Rinaldo B. Schinazi

Probability with Statistical Applications

Second Edition

Birkhäuser

Rinaldo B. Schinazi

Probability with Statistical Applications

Second Edition

Birkhäuser

Rinaldo B. Schinazi
Department of Mathematics
University of Colorado
Colorado Springs, CO 80933-7150
USA
rschinaz@uccs.edu

ISBN 978-0-8176-8249-1 e-ISBN 978-0-8176-8250-7
DOI 10.1007/978-0-8176-8250-7
Springer New York Dordrecht Heidelberg London

Library of Congress Control Number: 2011943352

Mathematics Subject Classification (2010): 60-01, 62-01

Printed on acid-free paper

Springer is part of Springer Science+Business Media
(www.birkhauser-science.com)

Preface to the Second Edition

Compared to the first edition, we have made few changes to the first six chapters. They are intended for a first one semester course in probability with some statistics. It is assumed that the reader has had a calculus course but the book is written so that the calculus difficulties of the students do not obscure the probability content. Since probability concepts are not easy to grasp, we drastically limit the number of topics and concentrate on a few concepts that every student should thoroughly understand in a first probability and statistics course. Statistics are introduced as early as possible in the book in order to provide interesting and useful applications of probability.

The main difference with the first edition is the addition of Sect. 8.4, Chaps. 9 and 10. Chapters 7–10 (with supplements from previous chapters) are now intended for a course in Mathematical Statistics. These last chapters rely heavily on calculus of one and several variables. Chapter 10 requires linear algebra. In Chap. 7, moment generating functions are introduced and used to study sums of random variables and convergence of sequences of random variables. Chapter 8 deals with transformations of random variables (using distribution function). Random vectors are introduced and are used to prove a number of facts regarding expectation, variance, covariance, and normal samples. We added Sect. 8.4 to cover conditional distributions and conditional expectations. The first three sections of Chap. 9 deal with finding estimators (moments, maximum likelihood) and comparing estimators (sufficiency, Rao–Blackwell Theorem). We chose not to cover the most general cases, we instead concentrated on the exponential family of distributions. This provides many interesting examples and the theory applied to it is considerably simpler than the general theory. Section 9.4 provides an introduction to Bayesian statistics. Finally, in Chap. 10 we wrote a brief introduction to multiple regression in which we try to balance applications and theory.

Preface to the First Edition

This book is intended as a text for a first one semester course in probability with some statistics. It is assumed that the reader has had a calculus course. At the University of Colorado, we teach this course to a number of majors, including computer science, electrical engineering, mathematics, and physics. In the last few years, some engineering professional societies have suggested that statistics be taught to students and so we have included statistics in the traditional one semester probability course. My main motivation to write this book was that the many good books on probability and statistics are intended for 1-year courses and are very extensive. Anyone who has taught probability knows that it is a hard subject for most students. For this reason I have decided to drastically limit the number of topics and concentrate on a few concepts that I feel every student should thoroughly understand in a first probability and statistics course. I have also decided to introduce statistics as early as possible in the book in order to provide interesting and useful applications of probability.

I have tried to write this book so that the calculus difficulties of the students do not obscure the probability content. I have kept theory to a minimum and I have concentrated on interesting examples. Chapter 1 has the basic rules of probability and conditional probability with some interesting applications such as Bayes' rule and the birthday problem. In Chap. 2 discrete and continuous random variables, expectation, and variance are introduced. Chapter 2 is mostly computational with few probability concepts and many applications of calculus. In Chaps. 3 and 4, we get to the heart of the subject: binomial distribution, normal approximation to the binomial, Poisson distribution, Law of Large Numbers, and Central Limit Theorem. I also cover the Poisson approximation to the binomial (in a nonstandard way) and the Poisson scatter theorem. In Chap. 5, we apply some of the concepts of the preceding chapters to introduce statistics. We cover confidence intervals and hypothesis testing for large samples, we also introduce Student tests to deal with small samples and a nonparametric test. Finally, we test independence and goodness of fit using chi-square tests. Chapter 6 is a short introduction to linear regression. Chapters 7 and 8 rely heavily on calculus of one and several variables to study sums of random variables (via moment generating functions), transformations

of random variables (using distribution functions), and transformations of random vectors. In Chap. 8, we prove a number of facts regarding expectation, variance, and covariance that are used throughout the book. We also prove facts about normal samples that are useful in statistics.

There are at least two ways to use this book for a one semester course. In both ways, one should first cover the first four chapters. Then one might choose to do some statistical applications and cover Chaps. 5 and 6 or one might choose to concentrate on probability and cover Chaps. 7 and 8.

Contents

Chapter 1
Probability Space

1.1 The Axioms of Probability

The study of probability is concerned with the mathematical analysis of random experiments such as tossing a coin, rolling a die, or playing at the lottery. Each time we perform a random experiment there are a number of possible outcomes. We now define the notions of sample space and event.

Sample Space and Events

The sample space Ω of a random experiment is the collection of all possible outcomes of the random experiment.

An event is a subset of Ω.

Example 1. Toss a coin. There are only two possible outcomes and the sample space is $\Omega = \{H, T\}$. The event $A = \{H\}$ is equivalent to the event "the outcome was heads."

Example 2. Roll a die. This time the sample space is $\Omega = \{1, 2, 3, 4, 5, 6\}$. The event $B = \{1, 3, 5\}$ is equivalent to the event "the die showed an odd face."

Example 3. The birthday of someone. The sample space has 365 points (ignoring leap years).

Example 4. We count the number of rolls until we get a 6. Here $\Omega = \{1, 2, \ldots\}$. That is, the sample space consists of all strictly positive integers. Note that this sample space has infinitely many points.

We define next some useful relations among events.

If A is an event included in the sample space Ω then the event consisting of all the points of Ω not included in A is called the complement of A and is denoted by A^c.

R.B. Schinazi, *Probability with Statistical Applications*,
DOI 10.1007/978-0-8176-8250-7_1, © Springer Science+Business Media, LLC 2012

Assume that A and B are two events then the intersection of A and B is the set of points that are both in A and B. The intersection of A and B is denoted by AB or by $A \cap B$.

If A and B are two events then the union of A and B is the set of points that are in A or in B (they may be in both). The union of A and B is denoted by $A \cup B$.

The empty set is denoted by $\emptyset$. Two events are said to be disjoint or mutually exclusive if

$$AB = \emptyset.$$

More generally, a sequence of events $A_1, A_2, \ldots$ in Ω is said to be mutually exclusive if

$$A_i A_j = \emptyset \text{ for } i \neq j.$$

Example 5. Let A be the event that a student is female, B the event that a student takes french, and C the event that a student takes calculus.

What is the event "a student is female and takes calculus"? We want both A and C so the event is AC.

What is the event "a student does not attend calculus"? We want everything not in C so the event is C^c.

What is the event "a student takes french or calculus"? We want everything in A and everything in B so the event is $A \cup B$.

We now state a few important set theories identities.

Set Identities

Let A and B be two events. We have that

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

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

$$A = AB \cup AB^c$$

The identities above are not difficult to establish. For instance, for the first one we have that x belongs to $(A \cup B)^c$ if and only if x does not belong to $A \cup B$, this in turn is equivalent to x not belonging to A AND not belonging to B, which is equivalent to x belonging to A^c and to B^c and thus to $A^c B^c$.

We now give the rules of probability.

Axioms of Probability

(i) For any event A in Ω we have $0 \leq P(A) \leq 1$.
(ii) $P(\Omega) = 1$.
(iii) For a finite or infinite sequence of disjoint events A_i we have

$$P(\cup A_i) = \sum P(A_i).$$

Consequences

C1. If $AB = \emptyset$ then by (iii)

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

C2. $P(A^c) = 1 - P(A)$.
We now prove it. Note that

$$A \cup A^c = \Omega.$$

Hence,

$$P(A \cup A^c) = 1.$$

By C1

$$P(A \cup A^c) = P(A) + P(A^c).$$

Hence, $P(A^c) = 1 - P(A)$ and C2 is proved.

C3. $P(\emptyset) = 0$.
Observe that $\Omega^c = \emptyset$ and by C2

$$P(\Omega^c) = 1 - P(\Omega) = 1 - 1 = 0.$$

C4. Using that $AB \cap AB^c = \emptyset$ and that $A = AB \cup AB^c$ we get by (iii) that

$$P(A) = P(AB) + P(AB^c).$$

C5. Using that $A \cup B = AB^c \cup B$ and that the last two events are disjoint we get by C1 that $P(A \cup B) = P(AB^c) + P(B)$. Now using C4 we know that $P(AB^c) = P(A) - P(AB)$. Thus, for any two events A and B (in particular they do not need to be disjoint) we have

Union of Two Events

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

Example 6. We pick at random a person in a certain population. Let A be the event that the person selected attends college. Let B be the event that the person selected speaks french. Assume that the proportion of persons attending college and speaking french in the population are 0.1 and 0.02, respectively. Then it makes sense to define $P(A) = 0.1$ and $P(B) = 0.02$. Assume also that the proportion of people attending college and speaking french is 0.01. That is, $P(AB) = 0.01$.

What is the probability that a person picked at random does not attend college?

This is the event A^c. By C2 we have

$$P(A^c) = 1 - P(A) = 0.9.$$

What is the probability that a person picked at random speaks french or attends college?

This is the event $A \cup B$. By C5 we have

$$P(A \cup B) = P(A) + P(B) - P(AB) = 0.1 + 0.02 - 0.01 = 0.11.$$

What is the probability that a person speaks french and does not attend college?

This is the event $A^c B$. According to C4 we have

$$P(A^c B) = P(B) - P(AB) = 0.02 - 0.01 = 0.01.$$

1.1.1 Equally Likely Outcomes

We start by considering an example.

Example 7. Roll a fair die. Then $\Omega = \{1, 2, 3, 4, 5, 6\}$. If all the outcomes are equally likely we define

$$P(i) = \frac{1}{6} \text{ for } i = 1, \ldots, 6.$$

What is the interpretation of the statement $P(1) = 1/6$? If we roll the die many times the frequency of observed 1's (i.e., the observed number of 1's divided by the total number of rolls) should be close to 1/6.

What is the probability of the die showing an odd face? By axiom of probability (iii) we have that

$$P(\text{odd}) = P(\{1, 3, 5\}) = P(\{1\}) + P(\{3\}) + P(\{5\}) = \frac{3}{6}.$$

More generally, we have the following.

Equally Likely Outcomes

Consider a finite sample space Ω with finitely many outcomes assumed to be equally likely. Let $|A|$ be the number of elements in A. Then

$$P(A) = \frac{|A|}{|\Omega|}$$

for every event A.

It is easy to check that P defined by the formula above satisfies the three axioms of probability and thus is a probability on Ω.

Example 8. Toss two fair coins. This time we have four equally likely outcomes $\Omega = \{HH, HT, TH, TT\}$.

$$P(\text{at least 1 head}) = \frac{|\{HH, HT, TH\}|}{4} = \frac{3}{4}.$$

Example 9. Roll two dice. What is the probability that the sum is 11? The most natural sample space is all the possible sums: so all integers from 2 to 12. But these outcomes are not equally likely so it is not a good choice. Instead we pick for Ω the collection of all ordered pairs: $\{(1,1),(1,2),\ldots,(2,1),(2,2),\ldots,(6,5),(6,6)\}$. There are 36 equally likely outcomes in Ω.

$$P(\text{sum is11}) = \frac{|\{(5,6),(6,5)\}|}{36} = \frac{2}{36}.$$

Example 10. Roll two dice. What is the probability that the sum of the two dice is 4 or more? It is quicker to compute the probability that the sum is 3 or less which is the complement of the event we want.

$$P(\text{sum is3or less}) = \frac{|\{(1,1),(1,2),(2,1)\}|}{36} = \frac{3}{36}.$$

Therefore,

$$P(\text{sum is4or more}) = 1 - \frac{3}{36} = \frac{33}{36}.$$

Exercises 1.1

1. Let A be the event that a person attends college and B be the event that a person speaks french. Using intersections, unions or complements describe the following events.

(a) A person does not speak french.
(b) A person speaks french and does not attend college.
(c) A person is either in college or speaks french.
(d) A person is either in college or speaks french but not both.

2. Let A and B be events such that $P(A) = 0.6$, $P(B) = 0.3$, and $P(AB) = 0.1$.

(a) Find the probability that A or B occurs.
(b) Find the probability that exactly one of A or B occurs.
(c) Find the probability that at most one of the two events A and B occurs.
(d) Find the probability that neither A nor B occurs.

3. Toss three fair coins.

(a) What is the probability of having at least one head?
(b) What is the probability of having exactly one head?

4. Roll two fair dice.

(a) What is the probability that they do not show the same face?
(b) What is the probability that the sum is 7?
(c) What is the probability that the maximum of the two faces is at least 3?

5. In a college it is estimated that 1/4 of the students drink, 1/8 of the students smoke, and 1/10 smoke and drink. Picking a student at random,

(a) What is the probability that the student does not drink nor smoke?
(b) What is the probability that a student smokes or drinks?

