r at a time is

$$_nP_r = \frac{n!}{(n-r)!}$$

Example 2.11.

From among ten employees, three are to be selected to travel to three out-of-town plants A, B, and C, one to each plant. Since the plants are located in different cities, the order in which the employees are assigned to the plants is an important consideration. In how many ways can the assignments be made?

Solution. Because order is important, the number of possible distinct assignments is

$$_{10}P_3 = \frac{10!}{7!} = 10(9)(8) = 720.$$

In other words, there are ten choices for plant A, but then only nine for plant B, and eight for plant C. This gives a total of 10(9)(8) ways of assigning employees to the plants. □

The term *combination* refers to the arrangement of objects when order does not matter. For example, choosing 4 books to buy at the store in any order will leave you with the same set of books.

Theorem 2.5. Combinations

The number of distinct subsets or combinations of size r that can be selected from n distinct objects, $(r \leq n)$, is given by

$$\binom{n}{r} = \frac{n!}{r!\,(n-r)!} \tag{2.3}$$

Proof. Start with picking ordered sets of size r. This can be done in $_nP_r = \frac{n!}{(n-r)!}$ ways. However, many of these are the re-orderings of the same basic set of objects. Each distinct set of r objects can be re-ordered in $_rP_r = r!$ ways. Therefore, we need to divide the number of permutations $_nP_r$ by $r!$, thus arriving at the equation (2.3). □

Example 2.12.

In the previous example, suppose that three employees are to be selected from among the ten available to go to the same plant. In how many ways can this selection be made?

Solution. Here, order is not important; we want to know how many subsets of size $r = 3$ can be selected from $n = 10$ people. The result is

$$\binom{10}{3} = \frac{10!}{3!\,7!} = \frac{10(9)(8)}{3(2)(1)} = 120$$

□

Example 2.13.

A package of six light bulbs contains 2 defective bulbs. If three bulbs are selected for use, find the probability that none of the three is defective.

Solution. P(none are defective) =

$$= \frac{\text{number of ways 3 nondefectives can be chosen}}{\text{total number of ways a sample of 3 can be chosen}} = \frac{\binom{4}{3}}{\binom{6}{3}} = \frac{1}{5}$$

□

Example 2.14.
In a poker hand consisting of 5 cards, find the probability of holding 2 aces and 3 jacks.

Solution. The number of ways of being dealt 2 aces from 4 is $\binom{4}{2} = 6$ and the number of ways of being dealt 3 jacks from 4 is $\binom{4}{3} = 4$.
The total number of 5-card poker hands, all of which are equally likely is

$$\binom{52}{5} = 2,598,960$$

Hence, the probability of getting 2 aces and 3 jacks in a 5-card poker hand is

$$P(C) = \frac{(6)(4)}{2,598,960}$$

□

Example 2.15.

A university warehouse has received a shipment of 25 printers, of which 10 are laser printers and 15 are inkjet models. If 6 of these 225 are selected at random to be checked by a particular technician, what is the probability that exactly 3 of these selected are laser printers? At least 3 inkjet printers?

Solution. First choose 3 of the 15 inkjet and then 3 of the 10 laser printers. There are $\binom{15}{3}$ and $\binom{10}{3}$ ways to do it, and therefore

$$P(\text{exactly 3 of the 6}) = \frac{\binom{15}{3}\binom{10}{3}}{\binom{25}{6}} = 0.3083$$

(b) P(at least 3)

$$= \frac{\binom{15}{3}\binom{10}{3}}{\binom{25}{6}} + \frac{\binom{15}{4}\binom{10}{2}}{\binom{25}{6}} + \frac{\binom{15}{5}\binom{10}{1}}{\binom{25}{6}} + \frac{\binom{15}{6}\binom{10}{0}}{\binom{25}{6}} = 0.8530$$

□

Theorem 2.6. Partitions

The number of ways of partitioning n distinct objects into k groups containing $n_1, n_2, \ldots, n_k$ objects respectively, is

$$\frac{n!}{n_1!\, n_2! \ldots n_k!}$$

where $\sum_{i=1}^{k} n_i = n$.

Note that when there are $k = 2$ groups, we will obtain combinations.

Example 2.16.
Consider 10 people to be split into 3 groups to be assigned to 3 plants. If we are to send 5 people to Plant A, 3 people to Plant B, and 2 people to Plant C, then the total number of assignments is

$$\frac{10!}{5!\, 3!\, 2!} = 2520 \quad \square$$

Exercises

2.9.
An incoming lot of silicon wafers is to be inspected for defectives by an engineer in a microchip manufacturing plant. Suppose that, in a tray containing 20 wafers, 4 are defective. Two wafers are to be selected randomly for inspection. Find the probability that neither is defective.

2.10.
A person draws 5 cards from a shuffled pack of cards. Find the probability that the person has at least 3 aces. Find the probability that the person has at least 4 cards of the same suit.

2.11.
Three people enter the elevator on the basement level. The building has 7 floors. Find the probability that all three get off at different floors.

2.12.
In a group of 7 people, each person shakes hands with every other person. How many handshakes did occur?

2.13.

A marketing director considers that there's "overwhelming agreement" in a 5-member focus group when either 4 or 5 people like or dislike the product.[1] If, in fact, the product's popularity is 50% (so that all outcomes are equally likely), what is the probability that the focus group will be in "overwhelming agreement" about it? Is the marketing director making a judgement error in declaring such agreement "overwhelming"?

2.14.

A die is tossed 5 times. Find the probability that we will have 4 of a kind.

2.15.

In a lottery, 6 numbers are drawn out of 45. You hit a jackpot if you guess all 6 numbers correctly, and get $400 if you guess 5 numbers out of 6. What are the probabilities of each of those events?

2.16.

A fuse box contains 20 fuses of which 5 are defective. You grab 2 fuses at random. What is the probability that both fuses are defective?

2.17.

How many distinct ways are there to permute the letters in the word PROBABILITY?

2.18.

Eight tires of different brands are ranked 1 to 8 (best to worst) according to mileage performance. If four of these tires are chosen at random by a customer, find the probability that the best tire among the four selected by the customer is actually ranked third among the original eight.

Pascal's triangle and binomial coefficients

Long before Pascal, this triangle has been described by several Oriental scholars. It was used in the budding discipline of probability theory by the French mathematician Blaise Pascal (1623-1662). The construction begins by writing 1's along the sides of a triangle and then filling it up row by row so that each number is a sum of the two numbers immediately above it.

1
1 1
1 2 1
1 3 3 1
1 4 6 4 1
1 5 10 10 5 1

A step in construction

The number in each cell represents the number of downward routes from the vertex to that point (can you explain why?). It is also a number of ways to choose r objects out of n (can you explain why?), that is, $\binom{n}{r}$.

1
1 1
1 2 1
1 3 3 1
1 4 6 4 1
1 5 10 10 5 1
1 6 15 20 15 6 1
1 7 21 35 35 21 7 1
1 8 28 56 70 56 28 8 1
1 9 36 84 126 126 84 36 9 1
1 10 45 120 210 252 210 120 45 10 1

The first 10 rows

The combinations numbers are also called *binomial coefficients* and are seen in Calculus. Namely, they are the terms in the expansion

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

Note that, if you let $a = b = 1/2$, then on the right-hand side of the sum you will get the probabilities

$$P(a \text{ is chosen } r \text{ times and } b \text{ is chosen } n-r \text{ times}) = \frac{\binom{n}{r}}{2^n}$$

and on the left-hand side you will have 1 (the total of all probabilities).

2.6 Conditional probability and independence

Humans often have to act based on incomplete information. If your boss has looked at you gloomily, you might conclude that something's wrong with your job performance. However, if you know that she just suffered some losses in the stock market, this extra information may change your assessment of the situation. Conditional probability is a tool for dealing with additional information like this.

Conditional probability is the probability of an event occurring given the knowledge that another event has occurred. The conditional probability of event A occurring, given that event B has occurred is denoted by $P(A|B)$ and is read "probability of A given B".

Definition 2.9. Conditional probability

The conditional probability of event A given B is

$$P(A \,|\, B) = \frac{P(A \cap B)}{P(B)} \text{ for } P(B) > 0 \tag{2.4}$$

Reduced sample space approach

In case when all the outcomes are equally likely, it is sometimes easier to find conditional probabilities directly, without having to apply equation (2.4). If we already know that B has happened, we need only to consider outcomes in B, thus reducing our sample space to B. Then,

$$P(A \,|\, B) = \frac{\text{Number of outcomes in } AB}{\text{Number of outcomes in } B}$$

For example, $P(\text{a die is } 3 \,|\, \text{ a die is odd}) = 1/3$ and
$P(\text{a die is } 4 \,|\, \text{ a die is odd}) = 0$.

Example 2.17.

Let $A = \{\text{a family has two boys}\}$ and $B = \{\text{a family of two has at least one boy}\}$
Find $P(A \,|\, B)$.

Solution. The event B contains the following outcomes: (B, B), (B, G) and (G, B). Only one of these is in A. Thus, $P(A \mid B) = 1/3$.

However, if I know that the family has two children, and I see one of the children and it's a boy, then the probability suddenly changes to 1/2. There is a subtle difference in the language and this changes the conditional probability![1] □

Statistical reasoning

Suppose I pick a card at random from a pack of playing cards, without showing you. I ask you to guess which card it is, and you guess the five of diamonds. What is the probability that you are right? Since there are 52 cards in a pack, and only one five of diamonds, the probability of the card being the five of diamonds is 1/52.

Next, I tell you that the card is red, not black. Now what is the probability that you are right? Clearly you now have a better chance of being right than you had before. In fact, your chance of being right is twice as big as it was before, since only half of the 52 cards are red. So the probability of the card being the five of diamonds is now 1/26. What we have just calculated is a conditional probability–the probability that the card is the five of diamonds, given that it is red.

If we let A stand for the card being the five of diamonds, and B stand for the card being red, then the conditional probability that the card is the five of diamonds given that it is red is written $P(A|B)$.

In our case, $P(A \cap B)$ is the probability that the card is the five of diamonds and red, which is 1/52 (exactly the same as P(A), since there are no black fives of diamonds!). P(B), the probability that the card is red, is 1/2. So the definition of conditional probability tells us that $P(A|B) = 1/26$, exactly as it should. In this simple case we didn't really need to use a formula to tell us this, but the formula is very useful in more complex cases.

If we rearrange the definition of conditional probability, we obtain the *multiplication rule* for probabilities:

$$P(A \cap B) = P(A|B)P(B) \tag{2.5}$$

The next concept, *statistical independence* of events, is very important.

[1] Always read the fine print!

Definition 2.10. Independence

The events A and B are called (statistically) independent if

$$P(A \cap B) = P(A)P(B) \tag{2.6}$$

Another way to express independence is to say that the knowledge of B occurring does not change our assessment of $P(A)$. This means that $P(A|B) = P(A)$. (The probability that a person is female given that he or she was born in March is just the same as the probability that the person is female.)

Equation (2.6) is often called *simplified multiplication rule* because it can be obtained from (2.5) by substituting $P(A|B) = P(A)$.

Example 2.18.

For a coin tossed twice, denote H_1 the event that we got Heads on the first toss, and H_2 is the Heads on the second. Clearly, $P(H_1) = P(H_2) = 1/2$. Then, counting the outcomes, $P(H_1H_2) = 1/4 = P(H_1)P(H_2)$, therefore H_1 and H_2 are independent events. This agrees with our intuition that the result of the first toss should not affect the chances for H_2 to occur. □

The situation of the above example is very common for repeated experiments, like rolling dice, or looking at random numbers etc.

Definition 2.10 can be extended to more than two events, but it's fairly difficult to describe.[2] However, it is often used in this context:

If events $A_1, A_2, ..., A_k$ are independent, then

$$P(A_1A_2...A_k) = P(A_1) \times P(A_2) \times ... \times P(A_k) \tag{2.7}$$

For example, if we tossed a coin 5 times, the probability that all are Heads is $P(H_1) \times P(H_2) \times ... \times P(H_5) = (1/2)^5 = 1/32$. However, this calculation also extends to outcomes with unequal probabilities.

Example 2.19.

Three bits (0 or 1 digits) are transmitted over a noisy channel, so they will be flipped independently with probability 0.1 each. What is the probability

[2] For example, the relation $P(ABC) = P(A)P(B)P(C)$ does not guarantee that the events A, B, C are independent.

that
a) At least one bit is flipped
b) *Exactly* one bit is flipped?

Solution. a) Using the complement rule, $P(\text{at least one}) = 1 - P(\text{none})$. If we denote F_k the event that kth bit is flipped, then $P(\text{no bits are flipped}) = P(F_1'F_2'F_3') = (1-0.1)^3$ due to independence. Then,

$$P(\text{at least one}) = 1 - 0.9^3 = 0.271$$

b) Flipping exactly one bit can be accomplished in 3 ways:

$$P(\text{exactly one}) = P(F_1F_2'F_3') + P(F_1'F_2F_3') + P(F_1'F_2'F_3) = 3(0.1)(1-0.1)^2 = 0.243$$

It is slightly smaller than the one in part (a). □

Self-test questions

Suppose you throw two dice, one after the other.

a) What is the probability that the first die shows a 2?

b) What is the probability that the second die shows a 2?

c) What is the probability that both dice show a 2?

d) What is the probability that the dice add up to 4?

e) What is the probability that the dice add up to 4 given that the first die shows a 2?

f) What is the probability that the dice add up to 4 and the first die shows a 2?

Answers:

a) The probability that the first die shows a 2 is 1/6.

b) The probability that the second die shows a 2 is 1/6.

c) The probability that both dice show a 2 is $(1/6)(1/6) = 1/36$ (using the special multiplication rule, since the rolls are independent).

d) For the dice to add up to 4, there are three possibilities–either both dice show a 2, or the first shows a 3 and the second shows a 1, or the first shows a 1 and the second shows a 3. Each of these has a probability of (1/6)(1/6) 3= 1/36 (using the special multiplication rule, since the rolls are independent). Hence the probability that the dice add up to 4 is $1/36 + 1/36 + 1/36 = 3/36 = 1/12$ (using the special addition rule, since the outcomes are mutually exclusive).

e) If the first die shows a 2, then for the dice to add up to 4 the second die must also show a 2. So the probability that the dice add up to 4 given that the first shows a 2 is 1/6.

f) Note that we cannot use the simplified multiplication rule here, because the dice adding up to 4 is not independent of the first die showing a 2. So we need to use the full multiplication rule. This tells us that probability that the first die shows a 2 and the dice add up to 4 is given by the probability that the first die shows a 2, multiplied by the probability that the dice add up to 4 given that the first die shows a 2. This is $(1/6)(1/6) = 1/36$.
Alternatively, see part (c). □

Example 2.20. Trees in conditional probability
Suppose we are drawing marbles from a bag that initially contains 7 red and 3 green marbles. The drawing is without replacement, that is after we draw the first marble, we do not put it back. Let's denote the events

$$R_1 = \{ \text{ the first marble is red } \} \qquad R_2 = \{ \text{ the second marble is red } \}$$

$$G_1 = \{ \text{ the first marble is green } \} \qquad \text{and so on.}$$

Let's fill out the tree representing the consecutive choices. See Figure 2.2. The conditional probability $P(R_2 \mid R_1)$ can be obtained directly from reasoning that after we took the first red marble, there remain 6 red and 3 green marbles. On the other hand, we could use the formula (2.4) and get

$$P(R_2 \mid R_1) = \frac{P(R_2 R_1)}{P(R_1)} = \frac{42/90}{7/10} = \frac{2}{3}$$

where the probability $P(R_2 R_1)$ – same as $P(R_1 R_2)$ – can be obtained from counting the outcomes

$$P(R_1 R_2) = \frac{\binom{7}{2}}{\binom{10}{2}} = \frac{\frac{7*6}{2*1}}{\frac{10*9}{2*1}} = \frac{42}{90} = \frac{7}{15}$$

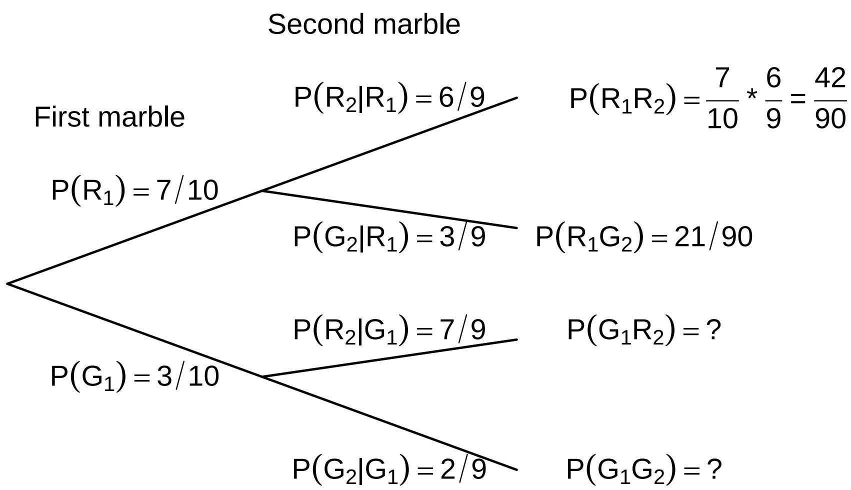


Figure 2.2: Tree diagram for marble choices

Now, can you tell me what $P(R_2)$ and $P(R_1 \mid R_2)$ are? Maybe you know the answer already. However, we will get back to this question in Section 2.7. □

Example 2.21.

Suppose that of all individuals buying a certain digital camera, 60% include an optional memory card in their purchase, 40% include a set of batteries, and 30% include both a card and batteries. Consider randomly selecting a buyer and let A={memory card purchased} and B= {battery purchased}. Then find $P(A|B)$ and $P(B|A)$.

Solution. From given information, we have $P(A) = 0.60, P(B) = 0.40$, and P(both purchased) = $P(A \cap B)$ =0.30. Given that the selected individual purchased an extra battery, the probability that an optional card was also purchased is

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{0.30}{0.40} = 0.75$$

That is, of all those purchasing an extra battery, 75% purchased an optional memory card. Similarly

$$P(\text{battery}\,|\,\text{memory card}) = P(B|A) = \frac{P(B \cap A)}{P(A)} = \frac{0.30}{0.60} = 0.50$$

Notice that $P(A|B) \neq P(A)$ and $P(B|A) \neq P(B)$, that is, the events A and B are dependent. □

Exercises

2.19.

A year has 53 Sundays. What is the conditional probability that it is a leap year?

2.20.

Let events A, B have positive probabilities. Show that, if $P(A\,|\,B) = P(A)$ then also $P(B\,|\,A) = P(B)$.

2.21.

The level of college attainment of US population by racial and ethnic group in 1998 is given in the following table[2]

Racial or Ethnic Group	Number of Adults (Millions)	Percentage with Associate's Degree	Percentage with Bachelor's Degree	Percentage with Graduate or Professional Degree
Native Americans	1.1	6.4	6.1	3.3
Blacks	16.8	5.3	7.5	3.8
Asians	4.3	7.7	22.7	13.9
Hispanics	11.2	4.8	5.9	3.3
Whites	132.0	6.3	13.9	7.7

The percentages given in the right three columns are conditional percentages.

a) How many Asians have had a graduate or professional degree in 1998?

b) What percent of all adult Americans has had a Bachelor's degree?

c) Given that the person had an Associate's degree, what is the probability that the person was Hispanic?

2.22.

The dealer's lot contains 40 cars arranged in 5 rows and 8 columns. We pick one car at random. Are the events $A = \{$the car comes from an odd-numbered row$\}$ and $B = \{$the car comes from one of the last 4 columns$\}$ independent? Prove your point of view.

2.23.

You have sent applications to two colleges. If you are considering your chances to be accepted to either college as 60%, and believe the results are statistically independent, what is the probability that you'll be accepted to at least one?

How will your answer change if you applied to 5 colleges?

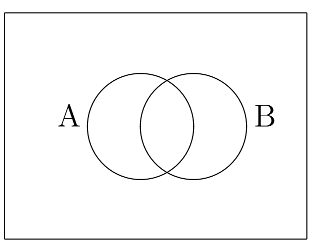
2.24.

Show that, if the events A and B are independent, then so are A' and B'.

2.25.

In a high school class, 50% of the students took Spanish, 25% took French and 30% of the students took neither.

Let A = event that a randomly chosen student took Spanish, and B = event that a student took French. Fill in either the Venn diagram or a 2-way table and answer the questions:



	B	B'	
A			
A'			

a) Describe in words the meaning of the event AB'. Find the probability of this event.

b) Are the events A, B independent? Explain with numbers why or why not.

c) If it is known that the student took Spanish, what are the chances that she also took French?

2.26.

One half of all female physicists are married. Among those married, 50% are married to other physicists, 29% to scientists other than physicists and 21% to nonscientists. Among male physicists, 74% are married. Among them, 7% are married to other physicists, 11% to scientists other than physicists and 82% to nonscientists.[3] What percent of all physicists are female? [**Hint:** This problem can be solved as is, but if you want to, assume that physicists comprise 1% of all population.]

2.27.

Error-correcting codes are designed to withstand errors in data being sent over communication lines. Suppose we are sending a binary signal (consisting of a sequence of 0's and 1's), and during transmission, any bit may get flipped with probability p, independently of any other bit. However, we might choose to repeat each bit 3 times. For example, if we want to send a sequence 010, we will code it as 000111000. If one of the three bits flips, say, the receiver gets the sequence 001111000, he will still be able to decode it as 010 by majority voting. That is, reading the first three bits, 001, he will interpret it as an attempt to send 000. However, if two of the three bits are flipped, for example 011, this will be interpreted as an attempt to send 111, and thus decoded incorrectly.

What is the probability of a bit being decoded incorrectly under this scheme?[4]

2.28. $\star$

Give an example of events A, B, C such that they are pairwise independent (i.e. $P(AB) = P(A)P(B)$ etc.) but $P(ABC) \neq P(A)P(B)P(C)$. [**Hint:** You may build them on a sample space with 4 elementary outcomes.]

2.7 Bayes Rule

Events $B_1, B_2, \ldots, B_k$ are said to be a **partition** of the sample space $\mathcal{S}$ if the following two conditions are satisfied.

a) $B_i B_j = \emptyset$ for each pair i, j

b) $B_1 \cup B_2 \cup \cdots \cup B_k = \mathcal{S}$

Consider a case when $k = 2$:
The event A can be written as the union of mutually exclusive events AB_1 and AB_2, that is

$$A = AB_1 \cup AB_2$$

$P(A) = P(AB_1) + P(AB_2)$
If the conditional probabilities of $P(A|B_1)$ and $P(A|B_2)$ are known, that is

$$P(A|B_1) = \frac{P(AB_1)}{P(B_1)} \quad \text{and} \quad P(A|B_2) = \frac{P(AB_2)}{P(B_2)},$$

then

$$P(A) = P(A|B_1)P(B_1) + P(A|B_2)P(B_2).$$

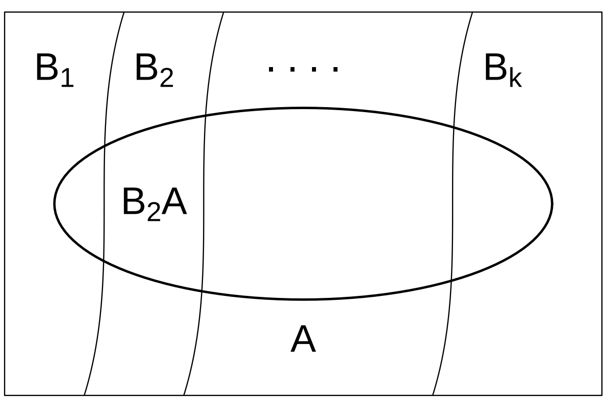


Figure 2.3: Partition $B_1, B_2, \ldots, B_k$ and event A (inside of the oval).

Suppose we want to find probability of the form $P(B_1|A)$, which can be written as

$$P(B_1|A) = \frac{P(AB_1)}{P(A)} = \frac{P(A|B_1)\ P(B_1)}{P(A)},$$

therefore

$$P(B_1|A) = \frac{P(B_1)P(A|B_1)}{P(B_1)P(A|B_1) + P(B_2)P(A|B_2)}$$

This calculation generalizes to $k > 2$ events as follows.

Theorem 2.7. Bayes Rule

If $B_1, B_2, \ldots, B_k$ form a partition of the sample space $\mathcal{S}$ such that $P(B_i) \neq 0$ for $i = 1, 2, \ldots, k$, then for any event A of $\mathcal{S}$,

$$P(A) = \sum_{i=1}^{k} P(B_i \cap A) = \sum_{i=1}^{k} P(B_i)P(A|B_i) \tag{2.8}$$

Subsequently,

$$P(B_j|A) = \frac{P(B_j)P(A|B_j)}{P(A)} \tag{2.9}$$

The equation (2.8) is often called *Law of Total Probability.*

Example 2.22.
At a certain assembly plant, three machines make 30%, 45%, and 25%, respectively, of the products. It is known from the past experience that 2%, 3%, and 2% of the products made by each machine, respectively, are defective. Now, suppose that a finished product is randomly selected.

a) What is the probability that it is defective?
b) If a product were chosen randomly and found to be defective, what is the probability that it was made by machine 3?

Solution. Consider the following events:

A: the product is defective

B_1 : the product is made by machine 1,

B_2 : the product is made by machine 2,

B_3 : the product is made by machine 3.

Applying additive and multiplicative rules, we can write

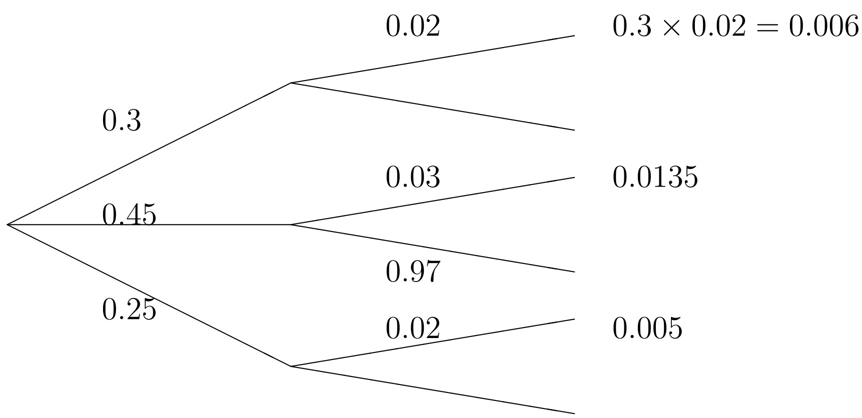
(a) $P(A) = P(B_1)P(A|B_1) + P(B_2)P(A|B_2) + P(B_3)P(A|B_3) =$

$= (0.3)(0.02) + (0.45)(0.03) + (0.25)(0.02) = 0.006 + 0.0135 + 0.005 = 0.0245$

(b) Using Bayes' rule

$$P(B_3|A) = \frac{P(B_3)P(A|B_3)}{P(A)} = \frac{0.005}{0.0245} = 0.2041$$

This calculation can also be represented using a tree



Here, the first branching represents probabilities of the events B_i, and the second branching represents conditional probabilities $P(A \mid B_i)$. The probabilities of intersections, given by the products, are on the right. $P(A)$ is their sum. □

Example 2.23.

A rare genetic disease (occuring in 1 out of 1000 people) is diagnosed using a DNA screening test. The test has *false positive rate* of 0.5%, meaning that $P(\text{test positive} \mid \text{no disease}) = 0.005$. Given that a person has tested positive, what is the probability that this person actually has the disease?
First, guess the answer, then read on.

Solution. Let's reason in terms of actual numbers of people, for a change. Imagine 1000 people, 1 of them having the disease. How many out of 1000 will test positive? One that actually has the disease, and about 5 disease-free people who would test false positive.[3] Thus, $P(\text{disease} \mid \text{test positive}) \approx 1/6$.

[3]a) Of course, of any actual 1000 people, the number of people having the disease and the number of people who test positive will vary randomly, so our calculation only makes sense when considering averages in a much larger population. b) There's also a possibility of a false negative, i.e. person having the disease and the test coming out negative. We will neglect this, quite rare, event.

It is left as an exercise for the reader to write down the formal probability calculation. □

Exercises

2.29.

Lucy is undecided as to whether to take a Math course or a Chemistry course. She estimates that her probability of receiving an A grade would be $\frac{1}{2}$ in a math course, and $\frac{2}{3}$ in a chemistry course. If Lucy decides to base her decision on the flip of a fair coin, what is the probability that she gets an A?

2.30.

Of the customers at a gas station, 70% use regular gas, and 30% use diesel. Of the customers who use regular gas, 60% will fill the tank completely, and of those who use diesel, 80% will fill the tank completely.

a) What percent of all customers will fill the tank completely?

b) If a customer has filled up completely, what is the probability it was a customer buying diesel?

2.31.

In 2004, 57% of White households directly and/or indirectly owned stocks, compared to 26% of Black households and 19% of Hispanic households.[5] The data for Asian households is not given, but let's assume the same rate as for Whites. Additionally, 77% of households are classified as either White or Asian, 12% as African American, and 11% as Hispanic.

a) What proportion of all families owned stocks?

b) If a family owned stock, what it the probability it was White/Asian?

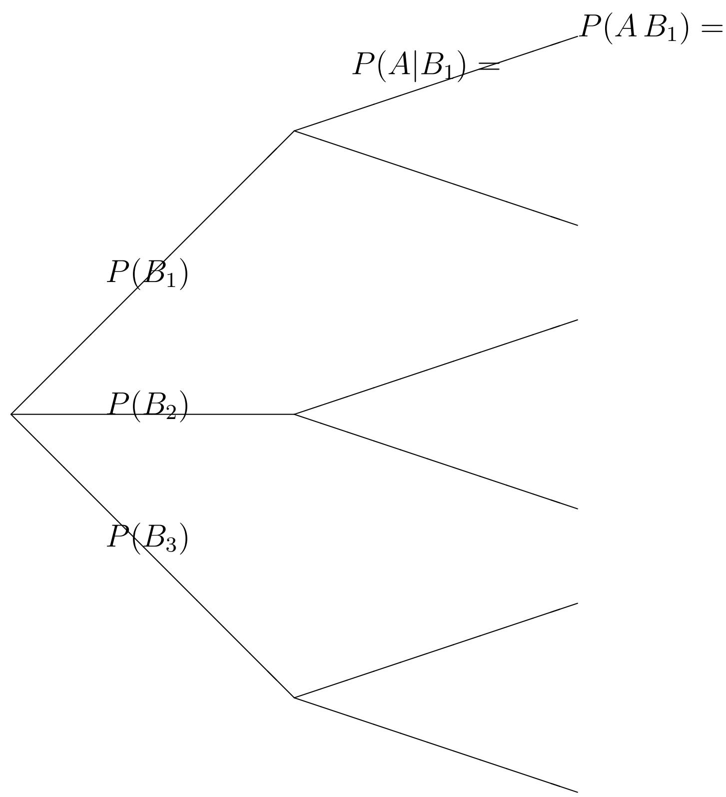
2.32.

For an on-line electronics retailer, 5% of customers who buy Zony digital cameras will return them, 3% of customers who buy Lucky Star digital cameras will return them, and 8% of customers who buy any other brand will return them. Also, among all digital cameras bought, there are 20% Zony's and 30% Lucky Stars.
Fill in the tree diagram and answer the questions.

(a) What percent of **all** cameras are returned?

(b) If the camera was just returned, what is the probability it is a Lucky Star?

(c) What percent of all cameras sold were Zony **and** were not returned?



2.33. ⋆

This is the famous *Monty Hall problem.*[6] A contestant on a game show is asked to choose among 3 doors. There is a prize behind one door and nothing behind the other two. You (the contestant) have chosen one door. Then, the host is flinging one other door open, and there's nothing behind it. What is the best strategy? Should you switch to the remaining door, or just stay with the door you have chosen? What is your probability of success (getting the prize) for either strategy?

2.34. ⋆

There are two children in a family. We overheard about one of them referred to as a boy.

a) Find the probability that there are 2 boys in the family.

b) Suppose that the oldest child is a boy. Again, find the probability that there are 2 boys in the family.[7] [Why is it different from part (a)?]

Chapter exercises

2.35.

At a university, two students were doing well for the entire semester but failed to show up for a final exam. Their excuse was that they traveled out of state and had a flat tire. The professor gave them the exam in separate rooms, with one question worth 95 points: "which tire was it?". Find the probability that both students mentioned the same tire.[8]

2.36.

In firing the company's CEO, the argument was that during the six years of her tenure, for the last three years the company's market share was lower than for the first three years. The CEO claims bad luck. Find the probability that, given six random numbers, the last three are the lowest among six.

Notes

[1] Taken from Leonard Mlodinow, *The Drunkard's Walk*

[2]Source: US Department of Education, National Center for Education Statistics, as reported in *Chronicle of Higher Education Almanac, 1998-1999*, 2000.

[3]Laurie McNeil and Marc Sher. *The dual-career-couple problem.* Physics Today, July 1999.