6. A roulette has 38 pockets, 18 are red, 18 are black, and 2 are green. I bet on red, you bet on black.

(a) What is the probability that I win?
(b) What is the probability that at least one of us wins?
(c) What is the probability that at least one of us loses?

7. Roll 3 dice.

(a) What is the probability that you get 3 7's?
(b) What is the probability that you get a triplet?
(c) What is the probability that you get a pair?

8. I buy many items at a grocery store. What is the probability that the bill be a whole number?

9. If $A \subset B$, show that $P(A^c B) = P(B) - P(A)$.

10. Show that for any three events A, B, C we have

$$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(AB) - P(AC) - P(BC) + P(ABC).$$

Can you guess what the formula is for the union of four events?

1.2 Conditional Probabilities and Bayes' Formula

Example 1. Roll two dice successively and observe the sum. As we observed before we should take for our sample space the 36 ordered pairs. Let A be the event "the sum is 11." Since all the outcomes are equally likely we have that

$$P(A) = \frac{|\{(5,6),(6,5)\}|}{36} = \frac{1}{18}.$$

Let B be the event "the first die shows a 6." We are now interested in the following question: if we observe the first die and it shows a 6, how does this affect the probability of observing a sum of 11? In other words, given B, what is the probability of A? Observe that for this question our sample space is B. The notation for the preceding probability is

$$P(A|B)$$

and is read "probability of A given B." Given that the first die shows a 6 there is only one possibility for the sum to be 11. The second die needs to show 5. The probability of this event is 1/6. Thus,

$$P(A|B) = \frac{1}{6}.$$

More generally, we have the following definition.

Conditional Probability

Assume that $P(B) > 0$. The probability of A given B is defined by

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

In the case of equally likely outcomes the formula becomes

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

By using the definition above it is easy to see that the rules of probability apply to conditional probabilities. In particular,

$$P(A \cup B|C) = P(A|C) + P(B|C) \text{if } A \text{ and } B \text{ are disjoint}$$

and

$$P(A^c|B) = 1 - P(A|B).$$

Example 2. We pick at random a person in a certain population. Let A be the event that the person selected attends college. Let B be the event that the person selected speaks french. Assume that the proportion of persons attending college and speaking french in the population are 0.1 and 0.02, respectively. Assume also that the proportion of people attending college and speaking french is 0.01. Given that the person we picked speaks french, what is the probability that this person attends college? We want

$$P(A|B) = \frac{P(AB)}{P(B)} = \frac{0.01}{0.02} = \frac{1}{2}.$$

Given that the selected person attends college, what is the probability that this person speaks french? This time we want

$$P(B|A) = \frac{P(AB)}{P(A)} = \frac{0.01}{0.1} = 0.1.$$

Given that the selected person attends college, what is the probability that this person does not speak french?

$$P(B^c|A) = 1 - P(B|A) = 1 - 0.1 = 0.9.$$

The previous two examples show how to compute conditional probabilities by using unconditional probabilities. In many situations, as we are going to see next, it is the reverse that is useful: the conditional probabilities are easy to compute and we use them to compute unconditional probabilities. Note first that the definition of conditional probability is equivalent to the following rule.

Multiplication Rule

$$P(AB) = P(A|B)P(B).$$

Example 3. A factory has an old (O) and a new (N) machine. The new machine produces 70% of the products and 1% of these products are defective. The old machine produces the remainder 30% of the products and of those 5% are defective. All products are randomly mixed. What is the probability that a product picked at random is defective and produced by the new machine?

Let D be the event that the product picked at random is defective. Note that the following probabilities are given.

$$P(N) = 0.7, \quad P(O) = 0.3, \quad P(D|N) = 0.01 \text{and} P(D|O) = 0.05.$$

We want the probability of DN. By the multiplication rule we have

$$P(DN) = P(D|N)P(N) = 0.01(0.7) = 0.007.$$

Assume now that we are interested in the probability that a product picked at random is defective. We can write

$$P(D) = P(DN) + P(DO).$$

That is, a defective product may come from the new or the old machine. Now we use the multiplication rule twice to get

$$P(D) = P(D|N)P(N) + P(D|O)P(O) = 0.01(0.7) + 0.05(0.3) = 0.022.$$

That is, we get the overall defective proportion by taking the weighted average of the defective proportions. This is a very useful way of proceeding and we now state the rule in its general form.

Rule of Average

For any events A and B we have

$$P(A) = P(A|B)P(B) + P(A|B^c)P(B^c).$$

More generally, if the events $B_1, B_2, \ldots, B_n$ are mutually exclusive and if their union is the whole sample space Ω then

$$P(A) = P(A|B_1)P(B_1)+P(A|B_2)P(B_2)+\cdots+P(A|B_n)P(B_n).$$

We now apply the rule of average to another example.

Example 4. We have three boxes labeled 1, 2, and 3. Box 1 has one white ball and two black balls, Box 2 has two white balls and one black ball, and Box 3 has three white balls. One of the three boxes is picked at random and then a ball is picked from this box. What is the probability that the ball picked is white?

Let W be the event "the ball picked is white." We use the rule of average and get

$$P(W) = P(W|1)P(1) + P(W|2)P(2) + P(W|3)P(3).$$

The conditional probabilities above are easy to compute. We have

$$P(W|1) = \frac{1}{3}, \quad P(W|2) = \frac{2}{3}, \quad P(W|3) = 1.$$

Thus,

$$P(W) = \frac{1}{3} \times \frac{1}{3} + \frac{2}{3} \times \frac{1}{3} + 1 \times \frac{1}{3} = \frac{2}{3}.$$

As we have just seen the conditional probability $P(W|1)$ is easy to compute. What about $P(1|W)$? That is, given that we picked a white ball what is the probability that it came from box 1?

In order to answer this question we start by using the definition of conditional probability.

$$P(1|W) = \frac{P(1W)}{P(W)}.$$

Now we use the multiplication rule for the numerator and the average rule for the denominator. We get

$$P(1|W) = \frac{P(W|1)P(1)}{P(W|1)P(1) + P(W|2)P(2) + P(W|3)P(3)}.$$

Numerically, we have

$$P(1|W) = \frac{1/3 \times 1/3}{2/3} = \frac{1}{6}.$$

Note that $P(1|W)$ is twice less likely than $P(1)$. That is, given that the ball drawn is white box 1 is less likely to have been picked than boxes 2 and 3. Since box 1 has less white balls than the other boxes this is not surprising. The preceding method applies each time we want the conditional probability $P(A|B)$ but what is readily available is the conditional probability $P(B|A)$. We now state the general form of this useful formula.

Bayes' Formula

For any events A and B we have

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

More generally, if the events $B_1, B_2, \ldots, B_n$ are disjoint and if their union is the whole sample space Ω then for every $i = 1, \ldots, n$

$$P(B_i|A) = \frac{P(A|B_i)P(B_i)}{P(A|B_1)P(B_1) + P(A|B_2)P(B_2) + \cdots + P(A|B_n)P(B_n)}.$$

As observed in Example 4, Bayes' formula is an easy consequence of the definition of conditional probabilities and the rule of average. Rather than memorizing it the reader should get familiar with the way to derive it. Next we give another example of use of the Bayes's rule.

Example 5. It is estimated that 10% of the population has a certain disease. A diagnostic test is available but is not perfect. There are two possible misdiagnoses. A healthy person may be misdiagnosed as sick with a probability of 5%. A person with the disease may be misdiagnosed as healthy with a probability of 1%. Given that a person picked at random is diagnosed with the disease, what is the probability that this person is actually sick?

Let D be the event that the person has the disease and + be the event that the person is diagnosed as having the disease. We are asked to compute the conditional

probability $P(D|+)$. Note that $P(+|D) = 1 - 0.01 = 0.99$ but $P(D|+)$ is not as readily available so we use Bayes' formula.

$$P(D|+) = \frac{P(D+)}{P(+)} = \frac{P(+|D)P(D)}{P(+|D)P(D) + P(+|D^c)P(D^c)}.$$

We know that $P(D) = 0.1$ so $P(D^c) = 0.9$. As observed before $P(+|D) = 0.99$ and $P(+|D^c) = 0.05$. Thus,

$$P(D|+) = \frac{0.99 \times 0.1}{0.99 \times 0.1 + 0.05 \times 0.9} = 0.69.$$

So given that the person has tested positive the probability that this person actually has the disease is only 0.69.

1.2.1 Symmetry

It is sometimes possible to avoid lengthy computations by invoking symmetry in a problem. We give next such an example.

Example 6. You are dealt two cards from a deck of 52 cards. What is the probability that the second card is black?

One way to answer the preceding question is to condition on whether the first card is black. Let B and R be the events "the first card is black" and the "first card is red," respectively. Let A be the event "the second card is black." We have

$$\begin{aligned} P(A) &= P(AR) + P(AB) = P(A|R)P(R) + P(A|B)P(B) \\ &= \left(\frac{26}{51}\right)\left(\frac{1}{2}\right) + \left(\frac{25}{51}\right)\left(\frac{1}{2}\right) = \frac{1}{2}. \end{aligned}$$

Now we show how a symmetry argument yields this result. By symmetry we have

$$P(\text{the second card is red}) = P(\text{the second card is black}).$$

Since

$$P(\text{the second card is red}) + P(\text{the second card is black}) = 1.$$

We get that

$$P(\text{the second card is black}) = \frac{1}{2}.$$

Exercises 1.2

1. Consider the student population in a college campus. Assume that 55% of the students are female. Assume that 20% of the male drink and 10% of the female drink.

(a) Pick a female student at random, what is the probability that she does not drink?
(b) Pick a student at random, what is the probability that the student does not drink?
(c) Pick a student at random, what is the probability that this student is male and drinks?

2. A company has two factories A and B. Assume that factory A produces 80% of the products and B the remaining 20%. The proportion of defectives are 0.05 for A and 0.01 for B.

(a) What is the probability that a product picked at random comes from A and is not defective?
(b) What is the probability that a product picked at random is defective?

3. Consider two boxes labeled 1 and 2. In box 1 there are two black balls and three white balls. In box 2 there are three black balls and two white balls. We pick box 1 with probability 1/3 and box 2 with probability 2/3. Then we draw a ball in the box we picked.

(a) Given that we pick box 2 what is the probability of drawing a white ball?
(b) Given that we draw a white ball what is the probability that we picked box 1?
(c) What is the probability of picking a black ball?

4. Consider an electronic circuit with components $C1$ and $C2$. The probability that $C1$ fails is 0.1. If $C1$ fails the probability that $C2$ fails is 0.15. If $C1$ works the probability that $C2$ fails is 0.05.

(a) What is the probability that both components fail?
(b) What is the probability that at least one component works?
(c) What is the probability that $C2$ works?

5. Suppose five cards are dealt from a deck of 52 cards.

(a) What is the probability that the second card is a queen?
(b) What is the probability that the fifth card is a heart?

6. Two cards are dealt from a deck of 52 cards. Given that the first card is red what is the probability that the second card is a heart?

7. A factory tests all its products. The proportion of defective items is 0.01. The probability that the test will catch a defective product is 0.95. The test will also reject nondefective products with probability 0.01.

(a) Given that a product passes the test, what is the probability that it is defective?
(b) Given that the product does not pass the test, what is the probability that the product is defective?

8. Consider the following game. There are three balls in a box, two are white and one is black. You win the game if you pick a white ball. You draw a ball but you do not see the color of the ball. Then someone takes out of the box a white ball. So at this point there is only one ball left in the box. At this point the rules of the game allow you to switch your ball with the one remaining in the box. What is the best strategy: to switch balls or not? In order to decide, compute the probability of winning for each strategy.

9. Two cards are randomly selected from a 52 cards deck. The two cards are said to form a blackjack if one of the cards is an ace and the other is either a ten, a jack, a queen, or a king. What is the probability that the two cards form a blackjack?

10. Two dice are rolled. Given that the sum is 9, what is the probability that at least one die showed 6?

11. Assume that 1% of men and 0.01% of women are color blind. A color blind person is chosen at random. What is the probability that this person is a man?

12. Hemophilia is a genetic disease that is caused by a recessive gene on the X chromosome. A woman is said to be a carrier of the disease if she has the hemophilia gene on one X chromosome and the healthy gene on the other X chromosome. A woman carrier has probability 1/2 of transmitting the disease to each son since a son will get an X chromosome from the mother and a Y chromosome from the father. Because of her family history a woman is thought to have a 50% chance of being a carrier before having children. Given that this woman has three healthy sons, what is the probability that she is a carrier?

1.3 Independent Events

We start with Example 4 of the preceding section.

Example 1. We have three boxes labeled 1, 2, and 3. Box 1 has one white ball and two black balls, Box 2 has two white balls and one black ball, and Box 3 has three white balls. One of the three boxes is picked at random and then a ball is picked from this box. Given that we draw a white ball what is the probability that we have picked box 1?