[4]see David MacKay, *Information Theory, Inference, and Learning Algorithms*, 640 pages, Published September 2003.
Downloadable from `http://www.inference.phy.cam.ac.uk/itprnn/book.html`

[5]According to `"http://www.highbeam.com/doc/1G1-167842487.html",`Consumer Interests Annual, January 1, 2007 by Hanna, Sherman D.; Lindamood, Suzanne

[6]There are some interesting factoids about this in Mlodinow's book, including Marylin vos Savant's column and scathing replies from academics, who believed that the probability was 50%.

[7]Puzzle cited by Martin Gardner, mentioned in *Math Horizons*, Sept. 2010. See also the discussion at `http://www.stat.columbia.edu/~cook/movabletype/archives/2010/05/hype_about_cond.html`

[8]This example is also from Mlodinow's book.

Chapter 3

Discrete probability distributions

3.1 Discrete distributions

In this chapter, we will consider random quantities that are usually called **random variables**.

Definition 3.1. Random variable

A random variable (RV) is a number associated with each outcome of some random experiment.

One can think of the shoe size of a randomly chosen person as a random variable. We have already seen the example when a die was rolled and a number was recorded. This number is also a random variable.

Example 3.1.

Toss two coins and record the number of heads: 0, 1 or 2. Then the following outcomes can be observed.

Outcome	TT	HT	TH	HH
Number of heads	0	1	1	2

The random variables will be denoted with capital letters $X, Y, Z, ...$ and the lowercase x would represent a particular value of X. For the above example, $x = 2$ if heads comes up twice. Now we want to look at the probabilities of the outcomes. For the probability that the random variable X has the value x, we write $P(X = x)$, or just $p(x)$.

For the coin flipping random variable X, we can make the table:

x	0	1	2
$p(x)$	1/4	1/2	1/4

This table represents the *probability distribution* of the random variable X.

Definition 3.2. Probability mass function

A random variable X is said to be **discrete** if it can take on only a finite or countable number of possible values x. In this case,

a) $P(X = x) = p_X(x) \geq 0$

b) $\sum_x P(X = x) = 1$, where the sum is over all possible x

The function $p_X(x)$ or simply $p(x)$ is called *probability mass function* (PMF) of X.

What does this actually mean? A discrete probability function is a function that can take a discrete number of values (not necessarily finite). This is most often the non-negative integers or some subset of the non-negative integers. There is no mathematical restriction that discrete probability functions only be defined at integers, but we will use integers in many practical situations. For example, if you toss a coin 6 times, you can get 2 heads or 3 heads but not 2.5 heads.

Each of the discrete values has a certain probability of occurrence that is between zero and one. That is, a discrete function that allows negative values or values greater than one is not a PMF. The condition that the probabilities add up to one means that one of the values has to occur.

Example 3.2.

A shipment of 8 similar microcomputers to a retail outlet contains 3 that are defective. If a school makes a random purchase of 2 of these computers, find the probability mass function for the number of defectives.

Solution. Let X be a random variable whose values x are the possible numbers of defective computers purchased by school. Then x must be $0, 1$ or 2.

Then,

$$P(X=0) = \frac{\binom{3}{0}\binom{5}{2}}{\binom{8}{2}} = \frac{10}{28}$$

$$P(X=1) = \frac{\binom{3}{1}\binom{5}{1}}{\binom{8}{2}} = \frac{15}{28}$$

$$P(X=2) = \frac{\binom{3}{2}\binom{5}{0}}{\binom{8}{2}} = \frac{3}{28}$$

Thus, the probability mass function of X is

x	0	1	2
$p(x)$	$\frac{10}{28}$	$\frac{15}{28}$	$\frac{3}{28}$

□

Definition 3.3. Cumulative distribution function

The cumulative distribution function (CDF) $F(x)$ for a random variable X is defined as

$$F(x) = P(X \leq x)$$

If X is discrete,

$$F(x) = \sum_{y \leq x} p(y)$$

where $p(x)$ is the probability mass function.

Properties of discrete CDF

a) $\lim_{x \to -\infty} F(x) = 0$

b) $\lim_{x \to \infty} F(x) = 1$

c) $F(x)$ is non-decreasing

d) $p(x) = F(x) - F(x-) = F(x) - \lim_{y \uparrow x} F(y)$

In words, CDF of a discrete RV is a step function, whose jumps occur at the values x for which $p(x) > 0$ and are equal in size to $p(x)$. It ranges from 0 on the left to 1 on the right.

Example 3.3.

Find the CDF of the random variable from Example 3.2. Using $F(x)$, verify that $P(X = 1) = 15/28$.

Solution. The CDF of the random variable X is:
$F(0) = p(0) = \frac{10}{28}$
$F(1) = p(0) + p(1) = \frac{25}{28}$
$F(2) = p(0) + p(1) + p(2) = \frac{28}{28} = 1.$

Hence,

$$F(x) = \begin{cases} 0 & \text{for } x < 0 \\ 10/28 & \text{for } 0 \leq x < 1 \\ 25/28 & \text{for } 1 \leq x < 2 \\ 1 & \text{for } x \geq 2 \end{cases} \tag{3.1}$$

Now, $P(X = 1) = p(1) = F(1) - F(0) = \frac{25}{28} - \frac{10}{28} = \frac{15}{28}$. □

Graphically, $p(x)$ can be represented as a *probability histogram* where the heights of the bars are equal to $p(x)$.

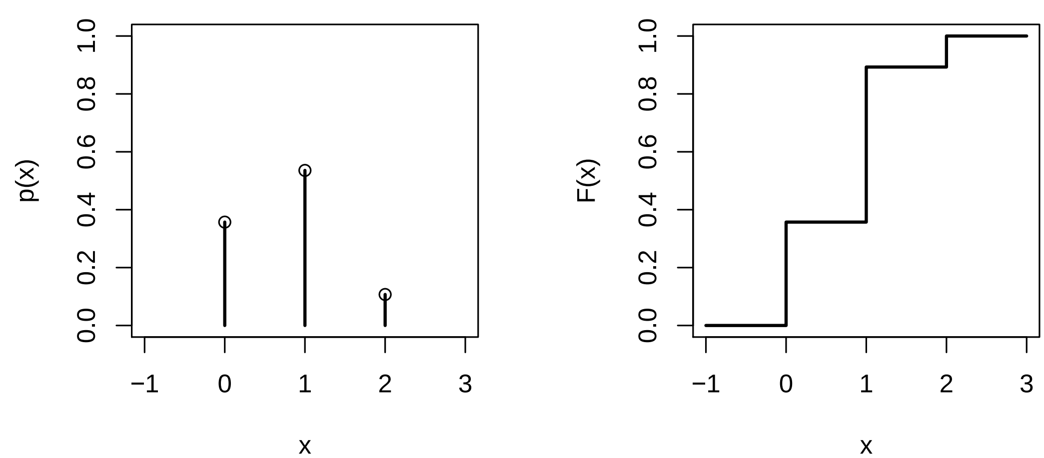


Figure 3.1: PMF and CDF for Example 3.3

Exercises

3.1.

Suppose that two dice are rolled independently, with outcomes X_1 and X_2. Find the distribution of the random variable $Y = X_1 + X_2$. [**Hint:** It's easier to visualize all the ourcomes if you make a two-way table.]

3.2.

What constant c makes $p(x)$ a valid PMF?

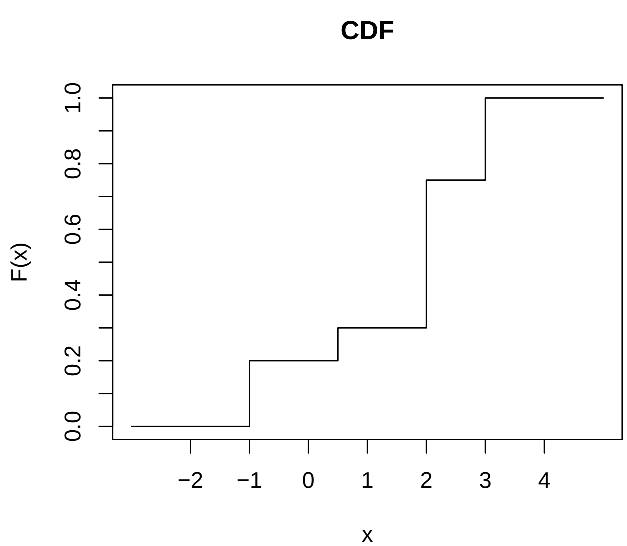
a) $p(x) = c$ for $x = 1, 2, ..., 5$.

b) $p(x) = c(x^2 + 1)$ for $x = 0, 1, 2, 3$.

c) $p(x) = cx \binom{3}{x}$ for $x = 1, 2, 3$.

3.3.

The CDF of a discrete random variable X is shown in the plot below.



Find the probability mass function $p_X(x)$ (make a table)

3.4.

For an on-line electronics retailer, X = the number of Zony digital cameras returned per day follows the distribution given by

x	0	1	2	3	4	5
$p(x)$	0.05	0.1	?	0.2	0.25	0.1

(a) Fill in the “?”
(b) Find $P(X > 3)$
(c) Find the CDF of X (make a table).

3.2 Expected values of Random Variables

One of the most important things we’d like to know about a random variable is: what value does it take on average? What is the average price of a computer? What is the average value of a number that rolls on a die?

Definition 3.4. Expected value (mean)

The mean or expected value of a discrete random variable X with probability mass function $p(x)$ is given by

$$\mathbb{E}(X) = \sum_x x\, p(x)$$

We will sometimes use the notation $\mathbb{E}(X) = \mu$.

Theorem 3.1. Expected value of a function

If X is a discrete random variable with probability mass function $p(x)$ and if $g(x)$ is a real valued function of x, then

$$\mathbb{E}[g(X)] = \sum_x g(x)p(x).$$

Definition 3.5. Variance

The variance of a random variable X with expected value μ is given by

$$V(X) = \sigma^2 = \mathbb{E}\,(X-\mu)^2 = \mathbb{E}\,(X^2) - \mu^2,$$

where

$$\mathbb{E}\,(X^2) = \sum_x x^2\, p(x).$$

The variance defines the average (or expected) value of the squared difference from the mean.
If we use $V(X) = \mathbb{E}\,(X-\mu)^2$ as a definition, we can see that

$$V(X) = \mathbb{E}\,(X-\mu)^2 = \mathbb{E}\,(X^2-2\mu X+\mu^2) = \mathbb{E}\,(X^2)-2\mu\mathbb{E}\,(X)+\mu^2 = \mathbb{E}\,(X^2)-\mu^2$$

due to the linearity of expectation (see Theorem 3.2 below).

Definition 3.6. Standard deviation

The standard deviation of a random variable X is the square root of the variance, and is given by

$$\sigma = \sqrt{\sigma^2} = \sqrt{\mathbb{E}\,(X-\mu)^2}$$

The mean describes the *center* of the probability distribution, while standard deviation describes the *spread*. Larger values of σ signify a distribution with larger variation. This will be undesirable in some situations, e.g. industrial process control, where we would like the manufactured items to have identical characteristics. On the other hand, a *degenerate* random variable X that has $P(X = a) = 1$ for some value of a is not random at all, and it has the standard deviation of 0.

Example 3.4.
The number of fire emergencies at a rural county in a week, has the following distribution

x	0	1	2	3	4
$P(X = x)$	0.52	0.28	0.14	0.04	0.02

Find $\mathbb{E}\,(X)$, $V(X)$ and σ.

Solution. From Definition 3.4, we see that

$$\mathbb{E}\,(X) = 0(0.52) + 1(0.28) + 2(0.14) + 3(0.04) + 4(0.02) = 0.76 = \mu$$

and from definition of $\mathbb{E}(X^2)$, we get

$$\mathbb{E}(X^2) = 0^2(0.52) + 1^2(0.28) + 2^2(0.14) + 3^2(0.04) + 4^2(0.02) = 1.52$$

Hence, from Definition 3.5, we get

$$V(X) = \mathbb{E}(X^2) - \mu^2 = 1.52 - (0.76)^2 = 0.9424$$

Now, from Definition 3.6, the standard deviation $\sigma = \sqrt{0.8456} = 0.9708$. $\square$

Theorem 3.2. Linear functions

For any random variable X and constants a and b,

a) $\mathbb{E}(aX + b) = a\mathbb{E}(X) + b$

b) $V(aX + b) = a^2\, V(X) = a^2\, \sigma^2$

c) $\sigma_{aX+b} = |a|\, \sigma$.

d) For several RV's, $X_1, X_2, ..., X_k$,

$$\mathbb{E}(X_1 + X_2 + ... + X_k) = \mathbb{E}(X_1) + \mathbb{E}(X_2) + ... + \mathbb{E}(X_k)$$

Example 3.5.

Let X be a random variable having probability mass function given in Example 3.4. Calculate the mean and variance of $g(X) = 4X + 3$.

Solution. In Example 3.4, we found $\mathbb{E}(X) = \mu = 0.88$ and $V(X) = 0.8456$. Now, using Theorem 3.2,

$$\mathbb{E}(g(X)) = 4\mathbb{E}(X) + 3 = 4(0.88) + 3 = 3.52 + 3 = 6.52$$

$$\text{and} \qquad V(g(X)) = 4^2\, V(X) = 16(0.8456) = 13.5296$$

$\square$

Theorem 3.3. Chebyshev Inequality

> Let X be a random variable with mean μ and a variance σ^2. Then for any positive k,
> $$P(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}$$

The inequality in the statement of the theorem is equivalent to

$$P(\mu - k\sigma < X < \mu + k\sigma) > 1 - \frac{1}{k^2}$$

To interpret this result, let $k = 2$, for example. Then the interval from $\mu - 2\sigma$ to $\mu + 2\sigma$ must contain at least $1 - \frac{1}{k^2} = 1 - \frac{1}{4} = \frac{3}{4}$ of the probability mass for the random variable.

Chebyshev inequality is useful when the mean and variance of a RV are known and we would like to calculate estimates of some probabilities. However, these estimates are usually quite crude.

Example 3.6.
The performance period of a certain car battery is known to have a mean of 30 months and standard deviation of 5 months.
a) Estimate the probability that a car battery will last at least 18 months.
b) Give a range of values to which at least 90% of all batteries' lifetimes will belong.

Solution. (a) Let X be the battery performance period. Calculate k such that the value of 18 is k standard deviations below the mean: $18 = 30 - 5k$, therefore $k = (30 - 18)/5 = 2.4$. From Chebyshev's theorem we have

$$P(30 - 5k < X < 30 + 5k) > 1 - 1/k^2 = 1 - 1/2.4^2 = 0.826$$

Thus, at least 82.6% of batteries will make it to 18 months. (However, in reality this percentage could be much higher, depending on distribution.)
(b) From Chebyshev's theorem we have

$$P(\mu - k\sigma < X < \mu + k\sigma) > 1 - \frac{1}{k^2}$$

According to the problem set $1 - \frac{1}{k^2} = 0.90$ and solve for k, we get $k = \sqrt{10} = 3.16$. Hence, the desired interval is between $30 - 3.16(5)$ and $30 + 3.16(5) =$ 14.2 to 45.8 months. □

Example 3.7.

The number of customers per day at a certain sales counter, X, has a mean of 20 customers and standard deviation of 2 customers. The probability distribution of X is not known. What can be said about the probability that X will be between 16 and 24 tomorrow?

Solution. We want $P(16 \leq X \leq 24) = P(15 < X < 25)$. From Chebyshev's theorem

$$P(\mu - k\sigma < X < \mu + k\sigma) \geq 1 - \frac{1}{k^2}$$

given $\mu = 20, \sigma = 2$ we set $\mu - k\sigma = 15$ and hence $k = 2.5$. Thus, $P(16 \leq X \leq 24) \geq 1 - \frac{1}{6.25} = 0.84$.
So, tomorrow's customer total will be between 16 and 24 with probability at least 0.84. □

Exercises

3.5.

Timmy is selling cholocates door to door. The probability distribution of X, the number of cholocates he sells in each house, is given by

x	0	1	2	3	4
$P(X = x)$	0.45	0.25	0.15	0.1	0.05

Find the expected value and standard deviation of X.

3.6.

In the previous exercise, suppose that Timmy earns 50 cents for school from each purchase. Find the expected value and standard deviation of his earnings per house.

3.7.

A dollar coin, a quarter, a nickel and a dime are tossed. I get to pocket all the coins that came up heads. What are my expected winnings?

3.8.

Consider X with the distribution of a random digit, $p(x) = 1/10$, $x = 0, 1, 2, ..., 9$

a) Find the mean and standard deviation of X.

b) According to Chebyshev's inequality, estimate the probability that a random digit will be between 1 and 8, inclusive. Compare to the actual probability.

3.9.

In the *Numbers* game, two players choose a random number between 1 and 6, and compute the absolute difference.

That is, if Player 1 gets the number Y_1, and Player 2 gets Y_2, then they find

$$X = |Y_1 - Y_2|$$

a) Find the distribution of the random variable X (make a table). [Hint: consider all outcomes (y_1, y_2).]

b) Find the expected value and variance of X, and $\mathbb{E}\left(X^3\right)$

c) If Player 1 wins whenever the difference is 3 or more, and Player 2 wins whenever the difference is 2 or less, who is more likely to win?

d) If Player 1 bets \$1, what is the value that Player 2 should bet to make the game fair?

3.10. "Baker's problem" ⋆

A shopkeeper is selling the quantity X (between 0 and 3) of a certain item per week, with a given probability distribution:

x	0	1	2	3
$p(x)$	0.05	0.2	0.5	0.25

For each item bought, the profit is \$50. On the other hand, if the item is stocked, but was not bought, then the cost of upkeep, insurance etc. is \$20. At the beginning of the week, the shopkeeper stocks a items.

For example, if 3 items were stocked, then the expected profit can be calculated from the following table:

Y = Profit

y	$-\$60$	\$10	\$80	\$150
$p(y)$	0.05	0.2	0.5	0.25

a) What is the expected profit if the shopkeeper stocked $a = 3$ items?

b) What is the expected profit if the shopkeeper stocked $a = 1$ and $a = 2$ items? [You'll need to produce new tables for Y first.]

c) Which value of a maximizes the expected profit?

3.3 Bernoulli distribution

Let X be the random variable denoting the condition of the inspected item. Agree to write $X = 1$ when the item is defective and $X = 0$ when it is not. (This is a convenient notation because, once we inspect n such items, $X_1, X_2, ..., X_n$ denoting their condition, the total number of defectives will be given by $X_1 + X_2 + ... + X_n$.)

Let p denote the probability of observing a defective item. The probability distribution of X, then, is given by

x	0	1
$p(x)$	$q = 1 - p$	p

Such a random variable is said to have a *Bernoulli distribution*. Note that

$$\mathbb{E}(X) = \sum xp(x) = 0 \times p(0) + 1 \times p(1) = 0(q) + 1(p) = p \qquad \text{and}$$

$$\mathbb{E}(X^2) = \sum x^2\, p(x) = 0(q) + 1(p) = p.$$

Hence, $V(X) = \mathbb{E}(X^2) - (\mathbb{E}\, X)^2 = p - p^2 = pq$.

3.4 Binomial distribution

Now, let us inspect n items and count the total number of defectives. This process of repeating an experiment n times is called **Bernoulli trials**. The Bernoulli trials are formally defined by the following properties:

a) The result of each trial is either a success or a failure

b) The probability of success p is constant from trial to trial.

c) The trials are independent

d) The random variable X is defined to be the number of successes in n repeated trials

This situation applies to many random processes with just two possible outcomes: a heads-or-tails coin toss, a made or missed free throw in basketball etc[1]. We arbitrarily call one of these outcomes a success and the other a failure.

[1] However, we have to make sure that the probability of success remains constant. Thus, for example, wins or losses in a series of football games may not be a Bernoulli experiment!

Definition 3.7. Binomial RV

Assume that each Bernoulli trial can result in a success with probability p and a failure with probability $q = 1 - p$. Then the probability distribution of the binomial random variable X, the number of successes in n independent trials, is

$$P(X = k) = \binom{n}{k} p^k q^{n-k}, \qquad k = 0, 1, 2, \ldots, n.$$

The mean and variance of the binomial distribution are

$$\mathbb{E}(X) = \mu = np \quad \text{and} \quad V(X) = \sigma^2 = npq.$$

We can notice that the mean and variance of the Binomial are n times larger than those of the Bernoulli random variable.

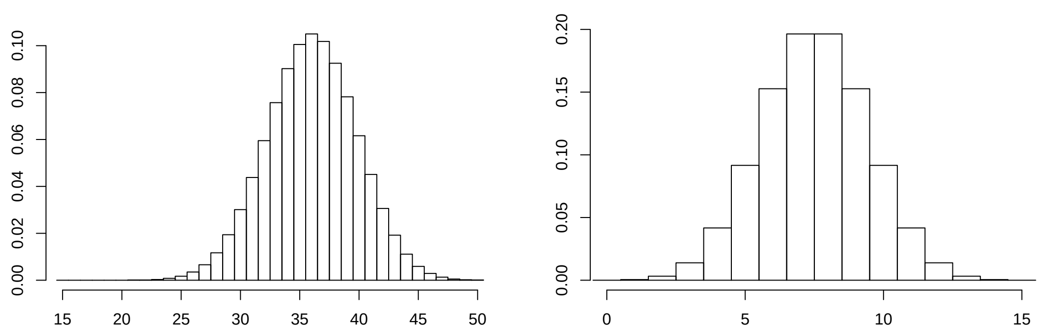


Figure 3.2: Binomial PMF: left, with $n = 60$, $p = 0.6$; right, with $n = 15$, $p = 0.5$

Note that Binomial distribution is symmetric when $p = 0.5$. Also, two Binomials with the same n and $p_2 = 1 - p_1$ are mirror images of each other.

Example 3.8.

The probability that a certain kind of component will survive a shock test is 0.75. Find the probability that

a) exactly 2 of the next 8 components tested survive,

b) at least 2 will survive,

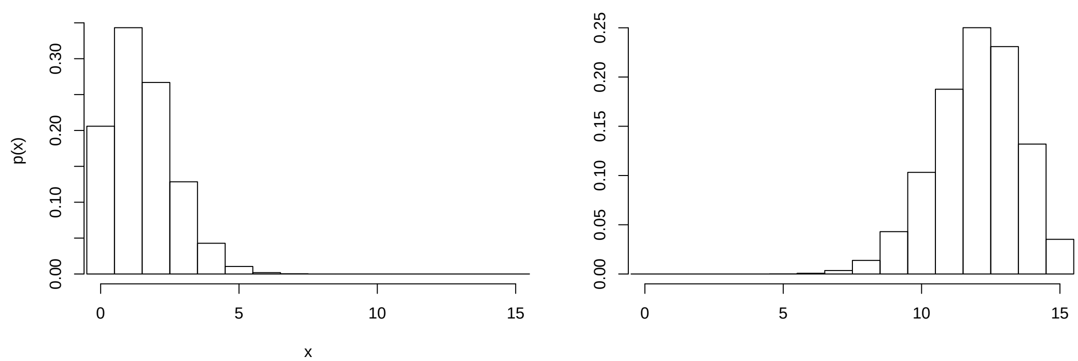


Figure 3.3: Binomial PMF: left, with $n = 15$, $p = 0.1$; right, with $n = 15$, $p = 0.8$

c) at most 6 will survive.

Solution. (a) Assuming that the tests are independent and $p = 0.75$ for each of the 8 tests, we get

$$P(X = 2) = \binom{8}{2}(0.75)^2\,(0.25)^{8-2} = \frac{8!}{2!\,(8-2)!}\,0.75^2\,0.25^6 =$$

$$= \frac{40320}{2 \times 720}\,(0.5625)(0.000244) = 0.003843$$

(b)

$$P(X \geq 2) = 1 - P(X \leq 1) = 1 - [P(X = 1) + P(X = 0)]$$

$$= 1 - [8(0.75)(0.000061) + 0.000002] = 1 - 0.000386 \approx 0.9996$$

(c)

$$P(X \leq 6) = 1 - P(X \geq 7) = 1 - [P(X = 7) + P(X = 8)]$$

$$= 1 - [0.2669 + 0.1001] = 1 - 0.367 = 0.633$$

□

Example 3.9.
It has been claimed that in 60% of all solar heating installations the utility bill is reduced by at least one-third. Accordingly, what are the probabilities that the utility bill will be reduced by at least one-third in

(a) four of five installations;

(b) at least four of five installations?

Solution.

$$\text{(a)} \quad P(X = 4) = \binom{5}{4}(0.60)^4\,(0.4)^{5-4} = 5(0.1296)(0.4) = 0.2592$$

$$\text{(b)} \quad P(X = 5) = \binom{5}{5}(0.60)^5\,(0.40)^{5-5} = 0.60^5 = 0.0777$$

Hence, $P(\text{reduction for at least four}) = P(X \geq 4) = 0.2592 + 0.0777 = 0.3369$ □

Exercises

3.11.

There's 50% chance that a mutual fund return on any given year will beat the industry's average. What proportion of funds will beat the industry average for at least 4 out of 5 last years?

3.12.

Biologists would like to catch Costa Rican glass frogs for breeding. There is 75% probability that a glass frog they catch is male. If 10 glass frogs of a certain species are caught, what are the chances that they will have at least 2 male and 2 female frogs? What is the expected value of the number of female frogs caught?

3.13.

A 5-member focus group are testing a new game console. Suppose that there's 50% chance that any given group member approves of the new console, and their opinions are independent of each other.

a) Calculate and fill out the probability distribution for X = number of group members who approve of the new console.

b) Calculate $P(X \geq 3)$.

c) How does your answer in part (b) change when there's 70% chance that any group member approves of the new console?

3.5 Geometric distribution

In the case of Binomial distribution, the number of trials was a fixed number n, and the variable of interest was the number of successes. It is sometimes of interest to count instead how many trials are required to achieve a specified number of successes.

The number of trials Y required to obtain the first success is called a *Geometric random variable* with parameter p.

Theorem 3.4. Geometric RV

The probability mass function for a Geometric random variable is

$$g(y; p) := P(Y = y) = (1 - p)^{y-1}p, \quad y = 1, 2, 3, \ldots$$

Its CDF is

$$F(y) = 1 - q^y, \quad y = 1, 2, 3, \ldots, \quad q = 1 - p$$

Its mean and variance are

$$\mu = \frac{1}{p} \quad \text{and} \quad \sigma^2 = \frac{1-p}{p^2}$$

Proof. To achieve the first success on yth trial means to have the first $y - 1$ trials to result in failures, and the last yth one a success, and then by independence of trials,

$$P(FF...FS) = q^{y-1}p$$

Now the CDF

$$F(y) = P(Y \leq y) = 1 - P(Y > y)$$

The latter means that all the trials up to and including the yth one, resulted in failures, which equals $P(y \text{ failures in a row}) = q^y$ and we get the CDF subtracting this from 1.

The mean $\mathbb{E}(Y)$ can be found by differentiating a geometric series:

$$\mathbb{E}(Y) = \sum_{i=1}^{\infty} yp(y) = \sum_{i=1}^{\infty} yp(1-p)^{y-1} = p\sum_{i=1}^{\infty} y(1-p)^{y-1} =$$

$$= p\sum_{i=1}^{\infty} \frac{d}{dq}q^y = p\frac{d}{dq}\sum_{i=1}^{\infty} q^y = p\left[\frac{d}{dq}(1 + q + q^2 + q^3 + \cdots - 1)\right] =$$

$$= p \left\{ \frac{d}{dq} \left[(1-q)^{-1} \right] - \frac{d}{dq}(1) \right\} = \frac{p}{(1-q)^2} = \frac{1}{p}.$$

The variance can be calculated by differentiating a geometric series twice:

$$\mathbb{E}\{Y(Y-1)\} = \sum_{i=1}^{\infty} p\, q^{y-1} = pq \sum_{i=1}^{\infty} \frac{d^2}{dq^2}(q^y) =$$

$$= pq \frac{d^2}{dq^2}(1-q)^{-1} = pq \frac{2}{(1-q)^3} = \frac{2q}{p^2}$$

Hence $\quad \mathbb{E}(Y^2) = \frac{2q}{p^2} + \frac{1}{p} \quad$ and $\quad V(Y) = \frac{2q}{p^2} + \frac{1}{p} - \frac{1}{p^2} = \frac{q}{p^2}$

□

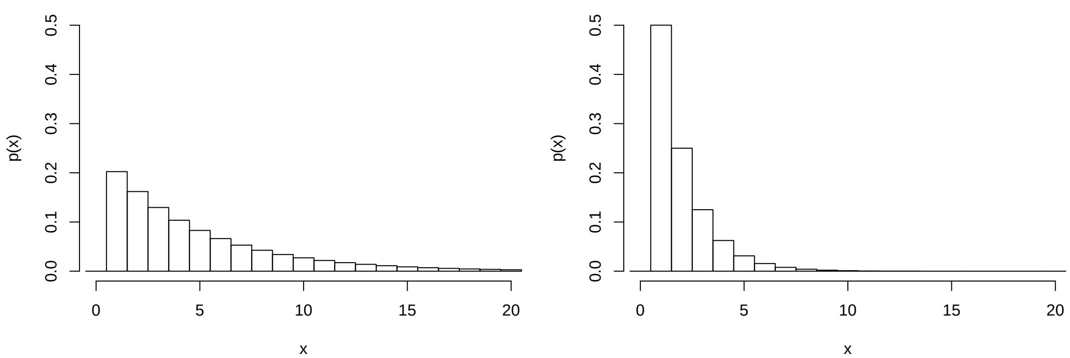


Figure 3.4: Geometric PMF: left, with $p = 0.2$; right, with $p = 0.5$

Example 3.10.

For a certain manufacturing process it is known that, on the average, 1 in every 100 items is defective. What is the probability that the first defective item found is the fifth item inspected? What is the average number of items that should be sampled before the first defective is found?

Solution. Using the geometric distribution with $x = 5$ and $p = 0.01$, we have $g(5; 0.01) = (0.01)(0.99)^4 = 0.0096$.

Mean number of items needed is $\mu = 1/p = 100$. □

Example 3.11.

If the probability is 0.20 that a burglar will get caught on any given job, what is the probability that he will get caught no later than on his fourth job?

Solution. Substituting $y = 4$ and $p = 0.20$ into the geometric CDF, we get $P(Y \leq 4) = 1 - 0.8^4 = 0.5904$ □

Exercises

3.14.

The probability to be caught while running a red light is estimated as 0.1. What is the probability that a person is first caught on his 10th attempt to run a red light? What is the probability that a person runs a red light at least 10 times without being caught?

3.6 Negative Binomial distribution

Let Y denote the number of the trial on which the rth success occurs in a sequence of independent Bernoulli trials, with p the probability of success. Such Y is said to have *Negative Binomial distribution*. When $r = 1$, we will of course obtain the Geometric distribution.

Theorem 3.5. Negative Binomial RV

The PMF of the Negative Binomial random variable Y is

$$nb(y; r, p) := P(Y = y) = \binom{y-1}{r-1} p^r\, q^{y-r}, \qquad y = r, r+1, \ldots$$

The mean and variance of Y are:

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

Proof. We have $P(Y = y) =$

$$= P[\text{First } y-1 \text{ trials contain } r-1 \text{ successes and } y\text{th trial is a success}] =$$

$$= \binom{y-1}{r-1} p^{r-1}\, q^{y-r} \times p = \binom{y-1}{r-1} p^r\, q^{y-r}, \qquad y = r, r+1, r+2, \ldots$$

The proof for the mean and variance uses the properties of the independent sums to be discussed in Section 5.4. However, note at this point that both μ and σ^2 are r times larger than those of the Geometric distribution.

$\square$

Example 3.12.
In an NBA championship series, the team which wins four games out of seven will be the winner. Suppose that team A has probability 0.55 of winning over the team B, and the teams A and B face each other in the championship games.
(a) What is the probability that team A will win the series in six games?
(b) What is the probability that team A will win the series?

Solution.
(a) $nb(6; 4, 0.55) = \binom{5}{3}\ (0.55)^4\ (1 - 0.55)^{6-4} = 0.1853.$
(b) P(team A wins the championship series) =

$$= nb(4; 4, 0.55) + nb(5; 4, 0.55) + nb(6; 4, 0.55) + nb(7; 4, 0.55) =$$

$$= 0.0915 + 0.1647 + 0.1853 + 0.1668 = 0.6083$$

$\square$

Example 3.13.
A pediatrician wishes to recruit 5 couples, each of whom is expecting their first child, to participate in a new childbirth regimen. She anticipates that 20% of all couples she asks will agree. What is the probability that 15 couples must be asked before 5 are found who agree to participate?

Solution. Substituting $x = 15, p = 0.2, r = 5$, we get

$$nb(15; 5, 0.2) = \binom{14}{4} (0.2)^5 (0.8)^{15-5} = 0.034$$

$\square$

Exercises

3.15.
Biologists catch Costa Rican glass frogs for breeding. There is 75% probability that a glass frog they catch is male. Biologists would like to have at least 2 female frogs. What is the expected value of the total number of frogs caught, until they reach their goal? What is the probability that they will need exactly 6 frogs to reach their goal?

3.16.

In the best of 5 series, Team A has 60% chance to win any single game, and the outcomes of the games are independent. Find the probability that Team A will win the series (i.e. will win the majority of the games).

3.7 Poisson distribution

It is often useful to define a random variable that counts the number of events that occur within certain specified boundaries. For example, the average number of telephone calls received by customer service within a certain time limit. The Poisson distribution is often appropriate to model such situations.

Definition 3.8. Poisson RV

A random variable X with a Poisson distribution takes the values $x = 0, 1, 2, \ldots$ with a probability mass function

$$pois(x; \lambda) := P(X = x) = \frac{e^{-\lambda}\,\lambda^x}{x!}$$

where λ is the parameter of the distribution.

Theorem 3.6. Mean and variance of Poisson RV

For Poisson RV with parameter λ,

$$\mathbb{E}\,(X) = V(X) = \lambda.$$

Proof. Recall the Taylor series expansion of e^x:

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \ldots$$

Now,

$$\mathbb{E}\,(X) = \sum x * pois(x) = \sum_{x=0}^{\infty} x\,\frac{e^{-\lambda}\,\lambda^x}{x!} = \sum_{x=1}^{\infty} \frac{x\,e^{-\lambda}\,\lambda\,\lambda^{x-1}}{x(x-1)!} =$$

$$= \lambda\,e^{-\lambda} \sum_{x=1}^{\infty} \frac{\lambda^{x-1}}{(x-1)!} = \lambda\,e^{-\lambda}\left[1 + \frac{\lambda}{1!} + \frac{\lambda^2}{2!} + \frac{\lambda^3}{3!} \ldots\right] = \lambda\,e^{-\lambda}\,e^{\lambda} = \lambda$$

To find $\mathbb{E}(X^2)$, let us consider the factorial expression $\mathbb{E}[X(X-1)]$.

$$\mathbb{E}[X(X-1)] = \sum_{x=0}^{\infty} x(x-1)\,\frac{e^{-\lambda}\,\lambda^x}{x!} = \sum_{x=2}^{\infty} x(x-1)\,\frac{\lambda^2\,e^{-\lambda}\,\lambda^{x-2}}{x(x-1)(x-2)!}$$

$$= \lambda^2\,e^{-\lambda}\sum_{x=2}^{\infty}\frac{\lambda^{x-2}}{(x-2)!} = \lambda^2\,e^{-\lambda}\,e^{\lambda} = \lambda^2$$

Therefore, $\mathbb{E}[X(X-1)] = \mathbb{E}(X^2) - \mathbb{E}(X) = \lambda^2$. Now we can solve for $\mathbb{E}(X^2)$ which is $\mathbb{E}(X^2) = \mathbb{E}[X(X-1)] + \mathbb{E}(X) = \lambda^2 + \lambda$.
Thus,

$$V(X) = \mathbb{E}(X^2) - [\mathbb{E}(X)]^2 = \lambda^2 + \lambda - \lambda^2 = \lambda.$$

□

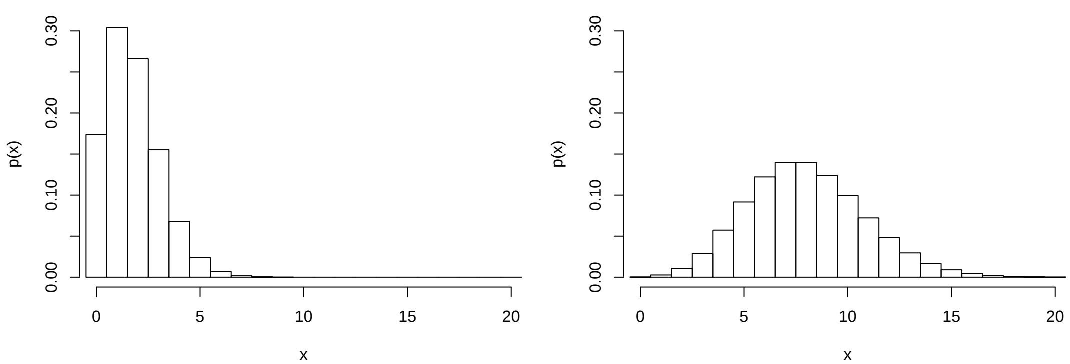


Figure 3.5: Poisson PMF: left, with $\lambda = 1.75$; right, with $\lambda = 8$

Example 3.14.

During World War II, the Nazis bombed London using V-2 missiles. To study the locations where missiles fell, the British divided the central area of London into 576 half-kilometer squares.[9] The following is the distribution of counts per square

Number of missiles in a square	Number of squares	Expected (Poisson) Number of squares
0	229	227.5
1	211	211.3
2	93	98.1
3	35	30.4
4	7	7.1
5 and over	1	1.6
Total	576	576.0

Are the counts suggestive of Poisson distribution?

Solution. The total number of missiles is $1(211) + 2(93) + 3(35) + 4(7) + 5(1) = 535$ and the average number per square, $\lambda = 0.9288$. If the Poisson distribution holds, then the expected number of 0 squares (out of 576) will be

$$576 \times P(X = 0) = 576 \times \frac{e^{-0.9288}\, 0.9288^0}{0!} = 227.5$$

The same way, fill out the rest of the expected counts column. As you can see, the data match the Poisson model very closely!

Poisson distribution is often mentioned as a distribution of *spatial randomness*. As a result, British command were able to conclude that the missiles were unguided. □

Using the CDF

Knowledge of CDF (cumulative distribution function) is useful for calculating probabilities of the type $P(a \le X \le b)$. In fact, when X is integer-valued, then

$$P(a < X \le b) = F_X(b) - F_X(a) \tag{3.2}$$

(you have to carefully watch strict and non-strict inequalities). We might use CDF tables to calculate such probabilities. Nowadays, CDF's of popular distributions are built into various software packages.

Example 3.15.

During a laboratory experiment, the average number of radioactive particles passing through a counter in one millisecond is 4. What is the probability that 6 particles enter the counter in a given millisecond? What is the probability of **at least** 6 particles?

Solution. Using the Poisson distribution with $x = 6$ and $\lambda = 4$, we get

$$pois(6; 4) = \frac{e^{-4}4^6}{6!} = 0.1042$$

Alternatively, using the CDF, $P(X = 6) = P(5 < X \leq 6) = F(6) - F(5)$. Using the Poisson table, $P(X = 6) = 0.8893 - 0.7851 = 0.1042$.

To find $P(X \geq 6)$, use $P(5 < X \leq \infty) = F(\infty) - F(5) = 1 - 0.7851 = 0.2149$ □

Poisson approximation for Binomial

Poisson distribution was originally derived as a limit of Binomial when $n \to \infty$ while $p = \lambda/n$, with fixed λ. We can use this fact to estimate Binomial probabilities for large n and small p.

Example 3.16.

At a certain industrial facility, accidents occur infrequently. It is known that the probability of an accident on any given day is 0.005 and the accidents are independent of each other. For a given period of 400 days, what is the probability that
(a) there will be an accident on only one day?
(b) there are at most two days with an accident?

Solution. Let X be a binomial random varaible with $n = 400$ and $p = 0.005$. Thus $\lambda = np = (400)(0.005) = 2$. Using the Poisson approximation,

a) $P(X = 1) = \frac{e^{-2}\,2^1}{1!} = 0.271$

b) $P(X \leq 2) = P(X = 0) + P(X = 1) + P(X = 2) = \frac{e^{-2}\,2^0}{0!} + \frac{e^{-2}\,2^1}{1!} + \frac{e^{-2}\,2^2}{2!}$
$= 0.1353 + 0.271 + 0.271 = 0.6766$

□

Exercises

3.17.

Number of cable breakages in a year is known to have Poisson distribution with $\lambda = 0.32$.

a) Find the mean and standard deviation of the number of cable breakages in a year.
b) According to Chebyshev's inequality, what is the upper bound for $P(X \geq 2)$?
c) What is the exact probability $P(X \geq 2)$, based on Poisson model?

3.18.

Poisson distribution can be derived by considering Binomial with n large and p small. Compare computationally

a) Binomial with $n = 20$, $p = 0.05$: find $P(X = 0), P(X = 1)$ and $P(X = 2)$.
b) Repeat for Binomial with $n = 200$, $p = 0.005$
c) Poisson with $\lambda = np = 1$ [Note that λ matches the expected value for both (a) and (b).]
d) Compare the standard deviations for distributions in (a)-(c)

3.19.

At a barber shop, expected number of customers per day is 8. What is a probability that, on a given day, between 5 and 10 customers (inclusive) show up ? At least 8 customers?

3.20.

Bolted assemblies on a hull of spacecraft may become loose with probabiity 0.005. There are 96 such assemblies on board. Assuming that assemblies behave statistically independently, find the probability that there is at most one loose assembly on board.

3.8 Hypergeometric distribution

Consider the **Hypergeometric experiment**, that is, one that possesses the following two properties:

a) A random sample of size n is selected *without replacement* from N items.

b) Of the N items overall, k may be classified as successes and $N-k$ are classified as failures.

We will be interested, as before, in the number of successes X, but now the probability of success is not constant (why?).

Theorem 3.7.

The PMF of the hypergeometric random variable X, the number of successes in a random sample of size n selected from N items of which k are labeled success and $N-k$ labeled failure, is

$$hg(x;N,n,k) = \frac{\binom{k}{x}\binom{N-k}{n-x}}{\binom{N}{n}}, \quad x = 0,1,...,\min(n,k)$$

The mean and variance of the hypergeometric distribution are $\mu = n\frac{k}{N}$ and $\sigma^2 = n\left(\frac{k}{N}\right)\left(1-\frac{k}{N}\right)\left(\frac{N-n}{N-1}\right)$

We have already seen such a random variable: see Example 3.2. Here are some more examples.

Example 3.17.
Lots of 40 components each are called unacceptable if they contain as many as 3 defectives or more. The procedure for sampling the lot is to select 5 components at random and to reject the lot if a defective is found. What is the probability that exactly 1 defective is found in the sample if there are 3 defectives in the entire lot?

Solution. Using the above distribution with $n=5, N=40, k=3$ and $x=1$, we can find the probability of obtaining one defective to be

$$hg(1;40,5,3) = \frac{\binom{3}{1}\binom{37}{4}}{\binom{40}{5}} = 0.3011$$

□

Example 3.18.
A shipment of 20 tape recorders contains 5 that are defective. If 10 of them are randomly chosen for inspection, what is the probability that 2 of the 10 will be defective?

Solution. Subsituting $x = 2, n = 10, k = 5$, and $N = 20$ into the formula, we get
$P(X = 2) = \dfrac{\binom{5}{2}\binom{15}{8}}{\binom{20}{10}} = \dfrac{10(6435)}{184756} = 0.348$ □

Note that, if we were sampling *with replacement*, we would have Binomial distribution (why?) with $p = k/N$. In fact, if N is much larger than n, then the difference between Binomial and Hypergeometric distribution becomes small.

Exercises

3.21.
Out of 10 construction facilities, 4 are in-state and 6 are out of state. Three facilities are earmarked as test sites for a new technology. What is the probability that 2 out of 3 are out of state?

3.22.
A box contains 8 diodes, among them 3 are of new design. If 4 diodes are picked randomly for a circuit, what is the probability that at least one is of new design?

3.23.
A small division, consisting of 6 women and 4 men, picks "employee of the month" for 3 months in a row. Suppose that, in fact, a random person is picked each month. Let X be the number of times a woman was picked. Calculate the distribution of X (make a table with all possible values), for the cases

a) No repetitions are allowed.

b) Repetitions are allowed (the same person can be picked again and again).

c) Compare the results.

3.9 Moment generating function

We saw in an earlier section that, if $g(Y)$ is a function of a random variable Y with PMF $p(y)$, then

$$\mathbb{E}\left[g(Y)\right] = \sum_y g(y)p(y)$$

The expected value of the exponential function e^{tY} is especially important.

Definition 3.9. Moment generating function

The moment generating function (MGF) of a random variable Y is

$$M(t) = \mathbb{E}\left(e^{tY}\right)$$

The expected values of powers of random variables are often called *moments*. For example, $\mathbb{E}\left(Y\right)$ is the first moment of Y, and $\mathbb{E}\left(Y^2\right)$ is the second moment of Y. When $M(t)$ exists, it is differentiable in a neighborhood of $t = 0$, and the derivatives may be taken inside the expectation. Thus,

$$M'(t) = \frac{dM(t)}{dt} = \frac{d}{dt}\mathbb{E}\left[e^{tY}\right] = \mathbb{E}\left[\frac{d}{dt}e^{tY}\right] = \mathbb{E}\left[Y\,e^{tY}\right]$$

Now if we set $t = 0$, we have $M'(0) = \mathbb{E}\,Y$. Going on the second derivative,

$$M''(t) = \mathbb{E}\left[Y^2 e^{tY}\right]$$

and hence $M''(0) = \mathbb{E}\left(Y^2\right)$. In general, $M^{(k)}(0) = \mathbb{E}\left(Y^k\right)$.

Theorem 3.8. Properties of MGF's

a) Uniqueness: Let X and Y be two random variables with moment generating functions $M_X(t)$ and $M_Y(t)$, respectively, If $M_X(t) = M_Y(t)$ for all values of t, in some neighborhood of 0, then X and Y have the same probability distribution.

b) $M_{X+b}(t) = e^{bt}M_X(t)$.

c) $M_{aX}(t) = M_X(at)$

d) If $X_1, X_2, \ldots, X_n$ are independent random variables with moment generating functions $M_1(t), M_2(t), \ldots, M_n(t)$, respectively, and $Y = X_1 + X_2 + \cdots + X_n$, then

$$M_Y(t) = M_1(t) \times M_2(t) \times \cdots \times M_n(t).$$

Example 3.19.
Evaluate the moment generating function for the geometric distribution

Solution. From definition,

$$M(t) = \sum_{x=1}^{\infty} e^{tx} p q^{x-1} = \frac{p}{q} \sum_{x=1}^{\infty} (qe^t)^x$$

On the right, we have an infinite geometric series with first term qe^t and the ratio qe^t. Its sum is $\sum_{x=1}^{\infty} (qe^t)^x = \frac{qe^t}{1-qe^t}$. We obtain

$$M(t) = p\,e^t \left[\frac{1}{1-qe^t}\right]$$

□

Exercises

3.24.

Find $M_X(t)$ for random variables X given by

a) $p(x) = 1/3,\ x = -1, 0, 1$

b) $p(x) = \left(\frac{1}{2}\right)^{x+1},\ x = 0, 1, 2, \ldots$

c) $p(x) = \frac{1}{8}\binom{3}{x},\ x = 0, 1, 2, 3$

3.25.

a) Find the MGF of the Bernoulli distribution.

b) Apply the property (d) of Theorem 3.8 to calculate the MGF of the Binomial distribution. [Hint: Binomial random variable Y with parameters n, p can be represented as $Y = X_1 + X_2 + ... + X_n$, where X's are independent and each has Bernoulli distribution with parameter p.]

3.26.

Apply the property (d) of Theorem 3.8 to calculate the MGF of Negative Binomial distribution.

3.27.

Use the derivatives of MGF to calculate the mean and variance of geometric distribution.

3.28.

Suppose that MGF of a random variable X was found equal to

$$M(t) = \frac{1}{1 - t^2}$$

Using the properties of MGF, find $\mathbb{E}(X)$ and $\mathbb{E}(X^2)$.

3.29. ⋆

a) Compute the MGF of Poisson distribution.

b) Using the property (d) of Theorem 3.8, describe the distribution of a sum of two independent Poissons, one with mean λ_1 and another with mean λ_2.

Chapter 4

Continuous probability distributions

4.1 Continuous random variables and their probability distributions

All of the random variables discussed previously were discrete, meaning they can take only a finite (or, at most, countable) number of values. However, many of the random variables seen in practice have more than a countable collection of possible values. For example, the proportions of impurities in ore samples may run from 0.10 to 0.80. Such random variables can take any value in an interval of real numbers. Since the random variables of this type have a continuum of possible values, they are called **continuous random variables**.

Even though the tools we will use to describe continuous RV's are different from the tools we use for discrete ones, practically there is not an enormous gulf between them. For example, a physical measurement of, say, wavelength may be continuous. However, when the measurements are recorded (either on paper or in computer memory), they will take a finite number of values. The number of values will increase if we keep more decimals in the recorded quantity. With rounding we can **discretize** the problem, that is, reduce a continuous problem to a discrete one, whose solution will hopefully be "close enough" to the continuous one. In order to see if we have discretized a problem in a right way, we still need to know something about the nature of continuous random variables.

Definition 4.1. Density (PDF)

The function $f(x)$ is a **probability density function** (PDF) for the continuous random variable X, defined over the set of real numbers $\mathcal{R}$, if

a) $f(x) \geq 0, \text{ for all } x$

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

c) $P(a \leq X \leq b) = \int_{a}^{b} f(x)\, dx.$

What does this actually mean? Since continuous probability functions are defined for an infinite number of points over a continuous interval, the probability at a single point is always zero. Probabilities are measured over intervals, not single points. That is, the area under the curve between two distinct points defines the probability for that interval. This means that the height of the probability function can in fact be greater than one. The property that the integral must equal one is equivalent to the property for discrete distributions that the sum of all the probabilities must equal one.

Probability mass function (PMF) vs. Probability Density Function (PDF)

Discrete probability functions are referred to as probability mass functions and continuous probability functions are referred to as probability density functions. The term probability functions covers both discrete and continuous distributions. When we are referring to probability functions in generic terms, we may use the term probability density functions to mean both discrete and continuous probability functions.

Example 4.1.

Suppose that the error in the reaction temperature, in $^{\circ}C$, for a controlled laboratory experiment is a continuous random variable X having the density

$$f(x) = \begin{cases} \frac{x^2}{3} & \text{for } -1 \leq x \leq 2 \\ 0 & \text{elsewhere} \end{cases}$$

(a) Verify condition (b) of Definition 4.1.
(b) Find $P(0 < X < 1)$.

Solution. (a) $\int_{-\infty}^{\infty} f(x)dx = \int_{-1}^{2} \frac{x^2}{3}dx = \frac{x^3}{9}|_{-1}^{2} = \frac{8}{9} + \frac{1}{9} = 1$

(b) $P(0 < X < 1) = \int_0^1 \frac{x^2}{3}dx = \frac{x^3}{9}|_0^1 = \frac{1}{9}$. □

Definition 4.2. CDF

The **cumulative distribution function (CDF)** $F(x)$ of a continuous random variable X, with density function $f(x)$, is

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)\,dt \tag{4.1}$$

As an immediate consequence of equation (4.1) one can write these two results:

(a) $P(a < X \leq b) = F(b) - F(a)$ [1]

(b) $f(x) = F'(x)$, if the derivative exists.

Example 4.2.

For the density function of Example 4.1, find $F(x)$ and use it to evaluate $P(0 < X < 1)$.

Solution. For $-1 < x < 2$, we have

$$F(x) = \int_{-\infty}^{x} f(t)dt = \int_{-1}^{x} \frac{t^2}{3}dt = \left.\frac{t^3}{9}\right|_{-1}^{x} = \frac{x^3+1}{9},$$

Therefore,

$$F(x) = \begin{cases} 0 & x \leq -1 \\ \frac{x^3+1}{9} & \text{for } -1 < x < 2 \\ 1 & x \geq 2. \end{cases}$$

Now, $P(0 < X < 1) = F(1) - F(0) = \frac{2}{9} - \frac{1}{9} = \frac{1}{9}$, which agrees with the result obtained using the density function in Example 4.1. □

[1]Note that the same relation holds for discrete RV's but in the continuous case $P(a \leq X \leq b)$, $P(a < X \leq b)$ and $P(a < X < b)$ are all the same. Why?

Example 4.3.

The time X in months until failure of a certain product has the PDF

$$f(x) = \begin{cases} \frac{3x^2}{64} \exp\left(-\frac{x^3}{64}\right) & \text{for } x > 0 \\ 0 & \text{elsewhere} \end{cases}$$

Find $F(x)$ and evaluate $P(2.84 < X < 5.28)$

Solution. $F(x) = 1 - \exp\left(-\dfrac{x^3}{64}\right)$, and $P(2.84 \leq X \leq 5.28) = 0.5988$

□

Example 4.4.

For each of the following functions,

(i) find the constant c so that $f(x)$ is a PDF of a random variable X, and

(ii) find the distribution function $F(x)$.

a) $$f(x) = \begin{cases} \frac{x^3}{4} & \text{for } 0 < x < c \\ 0 & \text{elsewhere} \end{cases}$$

b) $$f(x) = \begin{cases} \frac{3}{16}x^2 & \text{for } -c < x < c \\ 0 & \text{elsewhere} \end{cases}$$

c) $$f(x) = \begin{cases} 4x^c & \text{for } 0 < x < 1 \\ 0 & \text{elsewhere} \end{cases}$$

d) $$f(x) = \begin{cases} \frac{c}{x^{3/4}} & \text{for } 0 < x < 1 \\ 0 & \text{elsewhere} \end{cases}$$

Answers. a) $c = 2$ and $F(x) = \frac{x^4}{16}$, $0 < x < 2$.

b) $c = 2$ and $F(x) = \frac{x^3}{16} + \frac{1}{2}$, $-2 < x < 2$.

c) $c = 3$ and $F(x) = x^4$, $0 < x < 1$.

d) $c = \frac{1}{4}$ and $F(x) = x^{1/4}$, $0 < x < 1$.

□

Example 4.5.

The life length of batteries X (in hundreds of hours) has the density

$$f(x) = \begin{cases} \frac{1}{2}\, e^{-\frac{x}{2}} & \text{for } x > 0 \\ 0 & \text{elsewhere} \end{cases}$$

Find the probability that the life of a battery of this type is less than 200 or greater than 400 hours.

Solution. Let A denote the event that X is less than 2, and let B denote the event that X is greater than 4. Then

$$P(A \cup B) = P(A) + P(B) = \int_0^2 \frac{1}{2}\, e^{-\frac{x}{2}} dx + \int_4^\infty \frac{1}{2}\, e^{-\frac{x}{2}} dx$$

$$= (1 - e^{-1}) + (e^{-2}) = 1 - 0.368 + 0.135 = 0.767$$

□

Example 4.6.

Refer to Example 4.5. Find the probability that a battery of this type lasts more than 300 hours, given than it already has been in use for more than 200 hours.

Solution. We are interested in $P(X > 3|X > 2)$; and by the definition of conditional probability,

$$P(X > 3|X > 2) = \frac{P(X > 3, X > 2)}{P(X > 2)} = \frac{P(X > 3)}{P(X > 2)}$$

because the intersection of the events $(X > 3)$ and $(X > 2)$ is the event $(X > 3)$. Now

$$\frac{P(X > 3)}{P(X > 2)} = \frac{\displaystyle\int_3^\infty \frac{1}{2}\, e^{-x/2} dx}{\displaystyle\int_2^\infty \frac{1}{2}\, e^{-x/2} dx} = \frac{e^{-\frac{3}{2}}}{e^{-1}} = e^{-\frac{1}{2}} = 0.606$$

□

Exercises

4.1.

The lifetime of a vacuum cleaner, in years, is described by

$$f(x) = \begin{cases} x/4 & \text{for } 0 < x < 2 \\ (4-x)/4 & \text{for } 2 \leq x < 4 \\ 0 & \text{elsewhere} \end{cases}$$

Find the probability that the lifetime of a vacuum cleaner is
(a) less than 3 years
(b) between 1 and 2.5 years.

4.2.

The proportion of warehouse items claimed within 1 month is given by a random variable X with density

$$f(x) = \begin{cases} c(x+1) & \text{for } 0 < x < 1 \\ 0 & \text{elsewhere} \end{cases}$$

(a) Find c to make this a legitimate density function.
(b) Find the probability that the proportion of items claimed will be between 0.5 and 0.7.

4.3.

The demand for an antibiotic from a local pharmacy is given by a random variable X with CDF

$$f(x) = \begin{cases} 1 - \frac{2500}{(x+50)^2} & \text{for } x > 0 \\ 0 & \text{elsewhere} \end{cases}$$

a) Find the probability that the demand is at least 50 doses
b) Find the probability that the demand is between 40 and 80 doses
c) Find the density function of X.

4.4.
The waiting time, in minutes, between customers coming into a store is a continuous random variable with CDF

$$F(x) = \begin{cases} 0 & \text{for } x < 0 \\ 1 - \exp(-x/2) & \text{for } x \geq 0 \end{cases}$$

Find the probability of waiting less than 1.5 minutes between successive customers

a) using the cumulative distribution of X;

b) using the probability density function of X (first, you have to find it).

4.5.
A continuous random variable X that has a density function given by

$$f(x) = \begin{cases} \frac{1}{5} & \text{for } -1 < x < 4 \\ 0 & \text{elsewhere} \end{cases}$$

a) Show that the area under the curve is equal to 1.

b) Find $P(0 < X < 2)$.

c) Find c such that $P(X < c) = 1/2$. [This is called a *median* of the distribution.]

4.2 Expected values of continuous random variables

The expected values of continuous RV's are obtained using formulas similar to those of discrete ones. However, the summation is now replaced by integration.

Definition 4.3. Expected value

The **Expected Value** of the continuous random variable X that has a probability density function $f(x)$ is given by

$$\mu = \mathbb{E}(X) = \int_{-\infty}^{\infty} x\, f(x) dx$$

Theorem 4.1. Expected value of a function

If X is a continuous random variable with probability density function $f(x)$, and if $g(x)$ is any real-valued function of X, then

$$\mathbb{E}\left(g(X)\right) = \int_{-\infty}^{\infty} g(x)\, f(x)\, dx$$

Definition 4.4. Variance

Let X be a random variable with probability density function $f(x)$ and mean $\mathbb{E}\, X = \mu$. The variance of X is

$$\sigma^2 = \mathbb{E}\left[(X-\mu)^2\right] = \int_{-\infty}^{\infty} (x-\mu)^2 f(x) dx = \mathbb{E}\left(X^2\right) - \mu^2$$

Example 4.7.
Suppose that X has density function given by

$$f(x) = \begin{cases} 3x^2 & \text{for } 0 \leq x \leq 1 \\ 0 & \text{elsewhere} \end{cases}$$

(a) Find the mean and variance of X
(b) Find mean and variance of $u(X) = 4X + 3$.

Solution. (a) From the above definitions,

$$\mathbb{E}\left(X\right) = \int_{-\infty}^{\infty} x\, f(x) dx = \int_{0}^{1} x\,(3x^2) dx = \int_{0}^{1} 3x^3 dx = 3\frac{x^4}{4}\Big]_0^1 = \frac{3}{4} = 0.75$$

$$\text{Now,} \quad \mathbb{E}\left(X^2\right) = \int_{0}^{1} x^2\,(3x^2) dx = \int_{0}^{1} 3x^4 dx = 3\frac{x^5}{5}\Big]_0^1 = \frac{3}{5} = 0.6$$

Hence, $\sigma^2 = \mathbb{E}\left(X^2\right) - \mu^2 = 0.6 - (0.75)^2 = 0.6 - 0.5625 = 0.0375$

(b) From Theorem 3.2, we get

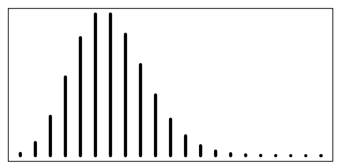
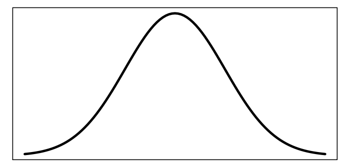
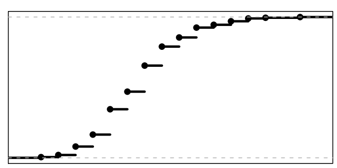
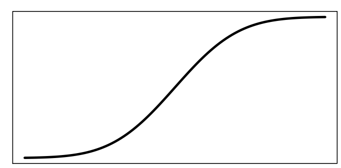
$$\mathbb{E}\left(u(X)\right) = \mathbb{E}\left(4X + 3\right) = 4\mathbb{E}\left(X\right) + 3 = 4(0.75) + 3 = 6$$

and

$$V(u(X)) = V(4X + 3) = 16[V(X)] + 0 \text{ (why?)} \;= 16(0.0375) = 0.6$$

$\square$

Discrete and Continuous random variables

	Discrete	Continuous
Probability	Probability function  $p(x) = P(X = x)$	Density  $f(x) = \frac{d}{dx}P(X \leq x) = F'(x)$ $P(X = x)$ is 0 for any x
CDF $F(x) = P(X \leq x)$ $P(a < X \leq b) =$ $= F(b) - F(a)$	Is a ladder function 	Is continuous 
Mean $\mathbb{E}(X) = \mu_X$	$\sum xp(x)$	$\int xf(x)\,dx$
Mean of a function $\mathbb{E}\,g(X)$	$\sum g(x)p(x)$	$\int g(x)f(x)\,dx$
Variance $\sigma_X^2 = \mathbb{E}(X^2) - \mu^2$	$\sum(x-\mu)^2 p(x)$	$\int (x-\mu)^2 f(x)\,dx$

Exercises

4.6.
For a random variable X with the density

$$f(x) = \begin{cases} \frac{1}{2\sqrt{x}} & \text{for } 0 < x < 1 \\ 0 & \text{elsewhere} \end{cases}$$

a) Find the mean of X

b) Find $V(X)$

c) Find $\mathbb{E}(X^4)$

4.7.
For a random variable X with the density

$$f(x) = \begin{cases} 2 - x & \text{for } 0 < x < c \\ 0 & \text{elsewhere} \end{cases}$$

a) Find c that makes f a legitimate density function

b) Find the mean of X

4.8.
For a random variable X with the CDF

$$F(x) = \begin{cases} x^3/8 & \text{for } 0 < x < 2 \\ 0, & x \le 0 \\ 1, & x \ge 2 \end{cases}$$

a) Find the mean of X

b) Find $V(X)$

4.9.
The waiting time X, in minutes, between successive customers coming into a store is given by

$$f(x) = \begin{cases} 0 & \text{for } x < 0 \\ 0.5 \exp(-x/2) & \text{for } x \ge 0 \end{cases}$$

a) Find the average time between customers

b) Find $\mathbb{E}(e^X)$

4.3 Uniform distribution

One of the simplest continuous distributions is the continuous uniform distribution. This distribution is characterized by a density function that is flat and thus the probability is uniform in a finite interval, say $[a, b]$. The density function of the continuous uniform random variable X on the interval $[a, b]$ is

$$f(x) = \begin{cases} \dfrac{1}{b-a} & \text{for } a < x < b \\ 0 & \text{elsewhere} \end{cases}$$

A graph here would be useful; possibly, double graph of PDF and CDF

The CDF of a uniformly distributed X is given by

$$F(x) = \int_a^x \frac{1}{b-a}\, dt = \frac{x-a}{b-a}, \quad a \le x \le b$$

The mean and variance of the uniform distribution are

$$\mu = \frac{b+a}{2} \quad \text{and} \quad \sigma^2 = \frac{(b-a)^2}{12}.$$

Example 4.8.
Suppose that a large conference room for a certain company can be reserved for no more than 4 hours. However, the use of the conference room is such that both long and short conferences occur quite often. In fact, it can be assumed that length X of a conference has a uniform distribution on the interval $[0, 4]$.

a) What is the probability density function of X?

b) What is the probability that any given conference lasts at least 3 hours?

Solution. (a) The appropriate density function for the uniformly distributed random variable X in this situation is

$$f(x) = \begin{cases} \frac{1}{4} & \text{for } 0 < x < 4 \\ 0 & \text{elsewhere} \end{cases}$$

(b)

$$P(X \ge 3) = \int_3^4 \frac{1}{4}\, dx = \frac{1}{4}.$$

□

Example 4.9.
The failure of a circuit board interrupts work by a computing system until a new board is delivered. Delivery time X is uniformly distributed over the interval of at least one but no more than four days. The cost C of this failure and interruption consists of a fixed cost C_0 for the new part and a cost that increases proportionally to X^2, so that

$$C = C_0 + C_1 X^2$$

(a) Find the probability that the delivery time is two or more days.
(b) Find the expected cost of a single failure, in terms of C_0 and C_1.

Solution. a)

$$f(x) = \begin{cases} \frac{1}{4} & \text{for } 1 \le x \le 5 \\ 0 & \text{elsewhere} \end{cases}$$

Thus,

$$P(X \ge 2) = \int_2^5 \frac{1}{4}\, dx = \frac{1}{4}(5-2) = \frac{3}{4}$$

b) We know that

$$\mathbb{E}(C) = C_0 + C_1 \mathbb{E}(X^2)$$

so it remains for us to find $\mathbb{E}(X^2)$. This value could be found directly from the definition or by using the variance and the fact that $\mathbb{E}(X^2) = V(X) + \mu^2$. Using the latter approach, we find

$$\mathbb{E}(X^2) = \frac{(b-a)^2}{12} + \left(\frac{a+b}{2}\right)^2 = \frac{(5-1)^2}{12} + \left(\frac{1+5}{2}\right)^2 = \frac{31}{3}$$

Thus, $\mathbb{E}(C) = C_0 + C_1\left(\frac{31}{3}\right)$. □

Exercises

4.10.
For a digital measuring device, rounding errors have Uniform distribution, between -0.05 and 0.05 mm.

a) Find the probability that the rounding error is between -0.01 and 0.03mm

b) Find the expected value and the standard deviation of the rouding error.

c) Calculate and plot the CDF of the rounding errors.