We have already computed this conditional probability and found it to be 1/6. On the other hand the (unconditional) probability of picking box 1 is 1/3. So the information that the ball drawn is white changes the probability of picking box 1. In this sense we say that the events $A = \{$ box 1 is picked $\}$ and $B = \{$ a white ball is drawn $\}$ are not independent. This leads to the following definition.

Independent Events

Two events A and B are said to be independent if

$$P(AB) = P(A)P(B).$$

We have the following consequences from this definition.

C1. Assume that $P(B) > 0$. By using the definition of conditional probability we see that if A and B are independent if and only if

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

C2. If A and B are independent so are A and B^c. In order to see this write that

$$P(A) = P(AB) + P(AB^c).$$

By using C1 we get

$$P(AB^c) = P(A) - P(A)P(B) = P(A)(1 - P(B)) = P(A)P(B^c)$$

and this shows that A and B^c are independent.

C3. If A and B are independent so are A^c and B^c.

Example 2. Consider again the three boxes of Example 1 but this time we put the same number of white balls in each box. For instance, assume that each box has two white balls and one black ball. Are the events $A = \{$ box 1 is picked $\}$ and $B = \{$ a white ball is drawn $\}$ independent?

By Bayes' formula we have

$$P(A|B) = \frac{1/3 \times 2/3}{1/3 \times 2/3 + 1/3 \times 2/3 + 1/3 \times 2/3} = \frac{1}{3} = P(A).$$

So this time A and B are independent. This should be intuitively clear: this time the fact the ball drawn is white does not yield additional information about which box was picked since all boxes have the same proportion of white balls.

Example 3. Assume that A and B are independent events such that $P(A) = 0.1$ and $P(B) = 0.3$. What is the probability that A or B occurs?

We want $P(A \cup B)$. Recall that

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

By C1 we have

$$P(A \cup B) = 0.1 + 0.3 - 0.1 \times 0.3 = 0.37.$$

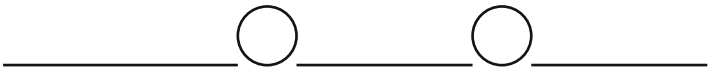
Example 4. Assume that A and B are independent, can they also be disjoint?

If A and B are disjoint then $AB = \emptyset$. Thus, $P(AB) = 0$. However, if A and B are also independent then

$$P(AB) = P(A)P(B) = 0.$$

Thus, $P(A) = 0$ or $P(B) = 0$. So if A and B are independent they may be disjoint if and only if one of these events has probability zero.

Example 5. Assume two components are in series as below.

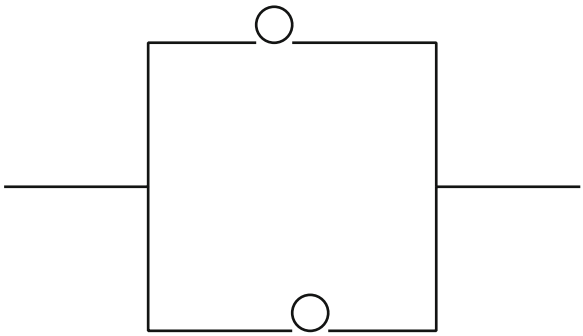


Assume that each component fails independently of the other with probability 0.01. What is the probability that the circuit fails?

In order for the circuit to fail we must have that one of the two components fails. Let A be the event that the left component fails and B be the event that the right component fails. So the probability of failure is

$$P(A \cup B) = P(A) + P(B) - P(AB) = 0.01 + 0.01 - 0.0001 = 0.0199.$$

Example 6. Assume two components are in parallel as below.



Assume they fail independently with probability 0.01. What is the probability that the circuit fails?

The circuit fails if both components fail.

$$P(AB) = P(A)P(B) = 0.0001.$$

As expected the reliability of a parallel circuit is superior to the reliability of a series circuit. However, it is the independence assumption that greatly increases the reliability. The independence assumption may or may not be realistic.

Exercises 1.3

1. Assume that A and B are independent events with $P(A) = 0.2$ and $P(B) = 0.5$.

(a) What is the probability that exactly one of the events A and B occurs?
(b) What is the probability that neither A nor B occurs?
(c) What is the probability that at least one of the events A or B occurs?

2. Two cards are successively dealt from a deck of 52 cards. Let A be the event "the first card is an ace" and B be event "the second card is a spade." Are these two events independent?

3. Two cards are successively dealt from a deck of 52 cards. Let A be the event "the first card is an ace" and B be event "the second card is an ace." Are these two events independent?

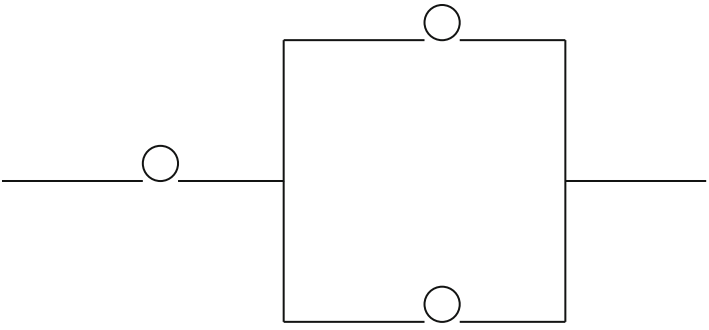
4. Roll two dice. Let A be the event "there is at least one 6" and B the event "the sum is 7." Are these two events independent?

5. Assume that the proportion of male students that drink is 0.2. Assume that there are 60% of male students and 40% of female students.

(a) Pick a student at random. What should the proportion of female drinkers be in order for the events "the student is male" and "the student drinks" be independent?
(b) Does your answer in (a) depend on the proportion of male students?

6. Show C3.

7. Assume that 3 components are as below.



Assume that each component fails independently of the others with probability p_i, for $i = 1, 2, 3$. Find the probability that the circuit fails in function of the p_i's.

8. (a) Roll one die 4 times. What is the probability of rolling at least one 6?
(b) Roll two dies 24 times. What is the probability of rolling at least one double 6?

9. Two cards are dealt from a 52 cards deck.

(a) What is the probability of getting a pair?
(b) What is the probability of getting two cards of the same suit?

1.4 Three or More Events

In this section we deal with probabilities involving several events. Our main tool is a generalization of the multiplication rule of Sect. 1.3. We now derive it for three events A, B, and C. We start by using the multiplication rule for the two events AB and C.

$$P(ABC) = P(C \cap (AB)) = P(C|AB)P(AB).$$

By the same multiplication rule

$$P(AB) = P(B|A)P(A).$$

Hence,

$$P(ABC) = P(C|AB)P(B|A)P(A) = P(A)P(B|A)P(C|AB).$$

The same computation can be done for an arbitrary number of events and yields the following.

Multiplication Rule for Three or More Events

Consider n events $A_1, A_2, \ldots, A_n$. The probability of the intersection of these n events can be written by using the following conditional probabilities.

$$P(A_1A_2 \ldots A_n) = P(A_1)P(A_2|A_1)P(A_3|A_1A_2) \ldots P(A_n|A_1A_2 \ldots A_{n-1}).$$

We now apply this formula to several examples.

Example 1. Deal four cards from a deck of 52 cards. What is the probability to get four aces?

Let A_1 be the event that the first card is an ace, let A_2 be the event that the second card is an ace and so on. We want to compute the probability of $A_1A_2A_3A_4$. We use the multiplication rule above to get.

$$P(A_1A_2A_3A_4) = P(A_1)P(A_2|A_1)P(A_3|A_1A_2)P(A_4|A_1A_2A_3).$$

The probability of A_1 is 4/52. Given that the first card is an ace the probability that the second card is an ace is 3/51 and so on. Thus,

$$P(A_1A_2A_3A_4) = \frac{4}{52} \times \frac{3}{51} \times \frac{2}{50} \times \frac{1}{49} = \frac{24}{6,497,400}.$$

A pretty slim chance to get four aces!

Example 2. We now deal with the famous birthday problem. Assume that there are 50 students in a class. What is the probability that at least two students have the same birthday?

It is easier to deal with the complement of this event. That is, we are going to compute the probability that all 50 students were born on different days. Assume that we are going through a list of the 50 birthdays in the class. Let B_2 be the event that the second birthday in the list is different from the first. Let B_3 be event that the third birthday on the list is different from the first two. More generally, let B_i be the event that the ith birthday on the list is different from the first $i-1$ birthdays for $i = 2, 3, \ldots, 50$. We want to compute the probability of $B_2B_3 \ldots B_{50}$. By the multiplication rule we have

$$P(B_2B_3 \ldots B_{50}) = P(B_2)P(B_3|B_2)P(B_4|B_2B_3) \ldots P(B_{50}|B_2B_3 \ldots B_{49}).$$

Ignoring the leap years, we assume that there are 365 days in a year. We also assume that all days are equally likely for birthdays. Note that $P(B_2) = 364/365$. Given that the first two birthdays are distinct the third birthday has only 363 choices in order to be distinct from the first two. So $P(B_3|B_2) = 363/365$. The same reasoning shows that $P(B_4|B_3B_2) = 362/365$. By doing the same type of computation for every term in the product above we get

$$P(B_2B_3 \ldots B_{50}) = \frac{364}{365} \times \frac{363}{365} \times \frac{362}{365} \times \cdots \times \frac{316}{365}.$$

The numerical computation gives a value of 0.96 for the probability of having at least two students having the same birthday in a class of 50! More generally, we have that

$$P(n \text{ people have } n \text{ distinct birthdays}) = \frac{364 \times 363 \times \cdots \times (365 - n + 1)}{365^{n-1}}.$$

The product above decreases rapidly to 0. If $n = 23$ we get that this product is about 0.50. For $n = 45$ it is about 0.05. Exercise 10 below will show how to approximate the product above by an exponential function. Note that if there are 365 people or more then the probability of having 365 or more distinct birthdays is zero.

1.4.1 Independence

We now consider the independence property for several events. We have the following definition.

Independent Events

Three events A, B and C are said to be independent if the following conditions hold

$$P(ABC) = P(A)P(B)P(C)$$

$$P(AB) = P(A)P(B), \quad P(AC) = P(A)P(C) \text{ and } P(BC) = P(B)P(C).$$

In general, n events are independent if for every integer k such that $2 \leq k \leq n$ and any choice of k events (among the n we are considering) the probability of the intersection of these k events is the product the probabilities of the k events.

The number of conditions to be checked grows rapidly with the number of events. It will be in general difficult to check that more than three events are independent. Typically, we will *assume* that events are independent (if that seems like a reasonable hypothesis) and then use the multiplication rule above to compute probabilities of interest. We illustrate this point next.

Example 3. Consider a class of 50 students. What is the probability that at least one of the students was born on December 25?

This is yet another case where it is easier to look at the complement of the event. We look at the list of birthdays in the class. Let A_i be the event that the ith student in the list was not born on December 25, for $1 \leq i \leq 50$. It is reasonable to assume that the A_i are independent: to know whether or not a certain student was born on December 25 does not give us additional information about the birthdays of other students (unless there are twins in the class and we assume that is not the case...). By the independence assumption we have

$$P(A_1 A_2 \ldots A_{50}) = P(A_1)P(A_2)\ldots P(A_{50}).$$

Note that each A_i has probability 364/365. Thus,

$$P(A_1 A_2 \ldots A_{50}) = \left(\frac{364}{365}\right)^{50} = 0.87.$$

That is, the probability that at least one student in a class of 50 was born on a certain fixed day is about 0.13. The reader should compare this value with the value in Example 2.

Example 4. How many students should we have in a class in order to have at least one birthday on December 25 with probability at least 0.5?

Let n be the minimum number of students that satisfies the condition above. We use the events A_i, for $1 \leq i \leq n$, defined in Example 3. We want

$$P(A_1 A_2 \dots A_n) \leq 0.5.$$

By independence we have that

$$\left(\frac{364}{365}\right)^n \leq 0.5.$$

We take logarithms on both sides of the inequality to get

$$n \ln\left(\frac{364}{365}\right) \leq \ln(0.5).$$

Recall that $\ln x < 0$ if $x < 1$. Thus,

$$n \geq \frac{\ln(0.5)}{\ln(364/365)}.$$

Numerically we get that n needs to be at least 253.

Exercises 1.4

1. Assume that three friends are randomly assigned to five classes. What is the probability that they are all in distinct classes?

2. Five cards are dealt from a 52 cards deck.

(a) What is the probability that the five cards are all hearts?
(b) What is the probability of a flush (all cards of the same suit)?

3. Roll 5 fair dice. What is the probability that at least two dice show the same face?

4. What is the probability of getting at least one 6 in 10 rolls of a fair die?

5. Assume that the chance to win at the lottery with one ticket is 1/1,000,000. Assume that you buy one ticket per week. How many weeks should you play to have at least 0.5 probability of winning at least once?

6. Three electric components are in parallel. Each component fails independently of the others with probability p_i, $i = 1, 2, 3$. What is the probability that the circuit fails?