4.4 Exponential distribution

Definition 4.5. Exponential distribution

The continuous random variable X has an exponential distribution, with parameter β, if its density function is given by

$$f(x) = \begin{cases} \frac{1}{\beta}\, e^{-\frac{x}{\beta}} & \text{for } x > 0 \\ 0 & \text{elsewhere} \end{cases}$$

The mean and variance of the exponential distribution are

$$\mu = \beta \text{ and } \sigma^2 = \beta^2.$$

The distribution function for the exponential distribution has the simple form:

$$F(t) = P(X \le t) = \int_0^t \frac{1}{\beta}\, e^{-\frac{x}{\beta}} dx = 1 - e^{-\frac{t}{\beta}} \qquad \text{for } t \ge 0$$

The **failure rate function r(t)** is defined as

$$r(t) = \frac{f(t)}{1 - F(t)}, \qquad t > 0 \tag{4.2}$$

Suppose that X, with density f, is a lifetime of an item. Consider the proportion of items currently alive (at the time t) that will fail in the next time interval $(t, t + \Delta t]$, where Δt is small. Thus, by the conditional probability formula,

$$P\{\text{die in the next } (t, t + \Delta t] \mid \text{ currently alive}\} =$$
$$= \frac{P\{X \in (t, t + \Delta t]\}}{P(X > t)} \approx \frac{f(t)\Delta t}{1 - F(t)} = r(t)\Delta t$$

so the rate at which the items fail is $r(t)$.

For the exponential case,

$$r(t) = \frac{f(t)}{1 - F(t)} = \frac{1/\beta\, e^{-t/\beta}}{e^{-t/\beta}} = \frac{1}{\beta}$$

Note that the failure rate $\lambda = \frac{1}{\beta}$ of an item with exponential lifetime does not depend on the item's age. This is known as the *memoryless property* of exponential distribution. The exponential distribution is the only continuous distribution to have a constant failure rate.

In reliability studies, the mean of a positive-valued distribution, is also called *Mean Time To Fail* or MTTF. So, we have exponential MTTF $= \beta$.

Relationship between Poisson and exponential distributions

Suppose that certain events happen at the rate λ, so that the average (expected) number of events on the interval $[0, t]$ is $\mu = \lambda t$. If we assume that the number of events on $[0, t]$ has Poisson distribution, then the probability of no events up to time t is given by

$$pois(0, \lambda t) = \frac{e^{-\lambda t}(\lambda t)^0}{0!} = e^{-\lambda t}.$$

Thus, if the time of first failure is denoted X, then

$$P(X \leq t) = 1 - P(X > t) = 1 - e^{-\lambda t}$$

We see that $P(X \leq t) = F(t)$, the CDF for X, has the form of an exponential CDF. Here, $\lambda = \frac{1}{\beta}$ is again the failure rate. Upon differentiating, we see that the density of X is given by

$$f(t) = \frac{dF(t)}{dt} = \frac{d(1 - e^{-\lambda t})}{dt} = \lambda e^{-\lambda t} = \frac{1}{\beta} e^{-t/\beta}$$

and thus X has an exponential distribution.

Some natural phenomena have a constant failure rate (or occurrence rate) property; for example, the arrival rate of cosmic ray alpha particles or Geiger counter ticks. The exponential model works well for interarrival times (while the Poisson distribution describes the total number of events in a given period).

Example 4.10.
A downtime due to equipment failure is estimated to have Exponential distribution with the mean $\beta = 6$ hours. What is the probability that the next downtime will last between 5 and 10 hours?

Solution. $P(5 < X < 10) =$

$= F(10) - F(5) = 1 - \exp(-10/6) - [1 - \exp(-5/6)] = 0.2457$ □

Example 4.11.
The number of calls to the call center has Poisson distribution with parameter $\lambda = 4$ calls per minute. What is the probability that we have to wait more than 20 seconds for the next call?

Solution. The waiting time between calls, X, has exponential distribution with parameter $\beta = \frac{1}{\lambda} = \frac{1}{4}$. Then, $P(X > \frac{1}{3}) = 1 - F(\frac{1}{3}) = e^{-\frac{4}{3}} = 0.2636$ □

Exercises

4.11.
Prove another version of the memoryless property of the exponential distribution,

$$P(X > t + s \,|\, X > t) = P(X > s).$$

Thus, an item that is t years old has the same probabilistic properties as a brand-new item.

4.12.
The 1-hour carbon monoxide concentrations in a big city are found to have an exponential distribution with a mean of 3.6 parts per million (ppm).

(a) Find the probability that a concentration will exceed 9 ppm.

(b) A traffic control policy is trying to reduce the average concentration. Find the new target mean β so that the probability in part (a) will equal 0.01

(c) The *median* of probability distribution is defined as solution m to the equation (F is the CDF)

$$F(m) = 0.5$$

Find the median of the concentrations from part (a).

4.5 The Gamma distribution

The Gamma distribution derives its name from the well-known gamma function, studied in many areas of mathematics. This distribution plays an important role in both queuing theory and reliability problems. Time between arrivals at service facilities, and time to failure of component parts and electrical systems, often are nicely modeled by the Gamma distribution.

Definition 4.6. Gamma function

The **gamma function**, for $\alpha > 0$, is defined by

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

$\Gamma(k) = (k-1)!$ for integer k.

Definition 4.7. Gamma distribution

The continuous random variable X has a gamma distribution, with parameters α and β, if its density function is given by

$$f(x) = \begin{cases} \frac{1}{\beta^\alpha \Gamma(\alpha)}\, x^{\alpha-1} e^{-\frac{x}{\beta}} & \text{for } \; x > 0 \\ 0 & \text{elsewhere} \end{cases}$$

The mean and variance of the Gamma distribution are

$$\mu = \alpha\beta \;\; \text{and} \;\; \sigma^2 = \alpha\beta^2.$$

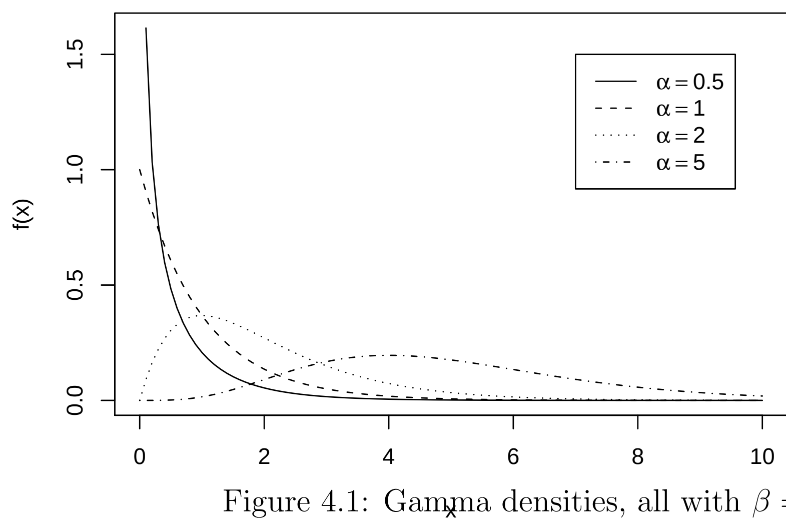


Figure 4.1: Gamma densities, all with $\beta = 1$

Note: When $\alpha = 1$, the Gamma reduces to the exponential distribution. Another well-known statistical distribution, chi-square, is also a special case of the gamma.

Uses of the Gamma Distribution Model

a) The gamma is a flexible life distribution model that may offer a good fit to some sets of failure data, or other data where positivity is enforced.

b) The gamma does arise naturally as the time-to-failure distribution for a system with standby exponentially distributed backups. If there are $n-1$ standby backup units and the system and all backups have exponential lifetimes with mean β, then the total lifetime has a Gamma distribution with $\alpha = n$. Note: when α is a positive integer, the Gamma is sometimes called *Erlang distribution*. The Erlang distribution is used frequently in queuing theory applications.

c) A simple and often used property of sums of identically distributed, independent gamma random variables will be stated, but not proved, at this point. Suppose that $X_1, X_2, \ldots, X_n$ represent independent gamma random variables with parameters α and β, as just used. If $Y = \sum_{i=1}^{n} X_i$ then Y also has a gamma distribution with parameters $n\alpha$ and β. Thus, we see that $\mathbb{E}(Y) = n\alpha\beta$, and $V(Y) = n\alpha\beta^2$.

Example 4.12.

The total monthly rainfall (in inches) for a particular region can be modeled using Gamma distribution with $\alpha = 2$ and $\beta = 1.6$. Find the mean and variance of the monthly rainfall.

Solution. $\mathbb{E}(X) = \alpha\beta = 3.2$, and variance $V(X) = \alpha\beta^2 = 1.6(2^2) = 6.4$ □

4.5.1 Poisson process

Following our discussion about Exponential distribution, the latter is a good model for the waiting times between randomly occurring events. Adding independent Exponential RV's will result in the *Poisson process*.

The Poisson process was first studied[2] in 1900's when modeling the observation times of radioactive particles recorded by Geiger counter. It has the property that the waiting time until kth particle appears has Gamma distribution with $\alpha = k$. As in Section 4.4, the average number of particles to appear during $[0, t)$ has Poisson distribution with the mean λt where the rate $\lambda = 1/\beta$.

[2]not by Poisson!

The Gamma CDF (for integer α) can be derived using this relationship. Suppose Y is the time to wait for kth event. Then it is Gamma ($\alpha = k$, β) random variable. On one hand, the probability that this event happens before time t is the CDF $F(t)$. On the other hand, this will happen if and only if there is a total of at least k events on the interval $[0, t]$. According to Poisson process, this will have Poisson distribution with the mean $\lambda t = t/\beta$. Thus,

$$P(Y \le t) = P(N \ge k) = 1 - P(N < k) = 1 - \sum_{i=0}^{k-1} e^{-t/\beta} \frac{(t/\beta)^i}{i!} \tag{4.3}$$

In particular, when $k = 1$, we get back the familiar exponential CDF, $F(t) = 1 - \exp(-t/\beta)$.

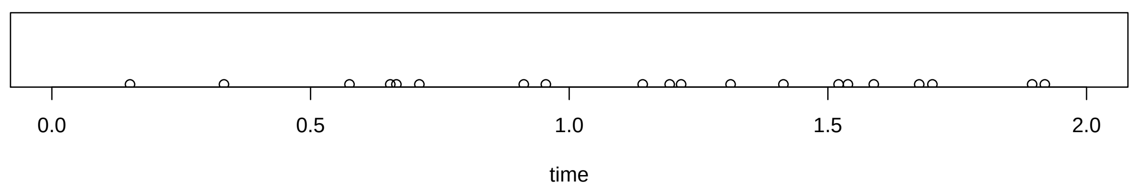


Figure 4.2: Events of a Poisson process

Example 4.13.

For the situation in Example 4.12, find the probability that the total monthly rainfall exceeds 5 inches.

Solution. $P(Y > 5) = 1 - F(5) = 1 - (1 - P(N < k)) = P(N < k)$ where $k = \alpha = 2$. Equation 4.3 yields $P(Y > 5) = e^{-5/1.6}(1 + 5/1.6) = 0.181$ $\square$

Exercises

4.13.

A truck has 2 spare tires. Under intense driving conditions, tire blowouts are determined to approximately follow a Poisson process with the intensity of 1.2 per 100 miles. Let X be the total distance the truck can go with 2 spare tires.

a) Find the expected value and the standard deviation of X

b) Find the probability that the truck can go at least 200 miles

4.14.

Differentiate Equation 4.3 for $k = 2$ to show that you indeed will get the Gamma density function with $\alpha = 2$.

4.6 Normal distribution

The most widely used of all the continuous probability distributions is the *normal distribution* (also known as Gaussian). It serves as a popular model for measurement errors, particle displacements under Brownian motion, stock market fluctuations, human intelligence and many other things. It is also used as an approximation for Binomial (for large n) and Gamma (for large α) distributions.

The normal density follows the well-known symmetric bell-shaped curve. The curve is centered at the mean value μ and its spread is, of course, measured by the standard deviation σ. These two parameters, μ and σ^2, completely determine the shape and center of the normal density function.

Definition 4.8.

The normal random variable X has the PDF

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[\frac{-(x-\mu)^2}{2\sigma^2}\right], \qquad \text{for } -\infty < x < \infty$$

It will be denoted as $X \sim \mathcal{N}(\mu, \sigma^2)$

The normal random variable Z with $\mu = 0$ and $\sigma = 1$ is said to have the *standard normal distribution*. Direct integration would show that $\mathbb{E}(Z) = 0$ and $V(Z) = 1$.

Usefulness of Z

We are able to transform the observations of any normal random variable X to a new set of observations of a standard normal random variable Z. This can be done by means of the transformation

$$Z = \frac{X - \mu}{\sigma}.$$

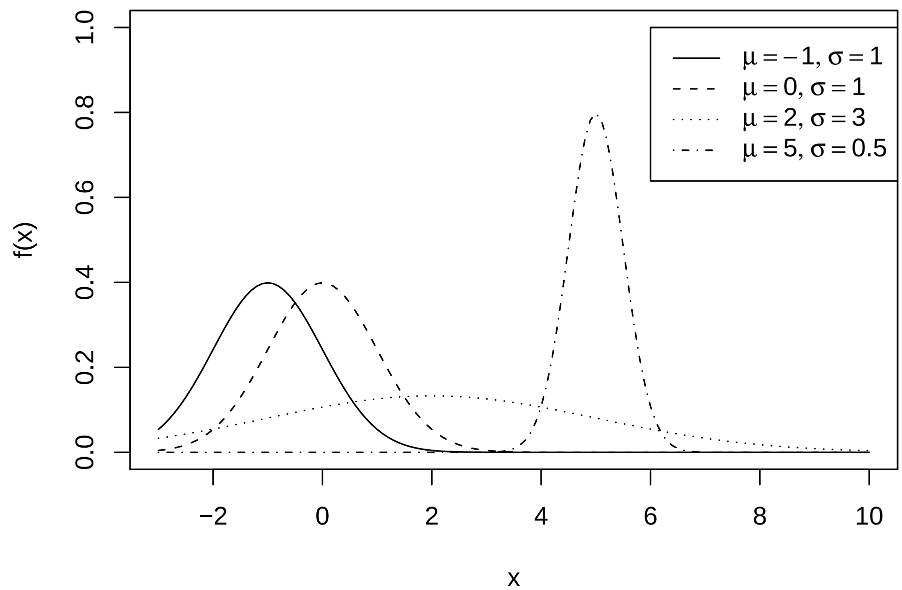


Figure 4.3: Normal densities

The values of the CDF of Z can be obtained from Table A. Namely,

$$F(z) = \begin{cases} 0.5 + \mathtt{TA}(z), & z \geq 0 \\ 0.5 - \mathtt{TA}(|z|), & z < 0 \end{cases}$$

where $\mathtt{TA}(z) = P(0 < Z < z)$ denotes table area of z. The second equation follows from the symmetry of the Z distribution.

Table A allows us to calculate probabilities and percentiles associated with normal random variables, as the direct integration of normal density is not possible.

Example 4.14.

If Z denotes a standard normal variable, find
(a) $P(Z \leq 1)$ (b) $P(Z > 1)$ (c) $P(Z < -1.5)$ (d) $P(-1.5 \leq Z \leq 0.5)$.
(e) Find a number, say z_0, such that $P(0 \leq Z \leq z_0) = 0.49$

Solution. This example provides practice in using Normal probability Table. We see that

a) $P(Z \leq 1) = P(Z \leq 0) + P(0 \leq Z \leq 1) = 0.5 + 0.3413 = 0.8413.$

b) $P(Z > 1) = 0.5 - P(0 \le Z \le 1) = 0.5 - 0.3413 = 0.1587$

c) $P(Z < -1.5) = P(Z > 1.5) = 0.5 - P(0 \le Z \le 1.5) = 0.5 - 0.4332 = 0.0668$.

d) $P(-1.5 \le Z \le 0.5) = P(-1.5 \le Z \le 0) + P(0 \le Z \le 0.5)$

$= P(0 \le Z \le 1.5) + P(0 \le Z \le 0.5) = 0.4332 + 0.1915 = 0.6247$.

e) To find the value of z_0 we must look for the given probability of 0.49 on the area side of Normal probability Table. The closest we can come is at 0.4901, which corresponds to a Z value of 2.33. Hence $z_0 = 2.33$.

□

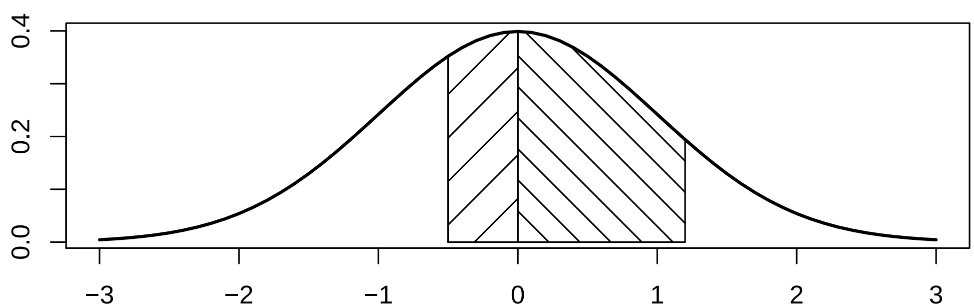


Figure 4.4: Splitting a normal area into two Table Areas

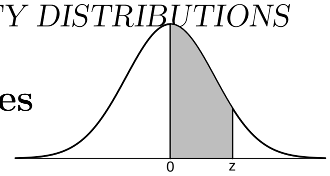
Example 4.15.
For $X \sim \mathcal{N}(50, 10^2)$, find the probability that X is between 45 and 62.

Solution. The Z- values corresponding to X = 45 and X = 62 are

$$Z_1 = \frac{45 - 50}{10} = -0.5 \quad \text{and} \quad Z_2 = \frac{62 - 50}{10} = 1.2.$$

Therefore, $P(45 \le X \le 62) = P(-0.5 \le Z \le 1.2) = \texttt{TA}(1.2) + \texttt{TA}(0.5) = 0.3849 + 0.1915 = 0.5764$ □

Table A: standard normal probabilities



z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
.0	.0000	.0040	.0080	.0120	.0160	.0199	.0239	.0279	.0319	.0359
.1	.0398	.0438	.0478	.0517	.0557	.0596	.0636	.0675	.0714	.0753
.2	.0793	.0832	.0871	.0910	.0948	.0987	.1026	.1064	.1103	.1141
.3	.1179	.1217	.1255	.1293	.1331	.1368	.1406	.1443	.1480	.1517
.4	.1554	.1591	.1628	.1664	.1700	.1736	.1772	.1808	.1844	.1879
.5	.1915	.1950	.1985	.2019	.2054	.2088	.2123	.2157	.2190	.2224
.6	.2257	.2291	.2324	.2357	.2389	.2422	.2454	.2486	.2517	.2549
.7	.2580	.2611	.2642	.2673	.2704	.2734	.2764	.2794	.2823	.2852
.8	.2881	.2910	.2939	.2967	.2995	.3023	.3051	.3078	.3106	.3133
.9	.3159	.3186	.3212	.3238	.3264	.3289	.3315	.3340	.3365	.3389
1.0	.3413	.3438	.3461	.3485	.3508	.3531	.3554	.3577	.3599	.3621
1.1	.3643	.3665	.3686	.3708	.3729	.3749	.3770	.3790	.3810	.3830
1.2	.3849	.3869	.3888	.3907	.3925	.3944	.3962	.3980	.3997	.4015
1.3	.4032	.4049	.4066	.4082	.4099	.4115	.4131	.4147	.4162	.4177
1.4	.4192	.4207	.4222	.4236	.4251	.4265	.4279	.4292	.4306	.4319
1.5	.4332	.4345	.4357	.4370	.4382	.4394	.4406	.4418	.4429	.4441
1.6	.4452	.4463	.4474	.4484	.4495	.4505	.4515	.4525	.4535	.4545
1.7	.4554	.4564	.4573	.4582	.4591	.4599	.4608	.4616	.4625	.4633
1.8	.4641	.4649	.4656	.4664	.4671	.4678	.4686	.4693	.4699	.4706
1.9	.4713	.4719	.4726	.4732	.4738	.4744	.4750	.4756	.4761	.4767
2.0	.4772	.4778	.4783	.4788	.4793	.4798	.4803	.4808	.4812	.4817
2.1	.4821	.4826	.4830	.4834	.4838	.4842	.4846	.4850	.4854	.4857
2.2	.4861	.4864	.4868	.4871	.4875	.4878	.4881	.4884	.4887	.4890
2.3	.4893	.4896	.4898	.4901	.4904	.4906	.4909	.4911	.4913	.4916
2.4	.4918	.4920	.4922	.4925	.4927	.4929	.4931	.4932	.4934	.4936
2.5	.4938	.4940	.4941	.4943	.4945	.4946	.4948	.4949	.4951	.4952
2.6	.4953	.4955	.4956	.4957	.4959	.4960	.4961	.4962	.4963	.4964
2.7	.4965	.4966	.4967	.4968	.4969	.4970	.4971	.4972	.4973	.4974
2.8	.4974	.4975	.4976	.4977	.4977	.4978	.4979	.4979	.4980	.4981
2.9	.4981	.4982	.4982	.4983	.4984	.4984	.4985	.4985	.4986	.4986
3.0	.4987	.4987	.4987	.4988	.4988	.4989	.4989	.4989	.4990	.4990

Example 4.16.

Given a random variable X having a normal distribution with $\mu = 300$ and $\sigma = 50$, find the probability that X is greater than 362.

Solution. To find $P(X > 362)$, we need to evaluate the area under the normal curve to the right of $x = 362$. This can be done by transforming $x = 362$ to the corresponding Z-value. We get

$$z = \frac{x - \mu}{\sigma} = \frac{362 - 300}{50} = 1.24$$

Hence $P(X > 362) = P(Z > 1.24) = P(Z < -1.24) = 0.5 - \texttt{TA}(1.24) = 0.1075$. □

Example 4.17.

A diameter X of a shaft produced has a normal distribution with parameters $\mu = 1.005, \sigma = 0.01$. The shaft will meet specifications if its diameter is between 0.98 and 1.02 cm. Which percent of shafts will not meet specifications?

Solution.

$$1 - P(0.98 < X < 1.02) = 1 - P\left(\frac{0.98 - 1.005}{0.01} < Z < \frac{1.02 - 1.005}{0.01}\right)$$

$= 1 - (0.4938 + 0.4332) = 0.0730$ □

4.6.1 Using Normal tables in reverse

Definition 4.9. Percentile

A pth *percentile* of a random variable X is the point q that leaves the area of $p/100\%$ to the left. That is, q is the solution for the equation

$$P(X \leq q) = p/100\%$$

For example, the median (introduced in Exercise 4.12) is the 50th percentile of a probability distribution.

We will discuss how to find percentiles of normal distribution. The previous two examples were solved by going first from a value of x to a z-value and then computing the desired area. In the next example we reverse the process and begin with a known area, find the z-value, and then determine x by rearranging the equation $z = \frac{x-\mu}{\sigma}$ to give

$$x = \mu + \sigma z$$

Using the Normal Table calculations, it's straightforward to show the following

The famous 68% - 95% rule

For a Normal population, 68% of all values lie in the interval $[\mu - \sigma, \mu + \sigma]$, and 95% lie in $[\mu - 2\sigma, \mu + 2\sigma]$.
In addition, 99.7% of the population lies in $[\mu - 3\sigma, \mu + 3\sigma]$.

Example 4.18.

Using the situation in Example 4.17, a diameter X of a shaft had $\mu = 1.005, \sigma = 0.01$. Give an interval that would contain 95% of all diameters.

Solution. The interval is $\mu \pm 2\sigma = 1.005 \pm 2(0.01)$, that is, from 0.985 to 1.025. □

Example 4.19.

The SAT Math exam is scaled to have the average of 500 points, and the standard deviation of 100 points. What is the cutoff score for top 10% of the SAT takers?

Solution. In this example we begin with a known area, find the z-value, and then find x from the formula $x = \mu + \sigma z$. The 90th percentile corresponds to the 90% area under the normal curve to the left of x. Thus, we also require a z-value that leaves 0.9 area to the left and hence, the Table Area of 0.4. From Table A, $P(0 < Z < 1.28) = 0.3997$. Hence

$$x = 500 + 100(1.28) = 628$$

Therefore, the cutoff for the top 10% is 628 points. □

Example 4.20.

Let X = monthly sick leave time have normal distribution with parameters $\mu = 200$ hours and $\sigma = 20$ hours.

a) What percentage of months will have sick leave below 150 hours?

b) What amount of time x_0 should be budgeted for sick leave so that the budget will not be exceeded with 80% probability?

Solution. (a) $P(X < 150) = P(Z < -2.25) = 0.5 - 0.4938 = 0.0062$
(b) $P(X < x_0) = P(Z < z_0) = 0.8$, which leaves a table area for z_0 of 0.3. Thus, $z_0 = 0.84$ and hence $x_0 = 200 + 20(0.84) = 216.8$ hours □

Quantile-Quantile (Q-Q) plots

If X is normal (μ, σ^2) distribution, then

$$X = \mu + \sigma Z$$

and there is a perfect linear relationship between X and Z. This is a graphical method for checking normality.

4.6.2 Normal approximation to Binomial

As another example of using the Normal distribution, consider the Normal approximation to Binomial distribution. This will be also used when discussing sample proportions.

Theorem 4.2. Normal approximation to Binomial

If X is a Binomial random variable with mean $\mu = np$ and variance $\sigma^2 = npq$, then the random variables

$$Z_n = \frac{X - np}{\sqrt{npq}}$$

approach the standard Normal as n gets large.

We already know one Binomial approximation (by Poisson). It mostly applies when the Binomial distribution in question has a skewed shape, that is, when p is close to 0 or 1. When the shape of Binomial distribution is close to symmetric, the Normal appoximation will work better.

Example 4.21.

Suppose X is Binomial with parameters $n = 15$, and $p = 0.4$, then $\mu = np = (15)(4) = 6$ and $\sigma^2 = npq = 15(0.4)(0.6) = 3.6$. Suppose we are interested in the probability that X assumes a value from 7 to 9 inclusive, that is, $P(7 \leq X \leq 9)$. The exact probability is given by

$$P(7 \leq X \leq 9) = \sum_{7}^{9} bin(x; 15, 0.4) = 0.1771 + 0.1181 + 0.0612 = 0.3564$$

For Normal approximation we find the area between $x_1 = 6.5$ and $x_2 = 9.5$ using z-values which are

$$z_1 = \frac{x_1 - np}{\sqrt{npq}} = \frac{x_1 - \mu}{\sigma} = \frac{6.5 - 6}{1.897} = 0.26,$$

and

$$z_2 = \frac{9.5 - 6}{1.897} = 1.85$$

Adding or removing 0.5 is called *continuity correction*. It arises when we try to approximate a distribution with integer values (here, Binomial) through the use of a continuous distribution (here, Normal). Shown in Fig.4.5, the sum over the discrete set $\{7 \leq X \leq 9\}$ is approximated by the integral of the continuous density from 6.5 to 9.5.

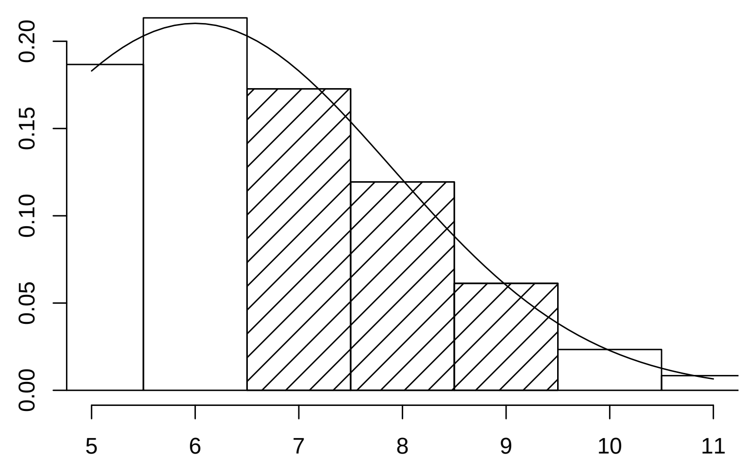


Figure 4.5: continuity correction

Now,

$$P(7 \leq X \leq 9) = P(0.26 < Z < 1.85) = 0.4678 - 0.1026 = 0.3652$$

therefore, the normal approximation provides a value that agrees very closely with the exact value of 0.3564. The degree of accuracy depends on both n and p. The approximation is very good when n is large and if p is not too near 0 or 1. To put it another way, if both np and nq are greater than or equal to 5, the approximation will be good. □

Example 4.22.
The probability that a patient recovers from a rare blood disease is 0.4. If 100 people are known to have contracted this disease, what is the probability that at most 30 survive?

Solution. Let the binomial variable X represent the number of patients that survive. Since $n = 100$ and $p = 0.4$, we have

$$\mu = np = (100)(0.4) = 40$$

and

$$\sigma^2 = npq = (100)(0.4)(0.6) = 24,$$

also $\sigma = \sqrt{\sigma^2} = 4.899$. To obtain the desired probability, we compute z-value for $x = 30.5$. Thus,

$$z = \frac{x-\mu}{\sigma} = \frac{30.5-40}{4.899} = -1.94,$$

and the probability of fewer than 30 of the 100 patients surviving is $P(X < 30) \approx P(Z < -1.94) = 0.5 - 0.4738 = 0.0262$. □

Example 4.23.
A fair coin ($p = 0.5$) is tossed 10,000 times, and the number of Heads X is recorded. What are the values that contain X with 95% certainty?

Solution.
We have $\mu = np = 10,000(0.5) = 5,000$ and $\sigma = \sqrt{10,000(0.5)(1-0.5)} = 50$. We need to find x_1 and x_2 so that $P(x_1 \le X \le x_2)$. Since the mean of X is large, we will neglect the continuity correction.

Since we will be working with Normal approximation, let's find z_1 and z_2 such that

$$P(z_1 \le Z \le z_2) = 0.95$$

The solution is not unique, but we can choose the values of $z_{1,2}$ that are symmetric about 0. This will mean finding z such that $P(0 < Z < z) = 0.475$.

Using Normal tables "in reverse" we will get $z = 1.96$. Thus, $P(-1.96 < Z < 1.96) = 0.95$.

Next, transforming back into X, use the formula $x = \mu + \sigma z$, so

$$x_1 = 5000 + 50(-1.96) = 4902 \quad \text{and} \quad x_2 = 5000 + 50(1.96) = 5098$$

Thus, with a large likelihood, our Heads count will be within 100 of the expected value of 5,000.

This is an example of the famous "2 sigma" rule. □

Exercises

4.15.

Given a normal distribution with $\mu = 30$ and $\sigma = 6$, find

a) the normal curve area to the right of $x = 17$

b) the normal curve area to the left of $x = 22$

c) the normal curve area between $x = 32$ and $x = 41$

d) the value of x that has 80% of the normal curve area to the left

e) the two values of x that contain the middle 75% of the normal curve area.

4.16.

Given the normally distributed variable X with mean 18 and standard deviation 2.5, find

a) $P(X < 15)$

b) the value of k such that $P(X < k) = 0.2236$

c) the value of k such that $P(X > k) = 0.1814$

d) $P(17 < X < 21)$.

4.17.

A soft drink machine is regulated so that it discharges an average of 200 milliliters (ml) per cup. If the amount of drink is normally distributed with a standard deviation equal to 15 ml,

a) what fraction of the cups will contain more than 224 ml?

b) what is the probability that a cup contains between 191 and 209 milliliters?

c) how many cups will probably overflow if 230 ml cups are used for the next 1000 drinks?

d) below what value do we get the smallest 25% of the drinks?

4.18.

A company pays its employees an average wage of \$15.90 an hour with a standard deviation of \$1.50. If the wages were approximately normally distributed and paid to the nearest cent,

a) What percentage of workers receive wages between \$13.75 and \$16.22 an hour?

b) What is the cutoff value for highest paid 5% of the employees?

4.19.

The likelihood that a job application will result in an interview is estimated as 0.1. A grad student has mailed 40 applications. Find the probability that she will get at least 3 interviews,

a) Using the Normal approximation.

b) Using the Poisson approximation.

c) Find the exact probability. Which approximation has worked better? Why?

4.20.

It is estimated that 33% of individuals in a population of Atlantic puffins have a certain recessive gene. If 90 individuals are caught, estimate the probability that there will be between 30 and 40 (inclusive) with the recessive gene.

4.7 Weibull distribution

The Weibull distribution is, like Gamma, another extension of the Exponential distribution. It has positive values and is, therefore, applied to model reliability and lifetimes, among other things.

The easiest way to look at the Weibull distribution is through its CDF

$$F(x) = 1 - \exp[-(x/\beta)^\gamma],\ x > 0 \tag{4.4}$$

Note: if $\gamma = 1$ then we get the Exponential distribution. The parameter β has the dimension of time and γ is dimensionless.

By differentiating the CDF, we get the Weibull density

Definition 4.10. Weibull distribution

The Weibull RV has the density function

$$f(x) = \frac{\gamma x^{\gamma-1}}{\beta^\gamma} \exp\left[-\left(\frac{x}{\beta}\right)^\gamma\right], \quad x > 0$$

and the CDF

$$F(x) = 1 - \exp[-(x/\beta)^\gamma], \; x > 0$$

Its mean is $\mu = \beta\,\Gamma\left(1 + \frac{1}{\gamma}\right)$

The Weibull distribution with $\gamma > 1$ typically has an asymmetric shape with a peak in the middle and the long right "tail". Shapes of Weibull density are shown in Fig. 4.6 for various values of γ.

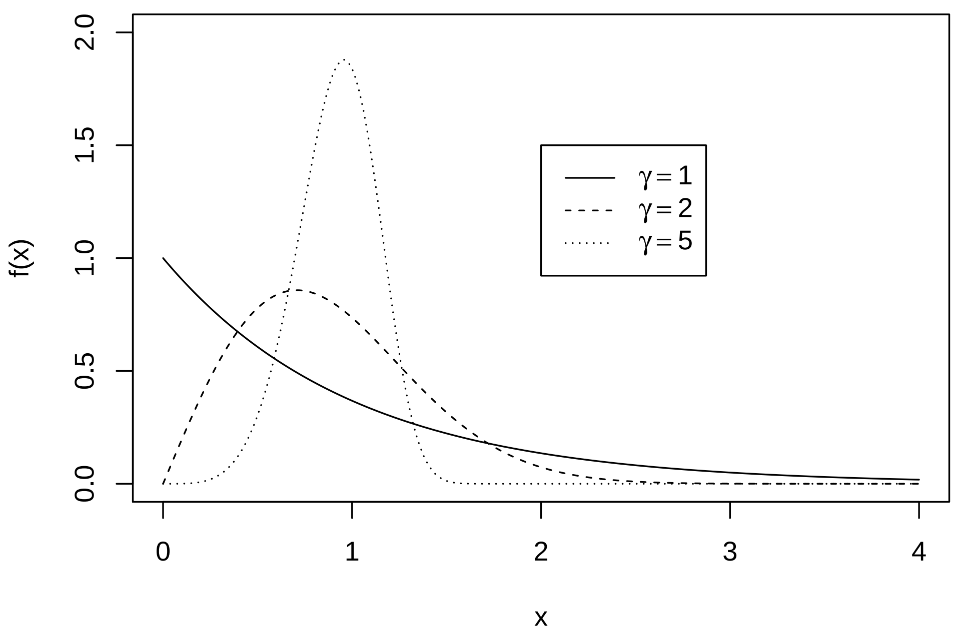


Figure 4.6: Weibull densities, all with $\beta = 1$

Regarding the computation of the mean: the Gamma function of non-integer parameter is, generally, not easy to find. Note only that $\Gamma(0.5) = \sqrt{\pi}$, and we can use the recursive relation $\Gamma(\alpha + 1) = \alpha\Gamma(\alpha)$ to compute the Gamma function for $\alpha = 1.5, 2.5$ etc.

Example 4.24.
The duration of subscription to the Internet services is modeled by the Weibull distribution with parameters $\gamma = 2$ and $\beta = 15$ months.

a) Find the average duration.

b) Find the probability that a subscription will last longer than 10 months.

Solution.
(a) $\mu = 15\,\Gamma(1.5) = 15(0.5)\Gamma(0.5) = 7.5\sqrt{\pi} = 13.29$

(b) $P(X > 10) = 1 - F(10) = \exp[-(10/15)^2] = 0.6412$ □

Exercises

4.21.
The time it takes for a server to respond to a request is modeled by the Weibull distribution with $\gamma = 2/3$ and $\beta = 15$ milliseconds.

a) Find the average time to respond.

b) Find the probability that it takes less than 12 milliseconds to respond.

c) Find the 70th percentile of the response times.

4.8 Moment generating functions for continuous case

The moment generating function of a continuous random variable X with a pdf of $f(x)$ is given by

$$M(t) = \mathbb{E}\,(e^{tX}) = \int_{-\infty}^{\infty} e^{tx} f(x)dx$$

when the integral exists. For the exponential distribution, this becomes

$$M(t) = \int_{0}^{\infty} e^{tx}\frac{1}{\beta}e^{-x/\beta}dx = (1 - \beta t)^{-1}$$

For properties of MGF's, see Section 3.9

Chapter 5

Joint probability distributions

5.1 Bivariate and marginal probability distributions

All of the random variables discussed previously were one dimensional, that is, we consider random quantities one at a time. In some situations, however, we may want to record the simultaneous outcomes of several random variables.

Examples:

a) We might measure the amount of precipitate A and volume V of gas released from a controlled chemical experiment, giving rise to a two-dimensional sample space.

b) A physician studies the relationship between exercise amount and pulse rate of his patients.

c) An educator studies the relationship between students' grades and time devoted to study.

If X and Y are two discrete random variables, the probability that X equals x while Y equals y is described by $p(x, y) = P(X = x, Y = y)$. That is, the function $p(x, y)$ describes the probability behavior of the pair X, Y.

Definition 5.1. Joint PMF

The function $p(x, y)$ is a joint probability mass function of the discrete random variables X and Y if

a) $p(x, y) \geq 0$ for all pairs (x, y),

b) $\sum_x \sum_y p(x, y) = 1$,

c) $P(X = x, Y = y) = p(x, y)$.

For any region A in the xy-plane, $P[(x, y) \text{ belongs to } A] = \sum_{(x,y)\in A} p(x, y)$.

Definition 5.2. Marginal PMF

The marginal probability functions of X and Y respectively are given by

$$p_X(x) = \sum_y p(x, y) \quad \text{and} \quad p_Y(y) = \sum_x p(x, y)$$

Example 5.1.
If two dice are rolled independently, then the numbers X and Y on the first and second die, respectively, will each have marginal PMF $p(x) = 1/6$ for $x = 1, 2, ..., 6$.

The joint PMF is $p(x, y) = 1/36$, so that $p(x) = \sum_{y=1}^{6} p(x, y)$ □

Example 5.2.
Consider X = person's age and Y = income.
The data are abridged from the US Current Population Survey, see
`http://www.census.gov/hhes/www/cpstables/032010/perinc/new01_001.htm`

For the purposes of this example, we replace the age and income groups by their midpoints. For example, the first row represents ages 25-34 and the first column represents incomes \$0-\$10,000.

	Y, income					
	5	20	40	60	85	Total
X, age 30	0.049	0.116	0.084	0.039	0.032	0.320
40	0.042	0.093	0.081	0.045	0.061	0.322
50	0.047	0.102	0.084	0.053	0.072	0.358
Total	0.139	0.310	0.249	0.137	0.165	1.000

Here, the joint PMF is given inside the table and the marginal PMF's of X and Y are row and column totals, respectively.

For example, $p(30,60) = 0.039$ and $p_Y(40) = 0.084 + 0.081 + 0.084 = 0.249$. $\square$

For continuous random variables, the PMF's turn to densities, and summation to integration.

Definition 5.3. Joint density, marginal densities

The function $f(x,y)$ is a **joint probability density function** for the continuous random variables X and Y if

a) $f(x,y) \geq 0,$ for all (x,y)

b) $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)\,dx\,dy = 1.$

c) $P[(X,Y) \in A] = \iint_A f(x,y)\,dx\,dy$ for any region A in the xy-plane.

The **marginal probability density functions** of X and Y are given by

$$f_X(x) = \int_{-\infty}^{\infty} f(x,y)\,dy \quad \text{and } f_Y(y) = \int_{-\infty}^{\infty} f(x,y)\,dx$$

When X and Y are continuous random variables, the joint density function $f(x,y)$ describes the likelihood that the pair (X,Y) belongs to the neighborhood of the point (x,y). It is visualized as a surface lying above the xy plane.

Example 5.3.

A certain process for producing an industrial chemical yields a product that contains two main types of impurities. Suppose that the joint probability distribution of the impurity concentrations (in mg/l) X and Y is given by

$$f(x,y) = \begin{cases} 2(1-x) & \text{for } 0 < x < 1,\ 0 < y < 1 \\ 0 & \text{elsewhere} \end{cases}$$

(a) Verify the condition (b) of Definition 5.3

(b) Find $P(0 < X < 0.5, 0.4 < Y < 0.7)$

(c) Find the marginal probability density functions for X and Y.

Solution. (b)

$$P(0 < X < 0.5,\ 0.4 < Y < 0.7) = \int_{0.4}^{0.7}\int_{0}^{0.5} 2(1-x)dx\,dy = 0.225$$

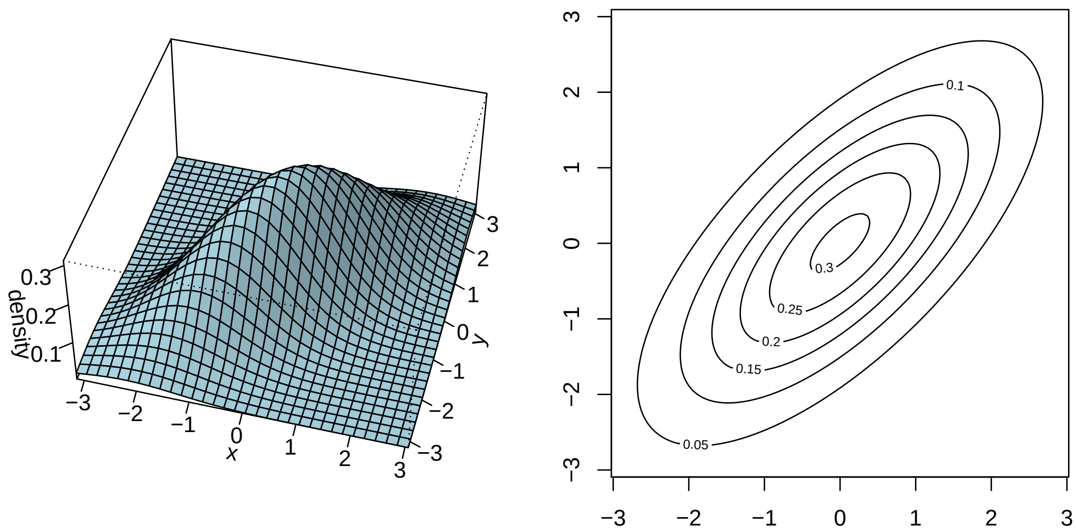


Figure 5.1: An example of a joint density function. Left: surface plot. Right: contour plot.

(c)

$$f_X(x) = \int_{-\infty}^{\infty} f(x,y)dy = \int_0^1 2(1-x)dy = 2(1-x), \quad 0 < x < 1$$

and $f_Y(y) = \int_{-\infty}^{\infty} f(x,y)dx = \int_0^1 2(1-x)dx = 1, \quad 0 < y < 1$ □

5.2 Conditional probability distributions

Definition 5.4. Conditional PMF or density

For a pair of **discrete** RV's, the conditional PMF of X given Y is

$$p(x \mid y) = \frac{p(x,y)}{p_Y(y)} \text{ for } y \text{ such that } p(y) > 0$$

For a pair of **continuous** RV's with joint density $f(x, y)$, the conditional density function of X given $Y = y$ is defined as

$$f(x|y) = \frac{f(x, y)}{f_Y(y)} \text{ for } y \text{ such that } f_Y(y) > 0$$

and the conditional density of Y given $X = x$ is defined by

$$f(y|x) = \frac{f(x, y)}{f_X(x)} \text{ for } x \text{ such that } f_X(x) > 0$$

For discrete RV's, the conditional probability distribution of X given Y fixes a value of Y. For example, conditioning on $Y = 0$, produces

$$P(X = 0 \,|\, Y = 0) = \frac{P(X = 0, Y = 0)}{P(Y = 0)}$$

Example 5.4.
Using the data from Example 5.2,

		Y, income					
		5	20	40	60	85	Total
X, age	30	0.049	0.116	0.084	0.039	0.032	0.320
	40	0.042	0.093	0.081	0.045	0.061	0.322
	50	0.047	0.102	0.084	0.053	0.072	0.358
Total		0.139	0.310	0.249	0.137	0.165	1.000

Calculate the conditional PMF of Y given $X = 30$.

Solution. Conditional PMF of Y given $X = 30$, will give the distribution of incomes in that age group. Divide all of the row $X = 30$ by its marginal and obtain

		Y, income		
		5	20	40
X, age	30	$0.049/0.320 = 0.153$	$0.116/0.320 = 0.362$	$0.084/0.320 = 0.263$

	[continued] 60	85	Total
	$0.039/0.320 = 0.122$	$0.032/0.32 = 0.1$	1

The conditional PMF's will add up to 1. □

Example 5.5.

The joint density for the random variables (X, Y), where X is the unit temperature change and Y is the proportion of spectrum shift that a certain atomic particle produces is

$$f(x,y) = \begin{cases} 10xy^2 & \text{for } 0 < x < y < 1 \\ 0 & \text{elsewhere} \end{cases}$$

(a) Find the marginal densities.
(b) Find the conditional densities $f(x|y)$ and $f(y|x)$.

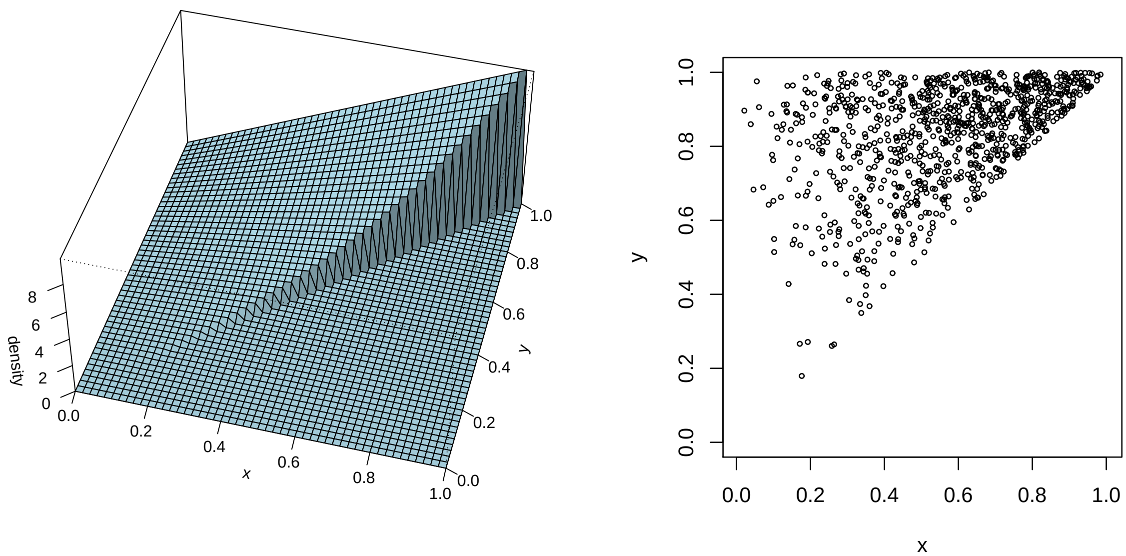


Figure 5.2: Left: Joint density from Example 5.5, right: a typical sample from this distribution

Solution. (a) By definition,

$$f_X(x) = \int_x^1 10xy^2 dy = \frac{10}{3}x(1 - x^3), \quad 0 < x < 1$$

$$f_Y(y) = \int_0^y 10xy^2 dx = 5y^4, \quad 0 < y < 1$$

(b) Now

$$f(y|x) = \frac{f(x,y)}{f_X(x)} = \frac{10xy^2}{(10/3)x(1-x^3)} = \frac{3y^2}{(1-x^3)},\ 0 < x < y < 1$$

and

$$f(x|y) = \frac{f(x,y)}{f_Y(y)} = \frac{10xy^2}{5y^4} = \frac{2x}{y^2},\ 0 < x < y < 1$$

For the last one, say, treat y as fixed (given) and x is the variable. □

5.3 Independent random variables

Definition 5.5. Independence

The random variables X and Y are said to be *statistically independent* iff

$$p(x,y) = p_X(x)p_Y(y) \text{ for discrete case}$$

and

$$f(x,y) = f_X(x)f_Y(y) \quad \text{for continuous case}$$

This definition of independence agrees with our definition for the events, $P(AB) = P(A)P(B)$. For example, if two dice are rolled independently, then the numbers X and Y on the first and second die, respectively, will each have PMF $p(x) = 1/6$ for $x = 1, 2, ..., 6$. The joint PMF will then be $p(x,y) = p_X(x)p_Y(y) = (1/6)^2 = 1/36$.

Example 5.6.
Show that the random variables in Example 5.3 are independent.

Solution. Here,

$$f(x,y) = \begin{cases} 2(1-x) & \text{for } 0 < x < 1 \text{ and } 0 < y < 1 \\ 0 & \text{elsewhere} \end{cases}$$

We have $f_X(x) = 2(1-x)$ and $f_Y(y) = 1$ from Example 5.3, thus

$$f_X(x)f_Y(y) = 2(1-x)(1) = 2(1-x) = f(x,y)$$

for $0 < x, y < 1$ and 0 elsewhere. Hence, X and Y are independent random variables. □

Exercises

5.1.

The joint distribution for the number of total sales $=X_1$ and number of electronic equipment sales $=X_2$ per hour for a wholesale retailer are given below

X_2	0	1	2	
$X_1 = 0$	0.1	0	0	
$X_1 = 1$	0.1	0.2	0	
$X_1 = 2$	0.1	?	0.15	

a) Fill in the "?"

b) Compute the marginal probability function for X_2.
(That is, find $P(X_2 = i)$ for every i.)

c) Find the probability that both $X_1 \leq 1$ and $X_2 \leq 1$.

d) Find the *conditional* probability distribution for X_2 given that $X_1 = 2$.
(That is, find $P(X_2 = i \,|\, X_1 = 2)$ for every i.)

e) Are X_1, X_2 independent? Explain.

5.2.

X and Y have the following joint density:

$$f(x,y) = \begin{cases} k & \text{for } \ 0 \leq x \leq y \leq 1 \\ 0 & \text{elsewhere} \end{cases}$$

a) Calculate the constant k that makes f a legitimate density.

b) Calculate the marginal densities of X and Y.

5.4 Expected values of functions

Definition 5.6. Expected values

Suppose that the discrete RV's (X, Y) have a joint PMF $p(x, y)$. If $g(x, y)$ is any real-valued function, then

$$\mathbb{E}\left[g(X, Y)\right] = \sum_{x} \sum_{y} g(x, y) p(x, y).$$

The sum is over all values of (x, y) for which $p(x, y) > 0$.

If (X, Y) are continuous random variables, with joint PDF $f(x, y)$, then

$$\mathbb{E}\left[g(X, Y)\right] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y)\, f(x, y) dx\, dy.$$

Definition 5.7. Covariance

The *covariance* between two random variables X and Y is given by

$$Cov(X, Y) = \mathbb{E}\left[(X - \mu_X)(Y - \mu_Y)\right],$$

where $\mu_X = \mathbb{E}(X)$ and $\mu_Y = \mathbb{E}(Y)$.

The covariance helps us assess the relationship between two variables. Positive covariance means *positive association* between X and Y meaning that, as X increases, Y also tends to increase. Negative covariance means *negative association*.

In Figure 5.3, positive covariance is achieved as pairs of x, y with positive products have higher densities than those with the negative products.

This definition also extends our notion of variance as $Cov(X, X) = V(X)$.

While covariance measures the direction of the association between two random variables, its value is not directly interpretable. *Correlation coefficient*, introduced below, measures the strength of the association and has some nice properties.

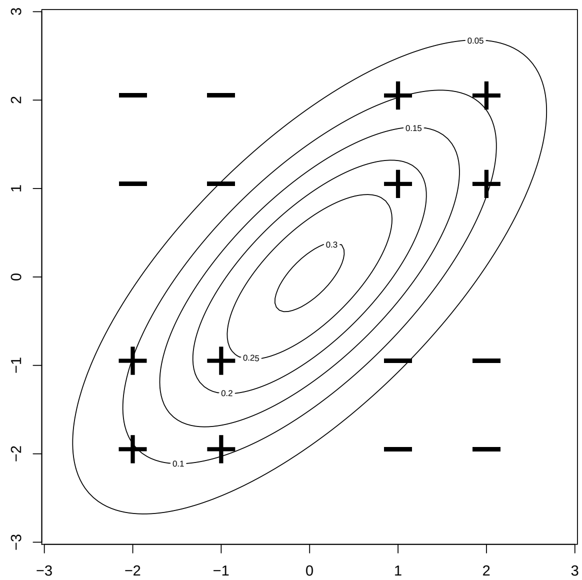


Figure 5.3: Explanation of positive covariance

Definition 5.8. Correlation

The *correlation coefficient* between two random variables X and Y is given by

$$\rho = \frac{Cov(X,Y)}{\sqrt{V(X)\ V(Y)}}$$

Properties of correlation:

- The correlation coefficient lies between -1 and $+1$.
- The correlation coefficient is dimensionless (while covariance has dimension of XY).
- If $\rho = +1$ or $\rho = -1$, then Y must be a linear function of X.
- The correlation coefficient does not change when X or Y are linearly transformed (e.g. when you change the units from miles to ångströms.)
- However, the correlation coefficient is not a good indicator of a *nonlinear* relationship.

The following Theorem simplifies the computation of covariance. Compare it to the variance identity $V(X) = \mathbb{E}(X^2) - (\mathbb{E}X)^2$.

Theorem 5.1. Covariance

$$Cov(X,Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)$$

Example 5.7.
The fraction X of male runners and the fraction Y of female runners who compete in marathon races is described by the joint density function

$$f(x,y) = \begin{cases} 8xy & \text{for } 0 \le x \le 1, \quad 0 \le y \le x \\ 0 & \text{elsewhere} \end{cases}$$

Find the covariance of X and Y.

Solution. We first compute the marginal density functions. They are

$$f(x) = \begin{cases} 4x^3 & \text{for } 0 \le x \le 1 \\ 0 & \text{elsewhere} \end{cases}$$

and

$$f(y) = \begin{cases} 4y(1-y^2) & \text{for } 0 \le y \le 1 \\ 0 & \text{elsewhere} \end{cases}$$

From the marginal density functions, we get

$$\mathbb{E}(X) = \int_0^1 4x^4 dx = \frac{4}{5} \quad \text{and } \mathbb{E}(Y) = \int_0^1 4y^2(1-y^2)dy = \frac{8}{15}$$

From the joint density functions given, we have

$$\mathbb{E}(XY) = \int_0^1 \int_y^1 8x^2y^2 dxdy = \frac{4}{9}.$$

Then

$$Cov(X,Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y) = \frac{4}{9} - \left(\frac{4}{5}\right)\left(\frac{8}{15}\right) = \frac{4}{225}$$

□

Theorem 5.2. Covariance and independence

If random variables X and Y are independent, then $Cov(X, Y) = 0$.

Proof. We will show the proof for the continuous case; the discrete case follows similarly.
For independent X, Y,

$$\mathbb{E}(XY) = \iint f(x,y)dx\,dy = \iint f_X(x)f_Y(y)dx\,dy =$$

$$= \left(\int f_X(x)dx\right)\left(\int f_Y(x)dy\right) = \mathbb{E}(X)\,\mathbb{E}(Y)$$

Therefore, $Cov(X, Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y) = 0$. □

Of course, if covariance is 0, then so is the correlation coefficient. Such random variables are called *uncorrelated*. The inverse of this Theorem is not true, meaning that **zero covariance does not necessarily imply independence**.

5.4.1 Variance of sums

The following Theorem simplifies calculation of variance in certain cases.

Theorem 5.3. Variance of sums

If X and Y are random variables and $U = aX + bY + c$, then

$$V(U) = V(aX + bY + c) = a^2V(X) + b^2V(Y) + 2ab\,Cov(X, Y)$$

If X and Y are independent then $V(U) = V(aX + bY) = a^2V(X) + b^2V(Y)$

Example 5.8.

If X and Y are random variables with variances $V(X) = 2$, $V(Y) = 4$, and covariance $Cov(X, Y) = -2$, find the variance of the random variable $Z = 3X - 4Y + 8$.

Solution. By Theorem 5.3,

$$V(Z) = \sigma_Z^2 = V(3X - 4Y + 8) = 9V(X) + 16V(Y) - 24\,Cov(X, Y)$$

so $V(Z) = (9)(2) + (16)(4) - 24(-2) = 130$. □

Corollary. If the random variables X and Y are independent, then

$$V(X + Y) = V(X) + V(Y)$$

Note. Theorem 5.3 and the above Corollary naturally extend to more than 2 random variables. If X_1, X_2, ..., X_n are all independent RV's, then

$$V(X_1 + X_2 + ... + X_n) = V(X_1) + V(X_2) + ... + V(X_n)$$

Example 5.9.
We have discussed in Chapter 3 that the Binomial random variable Y with parameters n, p can be represented as $Y = X_1 + X_2 + ... + X_n$. Here X_i are independent Bernoulli (0/1) random variables with $P(X_i = 1) = p$.

If was found that $V(X_i) = p(1-p)$. Then, using the above Note, $V(Y) = V(X_1) + V(X_2) + ... + V(X_n) = np(1-p)$, which agrees with the formula for Binomial variance in Section 3.4.

The same reasoning applies to Gamma RV's. If $Y = X_1 + X_2 + ... + X_n$, where X_i are independent Exponentials, each with mean β, then we know that $V(X_i) = \beta^2$ and Y has Gamma distribution with $\alpha = n$. Then, $V(Y) = V(X_1) + V(X_2) + ... + V(X_n) = n\beta^2$. □

Example 5.10.
A very important application of Theorem 5.3 is the calculation of variance of the **sample mean**

$$\overline{X} = \frac{X_1 + X_2 + ... + X_n}{n} = \frac{Y}{n}$$

where X_i are independent and identically distributed RV's (representing a sample of measurements), and Y denotes the total of all measurements.

Suppose that $V(X_i) = \sigma^2$ for each i. Then

$$V(\overline{X}) = \frac{V(Y)}{n^2} = \frac{V(X_1) + V(X_2) + ... + V(X_n)}{n^2} = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$$

This means that $\sigma_{\overline{X}} = \sigma/\sqrt{n}$, that is, **the mean of n independent measurements is $\sqrt{n}$ more precise than a single measurement.** □

Exercises

5.3.

Y	0	1	2	
$X = 0$	0.1	0	0	
$X = 1$	0.1	0.2	0	
$X = 2$	0.1	0.35	0.15	

Find the covariance and correlation between X and Y.

5.4.

X and Y have the following joint density:

$$f(x, y) = \begin{cases} 2 & \text{for } 0 \le x \le y \le 1 \\ 0 & \text{elsewhere} \end{cases}$$

a) Calculate $\mathbb{E}(X^2Y)$.

b) Calculate $\mathbb{E}(X/Y)$.

5.5.

Using the density in Problem 5.4, find the covariance and correlation between X and Y.

5.6.

Ten people get into an elevator. Assume that their weights are independent, with the mean 150 lbs and standard deviation 30 lbs.

a) Find the expected value and the standard deviation of their total weight.

b) Assuming Normal distribution, find the probability that their combined weight is less than 1700 pounds.

5.7.

While estimating speed of light in a transparent medium, an individual measurement X is determined to be unbiased (that is, the mean of X equals the unknown speed of light), but the measurement error, assessed as the standard deviation of X, equals 35 kilometers per second (km/s).

a) In an experiment, 20 independent measurements of the speed of light were made. What is the standard deviation of the mean of these measurements?

b) How many measurements should be made so that the error in estimating the speed of light (measured as $\sigma_{\overline{X}}$) will decrease to 5 km/s?

5.8.

A part is composed of two segments. One segment is produced with the mean length 4.2cm and standard deviation of 0.1cm, and the second segment is produced with the mean length 2.5cm and standard deviation of 0.05cm. Assuming that the production errors are independent, calculate the mean and standard deviation of the total part length.

5.9.

Random variables X and Y have means 3 and 5, and variances 0.5 and 2, respectively. Further, the correlation coefficient between X and Y equals -0.5. Find the mean and variance of $W = X - Y$.

5.10. $\star$

Find an example of uncorrelated, but not independent random variables. [**Hint:** Two discrete RV's with 3 values each are enough.]

5.5 Conditional Expectations

Definition 5.9. Conditional Expectation

If X and Y are any two random variables, the conditional expectation of X given that $Y = y$ is defined to be

$$\mathbb{E}\left(X \mid Y = y\right) = \int_{-\infty}^{\infty} x f(x|y) dx$$

if X and Y are jointly continuous, and

$$\mathbb{E}\left(X \mid Y = y\right) = \sum_{x} x\, p(x|y)$$

if X and Y are jointly discrete.

Note that $\mathbb{E}(X|Y=y)$ is a number depending on y. If now we allow y to vary randomly, we get a random variable denoted by $\mathbb{E}(X|Y)$. The concept of conditional expectation is useful when we have only a partial information about X, as in the following example.

Example 5.11.
Suppose that random variable X is the number rolled on a die, and $Y = 0$ when $X \leq 3$ and $Y = 1$ otherwise. Thus, Y carries partial information about X, namely, whether $X \leq 3$ or not.

a) Compute the conditional expectation $\mathbb{E}(X \mid Y = 0)$.
b) Describe the random variable $\mathbb{E}(X \mid Y)$.

Solution. (a) The conditional distributions of X are given by

$$P(X = x \mid Y = 0) = \frac{P(X = x, Y = 0)}{P(Y = 0)} = \frac{1/6}{1/2} = 1/3$$

for $x = 1, 2, 3$, and
$P(X = x \mid Y = 1) = 1/3$ for $x = 4, 5, 6$.

Thus, $\mathbb{E}(X \mid Y = 0) = (1/3)(1 + 2 + 3) = 2$ and
$\mathbb{E}(X \mid Y = 1) = (1/3)(4 + 5 + 6) = 5$

(b) $\mathbb{E}(X \mid Y)$ is 2 or 5, depending on Y. Each value may happen with probability 1/2. Thus, $P[\mathbb{E}(X \mid Y) = 2] = 0.5$ and $P[\mathbb{E}(X \mid Y) = 5] = 0.5$ □

Theorem 5.4. Expectation of expectation

Let X and Y denote random variables. Then

$$(a) \quad \mathbb{E}(X) = \mathbb{E}[\mathbb{E}(X|Y)]$$

$$(b) \quad V(X) = \mathbb{E}[V(X|Y)] + V[\mathbb{E}(X|Y)]$$

Proof. (Part (a) only.)
Let X and Y have joint density $f(x, y)$ and the marginal densities $f_X(x)$ and $f_Y(y)$, respectively. Then

$$\mathbb{E}(X) = \int_{-\infty}^{\infty} x f_X(x) dx = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x f(x, y) dx\, dy$$

$$= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x f(x|y) f(y) dx\, dy = \int_{-\infty}^{\infty} \left[\int_{-\infty}^{\infty} x f(x|y) dx \right] f(y) dy$$

$$= \int_{-\infty}^{\infty} \mathbb{E}\left(X|Y = y\right) f(y) dy = \mathbb{E}\left[\mathbb{E}\left(X|Y\right)\right]$$

□

Example 5.12.

Suppose that the total weight X of occupants in a car depends on how many there are, let the number of occupants equal Y, and each occupant weighs 150 lbs on average. Then $\mathbb{E}\left(X \mid Y = y\right) = 150y$. Suppose Y has the following distribution

y	1	2	3	4
$p(y)$	0.62	0.28	0.07	0.03
$150y$	150	300	450	600

Then $\mathbb{E}\left(X \mid Y\right)$ has the distribution with values given in the last row of the table, and probabilities identical to $p(y)$. We can verify by straightforward calculation that $\mathbb{E}\left(X \mid Y\right) = \mathbb{E}\left(150Y\right) = 226.5$. Then the Theorem says that $\mathbb{E}\left(X\right) = 226.5$ as well, so we don't even have to know the distribution of occupant weights, only its mean (150). □

Exercises

5.11.

For the random variables X and Y from Example 5.11, verify the identity in part (a) of the Theorem 5.4.

5.12.

Suppose that the number of lobsters caught in a trap follows the distribution

y	0	1	2	3
$p(y)$	0.5	0.3	0.15	0.05

and the average weight of lobster is 1.7 lbs, with variance 0.25 lbs^2. Find the expected value and the variance of the total catch in one trap.

Chapter 6

Functions of Random Variables

6.1 Introduction

At times we are faced with a situation where we must deal not with the random variable whose distribution is known but rather with some function of that random variable. For example, we might know the distribution of particle sizes, and would like to infer the distribution of particle weights.

In the case of a simple linear function, we have already asserted what the effect is on the mean and variance. What has been omitted was what actually happens to the distribution.