7. Three electric components are in series. Each component fails independently of the others with probability p_i, $i = 1, 2, 3$. What is the probability that the circuit fails?

8. Roll a die 4 times.

(a) What is the probability of getting 4 times the same face?
(b) What is the probability of getting 3 times the same face?

9. The probability of winning a certain game is $1/N$ for some fixed N. Show that you need to play the game approximately $\frac{2}{3}N$ times in order for the probability to win at least once be 0.5 or more. (Use that $\ln 2$ is approximately 2/3 and that $\ln(1-1/N)$ is approximately $-1/N$ for N large.)

10. In this exercise we are going to derive an approximate formula for the birthday problem (Example 2). Our starting point is that

$$p_n = P(n\text{people have } n \text{ distinct birthdays}) = \frac{364 \times 363 \times \cdots \times (365-n+1)}{365^{n-1}}.$$

(a) Show that $\ln(p_n) = \ln(1-1/365) + \ln(1-2/365) + \cdots + \ln(1-(n-1)/365)$.
(b) Use that $\ln(1-x)$ is approximately $-x$ for x near zero to show that $\ln(p_n)$ is approximately $-1/365 - 2/365 \cdots - (n-1)/365$.
(c) Show that $1+2+3+\cdots+n = n(n+1)/2$.
(d) Use (c) in (b) to show that $\ln(p_n)$ is approximately $\frac{-n(n-1)}{2\times 365}$.
(e) Show that p_n is approximately

$$e^{\frac{-n(n-1)}{2\times 365}}.$$

(f) Compute p_n for $n = 5, 10, 20, 30, 40, 50$ by using the exact formula and the approximation.

11. Take four persons at random. What is the probability that they are all born on different months?

Review Exercises for Chap. 1

1. Assume that $P(A) = 0.1$ and $P(AB) = 0.05$.

(a) What is the probability of A occurs and B does not occur?
(b) What is the probability that A or B do not occur?

2. I draw one card from a deck of 52 cards. Let A be the event "I draw a king" and let B be the event "I draw a heart." Are A and B independent?

3. Roll three dice. What is the probability of getting at least one 6?

4. I draw five cards from a deck of 52 cards.

(a) What is the probability that I get four kings?
(b) What is the probability that I get 4 of a kind?

5. (a) I roll a die until the first 6 appears. What is the probability that I need 6 or more rolls?
(b) How many times should I roll the die so that I get at least one 6 with probability at least 0.9?

6. I draw five cards from a deck of 52 cards.

(a) What is the probability that I get no spade.
(b) What is the probability that I get no black cards?

7. I draw cards from a deck until I get a spade.

(a) What is the probability that I need exactly seven draws?
(b) Given that six or more draws are required, what is the probability that exactly seven draws are required?

8. Box 1 contains two red balls and three black balls. Box 2 contains six red balls and b black balls. We pick one of the two boxes at random and draw a ball from that box. Find b so that the color of the ball is independent of which box is picked.

9. 0's and 1's are sent down a communication channel. Assume that $P(\text{receive } 0|\text{transmit } 0)=P(\text{receive } 1|\text{transmit } 1)=0.99$. Assuming that 0's and 1's are equally likely, what is the probability of a transmission error?

10. A student goes to class on a snowy day with probability 0.5 and on a nonsnowy day with probability 0.8. Assume that 10% of the days in January are snowy. Given that the student was in class on January 28, what is the probability that it snowed that day?

11. One die is biased and the probability of a 6 is 1/2. The other die is fair. You pick one die at random and roll it. Given that you got a 6, what is the probability that you picked the biased die?

12. Consider a placement test for Calculus. Assume that 80% of the students pass the placement test and that 70% of the students pass Calculus. Experience has shown that given that a student has failed the placement test there is a 90% probability that the student will fail Calculus. Pick a student at random. Let A be the event "the student passes the placement test," let B be event "the student passes Calculus."

(a) Show that
$$P(AB) = P(A) - P(AB^c).$$
(b) Use (a) to compute $P(AB)$.
(c) Given that a student passed the placement test what is the probability that the student will pass Calculus?

13. Consider a slot machine with three wheels, each marked with 20 symbols. On the central wheel, nine of 20 symbols are bells, on the left and right wheels there is

one bell. In order to win the jackpot one has to get three bells. Assume that the three wheels spin independently and that every symbol is equally likely.

(a) What is the probability of hitting the jackpot?
(b) What is the probability of getting exactly two bells?
(c) Can you think of another distribution of bells that does not change the probability of hitting the jackpot but decreases the probability of getting exactly two bells?

14. Assume that A, B, and C are independent events with probabilities 1/10, 1/5, and 1/2, respectively.

(a) Compute $P(ABC)$.
(b) Compute $P(A \cup B \cup C)$.
(c) What is the probability that exactly one of A, B, or C occurs?

15. A tosses one coin and B tosses two coins. The winner is the player who gets the most heads. In case of an equal number of heads A wins.

(a) Compute the probability that B wins given that A gets 0 heads.
(b) Compute the probability that B wins given that A gets 1 heads.
(c) Compute the probability that B wins.
(d) Change the game so that A tosses 2 coins and B tosses 3 coins. The winner is still the player who gets the most heads. In case of an equal number of heads A wins. Compute the probability that B wins in the new game.

16. A rolls one die and B rolls two dice. The winner is the player who gets the most 6's. In case of an equal number of 6's A wins. What is the probability that A wins?

Chapter 2
Random Variables

2.1 Discrete Random Variables

We start with an example.

Example 1. Toss two fair coins. Let X be the number of heads. X is a function from the sample space $\Omega = \{HH, HT, TH, TT\}$ into the set $\{0, 1, 2\}$. The *distribution* of X is given by the following table.

k	0	1	2
$P(X = k)$	1/4	1/2	1/4

More generally, we have the following definition.

Discrete Random Variables

A discrete random variable is a function from a sample space Ω into a countable set (usually the positive integers). The distribution of a random variable X is the sequence of probabilities $P(X = k)$ for all k in the range of X. We must have

$$P(X = k) \geq 0 \text{ for every } k \text{ and } \sum_k P(X = k) = 1.$$

The term *discrete* refers to the fact that the random variables, in this section, take values in countable sets. Next section deals with *continuous* random variables: random variables whose range include intervals of the real numbers. We now give several examples of important discrete random variables.

R.B. Schinazi, *Probability with Statistical Applications*,
DOI 10.1007/978-0-8176-8250-7_2,

2.1.1 Bernoulli Random Variables

These are the simplest possible random variables. Perform a random experiment with two possible outcomes: success or failure. Set $X = 1$ if the experiment is a success and $X = 0$ if the experiment is a failure. Such a 0–1 random variable is called a Bernoulli random variable. The usual notation is $P(X = 1) = p$ and $P(X = 0) = q = 1 - p$.

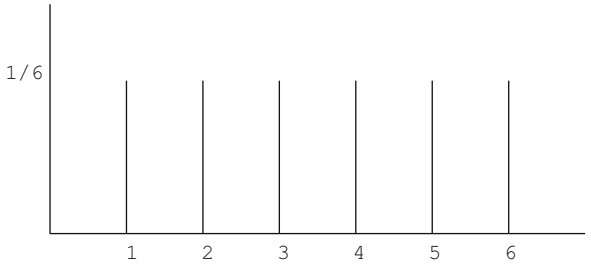
Example 2. Roll a fair die. We say that we have a success if we roll a 6. Thus, the probability of success is $P(X = 1) = 1/6$. We have $p = 1/6$ and $q = 5/6$.

2.1.2 Discrete Uniform Random Variables

Example 3. Roll a fair die. Let X be the face shown. The distribution of X is given by the following table.

k	1	2	3	4	5	6
$P(X = k)$	1/6	1/6	1/6	1/6	1/6	1/6

Below we graph this distribution.



This is called a uniform random variable. *Uniform* refers to the fact that all possible values of X are equally likely.

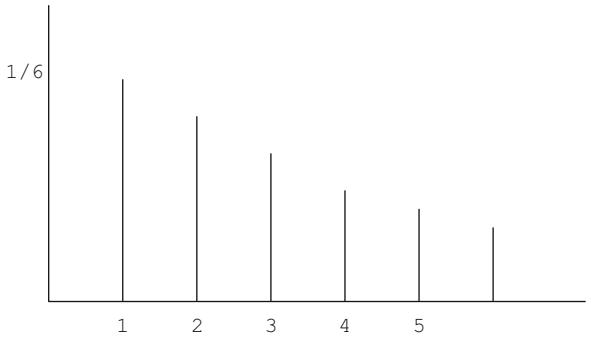
2.1.3 Geometric Random Variable

Example 4. Roll a fair die until you get a 6. Let X be the number of rolls to get the first 6. The possible values of X are all strictly positive integers. Note that $X = 1$ if and only if the first roll is a 6. So $P(X = 1) = 1/6$. In order to have $X = 2$ the first roll must be anything but 6 and the second one must be 6. By independence of the different rolls we get $P(X = 2) = 5/6 \times 1/6$. More generally, in order to have

$X = k$ the first $k-1$ rolls cannot yield any 6 and the kth roll must be a 6. Thus,

$$P(X = k) = \left(\frac{5}{6}\right)^{k-1} \times \frac{1}{6} \text{ for all } k \geq 1.$$

Next, we graph this distribution



Such a random variable is called geometric. More generally, we have the following.

Geometric Random Variables

Consider a sequence of independent identical trials. Assume that each trial can result in a success or a failure. Each trial has a probability p of success and $q = 1-p$ of failure. Let X be the number of trials up to and including the first success. Then X is called a geometric random variable. The distribution of X is given by

$$P(X = k) = q^{k-1} p \text{ for all } k \geq 1.$$

Note that a geometric random variable may be arbitrarily large since the above probabilities are never 0. In order to check that the sum of these probabilities is 1 we need the following fact about geometric series:

Geometric Series

$$\sum_{k \geq 0} x^k = \frac{1}{1-x} \text{ for all } x \in (-1, 1).$$

We have that

$$\sum_{k\geq 1} P(X=k) = \sum_{k\geq 1} q^{k-1}p = p\sum_{k\geq 0} q^k = \frac{p}{1-q} = 1.$$

Example 5. Toss a fair coin until you get tails. What is the probability that exactly three tosses were necessary?

In this example we have $p = q = 1/2$. So

$$P(X=3) = q^2 p = \frac{1}{8}.$$

What is the probability that three or more tosses were necessary?

Note that the event "three or more tosses are necessary" is the same as the event "the first two tosses are heads." Thus,

$$P(X \geq 3) = q^2 = \frac{1}{4}.$$

Example 6. Consider X a geometric random variable. What is the probability that X is strictly larger than r?

The event "$X > r$" is the same as the event "the first r trials are failures." Thus,

$$P(X > r) = q^r.$$

Example 7. Let X be a geometric random variable. Given that $X > r$ what is the probability that $X > r + s$?

We want

$$P(X > r+s | X > r) = \frac{P(\{X > r+s\} \cap \{X > r\})}{P(X > r)},$$

where the equality comes from the definition of a conditional probability. Note that the intersection $\{X > r+s\} \cap \{X > r\}$ is simply $\{X > r+s\}$. Thus,

$$P(X > r+s | X > r) = \frac{P(X > r+s)}{P(X > r)}.$$

By Example 6, we know that $P(X > r) = q^r$. So

$$P(X > r+s | X > r) = \frac{q^{r+s}}{q^r} = q^s = P(X > s).$$

That is, given that we had r failures the probability of getting an additional s failures is the same as getting s failures to start with. In this sense, the geometric distribution is said to have the memoryless property.

Example 8. Two players roll a die. If the die shows 6 then A wins if the die shows 1 or 2 then B wins. The die is rolled until A or B wins. What is the probability that A wins?

Let T be the number of times the die is rolled. Note that the events $\{T = n\}$ are disjoint. We have

$$P(A) = \sum_{n \geq 1} P(A \cap \{T = n\}).$$

The event "A wins in n rolls" is the same as the event "the first $n - 1$ rolls are draws and the nth roll is a 6." Note that the probability that a roll results in a draw is 3/6. Then

$$P(A \cap \{T = n\}) = \left(\frac{1}{2}\right)^{n-1} \times \frac{1}{6}.$$

Summing the geometric series we get

$$P(A) = \sum_{n \geq 1} \left(\frac{1}{2}\right)^{n-1} \times \frac{1}{6} = \frac{1}{3}.$$

Note that the probability that A wins is

$$P(A) = \frac{1}{3} = \frac{1/6}{1/6 + 2/6},$$

where 1/6 is the probability of A winning in 1 roll and 2/6 is the probability of B winning in 1 roll.

Exercises 2.1

1. Toss three fair coins. Let X be the number of heads.

(a) Find the distribution of X.
(b) Compute $P(X \geq 2)$.

2. Roll two dice. Let X be the sum of the faces. Find the distribution of X.

3. Recall that there are 38 pockets in a roulette and that 18 are red. I bet on red until I win. Let X be the number of bets I make.

(a) What is the probability that X is 2 or more?
(b) What is the probability that X is exactly 2?

4. I roll four dice. I win if I get at least one 6. What is the probability of winning?

5. Roll two fair dice. Let X be the largest of the two faces. What is the distribution of X?

6. I draw two cards from a deck of 52. Let X be the number of aces I draw. Find the distribution of X.

7. How many times should I toss a fair coin in order to get tails at least once with probability 90%?

8. In a lottery there are 100 tickets numbered from 1 to 100. Let X be the ticket drawn at random. What is the distribution of X?

9. I roll a die until I get a 6. Given that the first two rolls were not 6's, what is the probability I need 5 rolls or more in order to get a 6?

10. A and B roll a die. A wins if the die shows a 6 and B wins if the die shows a 1. The die is rolled until someone wins.

(a) What is the probability that A wins?
(b) What is the probability that B wins?
(c) Let T be the number of times the die is rolled. Find the distribution of T.

11. Let X be a discrete random variable.

(a) Show that

$$P(X = k) = P(X > k - 1) - P(X > k).$$

(b) Assume that for all $k \geq 1$ we have $P(X > k) = q^k$. Use (a) to show that X is a geometric random variable.

2.2 Continuous Random Variables

We start with the following definition.

Continuous Random Variables

A continuous random variable is a function from a sample space Ω to an interval of the real numbers. The distribution of a continuous random variable X is determined by its *density* function f as follows. For all $a < b$ we have that

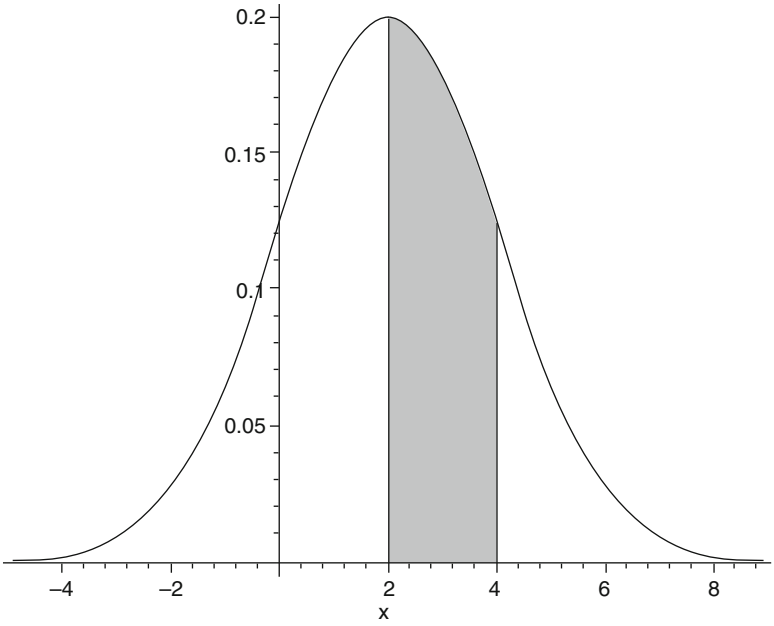
$$P(a < X < b) = \int_a^b f(x)\mathrm{d}x.$$

The function f is positive, continuous (except possibly at finitely many points) and

$$\int f(x)\mathrm{d}x = 1,$$

where the integral is taken on the largest interval on which f is strictly positive.

The shaded area below represents the probability that the random variable be between 2 and 4.



Note that for a continuous random variable X the following probabilities are all equal.

$$P(a \leq X < b) = P(a \leq X \leq b) = P(a < X \leq b) = P(a < X < b).$$

This is so because integrals of the type $\int_a^a f(x)\mathrm{d}x$ are always 0. In general, the above equalities do not hold for discrete random variables.

Example 1. Let X have density $f(x) = cx^2$ for x in $[-1,1]$ and $f(x) = 0$ elsewhere. Find c.

We must have

$$\int_{-1}^{1} cx^2 \mathrm{d}x = 1.$$

After integrating we get

$$c\left(\frac{2}{3}\right) = 1$$

and therefore $c = 3/2$.

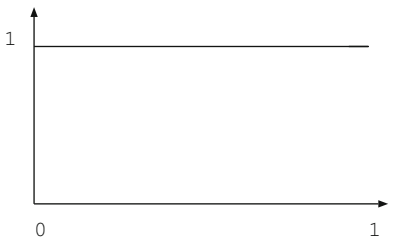
What is the probability that X is larger than 1/2?

$$P(X > 1/2) = \int_{1/2}^{1} f(x)\mathrm{d}x = \int_{1/2}^{1} \left(\frac{3}{2}\right) x^2 \mathrm{d}x = \frac{7}{16}.$$

We next give two examples of important continuous random variables.

2.2.1 Continuous Uniform Random Variables

Example 2. Let X be a random variable with density $f(x) = 1$ for x in $[0,1]$ and $f(x) = 0$ elsewhere. Since the density of X is flat on $[0,1]$, X is said to be uniform on $[0,1]$. Next we graph the density of X.



Note that

$$\int_0^1 f(x)\mathrm{d}x = \int_0^1 \mathrm{d}x = 1.$$

What is the probability of X to be in the interval (1/2, 3/4)?

We have that

$$P(1/2 < X < 3/4) = \int_{1/2}^{3/4} f(x)\mathrm{d}x = \frac{1}{4}.$$

What is the probability that X is larger than 1/2?

$$P(X > 1/2) = \int_{1/2}^{1} f(x)\mathrm{d}x = \frac{1}{2}.$$

More generally, we have the following.

Continuous Uniform Random Variables

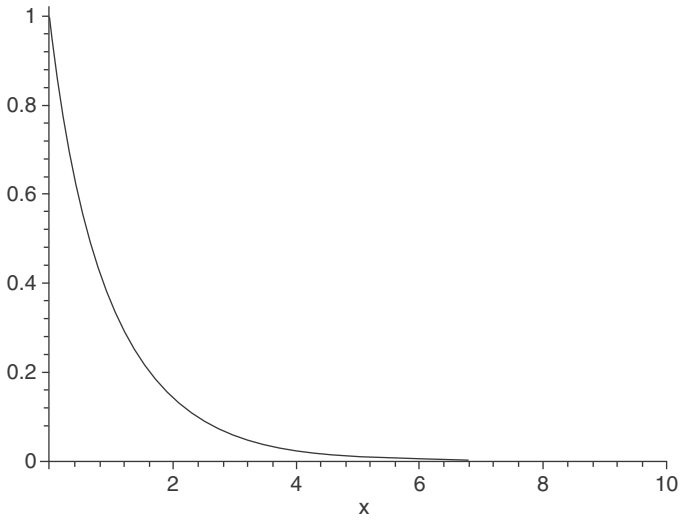
A continuous random variable X is uniform on the interval $[a, b]$ if the density of X is

$$f(x) = \frac{1}{b-a} \text{ for } x \in [a, b].$$

Note that the density of a uniform is always a constant on some interval and that the constant must be picked so that the area under the density is 1.

2.2.2 Exponential Random Variables

Example 3. Let T be a random variable with density $f(t) = \mathrm{e}^{-t}$ for $t \geq 0$. Below is the graph of f.



We first check that the area under the curve is 1.

$$\int_0^A \mathrm{e}^{-t}\mathrm{d}t = 1 - \mathrm{e}^{-A}.$$

By letting A go to infinity we get that the improper integral converges and that

$$\int_0^\infty \mathrm{e}^{-t}\mathrm{d}t = 1.$$

What is the probability that T is larger than 1?

$$P(T > 1) = \int_1^\infty \mathrm{e}^{-t}\mathrm{d}t = \mathrm{e}^{-1}.$$

What is the probability that T is less than 1?

$$P(T \leq 1) = 1 - P(T > 1) = 1 - \mathrm{e}^{-1}.$$

We next state the definition of an exponential random variable.

Exponential Random Variables

A random variable X with density $f(x) = a\mathrm{e}^{-ax}$ for $x \geq 0$ is said to be an exponential random variable with parameter (or rate) $a > 0$.

Example 4. Let T be an exponential random variable with parameter a. What is the probability that T is larger than s?

$$P(T > s) = \int_s^\infty a\mathrm{e}^{-at}\mathrm{d}t = \mathrm{e}^{-as}.$$

Example 5. Let T be an exponential random variable with parameter a. Given that T is larger than s, what is the probability that T is larger than $t + s$?

We want the conditional probability

$$P(T > t + s | T > s) = \frac{P(\{T > t + s\} \cap \{T > s\})}{P(T > s)}.$$

Note that the intersection of the events $T > t + s$ and $T > s$ is the event $T > t + s$. Thus,

$$P(T > t + s | T > s) = \frac{P(T > t + s)}{P(T > s)}.$$

By using the computation in Example 4 we get

$$P(T > t + s | T > s) = \frac{e^{-a(t+s)}}{e^{-as}} = e^{-at} = P(T > t).$$

So exactly as for the geometric distribution of the preceding section we say that the exponential distribution has the memoryless property.

Exercises 2.2

1. Let $f(x) = cx(1 - x)$ for x in [0,1] and $f(x) = 0$ elsewhere. Find c so that f is a density function.

2. Let the graph of the density f be a triangle for x in $[-1,1]$. Find f.

3. Let X be the density of an uniform random variable on $[-2,4]$. Find the density of X.

4. Let T be the waiting time for a bus. Assume that T has an exponential density with rate 3 per hour.

(a) What is the probability of waiting at least 20 min for the bus?
(b) Given that we have waited 20 min, what is the probability of waiting an additional 20 min for the bus?
(c) Under which conditions is the exponential model appropriate for this problem?

5. Let T be a waiting time for a bus. Assume that T has a uniform distribution on [0,40].

(a) What is the probability of waiting at least 20 min for the bus?
(b) Given that we have waited 20 min, what is the probability of waiting an additional 10 min for the bus?

6. Let Y have a density $g(y) = cye^{-2y}$ for $y \geq 0$. Find c.

7. Let X have density $f(x) = xe^{-x}$ for $x \geq 0$. What is the probability that X is larger than 3?

8. Let T have density $g(t) = 4t^3$ for t in [0,1].

(a) What is the probability that T is between 1/4 and 3/4?
(b) What is the probability that T is larger than 1/2?

9. (a) Show that for any random variable X we have

$$P(a < X < b) = P(X < b) - P(X \leq a).$$

(b) Assume that the random variable X is continuous and is such that $P(X > s) = e^{-2s}$. Use (a) to compute $P(a < X < b)$.
(c) Find the density of X.

2.3 Expectation

As we have seen in the preceding two sections knowing the distribution of a random variable entails knowing a lot of information. For a discrete random variable X the distribution is given by the sequence $P(X = k)$ for every k in the range of X. For a continuous random variable X the distribution is given by the density function f. For many problems it is enough to have a rough idea of the distribution and one tries to summarize the distribution by using a few numbers. The most important of these numbers is the *expectation* or the average value of the distribution. We first deal with discrete random variables.

Expectation of a Discrete Random Variable

The expectation (or mean) of the discrete random variable X is denoted by $E(X)$ and is given by

$$E(X) = \sum_k kP(X = k),$$

where the sum is taken over all the values in the range of X.

If a random variable may take infinitely many values then the computation of its expectation involves an infinite series. The expectation is defined only if the infinite series converges (see Exercise 18).

Note that the expectation of X is a measure of *location* of X.

Example 1. We perform an experiment with two possible outcomes: failure or success. If we have a success we set $X = 1$. If we have a failure we set $X = 0$. Let $P(X = 1) = p$. What is the expectation of this Bernoulli random variable?

$$E(X) = \sum_k kP(X = k) = 0 \times (1 - p) + 1 \times p = p.$$

Thus we can state,

Expectation of a Bernoulli Random Variable

Let X be a Bernoulli random variable with probability of success p. That is, X may take only values 0 and 1 and $P(X = 1) = p$. Then,

$$E(X) = p.$$

For instance, if we toss a fair coin and set $X = 1$ if we have heads and $X = 0$ if we get tails then $E(X) = 1/2$. What is the meaning of the value 1/2 since X can only take values 0 and 1?

The Law of Large Numbers that we will now (loosely) describe gives a physical meaning to the notion of expectation.

Law of Large Numbers

We make n independent and identical random experiments. Each experiment has a random outcome with the same distribution as the random variable X. The Law of Large Numbers states that as n goes to infinity the average over the n outcomes approaches $E(X)$.

We now come back to Example 1. The Law of Large Numbers states that if we toss a coin many times then the ratio of heads over the total number of tosses will approach 1/2. This gives a physical meaning to the expected value and also explains why this is a crucial notion.