We will discuss several methods of obtaining the distribution of $Y = g(X)$ from known distribution of X. The CDF method and the transformation method are most frequently used. The CDF method is all-purpose and flexible. The transformation method is typically faster (when it works).

6.1.1 Simulation

One use of these methods is to generate random variables with a given distribution. This is important in simulation studies. Suppose that we have a complex operation that involves several components. Suppose that each component is described by a random variable and that the outcome of the operation depends on the components in a complicated way. One approach to analyzing such a system is to simulate each component and calculate the outcome for the simulated values. If we repeat the simulation many times, then we can get an idea of the probability distribution of the outcomes.

6.2 Method of distribution functions (CDF)

The CDF method is straightforward and very versatile. The procedure is to derive the CDF for $Y = g(X)$ in terms of both the CDF of X, $F(x)$, and the function g, while also noting how the range of possible values changes. This is done by starting with the computation of $P(Y < y)$ and inverting this into a statement that can often be expressed in terms of the CDF of X.

If we also need to find the density of Y, we can do this by differentiating its CDF.

Example 6.1.
Suppose X has cdf given by $F(x) = 1 - e^{-\lambda x}$, so that X is Exponential with the mean $1/\lambda$. Let $Y = bX$ where $b > 0$. Note that the range of Y is the same as the range of X, namely $(0, \infty)$.

$$P(Y < y) = P(bX < y) = P(X < y/b) =$$

(Since $b > 0$, the inequality sign does not change.)

$$= 1 - e^{-\lambda y/b} = 1 - e^{-(\lambda/b)y}$$

The student should recognize this as CDF of the exponential distribution with the mean b/λ. We already knew that the mean would be b/λ, but we did not know that Y also has an exponential distribution. □

Example 6.2.
Suppose X has a uniform distribution on $[a, b]$ and $Y = cX + d$, with $c > 0$. Find the CDF of Y.

Solution. Recall that $F(t) = (t - a)/(b - a)$. Note that the range of Y is $[ca + d, cb + d]$. We have

$$P(Y < t) = P(cX + d < t) = P(X < (t - d)/c) = F((t - d)/c)$$

$$= ((t - d)/c - a)/(b - a) = (t - d - ac)/(c(b - a))$$

With a little algebra, this can be shown to be the uniform CDF on $[ca + d, cb + d]$. □

This example shows that certain simple transformations do not change the distribution type, only the parameters. Sometimes, however, the change is dramatic.

Example 6.3.
Show that if X has a uniform distribution on the interval $[0, 1]$ then $Y = -\ln(1 - X)$ has an exponential distribution with mean 1.

Solution. Recall that for the uniform distribution on $(0, 1)$, $P(X < x) = x$. Also, note that the range of Y is $(0, \infty)$.

$$P(Y < t) = P(-\ln(1 - X) < t) = P(\ln(1 - X) > -t) = \\ = P\left(1 - X > e^{-t}\right) = P\left(X < 1 - e^{-t}\right) = 1 - e^{-t}$$

Incidentally, note that if X has a uniform distribution on $(0, 1)$, then so does $W = 1 - X$. (See exercises.) □

Example 6.4.
The pdf of X is given by

$$f(x) = \begin{cases} 3x^2 & 0 \le x \le 1 \\ 0 & \text{elsewhere} \end{cases}$$

Find the pdf of $U = 40(1 - X)$.
Solution.

$$F(u) = P(U \le u) = P[40(1-X) \le u] = P\left(X > 1 - \frac{u}{40}\right) = 1 - P(X \le 1 - \frac{u}{40})$$

$$= 1 - F_X\left(1 - \frac{u}{40}\right) = 1 - \int_0^{1-u/40} f(x)dx = 1 - \left(1 - \frac{u}{40}\right)^3.$$

Therefore,

$$f(u) = F_U'(u) = \frac{3}{40}\left(1 - \frac{u}{40}\right)^2, \quad \text{for } \ 0 \le u \le 40$$

□

Exercises

6.1.
Show that if X has a uniform distribution on $[0, 1]$, then so does $1 - X$.

6.2.
Let X have a uniform distribution on $[0, 1]$. Let $Y = \sqrt{X}$.

a) Find the distribution of Y.

b) Find the mean of Y using the result in (a).

c) Find the mean of Y using the formula $\mathbb{E}\, g(X) = \int g(x) f(x)\, dx$.

6.3.

Using the CDF method, show that the Weibull random variable Y (with some parameter $\gamma > 0$, and $\beta = 1$) can be obtained from Exponential X (with the mean 1) as $Y = X^{1/\gamma}$.

6.4.

Suppose the radii of spheres R have a uniform distribution on $[2, 3]$. Find the mean volume. ($V = \frac{4}{3}\, \pi\, R^3$). Find the mean surface area. ($A = 4\pi\, R^2$).

6.5.

Suppose the radii of spheres have a normal distribution with mean 2.5 and variance $\frac{1}{12}$. Find the median volume and median surface area.

6.6.

Let X have a uniform distribution on $[0, 1]$. Show how you could define $H(x)$ so the $Y = H(X)$ would have a Poisson distribution with mean 1.3.

6.3 Method of transformations

Theorem 6.1. Transformations: discrete

Suppose that X is a discrete random variable with probability mass function $p(x)$. Let $U = h(X)$ define a one-to-one transformation between the values of X and U so that the equation $u = h(x)$ can be uniquely solved for x, say $x = w(u)$. Then the PMF of U is $g(u) = p[w(u)]$.

Example 6.5.

Let X be a geometric random variable with PMF

$$p(x) = \frac{3}{4}\left(\frac{1}{4}\right)^{x-1}, \quad x = 1, 2, 3, \ldots$$

Find the distribution of the random variable $U = X^2$.

Solution. Since the values of X are all positive, the transformation defines a one to one correspondence between the x and u values, $u = x^2$ and $x = \sqrt{u}$. Hence,

$$g(u) = p(\sqrt{u}) = \frac{3}{4}\left(\frac{1}{4}\right)^{\sqrt{u}-1}, \quad u = 1, 4, 9, ...$$

□

For continuous RV's, the transformation formula originates from the change of variable formula for integrals.

Theorem 6.2. Transformations: continuous

Suppose that X is a continuous random variable with density $f(x)$. Let $y = h(x)$ define a one-to-one transformation that can be uniquely solved for x, say $x = w(y)$. Then the density of $Y = h(X)$ is

$$f_Y(y) = f(x)\left|\frac{dx}{dy}\right| = f[w(y)] \times |J|$$

where $J = w'(y)$ is called the **Jacobian** of the transformation.

Example 6.6.

Let X be a continuous random variable with probability distribution

$$f(x) = \begin{cases} x/12 & \text{for } 1 \le x \le 5 \\ 0 & \text{elsewhere} \end{cases}$$

Find the probability distribution of the random variable $Y = 2X - 3$.

Solution. The inverse solution of $y = 2x-3$ yields $x = (y+3)/2$, from which we obtain $J = w'(y) = \frac{dx}{dy} = \frac{1}{2}$. Therefore, using the above Theorem 6.2, we find the density function of Y to be

$$f_Y(y) = \frac{1}{12}\left(\frac{y+3}{2}\right)\frac{1}{2} = \frac{y+3}{48}, \quad -1 < y < 7$$

□

Example 6.7.

Let X be Uniform$[0, 1]$ random variable. Find the distribution of $Y = X^5$.

Solution. Inverting, $x = y^{1/5}$, and $dx/dy = (1/5)y^{-4/5}$. Thus, we obtain

$$f_Y(y) = 1 \times (1/5)y^{-4/5} = (1/5)y^{-4/5}, \quad 0 < y < 1$$

□

Example 6.8.
Let X be a continuous random variable with density

$$f(x) = \begin{cases} \frac{x+1}{2} & \text{for } -1 \le x \le 1 \\ 0 & \text{elsewhere} \end{cases}$$

Find the density of the random variable $Y = X^2$.

Solution. The inversion of $y = x^2$ yields $x_{1,2} = \pm\sqrt{y}$, from which we obtain $J_1 = w_1'(y) = \frac{dx_1}{dy} = \frac{1}{2\sqrt{y}}$ and $J_2 = w_2'(y) = \frac{dx_2}{dy} = -\frac{1}{2\sqrt{y}}$. We cannot directly use Theorem 6.2 because the function $y = x^2$ is not one-to-one. However, we can split the range of X into two parts $(-1, 0)$ and $(0, 1)$ where the function is one-to-one. Then, we will just add the results.

Thus, we find the density function of Y to be

$$f_Y(y) = |J_1| f(\sqrt{y}) + |J_2| f(-\sqrt{y})$$

$$= \frac{1}{2\sqrt{y}} \left(\frac{\sqrt{y}+1}{2} + \frac{-\sqrt{y}+1}{2} \right) = \frac{1}{2\sqrt{y}} \text{ for } 0 \le y \le 1$$

□

Example 6.9. Location and Scale parameters
Suppose that X is some standard distribution (for example, Standard Normal, or maybe Exponential with $\beta = 1$) and $Y = a + bX$, or, solving for X,

$$X = \frac{Y - a}{b}$$

Then a is called **location (or shift) parameter** and b is **scale** parameter.

Let X have the density $f(x)$. Then the density of Y can be obtained from Theorem 6.2 as

$$f_Y(y) = f(x) \left| \frac{dx}{dy} \right| = \frac{1}{|b|} f\left(\frac{y-a}{b} \right) \tag{6.1}$$

For example, let X be Exponential with the mean 1, and $Y = bX$. Then $f(x) = e^{-x}$, $x > 0$, and (6.1) gives

$$f_Y(y) = (1/b)e^{-y/b}, \quad y > 0$$

That is, Y is Exponentially distributed with the mean b. This agrees with the result of Example 6.1.

Another example of location and scale parameters is provided by Normal distribution: if Z is standard Normal, then $Y = \mu + \sigma Z$ produces Y a Normal (μ, σ^2) random variable. Thus, μ is the location and σ is the scale parameter.

Formula (6.1) also provides a faster way to solve some of the above Examples. □

Exercises

6.7.

Let X be a continuous random variable with density

$$f(x) = \begin{cases} x + 1 & \text{for } 0 \leq x \leq 1 \\ 0 & \text{elsewhere} \end{cases}$$

Find the density of the random variable $Y = X^2$.

6.8.

Use the methods of this section to show that linear functions of normal random variables again have a normal distribution. Let $Y = a + bX$, where X is normal. How do the mean and variance of Y relate to those of X? Again, use the methods of this section.

6.9.

The so-called *Pareto* random variable X with parameters 10 and 2 has the density function

$$f(x) = \frac{10}{x^2}, \quad x > 10$$

Write down the density function of $Y = 4X - 20$ (do not forget the limits!)

6.10.

Re-do Example 6.4 using the transform (Jacobian) method.

6.11.

For the following distributions identify the parameters as location or scale parameters, or neither:

a) Weibull, parameter β.

b) Weibull, parameter γ.

c) Uniform on $[-\theta, \theta]$, parameter θ.

d) Uniform on $[b, b+1]$, parameter b.

6.4 Central Limit Theorem

Sample mean (average of all observations) plays a central role in statistics. We have discussed the variance of the sample mean in Section 5.4. Here are more facts about the behavior of the sample mean.

From the linear properties of the expectation, it's clear that

$$\mathbb{E}\left(\overline{X}\right) = \mathbb{E}\left(\frac{X_1 + X_2 + ... + X_n}{n}\right) = \frac{n\mu}{n} = \mu.$$

Summarizing the above, we obtain

Definition 6.1. Sample mean

A group of independent random variables from some distribution is called a *sample*, usually denoted as

$$X_1, X_2, ..., X_n.$$

Sample mean, denoted $\overline{X}$, is

$$\overline{X} = \frac{X_1 + X_2 + ... + X_n}{n}$$

If $\mathbb{E}\left(X_i\right) = \mu$ and $V(X_i) = \sigma^2$ for all i, then the mean and variance of sample mean are

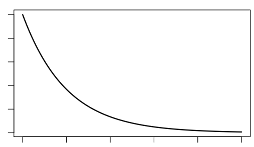
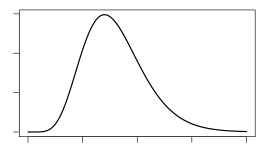
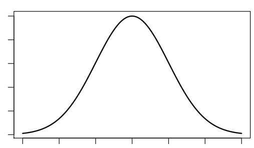
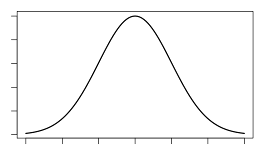
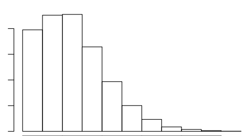
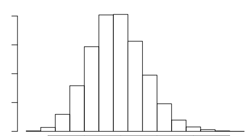
$$\mathbb{E}\left(\overline{X}\right) = \mu \quad \text{and} \quad V(\overline{X}) = \sigma^2/n$$

Example 6.10.

The average voltage of the batteries is 9.2V and standard deviation is 0.25V. Assuming normal distribution and independence, what is the distribution of total voltage $Y = X_1 + ... + X_4$? Find the probability that the total voltage is above 37.

Solution. The mean is $4 \times 9.2 = 36.8$. The variance is $4 \times 0.25^2 = 0.25$. Furthermore, Y itself will have a normal distribution.
Using z-scores, $P(Y > 37) = P(Z > (37 - 36.8)/0.5) = P(Z > 0.4) = 0.5 - 0.1554 = 0.345$ from Normal table, p. 92. □

Here we mention (without proof, which can be obtained using the moment generating functions) some properties of the sums of independent random variables.

	Distribution of X_i	Distribution of $Y = X_1 + X_2 + ... + X_n$ (indep.)
Exponential		$\mapsto$ Gamma 
Normal		$\mapsto$ Normal 
Poisson		$\mapsto$ Poisson 

What do these have in common?
The sum of independent Normal RV's is always Normal. The shape of the sum distribution for other independent RV's starts resembling Normal as n increases.

The Central Limit Theorem (CLT) ensures the similar property for most general distributions. However, it holds *in the limit*, that is, as n gets large (practically, $n > 30$ is usually enough). According to it, the sums of independent RV's approach normal distribution. The same holds for averages, since they are sums divided by n.

Theorem 6.3. CLT

Let $\overline{X}$ be the mean of a sample coming from some distribution with mean μ and variance σ^2. Then, for large n, $\overline{X}_n$ is approximately Normal with mean μ and variance σ^2/n.

If $n < 30$, the approximation is good only if the population distribution is not too different from a normal. If the population is normal, the sampling distribution of $\overline{X}$ will follow a normal distribution exactly, no matter how small the sample size.[1]

Example 6.11.
An electrical firm manufactures light bulbs with average lifetime equal to 800 hours and standard deviation of lifetimes equal 400 hours. Find the probability that a random sample of 16 bulbs will have an average life of less than 725 hours.

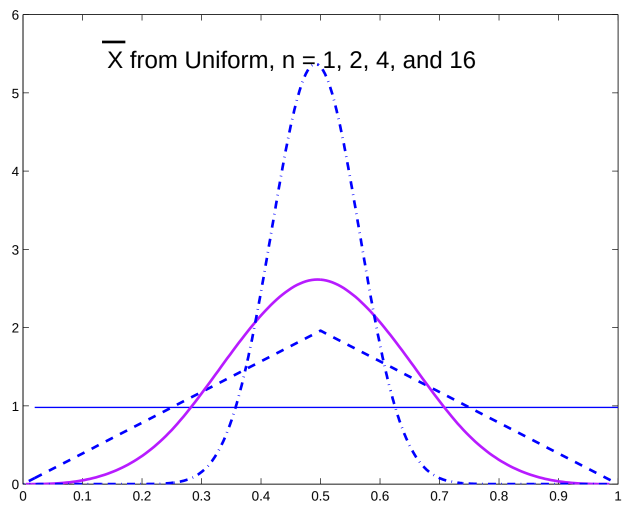
Solution. The sampling distribution of $\overline{X}$ will be approximately normal, with mean $\mu_{\overline{X}} = 800$ and $\sigma_{\overline{X}} = \frac{400}{\sqrt{16}} = 100$. Therefore,

$$P(\overline{X} < 725) \approx P\left(Z < \frac{725 - 800}{100}\right) = P(Z < -0.75) = 0.5 - 0.2734 = 0.2266$$

□

Dependence on n

As n increases, two things happen to the distribution of $\overline{X}$: it is becoming sharper (due to the variance decreasing) and also the shape is becoming more and more Normal. For example, if X_i are Uniform[0,1], then the density of $\overline{X}$ behaves as follows:



[1] There are some cases of the so-called "heavy-tailed" distributions for which the CLT does not hold, but they will not be discussed here.

Example 6.12.

The fracture strengths of a certain type of glass average 14 (thousands of pounds per square inch) and have a standard deviation of 2. What is the probability that the average fracture strength for 100 pieces of this glass exceeds 14.5?

Solution. By the central limit theorem the average strength $\overline{X}$ has approximately a normal distribution with mean= 14 and standard deviation, $\sigma = \frac{2}{\sqrt{100}} = 0.2$. Thus,

$$P(\overline{X} > 14.5) \approx P\left(Z > \frac{14.5 - 14}{0.2}\right) = P(Z > 2.5) = 0.5 - 0.4938 = 0.0062$$

from normal probability Table. □

6.4.1 CLT examples: Binomial

Historically, CLT was first discovered in case of Binomial distribution. Since Binomial Y is a sum of n independent Bernoulli RV's, CLT applies and says that $\overline{X} = Y/n$ is approximately Normal, mean p and variance $p(1-p)/n$. In this case, $\hat{p} := Y/n$ is called *sample proportion.* The Binomial Y itself is also approximately Normal with mean np and variance $np(1-p)$, as was discussed earlier in Section 4.6.2.

Example 6.13.

A fair ($p = 0.5$) coin is tossed 500 times.

a) What is the expected proportion of Heads?

b) What is the typical deviation from the expected proportion?

c) What is the probability that the sample proportion is between 0.46 and 0.54?

Solution. (a) We have $\mathbb{E}(\hat{p}) = p = 0.5$ and $\sigma_{\hat{p}} = \sqrt{p(1-p)/n} = \sqrt{0.25/500} = 0.0224$.

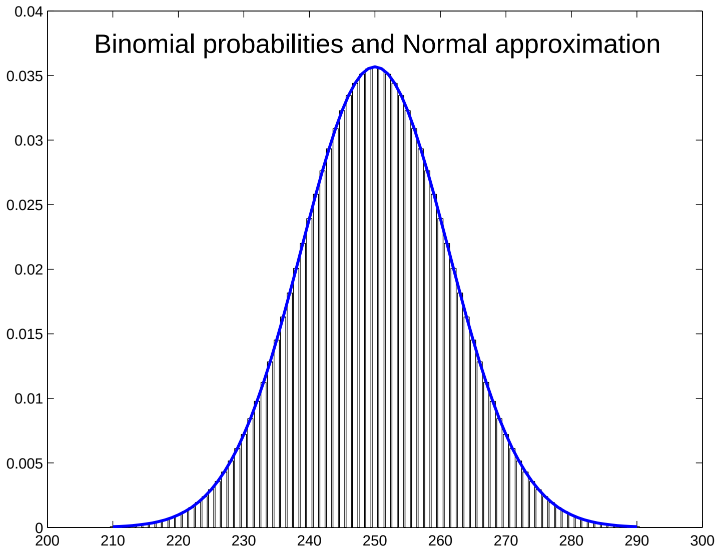
(b) For example, the *empirical rule* states that about 68% of a normal distribution is contained within one standard deviation of its mean. Here, the

68% interval is about 0.5 ± 0.0224, or 0.4776 to 0.5224.

(c)

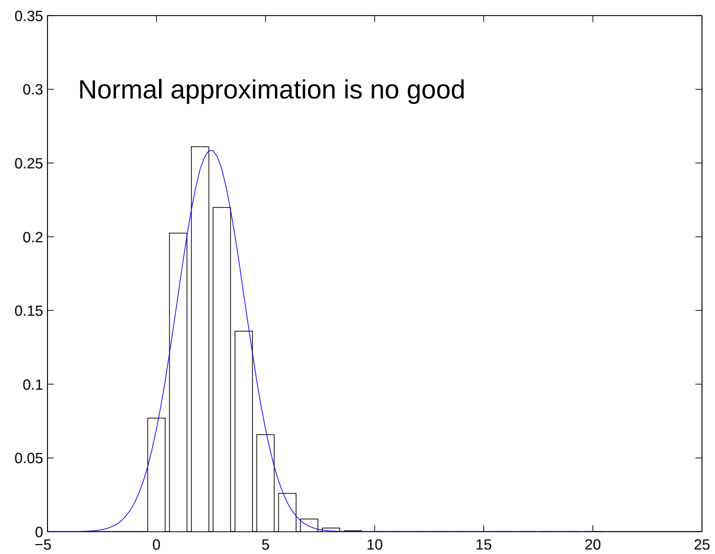
$$P(0.46 \leq Y \leq 0.54) \approx P\left(\frac{0.46-0.5}{0.0224} < Z < \frac{0.54-0.5}{0.0224}\right) =$$

$$= P(-1.79 < Z < 1.79) = 2P(0 < Z < 1.79) = 0.927$$ □



Normal approximation for $n = 500$ and $p = 0.5$

Normal approximation is not very good when np is small. Here's an example with $n = 50$ and $p = 0.05$:



Exercises

6.12.

The average concentration of potassium in county soils was determined as 85 ppm, with standard deviation 30 ppm. If $n = 20$ samples of soils are taken, find the probability that their average potassium concentration will be in the "medium" range (80 to 120 ppm).

6.13.

The heights of students have a mean of 174.5 centimeters (cm) and a standard deviation of 6.9 cm. If a random sample of 25 students is obtained, determine

a) the mean and standard deviation of $\overline{X}$;

b) the probability that the sample mean will fall between 172.5 and 175.8 cm;

c) the 70th percentile of the $\overline{X}$ distribution.

6.14.

A process yields 10% defective items. If 200 items are randomly selected from the process, what is the probability that the sample proportion of defectives

a) exceeds 13%?

b) is less than 8%?

Chapter 7

Descriptive statistics

The goal of statistics is somewhat complementary to that of the probability. Probability answers the question of what data are likely to be obtained from known probability distributions.
Statistics answers the opposite question: what kind of probability distributions are likely to have generated the data at hand?

Descriptive statistics are the ways to summarize the data set, to represent its tendencies in a concise form and/or describe them graphically.

7.1 Sample and population

We will usually refer to the given data set as a *sample* and denote its entries as $X_1, X_2, ..., X_n$. The objects whose measurements are represented by X_i are often called *experimental units* and are usually assumed to be sampled randomly from a larger *population* of interest. The probability distribution of X_i is then referred to as *population distribution*.

Definition 7.1. Population and sample

Population is the collection of all objects of interest. *Sample* is the collection of objects from the population picked for the study.

A *simple random sample* (SRS) is a sample for which each object in the population has the same probability to be picked as any other object, and is picked independently of any other object.

Example 7.1.

a) We would like to learn the public opinion regarding a tax reform. We set up phone interviews with $n = 1000$ people. Here, the population (which we really would like to learn about) is all U.S. adults, and the sample (which are the objects, or individuals we actually get), is the 1000 people contacted.

   For some really important matters, the U.S. Census Bureau tries to reach every single American, but this is practically impossible.

b) The gas mileage of a car is investigated. Suppose that we drive $n = 20$ times using a full tank of gas, until it's empty, and calculate the average gas mileage for each trip. Here, the population is *all potential trips* between fillups on this car to be made (under usual driving conditions) and the sample is the 20 trips actually made.

Usually, we require that our sample be a simple random sample (SRS) so that we can extend our findings to the entire population of interest. This means that no part of the population is preferentially selected for, or excluded from the study.

Bias often occurs when the sample is not an SRS. For example, *self-selection bias* occurs when subjects volunteer for the study. Medical studies that pay for participation may attract lower-income volunteers. A questionnaire issued by a website will represent only the people that visit that website etc.

The ideal way to implement an SRS is to create a list of all objects in a population, and then use a random number generator to pick the objects to be sampled. In practice, this is very difficult to accomplish.

In the future, we will always assume that we are dealing with an SRS, unless otherwise noted. Thus, we will obtain a sequence of independent and identically distributed (IID) random variables $X_1, X_2, ..., X_n$ from the population distribution we are studying.

7.2 Graphical summaries

The most popular graphical summary for a numeric data set is a *histogram*.

Definition 7.2.

> The histogram of the data set $X_1, X_2, \ldots, X_n$ is a bar chart representing the *classes* (or *bins*) on the x-axis and frequencies (or proportions) on the y-axis.

Bins should be of equal width so that all bars would visually be on the same level.[1] The construction of a histogram is easier to show by example.

Example 7.2.

Old Faithful is a famous geyser in Yellowstone National Park. The data recorded represent waiting times between eruptions (in minutes). There are $n = 272$ observations. The first ten observations are 79, 54, 74, 62, 85, 55, 88, 85, 51, 85. Using the bins 41-45, 46-50 etc we get

Bin	41-45	46-50	51-55	56-60	61-65	66-70	71-75
Count	4	22	33	24	14	10	27
Bin	76-80	81-85	86-90	91-95	96-100		
Count	54	55	23	5	1		

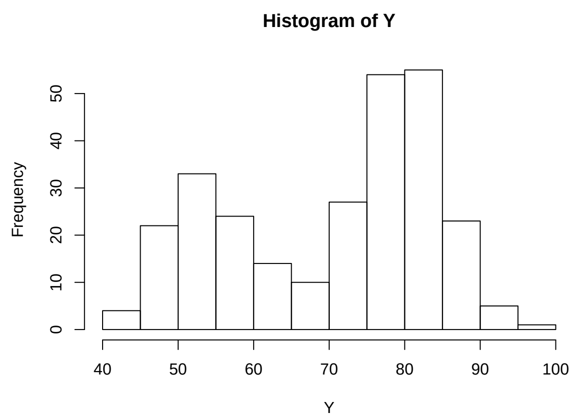


Figure 7.1: histogram of Old Faithful data

The choice of bins of course affects the appearance of a histogram. With too many bins, the graph becomes hard to read, and with too little bins, a lot of information is lost. We would generally recommend to use more bins

[1]Bins can be of unequal width but then some adjustment to their heights must be made.

for larger sample sizes; but not too many bins, so that the histogram keeps a smooth appearance. Some authors recommend the number of bins no higher than $\sqrt{n}$ where n is the sample size.

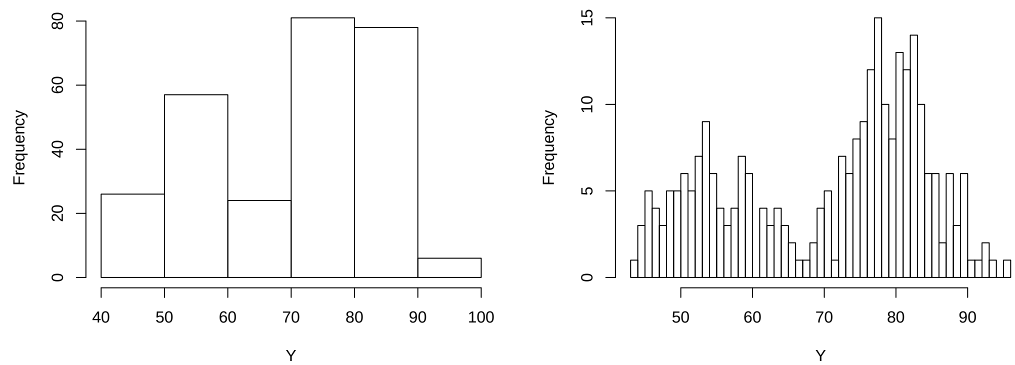


Figure 7.2: histograms of Old Faithful data: bins too wide, bins too narrow

Describing the shape of a histogram, we may note its features as being symmetric, or maybe skewed (left or right); having one "bulge" (*mode*) - that is, unimodal distribution, or two modes - that is, bimodal distribution etc. The Old Faithful data have bimodal shape. Some skewed histogram shapes are shown in Fig. 8.1

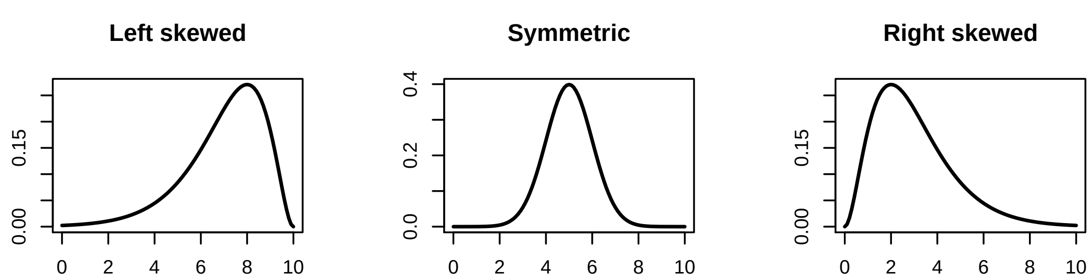


Figure 7.3: Symmetric and skewed shapes

7.3 Numerical summaries

7.3.1 Sample mean and variance

The easiest and most popular summary for a data set is its mean $\overline{X}$. The mean is a *measure of location* for the data set. We often need also a measure of spread. One such measure is the *sample standard deviation.*

Definition 7.3. Sample variance and standard deviation

The *sample variance* is denoted as S^2 and equals to

$$S^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1} = \frac{\sum_{i=1}^{n}(X_i^2) - n\overline{X}^2}{n-1} \tag{7.1}$$

Sample standard deviation S is the square root of S^2.

A little algebra may show that both expressions in the formula (7.1) are equivalent. Denominator in the formula is $n-1$ which is called *degrees of freedom*. A simple explanation is that the calculation starts with n numbers and is then constrained by finding $\overline{X}$, thus $n-1$ degrees of freedom are left. Note that if $n = 1$ then the calculation of sample variance is not possible.

The sample mean and standard deviation are counterparts of the mean and standard deviation of a probability distribution. Further we will use them as the estimates of the unknown mean and standard deviation of a probability distribution (or a population).

Example 7.3.

The heights of last 8 US presidents are (in cm)[10] : 185, 182, 188, 188, 185, 177, 182, 193. Find the mean and standard deviation of these heights.

Solution. The average height is $\overline{X} = 185$. To make the calculations more compact, let's subtract 180 from each number, as it will not affect the standard deviation: $0, 2, 8, 8, 5, -3, 2, 13$, and $\overline{X} = 5$. Then, $\sum X_i^2 = 364$ and we get $S^2 = \dfrac{364 - 5^2(8)}{8-1} = 23.43$ and $S = \sqrt{23.43} = 4.84$. □

7.3.2 Percentiles

Definition 7.4.

The pth percentile (or quantile) of a data set is a number q such that $p\%$ of the entire sample are below this number. It can be calculated as $r = ((n+1)p/100)$th smallest number in the sample.

The algorithm for calculating pth percentile is then as follows.[2]

a) Order the sample, from smallest to largest, denote these as $X_{(1)}, X_{(2)}, \dots, X_{(n)}$.

b) Calculate $r = (n+1)p/100$, let $k = \lfloor r \rfloor$ be the integer part of r.

c) If interpolation is desired, take $X_{(k)} + (r-k)[X_{(k+1)} - X_{(k)}]$,

   If interpolation is not needed, take $X_{(r^*)}$ where r^* is the rounded value of r.

Generally, if the sample size n is large, the interpolation is not needed.[3]

The 50-th percentile is known as *median.* It is, along with the mean, a measure of center of the data set.

Example 7.4.

Back to the example of US presidents: find the median and 22nd percentile of the presidents' heights.

Solution. The ordered data are 177, 182, 182, 185, 185, 188, 188, 193. For $n = 8$ we have two "middle observations": ranked 4th and 5th, these are both 185. Thus, the median is 185 (accidentally we have seen that $\overline{X} = 185$ also).

To find 22nd percentile, take $r = (n+1)p = 9(0.22) = 1.98$, round it to 2. Then, take 2nd ranked observation, which is 182. □

Mean and median

The mean and median are popular measures of center. For a symmetric data set, both give roughly the same result. However, for a skewed data set,

[2] see e.g. `http://www.itl.nist.gov/div898/handbook/prc/section2/prc252.htm`

[3] Software note: different books and software packages may have different ways to interpret the fractional value of $(n+1)p/100$, so the percentile results might vary.

they might produce fairly different results. For the right-skewed distribution, `mean > median`, and for the left-skewed, `mean < median`.

The median is *resistant to outliers*. This means that the unusually high or low observations do not greatly affect the median. The mean $\overline{X}$ is not resistant to outliers.

Exercises

7.1.

The temperature data one morning from different weather stations in the vicinity of Socorro were

71.9, 73.7, 72.3, 74.6, 72.8, 67.5, 72.0 (in °F)

a) Find the mean and standard deviation of temperatures

b) Find the median and 86th percentile.

c) Suppose that the last measurement came from Magdalena Ridge and became equal to 41.7 instead of 72.0. How will this affect the mean and the median, respectively?

d) Re-calculate the above answers if the temperature is expressed in Celcius. [**Hint:** you do not have to do it from scratch!]