Example 2. Roll a fair die. Let X be the face shown. We have $P(X = k) = 1/6$ for every $k = 1, 2, \ldots, 6$. Thus,

$$E(X) = \sum_k kP(X = k) = \sum_{k=1}^{6} \frac{k}{6} = \frac{7}{2}.$$

Example 3. The preceding example gave the expected value of a discrete uniform random variable in a particular case. We now treat the general case. Assume that X is a discrete uniform random variable on the set $\{1, 2, \ldots, n\}$. Thus, $P(X = k) = 1/n$ for $k = 1, 2, \ldots, n$. So

$$E(X) = \sum_{k=1}^{n} kP(X = k) = \frac{1}{n} \sum_{k=1}^{n} k.$$

Thus, we need to compute the sum of the first n integers. Let S_n be this sum and we write S_n in two different ways.

$$S_n = 1 + 2 + \cdots + (n-1) + n$$

$$S_n = n + (n-1) + \cdots + 2 + 1$$

We now add both equations to get

$$2S_n = (n+1) + (n+1) + \cdots + (n+1).$$

There are n terms equal to $n+1$ on the r.h.s. Thus,

$$2S_n = n(n+1)$$

and we get

$$S_n = \frac{n(n+1)}{2}.$$

Going back to the computation of the expected value we have:

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

Note that if we let $n = 6$ we get the particular case of Example 1.

Example 4. We now deal with geometric random variables. Let X be the number of independent and identical trials to get the first success. We denote by p the probability that a given trial be a success and $q = 1 - p$. The distribution of X is given by

$$P(X = k) = q^{k-1}p \text{for all} k = 1, 2, \ldots.$$

Thus,

$$E(X) = \sum_{k=1}^{\infty} kq^{k-1}p = p\sum_{k=1}^{\infty} kq^{k-1}.$$

Recall that

$$\sum_{k=0}^{\infty} x^k = \frac{1}{1-x} \text{for} x \in (-1, 1).$$

Recall also that power series are infinitely differentiable on their interval of convergence (except possibly at the boundary points). Thus, by taking derivatives on both sides of the preceding equality we get

$$\sum_{k=1}^{\infty} kx^{k-1} = \frac{1}{(1-x)^2} \text{for} x \in (-1, 1).$$

We plug $x = q$ and get for the expected value

$$E(X) = p\sum_{k=1}^{\infty} kq^{k-1} = p\frac{1}{(1-q)^2} = \frac{1}{p}.$$

Expectation of a Geometric Random Variable

Let X be the number of independent and identical trials up to and including the first success. We denote by p the probability that a given trial be a success and $q = 1 - p$. Then,

$$E(X) = \frac{1}{p}.$$

Example 5. Roll a die until you get a 6. What is the expected number of rolls?

Let T be the number rolls to get a 6. This is a geometric random variable with $p = 1/6$. Thus, $E(T) = 6$.

2.3.1 Continuous Random Variables

We start by defining the expected value for a continuous random variable.

Expectation of a Continuous Random Variable

Assume that X is a continuous random variable with density f. The expected value (or mean) of X is then

$$\int x f(x) \mathrm{d}x,$$

where the integral is taken on the largest interval on which f is strictly positive.

Example 6. Assume that X is uniformly distributed on $[a, b]$. What is its expected value?

Using that the density of X is $f(x) = \frac{1}{b-a}$ for x in $[a, b]$, we get

$$E(X) = \int_a^b x f(x) \mathrm{d}x = \frac{1}{b-a}\left(\frac{b^2}{2} - \frac{a^2}{2}\right) = \frac{a+b}{2}.$$

We get

Expectation of a Continuous Uniform Random Variable

Assume that X is uniformly distributed on $[a, b]$. Then

$$E(X) = \frac{a+b}{2}.$$

Example 7. Assume T is exponentially distributed with rate a. What is its expected value?

We integrate by parts to get

$$E(T) = \int_0^\infty t f(t) \mathrm{d}t = \int_0^\infty t a \mathrm{e}^{-at} \mathrm{d}t = -t\mathrm{e}^{-at}\Big]_0^\infty + \int_0^\infty \mathrm{e}^{-at} \mathrm{d}t = \frac{1}{a}.$$

Expectation of an Exponential Random Variable

Assume T is exponentially distributed with rate a. Then,

$$E(T) = \frac{1}{a}.$$

2.3.2 Other Measures of Location

To summarize the location of a distribution it is often a good idea to use more than one number. Besides the mean, there are two other important measures of location. The first one is the *median*.

Median of a Random Variable

A median m of a random variable X is a number m such that $P(X \leq m)$ and $P(X \geq m)$ are both at least 1/2.

As we will show in the exercises a median gives less weight to the extreme values of the distribution than the mean.

Example 8. Roll a die. Let X be the face shown. Note that $P(X \geq 3) = 2/3$ and $P(X \leq 3) = 1/2$. So 3 is a median. Observe that 4 is also a median and actually any number in [3,4] is a median. Recall that the mean in this case is 3.5. This example shows that a discrete random variable may have several medians.

Unlike what may happen for discrete random variables there is a unique median for continuous random variables. If the continuous variable X has density m then the median of X is such that

$$\int_m^\infty f(x) \mathrm{d}x = \int_{-\infty}^m f(x) \mathrm{d}x = \frac{1}{2}.$$

Example 9. Let T be an exponential random variable with rate 1. What is its median?

We solve the equation

$$P(T > m) = P(T \geq m) = \int_m^\infty \mathrm{e}^{-t}\mathrm{d}t = \mathrm{e}^{-m} = \frac{1}{2}.$$

Thus $m = \ln 2$. Note that $P(T < \ln 2) = 1 - P(T > \ln 2) = 1/2$. So $\ln 2$ is the unique median of this distribution.

An other measure of location, only defined for discrete random variables, is the *mode*.

Mode of a Discrete Random Variable

A mode M of a discrete random variable X is a number M such that $P(X = M)$ is maximum.

Example 10. For the uniform distribution on $\{1, 2, \ldots, 6\}$ there are 6 modes: $M = 1, 2, 3, 4, 5, 6$.

2.3.3 The Addition Rule

The following rule holds for any type (continuous or discrete) of random variables.

Addition Rule

Let X and Y be two random variables defined on the same sample space Ω. Then,

$$E(X + Y) = E(X) + E(Y).$$

More generally, if $X_1, X_2, \ldots, X_n$ are all defined on Ω we have

$$E(X_1 + X_2 + \cdots + X_n) = E(X_1) + E(X_2) + \cdots + E(X_n).$$

As the next examples will show this is a very important rule. Its proof involves the joint distribution of several random variables. We will prove this formula when we will see joint distributions in 8.3.

Example 11. I roll two dice. Let S be the sum of the two dice. What is the expected value of S?

Let X be the value of the first die and Y be the value of the second die. Then $S = X + Y$. According to the addition rule we have

$$E(S) = E(X) + E(Y).$$

But Example 1 tells us that $E(X) = E(Y) = 7/2$. Thus,

$$E(S) = 7.$$

We could have computed $E(S)$ by first computing the distribution of S and then averaging but this would have taken a lot longer.

2.3.4 Computing the Expectation By Breaking Up the Random Variable

In many cases the distribution of a given random variable is too involved to be computed. In some of those cases it is possible to break up a random variable into a sum of Bernoulli random variables. By using the addition rule we then get the mean of the random variable with the involved distribution without computing its distribution. We next give such an example.

Example 12. Assume that three people enter independently an elevator that goes to five floors. What is the expected number of stops S that the elevator is going to make?

Instead of computing the distribution of S we break S into a sum of 5 Bernoulli random variables as follows. Let $X_1 = 1$ if at least one person goes to floor 1, otherwise we set $X_1 = 0$. Likewise let $X_2 = 1$ if at least one person goes to floor 2, otherwise we set $X_2 = 0$. We do the same for the five possible choices. We have

$$S = X_1 + X_2 + \cdots + X_5.$$

Note that $X_1 = 0$ if none of the three people pick floor 1. Thus,

$$P(X_1 = 0) = \left(\frac{4}{5}\right)^3.$$

The probability of success for X_1 is $p = P(X_1 = 1) = 1 - (4/5)^3$. All the X_i have the same Bernoulli distribution. By the addition rule we have

$$E(S) = 5p = 5\left(1 - \left(\frac{4}{5}\right)^3\right) = \frac{61}{25} = 2.44.$$

We now compute $E(S)$ by using the distribution of S. The random variable S may only take values 1, 2, and 3. In order to have $S = 1$, the second and third person need to pick the same floor as the first person. Thus,

$$P(S = 1) = \left(\frac{1}{5}\right)^2.$$

To have $S = 2$, there are two possibilities: either the second person picks the same floor as the first one and the third a different floor (the probability of that is (1/5)(4/5)) or the second person picks a different floor from the first one and the third one picks one of the two floors that have already been picked (the probability of that is (4/5)(2/5)). Thus,

$$P(S = 2) = \left(\frac{1}{5}\right)\left(\frac{4}{5}\right) + \left(\frac{4}{5}\right)\left(\frac{2}{5}\right).$$

Finally, $S = 3$ happens only if the three persons pick distinct floors:

$$P(S = 3) = \left(\frac{4}{5}\right)\left(\frac{3}{5}\right).$$

Thus,

$$E(S) = 1 \times \frac{1}{25} + 2 \times \frac{12}{25} + 3 \times \frac{12}{25} = \frac{61}{25}.$$

So even in this very simple case (S has only three values after all) it is better to compute the expected value of S by breaking S in a sum of 0–1 random variables rather than compute the distribution of S.

Example 13. Let B be the number of distinct birthdays in a class of 50 students. What is the $E(B)$?

The distribution of B is clearly fairly involved. We are going to break B into a sum of Bernoulli random variables. Set $X_1 = 1$ if at least one student was born on January 1, otherwise set $X_1 = 0$. Set $X_2 = 1$ if at least one student was born on

January 2, otherwise set $X_2 = 0$. We define X_i like above for every one of the 365 days of the calendar. We claim that

$$B = X_1 + X_2 + \cdots + X_{365}.$$

This is so because the r.h.s. counts all the days on which at least one student has his birthday. Moreover, the X_i are Bernoulli random variables. In order for $X_1 = 0$ we must have that none of the 50 students was born on January 1. Thus,

$$P(X_1 = 0) = \left(\frac{364}{365}\right)^{50}.$$

The probability of success for X_1 is $p = 1 - \left(\frac{364}{365}\right)^{50}$. We do the same for every X_i and they all have the same p (which is also the expected value of a Bernoulli random variable). By the addition rule we have

$$E(B) = E(X_1) + E(X_2) + \cdots + E(X_{365}) = 365p = 365\left(1 - \left(\frac{364}{365}\right)^{50}\right).$$

Numerically, we get

$$E(B) = 46.79.$$

From Example 2 in Sect. 1.4 we know that the probability of having at least two students born on the same day is 0.96. However from the value of $E(B)$ we see that more than two students born on the same day or more than one set of students born on the same day are not that likely, otherwise $E(B)$ would be lower.

Example 14. The collector's problem. Assume that a certain brand of cereals has a cartoon character in each box. There are r different cartoon characters. What is the expected number of cereal boxes that need to be purchased in order to get all the cartoon characters?

Let T_1 be the number of boxes needed to get the first character. Obviously, $T_1 = 1$. Let T_2 be the number of boxes needed to get the second (different) character. Since we have already one character every time we buy a box there is a probability $\frac{1}{r}$ of getting the same character we already have and a probability $\frac{r-1}{r}$ to get a different one. Hence, T_2 is a geometric random variable with success probability $\frac{r-1}{r}$. More generally, let T_k be the number of boxes needed to get the kth different character given that we have already $k-1$ different characters. Since we have already $k-1$ characters every time we purchase a box the probability to get a kth different character is $\frac{r-(k-1)}{r}$. That is, T_k is a geometric random variable with probability of

success $\frac{r-k+1}{r}$ for $k = 2, \ldots, r$. The number of boxes needed to have a complete collection is therefore

$$T_1 + T_2 + \cdots + T_r.$$

Recall that the expected value of a geometric random variable with success probability p is $1/p$. Hence, the expected number of boxes needed to have the complete collection is:

$$E(T_1 + T_2 + \cdots + T_r) = 1 + \frac{r}{r-1} + \frac{r}{r-2} + \cdots + \frac{r}{2} + \frac{r}{1}.$$

It is convenient to rewrite the formula as

$$E(T_1 + T_2 + \cdots + T_r) = r\left(1 + \frac{1}{2} + \cdots + \frac{1}{r}\right).$$

As r goes to infinity one can show that

$$1 + \frac{1}{2} + \cdots + \frac{1}{r} \sim \ln r$$

in the sense that the ratio goes to 1. Hence, the expected number of boxes needed to complete the collection is approximately $r \ln r$.

2.3.5 *Fair Gambling*

Example 15. We roll a die. You pay me $\$b$ if the die shows 5 or 6. I pay you \$1 otherwise. Clearly, the probabilities of winning are not the same for both players. Can we pick b so that this is a fair game?

Assume we play this game many times. By the Law of Large Numbers my average winnings will be close to my expected winnings. We will say that the game is fair if the expected winnings (of each player) are 0. So that in the long run my average winnings will approach 0.