7.2.

The heights of the last 20 US presidents are, in cm: 185, 182, 188, 188, 185, 177, 182, 193, 183, 179, 175, 188, 182, 178, 183, 180, 182, 178, 170, 180.

a) Make a histogram of the heights, choosing bins wisely.

b) Calculate mean and the median, compare. How do these relate to the shape of the histogram?

Chapter 8

Statistical inference

8.1 Introduction

In previous sections we emphasized properties of the sample mean. In this section we will discuss the problem of estimation of population parameters, in general. A **point estimate** of some population parameter θ is a single value $\hat{\theta}$ of a statistic. For example, the value $\overline{X}$ is the point estimate of population parameter μ. Similarly, $\hat{p} = \dfrac{X}{n}$ is a point estimate of the true proportion p in a binomial experiment.

Statistical inference deals with the question: can we infer something about the unknown population parameters (e.g., μ, σ or p)? Two major tools for statistical inference are *confidence intervals* (they complement a point estimate with a margin of error) and *hypothesis tests* that try to prove some statement about the parameters.

8.1.1 Unbiased Estimation

What are the properties of desirable estimators? We would like the sampling distribution of $\hat{\theta}$ to have a mean equal to the parameter estimated. An estimator possessing this property is said to be **unbiased**.

Definition 8.1.

A statistic $\hat{\theta}$ is said to be an unbiased estimator of the parameter θ if

$$\mathbb{E}\,(\hat{\theta}) = \theta.$$

The unbiased estimators are correct "on average", while actual samples yield results higher or lower than the true value of the parameter

On the other hand, biased estimators would consistently overestimate or underestimate the target parameter.

Example 8.1.
One reason that the sample variance $S^2 = \sum(X_i - \overline{X})^2/(n-1)$ is divided by $n-1$ (instead of n) is the unbiasedness property. Indeed, it can be shown that $\mathbb{E}(S^2) = \sigma^2$. □

8.2 Confidence intervals

The **confidence interval (CI)** or **interval estimate** is an interval within which we would expect to find the "true" value of the parameter.

Interval estimates, say, for population mean, are often desirable because the point estimate $\overline{X}$ varies from sample to sample. Instead of a single estimate for the mean, a confidence interval generates a lower and an upper bound for the mean. The interval estimate provides a measure of uncertainty in our estimate of the true mean μ. The narrower the interval, the more precise is our estimate.

Confidence limits are evaluated in terms of a confidence level.[1] Although the choice of confidence level is somewhat arbitrary, in practice 90%, 95%, and 99% intervals are often used, with 95% being the most commonly used.

Theorem 8.1. CI for the mean

If $\overline{X}$ is the mean of a random sample of size n from a normal population with known variance σ^2, an approximate $(1-\alpha)100\%$ confidence interval for μ is given by

$$\overline{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \tag{8.1}$$

where $z_{\alpha/2}$ is the Z-value leaving an area of $\alpha/2$ to the right.

[1]On a technical note, a 95% confidence interval does not mean that there is a 95% probability that the interval contains the true mean. The interval computed from a given sample either contains the true mean or it does not. Instead, the level of confidence is associated with the method of calculating the interval. For example, for a 95% confidence interval, if many samples are collected and a confidence interval is computed for each, in the long run about 95% of these intervals would contain the true mean.

Proof. Central Limit Theorem (CLT) claims that, regardless of the initial distribution, the sample mean $\overline{X} = (X_1 + ... + X_n)/n$ will be approximately Normal:

$$\overline{X} \approx \text{ Normal } (\mu, \sigma^2/n)$$

for n reasonably large (usually $n \geq 30$ is considered enough).

Suppose that a confidence level $C = 100\%(1 - \alpha)$ is given. Then, find $z_{\alpha/2}$ such that

$$P(-z_{\alpha/2} < Z < z_{\alpha/2}) = 1 - \alpha, \quad Z \text{ is a standard Normal RV}$$

Due to the symmetry of Z-distribution, we need to find the z-value with the upper tail probability $\alpha/2$. That is, table area $\texttt{TA}(z_{\alpha/2}) = 0.5 - \alpha/2$.

Then, using CLT, $Z \approx \dfrac{\overline{X} - \mu}{\sigma/\sqrt{n}}$, therefore

$$P\left(-z_{\alpha/2} < \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} < z_{\alpha/2}\right) \approx 1 - \alpha$$

Solving for μ, we obtain the result. □

Notes:

(a) If σ is unknown, it can replaced by S, the sample standard deviation, with no serious loss in accuracy for the large sample case. Later, we will discuss what happens for small samples.

(b) This CI (and many to follow) has the following structure

$$\overline{X} \pm m$$

where m is called *margin of error.*

Example 8.2.
The drying times, in hours, of a certain brand of latex paint are

```
3.4  2.5  4.8  2.9  3.6  2.8  3.3  5.6
3.7  2.8  4.4  4.0  5.2  3.0  4.8
```

Compute the 95% confidence interval for the mean drying time. Assume that $\sigma = 1$.

Solution. We compute $\overline{X} = 3.79$ and $z_{\alpha/2} = 1.96$
($\alpha = 0.05$, upper-tail probability $= 0.025$, table area $= 0.5 - 0.025 = 0.475$)

Then, using (8.1), the 95% C.I. for the mean is

$$3.79 \pm 1(1.96)/\sqrt{15} = 3.79 \pm 0.51$$

□

Example 8.3.
The average zinc concentration recovered from a sample of zinc measurements in 36 different locations in the river is found to be 2.6 milligrams per liter. Find the 95% and 99% confidence intervals for the mean zinc concentration μ. Assume that the population standard deviation is 0.3.

Solution. The point estimate of μ is $\overline{X} = 2.6$. For 95% confidence, $z_{\alpha/2} = 1.96$. Hence, the 95% confidence interval is

$$2.6 - 1.96\frac{0.3}{\sqrt{36}} < \mu < 2.6 + 1.96\frac{0.3}{\sqrt{36}} = (2.50, 2.70)$$

For a 99% confidence, $z_{\alpha/2} = 2.575$ and hence the 99% confidence interval is

$$2.6 - 2.575\frac{0.3}{\sqrt{36}} < \mu < 2.6 + 2.575\frac{0.3}{\sqrt{36}} = (2.47, 2.73)$$

We see that a wider interval is required to estimate μ with a higher degree of confidence. □

Example 8.4.
An important property of plastic clays is the amount of shrinkage on drying. For a certain type of plastic clay 45 test specimens showed an average shrinkage percentage of 18.4 and a standard deviation of 1.2. Estimate the "true" average shrinkage μ for clays of this type with a 95% confidence interval.

Solution. For these data, a point estimate of μ is $\overline{X} = 18.4$. The sample standard deviation is $S = 1.2$. Since n is fairly large, we can replace σ by S. Hence, 95% confidence interval for μ is

$$18.4 - 1.96\frac{1.2}{\sqrt{45}} < \mu < 18.4 + 1.96\frac{1.2}{\sqrt{45}} = (18.05, 18.75)$$

Thus we are 95% confident that the true mean lies between 18.05 and 18.75.
□

Sample size calculations

In practice, another problem often arises: how many data should be collected to determine an unknown parameter with a given accuracy? That is, let m be the desired size of the margin of error, for a given confidence level $100\%(1-\alpha)$

$$m = \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \tag{8.2}$$

What is the sample size n to achieve this goal?

To do this, assume that some estimate of σ is available. Then, solving for n,

$$n = \left(\frac{z_{\alpha/2}\, \sigma}{m} \right)^2$$

Example 8.5.

We would like to estimate the pH of a certain type of soil to within 0.1, with 99% confidence. From past experience, we know that the soils of this type usually have pH in the 5 to 7 range. Find the sample size necessary to achieve our goal.

Solution. Let us take the reported 5 to 7 range as the $\pm 2\sigma$ range. This way, the crude estimate of σ is $(7-5)/4 = 0.5$. For 99% confidence, we find the upper tail area $\alpha/2 = (1-0.99)/2 = 0.005$, thus $z_{\alpha/2} = 2.576$, and $n = (2.576 \times 0.5/0.1)^2 \approx 166$ □

Exercises

8.1.

In a school district, they would like to estimate the average reading rate of first-graders. After selecting a random sample of $n = 65$ readers, they obtained sample mean of 53.4 words per minute (wpm), and standard deviation of 33.9 wpm.[11] Calculate a 98% confidence interval for the average reading rate of all first-graders in the district.

8.2.

A random sample of 200 calls initiated while driving had a mean duration of 3.5 minutes with standard deviation 2.2 minutes. Find a 99% confidence interval for the mean duration of telephone calls initiated while driving.

8.3.

a) Bursting strength of a certain brand of paper is supposed to have Normal distribution with $\mu = 150$ kPa and $\sigma = 15$ kPa. Give an interval that contains about 95% of all bursting strength values

b) Assuming now that the true μ and σ are unknown, the researchers collected a sample of $n = 100$ paper bags and measured their bursting strength. They obtained $\overline{X} = 148.4$ kPa and $S = 18.9$ kPa. Calculate the 95% C.I. for the mean bursting strength.

c) Sketch a Normal density curve with $\mu = 150$, $\sigma = 15$, with both of your intervals shown on the x-axis. Compare the intervals' widths.

8.4.

In determining the mean viscosity of a new type of motor oil, the lab needs to collect enough observations to approximate the mean within ± 0.2 SAE grade, with 96% confidence. The standard deviation typical for this type of measurement is 0.4. How many samples of motor oil should the lab test?

8.3 Statistical hypotheses

Definition 8.2.

> A Statistical hypothesis is an assertion or conjecture concerning one or more populations.

The goal of a statistical hypothesis test is to **make a decision** about an unknown parameter (or parameters). This decision is usually expressed in terms of rejecting or accepting a certain value of parameter or parameters.

Some common situations to consider:

- Is the coin fair? That is, we would like to test if $p = 1/2$ where $p = P(Heads)$.

- Is the new drug more effective than the old one? In this case, we would like to compare two parameters, e.g. the average effectiveness of the old drug versus the new one.

In making the decision, we will compare the statement (say, $p = 1/2$) with the available data and will reject the claim $p = 1/2$ if it contradicts the data. In the subsequent sections we will learn how to set up and test the hypotheses in various situations.

Null and alternative hypotheses

A statement like $p = 1/2$ is called the **Null hypothesis** (denoted by H_0). It expresses the idea that the parameter (or a function of parameters) is equal to some fixed value. For the coin example, it's

$$H_0 : p = 1/2$$

and for the drug example it's

$$H_0 : \mu_1 = \mu_2$$

where μ_1 is the mean effectiveness of the old drug compared to μ_2 for the new one. **Alternative hypothesis** (denoted by H_A) seeks to disprove the null. For example, we may consider *two-sided* alternatives

$$H_A : p \neq 1/2 \quad \text{or, in the drug case, } H_A : \mu_1 \neq \mu_2$$

8.3.1 Hypothesis tests of a population mean

A **null hypothesis** H_0 for the population mean μ is a statement that designates the value μ_0 for the population mean to be tested. It is associated with an **alternative hypothesis** H_A, which is a statement incompatible with the null. A **two-sided** hypothesis setup is

$$H_0 : \mu = \mu_0 \text{ versus } H_A : \mu \neq \mu_0$$

for a specified value of μ_0, and a **one-sided** hypothesis setup is either

$$H_0 : \mu = \mu_0 \quad \text{versus} \quad H_A : \mu > \mu_0 \qquad \text{(right-sided test)}$$

or

$$H_0 : \mu = \mu_0 \quad \text{versus} \quad H_A : \mu < \mu_0 \qquad \text{(left-sided test)}$$

Steps of a Hypothesis Test

a) Null Hypothesis $H_0 : \mu = \mu_0$

b) Alternative Hypothesis $H_A : \mu \neq \mu_0$, or $H_A : \mu > \mu_0$, or $H_A : \mu < \mu_0$.

c) Critical value: $z_{\alpha/2}$ for two-tailed or z_α for one-tailed test, for some chosen *significance level* α. (Here, α is the false positive rate, i.e. how often you will reject H_0 that is, in fact, true.)

d) Test Statistic $z = \dfrac{\sqrt{n}(\overline{X} - \mu_0)}{\sigma}$

e) Decision Rule: Reject H_0 if

$$\begin{array}{cc} |z| > z_{\alpha/2} & \text{for two-tailed} \\ z > z_\alpha & \text{for right-tailed} \\ z < -z_\alpha & \text{for left-tailed} \end{array}$$

or, using p-value (see below), Reject H_0 when p-value $< \alpha$

f) Conclusion in the words of the problem.

Definition 8.3. P-values

A data set can be used to measure the plausibility of a null hypothesis H_0 through the calculation of a p-value.[a] The smaller the p-value, the less plausible is the null hypothesis.

Rejection Rule: Given the *significance level* α,

Reject H_0 when p-value $< \alpha$

otherwise **Accept** H_0.

[a]Do not confuse p-value with notation for proportion p

Calculation of P-values

For the two-sided hypothesis, P-value $= 2 \times P(Z > |z|)$.

For the right-tailed hypothesis, $H_A : \mu > \mu_0$, P-value $= P(Z > z)$

For the left-tailed hypothesis, $H_A : \mu < \mu_0$, P-value $= P(Z < z)$

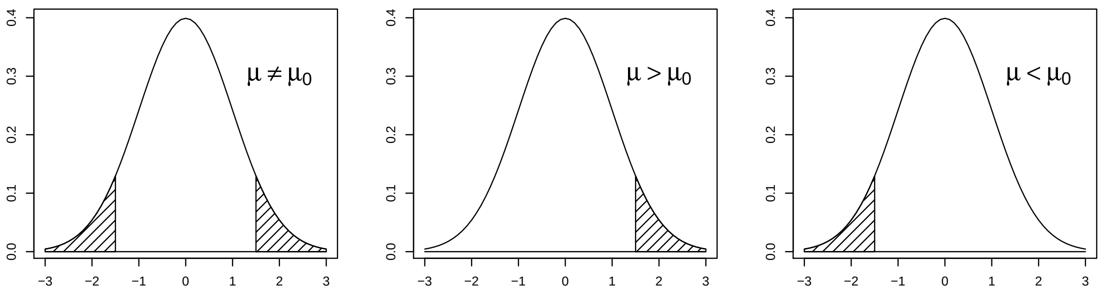


Figure 8.1: P-value calculation for different H_A

Example 8.6.

A manufacturer of sports equipment has developed a new synthetic fishing line that he claims has a mean breaking strength of 8 kg with a standard deviation of 0.5 kg. A random sample of 50 lines is tested and found to have a mean breaking strength of 7.8 kg. Test the hypothesis that $\mu = 8$ against the alternative that $\mu \neq 8$. Use $\alpha = 0.01$ level of significance.

Solution.

a) $H_0 : \mu = 8$

b) $H_A : \mu \neq 8$

c) $\alpha = 0.01$ and hence critical value $z_{\alpha/2} = 2.57$

d) Test statistic:

$$z = \frac{\sqrt{n}(\overline{X} - \mu_0)}{\sigma} = \frac{\sqrt{50}(7.8 - 8)}{0.5} = -2.83$$

e) Decision: reject H_0 since $|-2.83| > 2.57$.

f) Conclusion: there is evidence that the mean breaking strength is **not** 8 kg (in fact, it's lower).

Decision based on P-value:

Since the test in this example is two-sided, the p-value is double the area.

$$\text{P-value} = P(|Z| > 2.83) = 2\,[0.5 - \texttt{TA}(2.83)] = 2(0.5 - 0.4977) = 0.0046$$

which allows us to reject the null hypothesis that $\mu = 8$ kg at a level of significance smaller than 0.01. □

Example 8.7.

A random sample of 100 recorded deaths in the United States during the past year showed an average life span of 71.8 years. Assuming a population standard deviation of 8.9 years, does this seem to indicate that the mean life span today is greater than 70 years? Use a 0.05 level of significance.

Solution.

a) $H_0 : \mu = 70$ years.

b) $H_A : \mu > 70$ years.

c) $\alpha = 0.05$ and $z_\alpha = 1.645$

d) Test statistic:

$$z = \frac{\sqrt{n}(\overline{X} - \mu_0)}{\sigma} = \frac{\sqrt{100}(71.8 - 70)}{8.9} = 2.02$$

e) Decision: Reject H_0 if $2.02 > 1.645$, since $2.02 > 1.645$, we reject H_0.

f) Conclusion: We conclude that the mean life span today is greater than 70 years.

Decision based on P-value:
Since the test in this example is one-sided, the desired p-value is the area to the right of $z = 2.02$. Using Normal Table, we have

$$\text{P-value} = P(Z > 2.02) = 0.5 - 0.4783 = 0.0217.$$

Reject H_0 □

Example 8.8.

The nominal output voltage for a certain electrical circuit is 130V. A random sample of 40 independent readings on the voltage for this circuit gave a sample mean of 128.6V and a standard deviation of 2.1V. Test the hypothesis that the average output voltage is 130 against the alternative that it is less than 130. Use a 5% significance level.

Solution.

a) $H_0 : \mu = 130$

b) $H_A : \mu < 130$

c) $\alpha = 0.05$ and $z_\alpha = -1.645$

d) Test statistic:

$$z = \frac{\sqrt{n}(\overline{X} - \mu_0)}{\sigma} = \frac{\sqrt{40}(128.6 - 130)}{2.1} = -4.22$$

e) Decision: Reject H_0 since $-4.22 < -1.645$.

f) Conclusion: We conclude that the average output voltage is less than 130.

Decision based on p-value:

$$\text{P-value} = P(Z < -4.22) = (0.5 - 0.4990) = 0.001.$$

As a result, the evidence in favor of H_A is even stronger than that suggested by the 0.05 level of significance. (P-value is very small!) □

Exercises

8.5.
It is known that the average height of US adult males is about 173 cm, with standard deviation of about 6 cm.

Referring to Exercise 7.2, the average height of 20 last US presidents was 181.9 cm. Are the presidents taller than the average? Test at the level $\alpha = 0.05$ and also compute the p-value.

8.6.
Is it more difficult to reject H_0 when the significance level is smaller? Suppose that the p-value for a test was 0.023. Would you reject H_0 at the level $\alpha = 0.05$? At $\alpha = 0.01$?

8.4 The case of unknown σ

8.4.1 Confidence intervals

Frequently, we are attempting to estimate the mean of a population when the variance is unknown. Suppose that we have a random sample from a normal distribution, then the random variable

$$T = \frac{\overline{X} - \mu}{S/\sqrt{n}}$$

is said to have a (Student)[12] T-distribution with $n-1$ degrees of freedom. Here, S is the sample standard deviation.

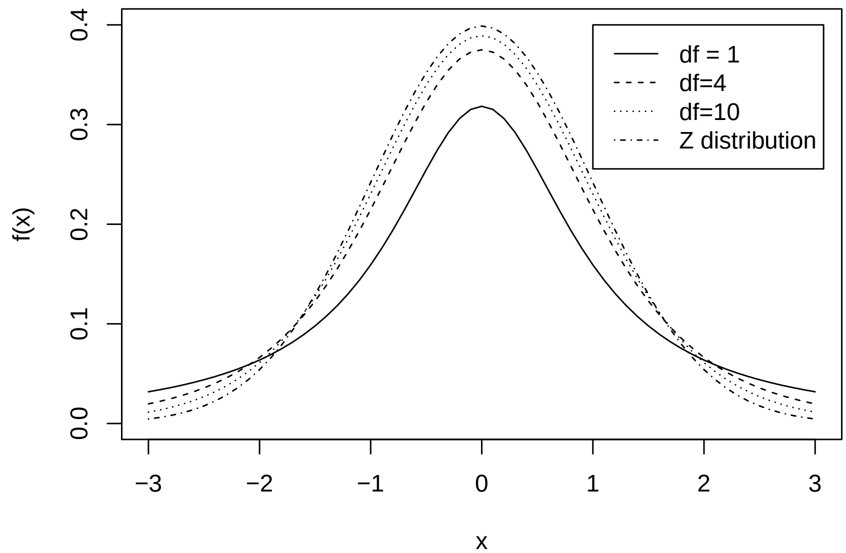


Figure 8.2: T distribution for different values of df = degrees of freedom

With σ unknown, T should be used instead of Z to construct a confidence interval for μ. The procedure is same as for known σ except that σ is replaced by S and the standard normal distribution is replaced by the T-distribution.

T-distribution is also symmetric, but has somewhat "heavier tails" than Z. This is because of extra uncertainty of not knowing σ.

Definition 8.4. CI for mean, unknown σ

If $\overline{X}$ and S are the mean and standard deviation of a random sample from a normal population with unknown variance σ^2, a $(1-\alpha)100\%$ confidence interval for μ is

$$\overline{X} - t_{\alpha/2}\frac{S}{\sqrt{n}} < \mu < \overline{X} + t_{\alpha/2}\frac{S}{\sqrt{n}},$$

where $t_{\alpha/2}$ is the t-value with $n-1$ degrees of freedom leaving an area of $\alpha/2$ to the right.

Normality assumption becomes more important as n gets smaller. As a practical rule, we will not trust the confidence intervals based on small samples (generally, $n < 30$) that are strongly skewed or have outliers.

On the other hand, we already noted that for large n we could simply use Z-distribution for the C.I. calculation. This is justified by the fact that $t_{\alpha/2}$ values approach $z_{\alpha/2}$ values as n gets larger.

Example 8.9.

The contents of 7 similar containers of sulfuric acid are 9.8, 10.2, 10.4, 9.8, 10.0, 10.2 and 9.6 liters. Find a 95% confidence interval for the mean volume of all such containers, assuming an approximate normal distribution.

Solution. The sample mean and standard deviation for the given data are $\overline{X} = 10.0$ and S = 0.283. Using the T-Table, we find $t_{0.025} = 2.447$ for 6 degrees of freedom. Hence the 95% confidence interval for μ is

$$10.0 - 2.447\frac{0.283}{\sqrt{7}} < \mu < 10.0 + 2.447\frac{0.283}{\sqrt{7}},$$

which reduces to $9.74 < \mu < 10.26$ □

Example 8.10.

A random sample of 12 graduates of a certain secretarial school typed an average of 79.3 words per minute (wpm) with a standard deviation of 7.8 wpm. Assuming a normal distribution for the number of words typed per minute, find a 99% confidence interval for the average typing speed for all graduates of this school.

Solution. The sample mean and standard deviation for the given data are $\overline{X} = 79.3$ and S = 7.8. Using the T-Table, we find $t_{0.005} = 3.106$ with 11 degrees of freedom. Hence the 95% confidence interval for μ is

$$79.3 - 3.106\frac{7.8}{\sqrt{12}} < \mu < 79.3 + 3.106\frac{7.8}{\sqrt{12}},$$

which reduces to $72.31 < \mu < 86.30$.
We are 99% confident that the interval 72.31 to 86.30 includes the true average typing speed for all graduates. □

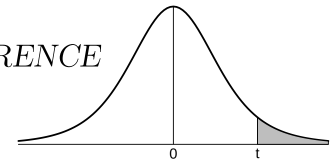


Table 2: Critical points of the t-distribution

	Upper tail probability						
Degrees of freedom	0.10	0.05	0.025	0.01	0.005	0.001	0.0005
1	3.078	6.314	12.706	31.821	63.657	318.309	636.619
2	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	1.319	1.714	2.069	2.500	2.807	3.485	3.768
24	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	1.316	1.708	2.060	2.485	2.787	3.450	3.725
30	1.310	1.697	2.042	2.457	2.750	3.385	3.646
40	1.303	1.684	2.021	2.423	2.704	3.307	3.551
60	1.296	1.671	2.000	2.390	2.660	3.232	3.460
120	1.289	1.658	1.980	2.358	2.617	3.160	3.373
∞	1.282	1.645	1.960	2.326	2.576	3.090	3.291

8.4.2 Hypothesis test

When sample sizes are small and population variance is unknown, use the test statistic

$$t = \frac{\sqrt{n}(\overline{X} - \mu_0)}{S},$$

with $n - 1$ degrees of freedom.

Steps of a Hypothesis Test

a) Null Hypothesis $H_0 : \mu = \mu_0$

b) Alternative Hypothesis $H_A : \mu \neq \mu_0$, or $H_A : \mu > \mu_0$, or $H_A : \mu < \mu_0$.

c) Critical value: $t_{\alpha/2}$ for two-tailed or t_α for one-tailed test.

d) Test Statistic $t = \dfrac{\sqrt{n}(\overline{X} - \mu_0)}{S}$ with $n - 1$ degrees of freedom

e) Decision Rule: Reject H_0 if

$$\begin{array}{cc} |t| > t_{\alpha/2} & \text{for two-tailed} \\ t > t_\alpha & \text{for right-tailed} \\ t < -t_\alpha & \text{for left-tailed} \end{array}$$

or, using p-value, Reject H_0 when p-value $< \alpha$

f) Conclusion.

Example 8.11.

Engine oil was stated to have the mean viscosity of $\mu_0 = 85.0$. A sample of $n = 25$ viscosity measurements resulted in a sample mean of $\overline{X} = 88.3$ and a sample standard deviation of $S = 7.49$. What is the evidence that the mean viscosity is not as stated? Use $\alpha = 0.1$.

Solution.

a) $H_0 : \mu = 85.0$

b) $H_A : \mu \neq 85.0$

c) $\alpha = 0.1$ and $t_{\alpha/2} = 1.711$ with 24 degrees of freedom.

d) Test statistic:

$$t = \frac{\sqrt{n}(\overline{X} - \mu_0)}{S} = \frac{\sqrt{25}(88.3 - 85.0)}{7.49} = 2.203$$

e) Decision: Reject H_0 since $2.203 > 1.711$.

f) Conclusion: We conclude that the average viscosity is not equal to 85.0

Decision based on P-value:
Since the test in this example is two sided, the desired p-value is twice the tail area. Therefore, using t-table with $df = 24$, we have

$$\text{P-value} = 2 \times P(T > 2.203) = 2(0.0187) = 0.0374,$$

which allows us to reject the null hypothesis that $\mu = 85$ at a level of significance smaller than 0.1.

Conclusion: In summary, we conclude that there is fairly strong evidence that the mean viscosity is not equal to 85.0 □

Example 8.12.
A sample of $n = 20$ cars driven under varying highway conditions achieved fuel efficiencies with a sample mean of $\overline{X} = 34.271$ miles per gallon (mpg) and a sample standard deviation of $S = 2.915$ mpg. Test the hypothesis that the average highway mpg is less than 35 with $\alpha = 0.05$.

Solution.
a) $H_0 : \mu = 35.0$

b) $H_A : \mu < 35.0$

c) $\alpha = 0.05$ and $t_\alpha = 1.729$ with 19 degrees of freedom.

d) Test statistic:

$$t = \frac{\sqrt{n}(\overline{X} - \mu_0)}{S} = \frac{\sqrt{20}(34.271 - 35.0)}{2.915} = -1.119$$

e) Decision: since $-1.119 > -1.729$, we do not reject H_0.

f) Conclusion: There is no evidence that the average highway mpg is any less than 35.0

Decision based on P-value:

$$\text{P-value} = P(T < -1.119) = P(T > 1.119) > 0.10,$$

(using $df = 19$), thus p-value $> \alpha = 0.05$, do not reject H_0. □

8.4.3 Connection between Hypothesis tests and C.I.'s

We can test a two-sided hypothesis

$$H_0 : \mu = \mu_0 \text{ vs. } H_A : \mu \neq \mu_0$$

at the level α, using a confidence interval with the confidence level $100\%(1-\alpha)$. If we found the $100\%(1-\alpha)$ C.I. for the mean μ, and μ_0 belongs to it, we accept H_0, otherwise we reject H_0.

This way, the C.I. is interpreted as the range of "plausible" values for μ. The false positive rate in this case will be equal to $\alpha = 1 - C/100\%$

Example 8.13.

Reconsider Example 8.11. There, we had to test $H_0 : \mu = 85.0$ with the data $n = 25$, $\overline{X} = 88.3$ and $S = 7.49$, at the level $\alpha = 0.1$. Is there evidence that the mean average viscosity is not 85.0?

Solution. If we calculate a 90% C.I. ($90\% = 100\%(1-\alpha)$), we get

$$88.3 \pm 1.711 \frac{7.49}{\sqrt{25}} = 88.3 \pm 2.6 \text{ or } (85.7, 90.9)$$

Since 85.0 does not belong to this interval, there is evidence that the "true" mean viscosity is **not** 85.0 (in fact, it's higher).
We arrived at the same conclusion as in Example 8.11. □

8.4.4 Statistical significance vs Practical significance

Statistical significance sometimes has little to do with practical significance. Statistical significance (i.e. a small p-value) is only concerned with the amount of evidence to reject H_0. It does not directly reflect the size of the effect itself. Confidence intervals are more suitable for that.

For example, in testing the effect of a new medication for lowering cholesterol, we might find that the confidence interval for the average decrease μ equals $(1.2, 2.8)$ units (mg/dL). Since the C.I. has positive values we proved $H_A : \mu > 0$. However, the decrease of 1.2 to 2.8 units might be too small in practical terms to justify developing this new drug.

Exercises

8.7.

In determining the gas mileage of a new model of hybrid car, the independent research company collected information from 14 randomly selected drivers. They obtained the sample mean of 38.4 mpg, with the standard deviation of 5.2 mpg. Obtain a 99% C.I. for μ.
What is the meaning of μ in this problem? What assumptions are necessary for your C.I. to be correct?

8.8.

This problem is based on the well-known Newcomb data set for the speed of light.[13] It contains the measurements (in nanoseconds) it took the light to bounce inside a network of mirrors. The numbers given are the time recorded minus $24,800$ ns. We will only use the first ten values.

```
28  26  33  24  34  -44  27  16  40  -2
```

Some mishaps in the experimental procedure led to the two unusually low values (-44 and -2). Calculate the 95% C.I.'s for the mean in case when

a) all the values are used

b) the two outliers are removed

Which of the intervals will you trust more and why?

8.9.

For the situation in Example 8.6 (fishing line strength), test the hypotheses using the C.I. approach.

8.5 C.I. and hypothesis tests for comparing two population means

Two-sample problems:

- The goal of inference is to compare the response in two groups.
- Each group is considered to be a sample from a distinct population.
- The responses in each group are independent of those in the other group.

We have two independent samples, from two distinct populations. Here is the notation that we will use to describe the two populations:

population	Variable	Mean	Standard deviation
1	X_1	μ_1	σ_1
2	X_2	μ_2	σ_2

We want to compare the two population means, either by giving a confidence interval for $\mu_1 - \mu_2$ or by testing the hypothesis of difference, $H_0 : \mu_1 = \mu_2$. Inference is based on two independent random samples. Here is the notation that describes the samples:

sample	sample size	sample mean	sample st.dev.
1	n_1	$\overline{X}_1$	S_1
2	n_2	$\overline{X}_2$	S_2

If independent samples of size n_1 and n_2 are drawn at random from two populations, with means μ_1 and μ_2 and variances σ_1^2 and σ_2^2, respectively, the sampling distribution of the differences of the means $\overline{X}_1 - \overline{X}_2$, is normally distributed with mean $\mu_{\overline{X}_1 - \overline{X}_2} = \mu_1 - \mu_2$ and variance $\sigma_D^2 = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$. Then, the two-sample Z statistic

$$Z = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sigma_D}$$

has the standard normal $N(0, 1)$ sampling distribution.

Usually, population standard deviations σ_1 and σ_2 are not known. We estimate them by using sample standard deviations S_1 and S_2. But then the Z-statistic will turn into (approximately) T-statistic, with degrees of freedom equal to the smaller of $n_1 - 1$ or $n_2 - 1$.

Further, if we are testing $H_0 : \mu_1 = \mu_2$, then $\mu_1 - \mu_2 = 0$. Thus, we obtain the confidence intervals and hypothesis tests for $\mu_1 - \mu_2$.

The 100%$(1 - \alpha)$ confidence interval for $\mu_1 - \mu_2$ is given by

$$(\overline{X}_1 - \overline{X}_2) \pm t_{\alpha/2} \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}} \qquad T \text{ has } \texttt{df} = \min(n_1, n_2) - 1$$

Steps of a Hypothesis Test

a) Null Hypothesis $H_0 : \mu_1 = \mu_2$

b) Alternative Hypothesis $H_A : \mu_1 \neq \mu_2$, or $H_A : \mu_1 > \mu_2$, or $H_A : \mu_1 < \mu_2$.

c) Critical value: $t_{\alpha/2}$ for two-tailed or t_α for one-tailed test, for some chosen *significance level* α.

d) Test Statistic $T = \dfrac{\overline{X}_1 - \overline{X}_2}{\sqrt{S_1^2/n_1 + S_2^2/n_2}}$

e) Decision Rule: Reject H_0 if

$$\begin{array}{cl} |t| > t_{\alpha/2} & \text{for two-tailed} \\ t > t_\alpha & \text{for right-tailed} \\ t < -t_\alpha & \text{for left-tailed} \end{array}$$

or, using p-value, Reject H_0 when p-value $< \alpha$.
P-value is calculated similarly to 1-sample T-test, but now with `df` $= \min(n_1, n_2) - 1$.

f) Conclusion in the words of the problem.