In this particular case let W be my winnings in 1 bet. We have that $W = b$ with probability 1/3 and $W = -1$ with probability 2/3. Thus,

$$E(W) = b \times \frac{1}{3} + (-1) \times \frac{2}{3}.$$

We want $E(W) = 0$. Solving for b we get $b = 2$. Since I am twice less likely to win than you are you should pay me twice as much when I win.

2.3.6 Expectation of a Function of a Random Variable

As we will see in the next section it is often necessary to compute $E(X^2)$ for a random variable X. This is NOT $E(X)^2$. We could compute the distribution of X^2 and use the distribution to compute the expected value. However, there is a quicker way to do things and it is contained in the following formula.

Expectation of a Function of X

Let X be a random variable and g be a real valued function. For instance, $g(x) = x^2$. Then if X is discrete we have

$$E(g(X)) = \sum_k g(k)P(X = k).$$

If X is continuous with density f then

$$E(g(X)) = \int g(x)f(x)\mathrm{d}x.$$

Example 16. Let X be a discrete random variable such that $P(X = -1) = 1/3$, $P(X = 0) = 1/2$, and $P(X = 2) = 1/6$. What is $E(X^2)$?

We use the formula above with $g(x) = x^2$ to get

$$E(X^2) = (-1)^2 \times \frac{1}{3} + 0^2 \times \frac{1}{2} + (2)^2 \times \frac{1}{6} = 1.$$

Example 17. Let X be uniformly distributed on [0,1]. What is $E(X^3)$?

This time we use the formula with $g(x) = x^3$. We get

$$E(X^3) = \int_0^1 x^3 f(x)\mathrm{d}x = \frac{1}{4}.$$

Another case which is of particular interest is when $g(x) = ax + b$. Assume that X is a discrete random variable. Then we use the formula above to get

$$E(aX + b) = \sum_k (ak + b)P(X = k) = a\sum_k kP(X = k)$$
$$+ b\sum_k P(X = k) = aE(X) + b.$$

The same formula may be derived for continuous random variables. We have the following for continuous and discrete random variables.

Expectation of a Linear Function of X

$$E(aX + b) = aE(X) + b.$$

The following observation gives the expectation without any computation provided we have a symmetric distribution.

Symmetric Case

Let f be the density of a continuous random variable X. Assume that there is $a \geq 0$ such that

$$f(a + x) = f(a - x)$$

for every x. Then, $E(X) = a$ (if $E(X)$ exists!).

We now show this property. Assume first that $a = 0$. That is,

$$f(x) = f(-x)$$

for every x. We have

$$E(X) = \int_{-b}^{+b} xf(x)\mathrm{d}x,$$

where b is a positive number or $+\infty$. Let $g(x) = xf(x)$, note that g is an odd function. That is,

$$g(-x) = -g(x)$$

for every x. Hence, $E(X)$ is the integral of an odd function on a symmetric interval. It is easy to see that this integral must be 0 and therefore $E(X) = 0$. We are done in the case $a = 0$.

If $a > 0$ let $Y = X - a$. One can check (this will be done in Sect. 8.1) that the density of Y is $f_Y(x) = f(a + x)$. This implies that

$$f_Y(x) = f(a + x) = f(a - x) = f_Y(-x).$$

That is, Y has a symmetry at 0 and therefore by the case $a = 0$ we know that $E(Y) = 0$. But

$$E(Y) = E(X) - a = 0.$$

Hence, $E(X) = a$ and we are done.

Exercises 2.3

1. What is the expected value of a random variable uniformly distributed on $\{-1,0,3\}$.

2. Toss two fair coins. What is the expected number of heads?

3. The probability of finding someone in favor of a certain initiative is 0.01. We interview people at random until we find a person in favor of the initiative. What is the expected number of interviews?

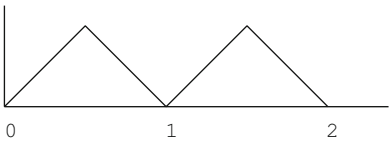
4. Roll two dice. What is the expected value of the maximum of the two dice?

5. Let X be exponentially distributed with mean 1/2. What is the density of X?

6. Let U be a random variable which is uniformly distributed on $[-1,2]$.

(a) Compute the mean of U.
(b) What is the median of U?

7. Let X have the following density.



(a) Find the expected value of X.
(b) How good is $E(X)$ as a measure of location of X?

8. Let $f(x) = 3x^2$ for x in $[0,1]$. Let X be a random variable with density f.

(a) What is $E(X)$?
(b) What is the median of X?

9. Let X be a random variable such that $P(X = 0) = 1/5$ and $P(X = 4) = 4/5$. Find the mean, medians, and modes.

10. Let T be exponentially distributed with rate a. Find the median of T in function of a.

11. Roll four dice. What is the expected value of the sum?

12. There are three components in a circuit. Each one of them fails with probability p. The failure of one component may influence the other components in a way that is not well understood. What is the expected number of working components?

13. Let B be the number of distinct birthdays in a class of 200 students. What is the $E(B)$?

14. There are eight people in a bus and five bus stops ahead. What is the expected number of stops the bus will have to make for these eight people?

15. I roll four dice. If there is at least one 6 you pay me \$1. If there are no 6's I pay you \$1.

(a) Is this a fair game?
(b) How would you make it into a fair game?

16. Let X be uniform on $\{1, 2, \ldots, 6\}$. What is $E(X^2)$?

17. Let X be exponentially distributed with rate 1. What is $E(X^2)$?

18. In this problem we give an example of a discrete random variable for which the expectation is not defined.

(a) Use the fact that

$$\sum_{k=1}^{\infty} \frac{1}{k^2} = \frac{\pi^2}{6}$$

to find c so that $P(X = k) = c/k^2$ is a probability distribution.
(b) Show that the expectation of the random variable defined above does not exist.

19. This problem gives an example of a continuous random variable that has no expectation.

(a) Show that $f(x) = \frac{2}{\pi(1+x^2)}$ for $x > 0$ is a density function.
(b) Show that a random variable with the density above has no expectation.

20. I roll a die repeatedly.

(a) What is the expected number of rolls to get three different faces?
(b) What is the expected number of rolls to get all six faces?

2.4 Variance

We have seen in Sect. 2.3 that the expectation is a measure of location for a distribution. Next we are going to define a measure of dispersion: the variance. A small variance will mean that the distribution is concentrated around the mean and that the mean is a good measure of location. A large variance will mean that the distribution is dispersed and that no value is really typical for this distribution.

Variance of a Random Variable

Let X be random variable with mean $E(X) = \mu$. The variance of X is denoted by $\text{Var}(X)$ and is defined by

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

The following formula for the variance is useful for computational purposes

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

Finally, the standard deviation of X is denoted by $SD(X)$ and is defined by

$$SD(X) = \sqrt{\text{Var}(X)}.$$

We now list the consequences of these definitions.

Consequences

C1. The variance of a random variable is ALWAYS positive or 0. This is so because the variance is the expected value of the positive random variable $(X - \mu)^2$.

C2. The variance of a random variable X is 0 if and only if X is a constant. For a discrete random variable this can be seen from the formula

$$E[(X - \mu)^2] = \sum_k (k - \mu)^2 P(X = k).$$

If this sum is 0 it means that every term must be 0 since these are all positive terms. But the sum of the $P(X = k)$ is 1 so at least some of these terms are nonzero. It is easy to see that for exactly one k $P(X = k)$ is not 0 and that corresponds to $k = \mu$. Thus, X is a constant equal to μ.

For a continuous random variable (that is a random variable whose density is strictly positive on some interval) one can show that the variance is always strictly positive.

C3. An easy consequence of the definition of variance is that

Properties of Variance

For any random variable X and constants a and b we have that

$$\text{Var}(aX + b) = a^2\text{Var}(X).$$

$$SD(aX + b) = |a|SD(X).$$

Observe that the translation by b has no effect on the variance of $aX + b$. Intuitively, this is clear since the variance measures the dispersion, not the location, of a random variable.

Example 1. We start with the Bernoulli distribution. Assume that X takes values 0 and 1. We denote $P(X = 1) = p$ and $P(X = 0) = 1 - p = q$. What is the variance of X?

We have that

$$E(X) = p.$$

We now compute

$$E(X^2) = 0^2 \times q + 1^2 \times p = p.$$

Thus,

$$\text{Var}(X) = E(X^2) - E(X)^2 = p - p^2 = pq.$$

Variance of a Bernoulli Random Variable

Assume that X takes values 0 and 1. We denote $P(X = 1) = p$ and $P(X = 0) = 1 - p = q$. Then,

$$\text{Var}(X) = pq.$$

Example 2. What is the variance of the discrete random variable uniformly distributed on $\{1, 2, 3, 4, 5, 6\}$?

We know that $E(X) = 7/2$.

We now compute

$$E(X^2) = 1^2 \times \frac{1}{6} + 2^2 \times \frac{1}{6} + \cdots + 6^2 \times \frac{1}{6} = \frac{91}{6}.$$

Thus,

$$\mathrm{Var}(X) = E(X^2) - E(X)^2 = \frac{91}{6} - \frac{49}{4} = \frac{35}{12}.$$

So the standard deviation is approximately 1.7. It is large for a distribution on $\{1, \ldots, 6\}$. But this is not surprising since the extreme values have the same weight as the middle values for this distribution.

Example 3. We now turn to the variance of a geometric random variable. We have independent identical trials that have a probability p of success. Let T be the number of trials to get the first success. The random variable T has a geometric distribution and we know that

$$E(T) = \frac{1}{p}.$$

As before we need to compute $E(T^2)$. In this case it is easier to compute $E(T(T-1))$ first. We need a new fact about geometric series. Recall that for every x in $(-1,1)$ we have

$$\sum_{k=0}^{\infty} x^k = \frac{1}{1-x}.$$

Power series are infinitely differentiable on their interval of convergence. We take derivatives twice in the formula above to get:

$$\sum_{k=2}^{\infty} k(k-1)x^{k-2} = \frac{2}{(1-x)^3}. \tag{2.2}$$

Now we compute

$$E(T(T-1)) = \sum_{k=1}^{\infty} k(k-1)P(T=k)$$

$$= \sum_{k=2}^{\infty} k(k-1)q^{k-1}p = pq\sum_{k=2}^{\infty} k(k-1)q^{k-2}.$$

We let $x = q$ in (2.2) to get

$$E(T(T-1)) = \frac{2pq}{(1-q)^3} = \frac{2q}{p^2}.$$

We have that

$$E(T^2) = E(T(T-1)) + E(T) = \frac{2q}{p^2} + \frac{1}{p}.$$

Finally,

$$\mathrm{Var}(T) = E(T^2) - E(T)^2 = \frac{2q}{p^2} + \frac{1}{p} - \frac{1}{p^2} = \frac{2q + p - 1}{p^2}.$$

Note that $p + q = 1$, so $2q + p - 1 = q$. Hence,

$$\mathrm{Var}(T) = \frac{q}{p^2}.$$

Variance of a Geometric Random Variable

Assume that we have independent identical trials that have a probability p of success. Let T be the number of trials to get the first success. Then,

$$\mathrm{Var}(T) = \frac{q}{p^2}.$$

We now compute variances for a few continuous random variables.

Example 4. Assume that X is uniformly distributed on $[a, b]$. Then

$$E(X) = \frac{a + b}{2}.$$

We compute $E(X^2)$.

$$E(X^2) = \int_a^b x^2 f(x)\mathrm{d}x = \frac{1}{b-a}\int_a^b x^2\mathrm{d}x = \frac{1}{3(b-a)}(b^3 - a^3) = \frac{b^2 + ab + a^2}{3}.$$

Thus,

$$\mathrm{Var}(X) = E(X^2) - E(X)^2 = \frac{b^2 - 2ab + a^2}{12} = \frac{(b-a)^2}{12}.$$

Variance of a Continuous Uniform Random Variable

Assume that X is uniformly distributed on $[a, b]$. Then

$$\mathrm{Var}(X) = \frac{(b-a)^2}{12}.$$

We now deal with exponential random variables.

Example 5. Assume that T is exponentially distributed with rate a. Then, $E(T) = 1/a$. We have

$$E(T^2) = \int_0^\infty t^2 f(t)\mathrm{d}t = \int_0^\infty t^2 a\mathrm{e}^{-at}\mathrm{d}t.$$

We do an integration by parts to get

$$E(T^2) = -t^2\mathrm{e}^{-at}\Big]_0^\infty + \int_0^\infty 2t\mathrm{e}^{-at}\mathrm{d}t = \frac{2}{a}\int_0^\infty ta\mathrm{e}^{-at}\mathrm{d}t = \frac{2}{a^2},$$

where we have used that $E(T) = 1/a$ to get the last equality. So

$$\mathrm{Var}(T) = E(T^2) - E(T)^2 = \frac{2}{a^2} - \frac{1}{a^2} = \frac{1}{a^2}.$$

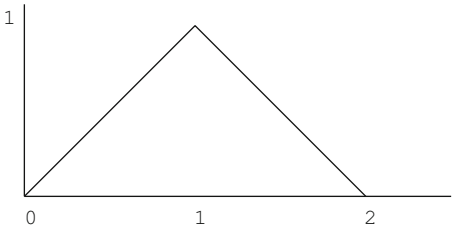
That is, the mean and the standard deviation are equal for an exponential distribution. This shows that exponential distributions are rather dispersed.