Example 8.14.

A study of iron deficiency among infants compared samples of infants following different feeding regimens. One group contained breast-fed infants, while the other group were fed a standard baby formula without any iron supplements. Here are the data on blood hemoglobin levels at 12 months of age:

Group	n	$\overline{X}$	s
Breast-fed	23	13.3	1.7
Formula	19	12.4	1.8

(a) Is there significant evidence that the mean hemoglobin level is higher among breast-fed babies?
(b) Give a 95% confidence interval for the mean difference in hemoglobin level between the two populations of infants.

Solution. (a) $H_0 : \mu_1 - \mu_2 = 0$ vs $H_A : \mu_1 - \mu_2 > 0$, where μ_1 is the mean of the Breast-fed population and μ_2 is the mean of the Formula population.

The test statistic is

$$t = \frac{13.3 - 12.4}{\sqrt{\frac{1.7^2}{23} + \frac{1.8^2}{19}}} = \frac{0.9}{0.544} = 1.654$$

with 18 degrees of freedom. The p-value is $P(T > 1.654) = 0.057$. This is not quite significant at 5% level.
(b) The 95% confidence interval is

$$0.9 \pm 2.101(0.544) = 0.9 \pm 1.1429 = (-0.2429, 2.0429)$$

□

8.5.1 Matched pairs

Sometimes, we are comparing data that come in pairs of matched observations. A good example of this are "before" and "after" studies. They present the measurement of some quantity for the same set of subjects before and after a certain treatment has been administered. Another example of this situation is twin studies for which pairs of identical twins are selected and one twin (at random) is given a treatment, while the other is serving as a *control* (that is, does not receive any treatment, or maybe receives a fake treatment, *placebo*, to eliminate psychological effects).

When the same subjects are used, we should not consider the measurements independent. In this case, we would compute `Difference` = `Before` − `After` or `Treatment` − `Control` and just do a one-sample test for the mean difference.

Example 8.15.

The following are the left hyppocampus volumes (in cm^3) for a group of twin pairs, one is affected by schizophrenia, and the other is not[14]

Pair number	1	2	3	4	5	6	7	8	9	10	11	12
Unaffected	1.94	1.44	1.56	1.58	2.06	1.66	1.75	1.77	1.78	1.92	1.25	1.93
Affected	1.27	1.63	1.47	1.39	1.93	1.26	1.71	1.67	1.28	1.85	1.02	1.34
Difference	0.67	-0.19	0.09	0.19	0.13	0.40	0.04	0.10	0.50	0.07	0.23	0.59

Is there evidence that the LH volumes for schizophrenia-affected people are different from the unaffected ones?

Solution. Since the twins' LH volumes are clearly not independent (if one is large the other is likely to be large, too – positive correlation!), we cannot use the 2-sample procedure.
However, we can just compute the differences (Unaffected – Affected) and test for the mean difference to be equal to 0. That is,

$$H_0 : \mu = 0 \text{ versus } H_A : \mu \neq 0$$

where μ is the "true" average difference, and $\overline{X}, S$ are computed for the sample of differences.

Given that $\overline{X} = 0.235$ and $S = 0.254$, let's test these hypotheses at $\alpha = 0.10$. We obtain $t = (0.235 - 0)/(0.254/\sqrt{12}) = 3.20$. From the T-table with $df = 11$ we get p-value between $2(0.005) = 0.01$ and $2(0.001) = 0.002$. At $\alpha = 0.05$, we Reject H_0, thus stating that there is a significant difference between LH volumes of normal and schizophrenic people. □

Exercises

8.10.
In studying how humans pick random objects, the subjects were presented a population of rectangles and have used two different sampling methods. They then calculated the average areas of the sampled rectangles for each method. Their results were

	mean	st.dev.	n
Method 1	10.8	4.0	16
Method 2	6.1	2.3	16

Calculate the 99% C.I. for the difference of "true" means by the two methods. Is there evidence that the two methods produce different results?

8.11.
The sports research lab studies the effects of swimming on maximal volume of oxygen uptake.
For 8 volunteers, the maximal oxygen uptake was measured before and after the 6-week swimming program. The results are as follows:

Before	2.1	3.3	2.0	1.9	3.5	2.2	3.1	2.4
After	2.7	3.5	2.8	2.3	3.2	2.1	3.6	2.9

Is there evidence that the swimming program has increased the maximal oxygen uptake?

8.6 Inference for Proportions

8.6.1 Confidence interval for population proportion

In this Chapter, we will consider estimating the proportion p of items of certain type, or maybe some probability p. The unknown population proportion p is estimated by the **sample proportion**

$$\hat{p} = \frac{X}{n}.$$

We know (from CLT, Section 6.4) that if the sample size is sufficiently large, $\hat{p}$ has approximately normal distribution, with mean $\mathbb{E}(\hat{p}) = p$ and standard deviation $\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$. Based on this, we obtain

Theorem 8.2. CI for proportion

For a random sample of size n from a large population with unknown proportion p of successes, the $(1-\alpha)100\%$ confidence interval for p is

$$\hat{p} \pm z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n}$$

8.6.2 Test for a single proportion

To test the hypothesis $H_0 : p = p_0$, use the z-statistic

$$z = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}}$$

In terms of a standard normal Z, the approximate p-value for a test of H_0 is
$P(Z > z)$ against $H_A : p > p_0$,
$P(Z < z)$ against $H_A : p < p_0$,
$2P(Z > |z|)$ against $H_A : p \neq p_0$.

In practice, Normal approximation works well when both X and $n - X$ are at least 10.

Example 8.16.
The French naturalist Count Buffon once tossed a coin 4040 times and obtained 2048 heads. Test the hypothesis that the coin was balanced.

Solution. To assess whether the data provide evidence that the coin was not balanced, we test $H_0 : p = 0.5$ versus $H_A : p \neq 0.5$.

The test statistic is

$$z = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}} = \frac{0.5069 - 0.50}{\sqrt{0.50(1-0.5)/4040}} = 0.88$$

From Z chart we find $P(Z < 0.88) = 0.8106$. Therefore, the p-value is $2(1 - 0.8106) = 0.38$. The data are compatible with balanced coin hypothesis.

Now we will calculate a 99% confidence interval for p. The $z_{\alpha/2} = 2.576$ from the normal table. Hence, the 99% CI for p is

$$\hat{p} = 0.5069 \pm 2.576\sqrt{\frac{(0.5069)(1-0.5069)}{4040}} = 0.5069 \pm (2.576)(0.00786)$$

$$= 0.5069 \pm 0.0202 = (0.4867, 0.5271)$$

□

8.6.3 Comparing two proportions*

We will call the two groups being compared Population 1 and Population 2, with population proportions of successes p_1 and p_2. Here is the notation we will use in this section:

population	Pop.prop.	sample	Successes	sample Prop.
1	p_1	n_1	X_1	$\hat{p}_1 = \frac{X_1}{n_1}$
2	p_2	n_2	X_2	$\hat{p}_2 = \frac{X_2}{n_2}$

To compare the two proportions, we use the difference between the two sample proportions: $\hat{p}_1 - \hat{p}_2$. Therefore, when n_1 and n_2 are large, $\hat{p}_1 - \hat{p}_2$ is approximately normal with mean $\mu = p_1 - p_2$ and standard deviation $\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$. Note that for unknown p_1 and p_2 we replace them by $\hat{p}_1$ and $\hat{p}_2$ respectively.

Definition 8.5. Inference for two proportions

The $(1-\alpha)100\%$ confidence interval for $p_1 - p_2$ is

$$\hat{p}_1 - \hat{p}_2 \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

To test the hypothesis $H_0 : p_1 - p_2 = 0$, we use the test statistic

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{SE_{\hat{p}}},$$

where $SE_{\hat{p}} = \sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$ and $\hat{p} = \frac{X_1+X_2}{n_1+n_2}$.

Example 8.17.

To test the effectiveness of a new pain relieving drug, 80 patients at a clinic were giving a pill containing the drug and 80 others were giving a placebo. At the 0.01 level of significance, what can we conclude about the effectiveness of the drug if the first group 56 of the patients felt a beneficial effect while 38 out of those who received placebo felt a beneficial effect?

Solution. $H_0 : p_1 - p_2 = 0$ and $H_A : p_1 - p_2 > 0$
$Z = 2.89$, where $\hat{p}_1 = \frac{56}{80} = 0.7$ and $\hat{p}_2 = \frac{38}{80} = 0.475$ and $\hat{p} = \frac{56+38}{80+80} = 0.5875$
P-value $= 1 - P(Z > 2.89) = 0.0019$
Since the p-value is less than 0.01, the null hypothesis must be rejected, so the drug is effective. □

Exercises

8.12.

Suppose that a nutritionist claims that at least 75% of the preschool children in a certain country have protein deficient diets, and that a sample survey reveals that this true for 206 preschool children in a sample of 300. Test the claim at the 0.02 level of significance. Also, compute a 98% confidence interval.

8.13.

In a survey of 200 office workers, 165 said they were interrupted three or more times an hour by phone messages, faxes etc. Find and interpret a

90% confidence interval for the population proportion of workers who are interrupted three or more times an hour.

8.14.

You would like to design a poll to determine what percent of your peers volunteer for charities. You have no clear idea of what the value of p is going to be like, and you'll be satisfied with the 90% margin of error equal to $\pm 10\%$. Find the sample size needed for your study.

8.15.

In random samples of 200 tractors from one assembly line and 400 tractors from another, there were, respectively, 16 tractors and 20 tractors which required extensive adjustments before they could be shipped. At the 5% level of significance, can we conclude that there is a difference in the quality of the work of the two assembly lines?

Chapter Exercises

For each of the questions involving hypothesis tests, state the null and alternative hypotheses, compute the test statistic, determine the p-value, make the decision and summarize the results in plain English. Use $\alpha = 0.05$ unless otherwise specified.

8.16.

Two brands of batteries are tested and their voltages are compared. The summary statistics are below. Find and interpret a 95% confidence interval for the true difference in means.

	mean	st.dev.	n
Brand 1	9.2	0.3	25
Brand 2	8.9	0.6	27

8.17.

You are studying yield of a new variety of tomato. In the past, yields of similar types of tomato have shown a standard deviation of 8.5 lbs per plant. You would like to design a study that will determine the average yield within a 90% error margin of ± 2 lbs. How many plants should you sample?

8.18.

A biologist knows that the average length of a leaf of a certain full-grown plant is 4 inches. A sample of 45 leaves from the plants that were given a new type of plant food had an average length of 4.2 inches, with the standard deviation of 0.6 inches. Is there reason to believe that the new plant food is responsible for a change in the average growth of leaves? Use $\alpha = 0.02$. Would your conclusion have changed if you used $\alpha = 0.05$?

8.19.

A job placement director claims that mean starting salary for nurses is \$38,000. A random sample of 10 nurses' salaries has a mean \$35,450 and a standard deviation of \$4,700. Is there enough evidence to reject the director's claim at $\alpha = 0.01$?

8.20.

College Board claims[15] that in 2010, public four-year colleges charged, on average, \$7,605 per year in tuition and fees for in-state students. A sample of 20 public four-year colleges collected in 2011 indicated a sample mean of \$8,039 and the sample standard deviation was \$1,950. Is there sufficient evidence to conclude that the average in-state tuition has increased?

8.21.

The weights of grapefruit follow a normal distribution. A random sample of 12 new hybrid grapefruit had a mean weight of 1.7 pounds with standard deviation 0.24 pounds. Find a 95% confidence interval for the mean weight of the population of the new hybrid grapefruit.

8.22.

The Mountain View Credit Union claims that the average amount of money owed on their car loans is \$ 7,500. Suppose a random sample of 45 loans shows the average amount owed equals \$8,125, with standard deviation \$4,930. Does this indicate that the average amount owed on their car loans is not \$7,500? Use a 1% level of significance.

8.23.

An overnight package delivery service has a promotional discount rate in effect this week only. For several years the mean weight of a package delivered by this company has been 10.7 ounces. However, a random sample of 12 packages mailed this week gave the following weights in ounces:

```
12.1  15.3   9.5  10.5  14.2  8.8
10.6  11.4  13.7  15.0   9.5  11.1
```

Use a 1% level of significance to test the claim that the packages are averaging more than 10.7 ounces during the discount week.

8.24.

An item in USA Today reported that 63% of Americans owned a mobile browsing device. A survey of 143 employees at a large school showed that 85 owned a mobile browsing device. At $\alpha = 0.02$, test the claim that the percentage is the same as stated in USA Today.

8.25.

A poll by CNN revealed that 47% of Americans approve of the job performance of the President. The poll was based on a random sample of 537 adults.

a) Find the 95% margin of error for this poll.

b) Based on your result in part (a), test the hypothesis $H_0 : p = 0.5$ where p is the proportion of all American adults that approve of the job performance of the President. *Do not compute the test statistic and p-value.*

c) Would you have also reached the same conclusion for $H_0 : p = 0.45$?

8.26.

Find a poll cited in a newspaper, web site or other news source, with a mention of the sample size and the margin of error. (For example, `rasmussenreports.com` frequently discuss their polling methods.) Confirm the margin of error presented by the pollsters, using your own calculations.

Chapter 9

Linear Regression

In science and engineering, there is often a need to investigate the relationship between two continuous random variables.

Suppose that, for every case observed, we record two variables, X and Y. The *linear* relationship between X and Y means that $\mathbb{E}(Y) = \beta_0 + \beta_1 X$.

X variable is usually called *predictor* or *independent variable* and Y variable is the *response* or *dependent variable*.

Example 9.1.

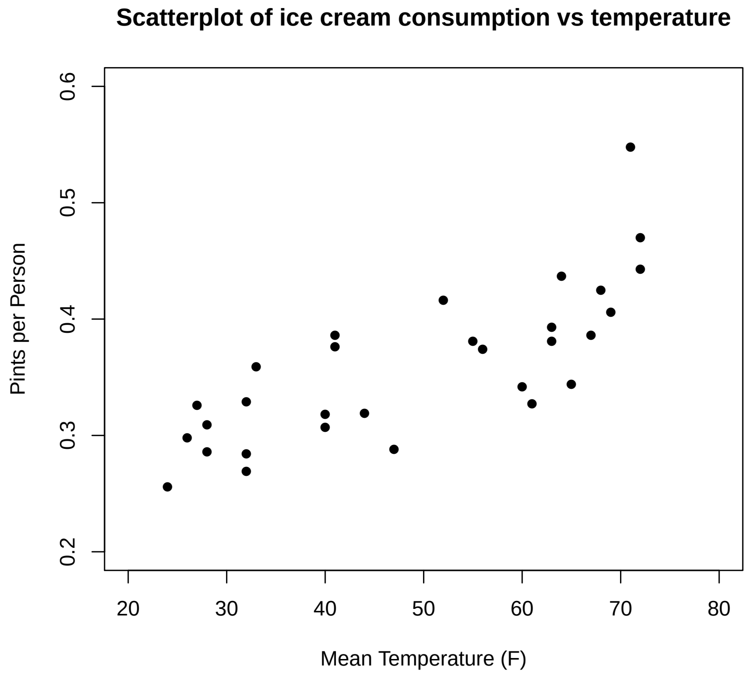
Imagine that we are opening an ice cream stand and would like to be able to predict how many customers we will have. We might use the temperature as a predictor. We decided to collect data over a 30-week period from March to July.[16]

Week	1	2	3	4	5	6	7	8	9	10
Mean temp	41	56	63	68	69	65	61	47	32	24
Consumption	0.386	0.374	0.393	0.425	0.406	0.344	0.327	0.288	0.269	0.256

Week	11	12	13	14	15	16	17	18	19	20
Mean temp	28	26	32	40	55	63	72	72	67	60
Consumption	0.286	0.298	0.329	0.318	0.381	0.381	0.47	0.443	0.386	0.342

Week	21	22	23	24	25	26	27	28	29	30
Mean temp	44	40	32	27	28	33	41	52	64	71
Consumption	0.319	0.307	0.284	0.326	0.309	0.359	0.376	0.416	0.437	0.548

The following *scatterplot* is made to graphically investigate the relationship.



There indeed appears to be a straight-line trend. We will discuss fitting the equation a little later.

9.1 Correlation coefficient

We already know the correlation coefficient between two random variables,

$$\rho = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}$$

Now, let's consider its sample analog, **sample correlation coefficient**

$$r = \frac{\sum_{i=1}^n (Y_i - \overline{Y})(X_i - \overline{X})}{\sqrt{\sum_{i=1}^n (X_i - \overline{X})^2 \ \sum_{i=1}^n (Y_i - \overline{Y})^2}} \equiv \frac{SS_{XY}}{\sqrt{SS_X SS_Y}}$$

You can recognize the summation on top as a discrete version of $Cov(X,Y)$ and the sums on the bottom as part of the computation for the sample variances of X, Y. These are

$$SS_{XY} = \sum XY - \frac{\sum X \sum Y}{n},$$

$$SS_X = \sum X^2 - \frac{\left(\sum X\right)^2}{n}, \quad SS_Y = \sum Y^2 - \frac{\left(\sum Y\right)^2}{n}$$

All sums are taken from 1 to n.
For example, the sample variance of X is $S_X^2 = SS_X/(n-1)$.

Let's review the properties of the correlation coefficient ρ and its sample estimate, r:

- the sign of r points to positive (when X increases, Y increases too) or negative (when one increases, the other decreases) relationship
- $-1 \le r \le 1$, with $+1$ being a perfect positive and -1 a perfect negative relationship
- $r \approx 0$ means no linear relationship between X and Y (caution: there can still be a non-linear relationship!)
- r is dimensionless, and it does not change when X or Y are linearly transformed.

9.2 Least squares regression line

The complete regression equation is

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad i = 1, ..., n$$

where the errors ε_i are assumed to be independent, $\mathcal{N}(0, \sigma^2)$.

To find the "best fit" line, we choose $\hat{\beta}_0$ and $\hat{\beta}_1$ that minimize the sum of squared residuals

$$SSE = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2$$

(SSE is for the Sum of Squared Errors, however the quantities $Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i$ are usually referred to as *residuals*.)

To find the minimum, we would calculate partial derivatives of SSE with respect to β_0, β_1. Solving the resulting system of equations, we get the following

Theorem 9.1. Least squares estimates

> The estimates for the regression equation $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad i = 1, ..., n$ are:
>
> $$\text{Slope } \hat{\beta}_1 = \frac{SS_{XY}}{SS_X} = r\frac{S_Y}{S_X} \qquad \text{and Intercept } \hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X}$$

Example 9.2.

To illustrate the computations, let's consider another data set. Here, X = amount of tannin in the larva food, and Y = growth of insect larvae.

X	0	1	2	3	4	5	6	7	8
Y	12	10	8	11	6	7	2	3	3

Estimate the regression equation and correlation coefficient.

Solution.
$\sum X = 36$, $\sum Y = 62$, $\sum X^2 = 204$, $\sum Y^2 = 536$, $\sum XY = 175$.

Therefore,

$\overline{X} = 36/9 = 4, \quad \overline{Y} = 62/9 = 6.89, \quad SS_X = 204 - 36^2/9 = 60,$
$SS_Y = 536 - 62^2/9 = 108.9, \quad SS_{XY} = 175 - 36(62)/9 = -73$

and finally,
$\hat{\beta}_1 = -73/60 = -1.22, \quad \hat{\beta}_0 = 6.89 - (-1.22)4 = 11.76, \quad r = -0.903$

Thus, we get the equation

$$\hat{Y} = 11.76 - 1.22X$$

that is interpretable as a prediction for *any* given value X. In practice, the accuracy of prediction depends on X, see the next Section. □

Example 9.3.

For the data in Example 9.1,

a) Calculate and plot the least squares regression line

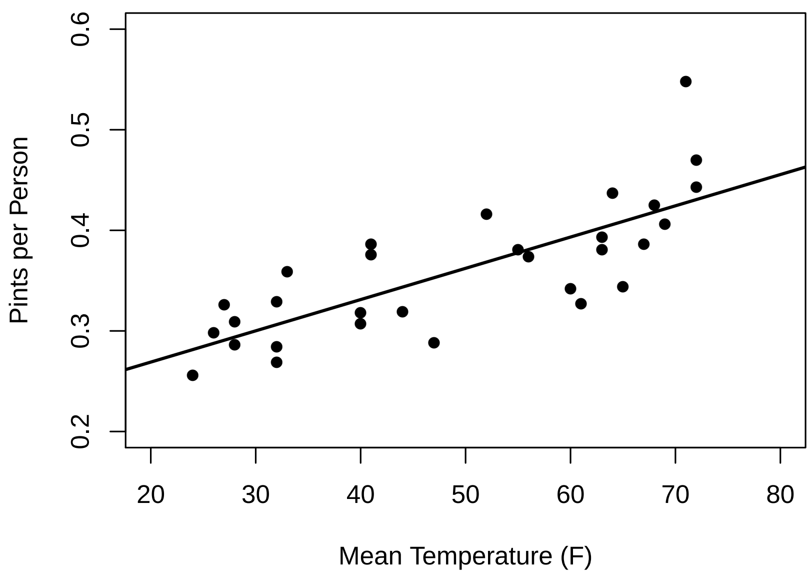
b) Predict the consumption when $X = 50°F$.

Solution.

(a) We may obtain the following estimates (usually done by a computer)

$$\hat{\beta}_0 = 0.2069, \quad \hat{\beta}_1 = 0.003107 \quad \text{and} \quad r = 0.776$$

These can be used to plot the regression line and make predictions



Can you interpret the slope and the intercept for this problem in plain English?

(b) $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X = 0.2069 + 0.003107(50) = 0.362$ pints per person. □

9.3 Inference for regression

The error variance σ^2 determines the amount of scatter of the Y-values about the line. That is, it reflects the uncertainty of prediction of Y using X.
Its sample estimate is

$$S^2 = \frac{SSE}{n-2} = \frac{\sum_{i=1}^{n} [Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i)]^2}{n-2}$$

we divide by $n-2$ because two degrees of freedom have been used up when estimating $\hat\beta_0$, $\hat\beta_1$. The estimate of S can be obtained by hand or using the computer output.

The values $\hat Y_i = \hat\beta_0 + \hat\beta_1 X_i$ are called *predicted* or *fitted* values of Y. The differences

$$\text{Actual} - \text{ Predicted} \equiv Y_i - \hat Y_i = e_i, \quad i = 1, ..., n$$

are called *residuals.*

The least squares esimates for slope and intercept are also the sample estimates for the "true" slope and intercept. We can apply the same methods we have done for, say, estimating the unknown mean μ. However, it is harder to compute the margins of error. We would typically use computer output to produce *standard errors* (that is, the estimates of standard deviations) of these estimates.
$100\%(1-\alpha)$ CI's for regresison parameters are then found as

$$\texttt{Estimate} \pm t_{\alpha/2}(\texttt{Std.Error}), \quad t \text{ has } df = n-2$$

Example 9.4.
Continuing the analysis of data from Example 9.1, let's examine a portion of computer output (done by R statistical package).

	Estimate	Std.Error	t-value	Pr(>|t|)
(Intercept)	0.2069	0.0247	8.375	4.13e-09
X	0.003107	0.000478	6.502	4.79e-07

We can calculate confidence intervals and hypothesis tests for the parameters β_0 and β_1.

The 95% C.I. for the slope β_1 is

$$0.003107 \pm 2.048(0.000478) = [0.002128, 0.004086]$$

To test the hypothesis $H_0 : \beta_1 = 0$ we could use the test statistic

$$t = \frac{\texttt{Estimate}}{\texttt{Std.Error}}$$

For the above data, we have $t = 0.003107/0.000478 = 6.502$, as reported in the table. The p-values for this test can be found using a T-table; they are also reported by the computer. Above, the reported p-value of `4.79e-07` is very small, meaning that the hypothesis $H_0 : \beta_1 = 0$ is strongly rejected. □

9.3.1 Hypothesis test for linear relationship

In terms of correlation r, the above test can be calculated more easily using the test statistic

$$t = r\sqrt{\frac{n-2}{1-r^2}}, \quad df = n-2$$

Strictly speaking, this is for testing

$$H_0 : \rho = 0 \text{ versus } H_A : \rho \neq 0$$

but $\rho = 0$ and $\beta_1 = 0$ are equivalent statements.

Example 9.5.

For a relationship between Population size and Divorce rate in $n = 20$ American cities the correlation of 0.28 was found. Is there a significant linear relationship between Population size and Divorce rate?

Solution.

$$t = 0.28\sqrt{\frac{20-2}{1-0.28^2}} = 1.23 \quad \text{with } df = 18$$

From T-table (comparing with table value $t = 1.33$), $\texttt{p} - \texttt{value} > 2(0.1) = 0.2$. Since p-value is larger than our default level $\alpha = 0.05$, do not reject H_0. Thus, we can claim **no significant evidence** of the linear relationship between Population size and Divorce rate. □

9.3.2 Confidence and prediction intervals

In addition to the C.I.'s for β_0 and β_1, we might be interested in the uncertainty of estimating Y-values given the particular value of X.

$100\%(1-\alpha)$ **confidence interval for mean response** $\mathbb{E}(\hat{Y})$ **given** $X = x^*$

$$(\hat{\beta}_0 + \hat{\beta}_1 x^*) \pm t_{\alpha/2} S \sqrt{\frac{1}{n} + \frac{(x^* - \overline{X})^2}{(n-1)S_X^2}}$$

$100\%(1-\alpha)$ **prediction interval for a future observation** Y **given** $X = x^*$

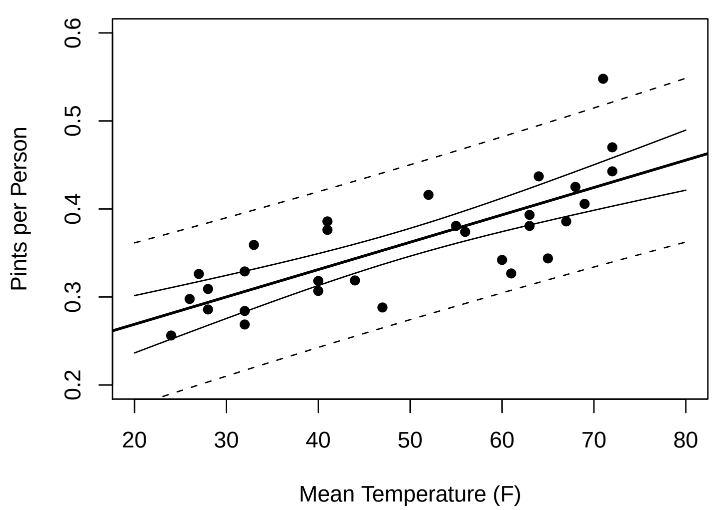
$$(\hat{\beta}_0 + \hat{\beta}_1 x^*) \pm t_{\alpha/2} S \sqrt{1 + \frac{1}{n} + \frac{(x^* - \overline{X})^2}{(n-1)S_X^2}}$$

What is the main difference between *confidence* and *prediction* intervals? Confidence interval is only concerned with the **mean** response $\mathbb{E}(Y)$. That is, it's trying to catch the regression line. Prediction interval is concerned with any future observation. Thus, it is trying to catch all the points in the scatterplot. As a consequence, prediction interval is typically much wider.

Note also that $(n-1)S_X^2 = SS_X$, and both intervals are narrowest when x^* is closest to $\overline{X}$, the center of all data. The least squares fit becomes less reliable as you move to values of X away from the center, especially the areas where there is no X-data.

Example 9.6.

Continuing the analysis of data from Example 9.1, calculate both 95% confidence and prediction intervals for the ice cream consumption when temperature is $70°F$



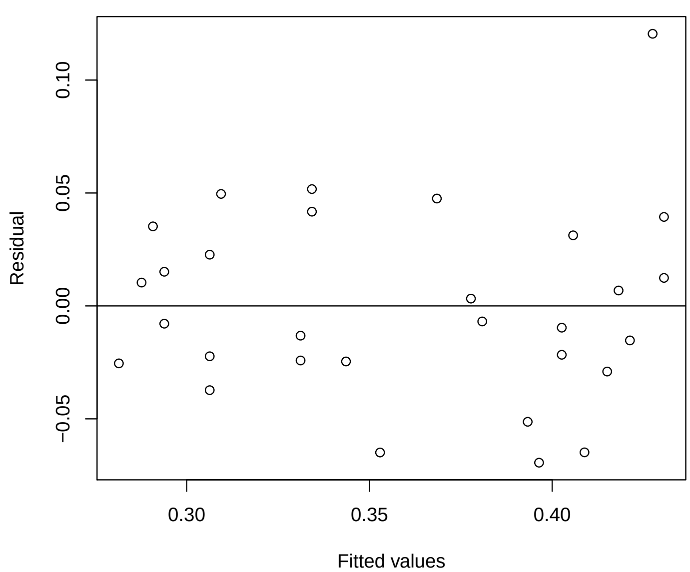
Solution. $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x^* = 0.2069 + (0.003107)70 = 0.4244$, and using the computer output in Example 9.4, we will get $t_{\alpha/2} = 2.048$,
$S = \sqrt{\texttt{Mean Sq Residuals}} = \sqrt{0.001786} = 0.0423$ and $\dfrac{1}{n} + \dfrac{(x^* - \overline{X})^2}{(n-1)S_X^2} = 0.0892.$
Then

$$\text{CI} \quad 0.4244 \pm 0.0259, \qquad \text{PI} \quad 0.4244 \pm 0.0904$$

For comparison, both intervals are plotted above for various values of x^*. Note that the 95% prediction interval contains all but one observation. □

9.3.3 Checking the assumptions

To check the assumption of linear relationship and the constant variance (σ^2) of the residuals, we might make a plot of Residuals $e_i = Y_i - \hat{Y}_i$ versus Fitted values. Here is such plot for the ice cream example



If there is any trend or pattern in the residuals, then the assumptions for linear regression are not met. We do not see any particular trend except possibly one unusually high value (an outlier) in the top right corner.

Exercises

9.1.

In the file `http://www.nmt.edu/~olegm/382/cars2010.csv`, there are some data on several 2010 compact car models. The variables are: engine displacement (liters), city MPG, highway MPG, and manufacturer's suggested price.

a) Is the car price related to its highway MPG?

b) Is there a relationship between city and highway MPG?

Use scatterplots, calculate and interpret the correlation coefficient, test to determine if there is a linear relationship.
For part (b), also compute and interpret the regression equation. Plot this line on the scatterplot. Plot the residuals versus predicted values. Does the model fit well?

9.2.

The following is an illustration of famous Moore's Law for computer chips. X = Year (minus 1900, for ease of computation), Y = number of transistors (in 1000)

X	71	79	83	85	90	93	95
Y	2.3	31	110	280	1200	3100	5500

a) Make a scatterplot of the data. Is the growth linear?

b) Let's try and fit the exponential growth model using a transformation:

$$\text{If} \qquad Y_i = a_0 e^{a_1 X_i} \qquad \text{then} \qquad \ln Y_i = \ln a_0 + a_1 X_i$$

That is, doing the linear regression analysis of $\ln Y$ on X will help recover the exponential growth. Make the regression analysis of $\ln Y$ on X. Does this model do a good job fitting the data?

c) Predict the number of transistors in the year 2005. Did this prediction come true?

9.3.

A head of a large Hollywood company has seen the following values of its market share in the last six years:[17]

```
11.4, 10.6, 11.3, 7.4, 7.1, 6.7
```

Is there statistical evidence of a downward trend in the company's market share?

9.4.

For the Old Faithful geyser, the durations of eruption (X) were recorded, with the interval to the next eruption (Y), both in minutes.

X	3.6	1.8	3.3	2.3	4.5	2.9	4.7	3.6	1.9
Y	79	54	74	62	85	55	88	85	51

Perform the regression analysis of Y on X. Interpret the slope and give a 95% confidence interval for the slope.

Notes

[9]see e.g. http://forgetomori.com/2009/skepticism/seeing-patterns/

[10]http://en.wikipedia.org/wiki/Heights_of_Presidents_of_the_United_States_and_presidential_candidates

[11]see http://www.readingonline.org/articles/bergman/wait.html

[12]"Student" [William Sealy Gosset] (March 1908). "The probable error of a mean". Biometrika 6 (1): 1-25.

[13]For example, see "http://www.stat.columbia.edu/~gelman/book/data/light.asc"

[14]example from "Statistical Sleuth"

[15]http://www.collegeboard.com/student/pay/add-it-up/4494.html

[16] Source: Kotswara Rao Kadilyala (1970). "Testing for the independence of regression disturbances" Econometrica, 38, 97-117. Appears in: *A Handbook of Small Data Sets*, D. J. Hand, et al, editors (1994). Chapman and Hall, London.

[17]Mlodinow again. The director, Sherry Lansing, was subsequently fired only to see several films developed during her tenure, including *Men In Black*, hit it big.