Variance of an Exponential Random Variable

Assume that T is exponentially distributed with rate a. Then,

$$\mathrm{Var}(T) = \frac{1}{a^2}.$$

Example 6. Consider Y with the following density.



What is the Var(Y)?

The density of Y is $f(y) = y$ for y in [0,1] and $f(y) = 2 - y$ for y in [1,2]. The mean of Y is 1 because of the symmetry of the density. We confirm this by computation.

$$E(Y) = \int_0^2 yf(y)\mathrm{d}y = \int_0^1 y^2\mathrm{d}y + \int_1^2 y(2-y)\mathrm{d}y.$$

Thus,

$$E(Y) = y^3/3\big]_0^1 + y^2\big]_1^2 - y^3/3\big]_1^2 = 1.$$

We now deal with $E(Y^2)$.

$$E(Y^2) = \int_0^2 y^2 f(y)\mathrm{d}y = \int_0^1 y^3\mathrm{d}y + \int_1^2 y^2(2-y)\mathrm{d}y.$$

So

$$E(Y^2) = y^4/4\big]_0^1 + 2y^3/3\big]_1^2 - y^4/4\big]_1^2 = \frac{7}{6}.$$

$$\mathrm{Var}(Y) = E(Y^2) - E(Y)^2 = \frac{7}{6} - 1 = \frac{1}{6}.$$

2.4.1 Independent Random Variables

We will need to compute the variance of sums of random variables. This turns out to a simple task only when the random variables in the sum are independent. We start by defining independence for random variables.

Independent Random Variables

Two discrete random variables X and Y are said to be independent if

$$P(\{X = x\} \cap \{Y = y\}) = P(X = x)P(Y = y) \text{for ALL} x, y.$$

Two continuous random variables X and Y are said to be independent if for ALL real numbers $a < b, c < d$ we have

$$P(\{a < X < b\} \cap \{c < Y < d\}) = P(a < X < b)P(c < Y < d).$$

We now examine two examples.

Example 7. Roll two dice. Let X be the face shown by the first die and S be the sum of the two dice. Are X and S independent?

Intuitively it is clear that the answer should be no. It is enough to find one x and one y such that

$$P(\{X = x\} \cap \{S = y\}) \neq P(X = x)P(S = y)$$

in order to show that X and S are not independent. For instance, take $x = 1$ and $y = 12$. Clearly if one die shows 1 the sum cannot be 12. So $P(\{X = 1\} \cap \{S = 12\}) = 0$. However, $P(X = 1)$ and $P(S = 12)$ are strictly positive so $P(\{X = 1\} \cap \{S = 12\}) \neq P(X = 1)P(S = 12)$. X and S are not independent.

Example 8. Toss two fair coins. Set $X = 1$ if the first coin shows heads, set $X = 0$ otherwise. Set $Y = 1$ if the second coin shows heads, set $Y = 0$ otherwise. Are X and Y independent?

Our sample space is $\Omega = \{(H, H), (H, T), (T, H), (T, T)\}$. We need to examine the 4 possible outcomes for (X, Y). Note that the event $\{X = 0\} \cap \{Y = 0\}$ is the event $\{(T, T)\}$ and that has probability 1/4. Note that $P(X = 0) = 2/4 = P(Y = 0)$. So the product rule holds for $x = 0$ and $y = 0$. We now examine $x = 0$ and $y = 1$. The event $\{X = 0\} \cap \{Y = 1\}$ is the event $\{(T, H)\}$. This has probability 1/4. Since $P(Y = 1) = 2/4$ the product rule holds in this case as well. The two remaining cases are symmetric to the cases we just examined. We may conclude that X and Y are independent.

2.4.2 Variance of a Sum of Random Variables

If X and Y are independent it is easy to compute the variance of $X + Y$.

Variance of a Sum of Independent Random Variables

Assume that X and Y are two INDEPENDENT random variables defined on the same sample space Ω. Then,

$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y).$$

More generally, if $X_1, X_2, \ldots, X_n$ are independent random variables then

$$\text{Var}(X_1, X_2 + \cdots + X_n) = \text{Var}(X_1) + \text{Var}(X_2) + \cdots + \text{Var}(X_n).$$

Example 9. Roll 2 dice. Let S be the sum of the two dice. What is the variance of S?

Let X and Y be the faces shown by each die. From Example 2 we know that $\mathrm{Var}(X) = \mathrm{Var}(Y) = 35/12$. Since X and Y are independent we get that

$$\mathrm{Var}(S) = \mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) = 2 \times \frac{35}{12} = \frac{35}{6}.$$

Example 10. Assume that X and Y are independent random variables with

$$\mathrm{Var}(X) = 2 \qquad \mathrm{Var}(Y) = 3.$$

What is the variance of $2X - 3Y$?

From the definition of independence it is to see that if X and Y are independent so are $2X$ and $-3Y$. Thus,

$$\mathrm{Var}(2X - 3Y) = \mathrm{Var}(2X) + \mathrm{Var}(-3Y) = 4\mathrm{Var}(X) + 9\mathrm{Var}(Y) = 35.$$

Exercises 2.4

1. What is the variance of a random variable uniformly distributed on $\{-1, 0, 3\}$?

2. Let X be a random variable such that $P(X = 0) = 1/5$ and $P(X = 4) = 4/5$. Find the variance of X.

3. The probability of finding someone in favor of a certain initiative is 0.01. We interview people at random until we find a person in favor of the initiative. What is the standard deviation of the number of interviews?

4. Roll two dice.

(a) What is the variance of the maximum of the two dice?
(b) Compare the result of (a) to the variance of a single roll obtained in Example 2.

5. Let X have density $f(x) = x^2 \mathrm{e}^{-x}/2$. What is the variance of X?

6. Let U be a random variable which is uniformly distributed on [−1,2]. What is the variance of U?

7. Consider the random variables X and Y with densities $f(x) = \frac{3}{2}x^2$ for x in $[-1, 1]$ and $g(x) = \frac{3}{4}(1 - x^2)$ for x in $[-1, 1]$, respectively.

(a) Sketch the graphs of f and g. Based on the graphs which random variable should have the largest variance?
(b) Compute the variances of X and Y.

8. Let $f(x) = 3x^2$ for x in [0,1]. Let X be a random variable with density f. What is the variance of X?

9. Let X have variance 2. What is the variance of $-3X + 1$?

10. Let X be a measure in cm and let Y be the measure of the same object in inches. How are $SD(X)$ and $SD(Y)$ related?

11. Roll two dice successively. Let X be the face of the first die and Y be the face of the second die.

(a) Find $\text{Var}(X - Y)$.
(b) Find $\text{Var}(|X - Y|)$.

12. A circuit has three components that work independently one of each other with probability p_i for $i = 1, 2, 3$. Let S be the number of components that work. Find the variance of S.

2.5 Normal Random Variables

We start by giving the following definition.

Normal Random Variables

The continuous random variable X is said to be a normal random variable with mean μ and standard deviation σ if it has the density

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2}.$$

There are several things to be checked here: that f is a density, that $E(X) = \mu$ and that $\text{Var}(X) = \sigma^2$. Since these computations involve calculus only they will be left as exercises.

The case $\mu = 0$ and $\sigma = 1$ is of particular interest. The density becomes

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}.$$

See Fig. 2.1 for the graph of f.

We also graph below the densities of two normal densities with $\mu = 2$, see Fig. 2.2. One has a standard deviation equal to 1 and the other one a standard

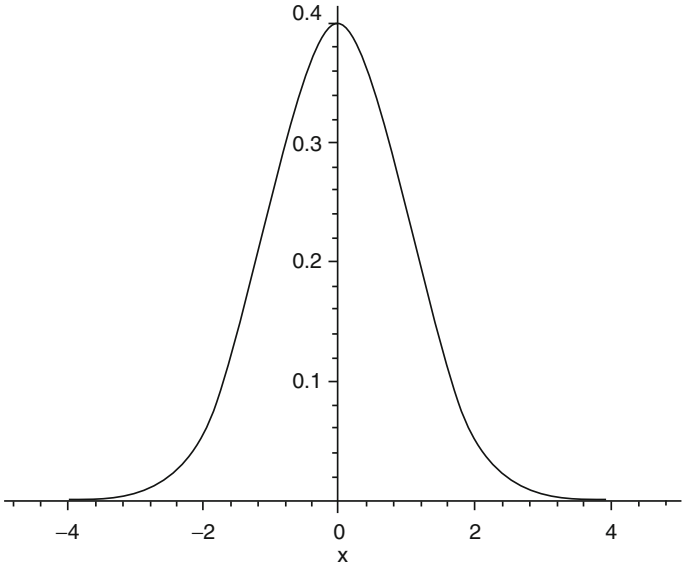


Fig. 2.1

deviation equal to 2. They both have the characteristic bell-shaped form. However, one can see below how much more spread out the curve with $\sigma = 2$ is compared to the one with $\sigma = 1$.

Standard Normal Random Variable

The continuous random variable Z is said to be a standard normal random variable if it has the density

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}.$$

That is, Z is a normal random variable with mean 0 and standard deviation 1.

The notation Z will be reserved to standard normal random variables. In order to compute probabilities involving Z we will need to integrate its density. Unfortunately, there is no explicit formula for antiderivatives of $\frac{1}{\sqrt{2\pi}} e^{-z^2/2}$. We will need to rely on a numerical table provided in the appendix. What is provided is a table for the function

$$\Phi(x) = P(0 < Z < x) = \int_0^x \frac{1}{\sqrt{2\pi}} e^{-z^2/2}.$$

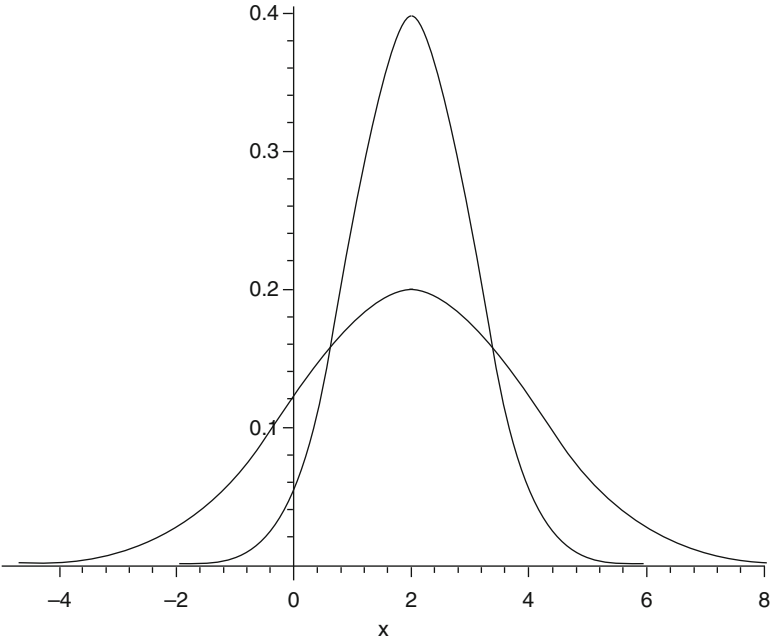


Fig. 2.2

Example 1. What is the probability that a standard normal random variable Z is larger than 1?

We have that

$$P(Z > 1) = 1/2 - \Phi(1) = 1/2 - 0.34 = 0.16.$$

Example 2. What is the probability that a standard normal random variable Z is larger than -1?

By symmetry of the distribution of Z we have

$$P(Z > -1) = P(Z < 1) = 0.84.$$

Example 3. What is the value below which a standard normal random variable is with probability 90%?

We want c such that

$$P(Z < c) = \frac{1}{2} + \Phi(c) = 0.9.$$

We see from the table that c is between 1.28 and 1.29. Since c is closer to 1.28, we take $c = 1.28$.

Example 4. What is the value below which a standard normal random variable is with probability 20%?

This time we want c such that

$$P(Z < c) = 0.2.$$

Note that c is negative. By symmetry we have that

$$P(Z < c) = P(Z > -c) = \frac{1}{2} - \Phi(-c) = 0.2.$$

Thus,

$$\Phi(-c) = 0.3.$$

We read in the table that $-c$ is approximately 0.84. Thus, we have $c = -0.84$.

Example 5. What is the probability that a standard normal random variable Z is between -2 and 2?

$$P(-2 < Z < 2) = 2P(0 < Z < 2) = 2\Phi(2) \sim 0.95.$$

So there is only a 5% chance that a standard normal distribution is larger than 2 or smaller than -2.

One of the nice properties of the normal distributions is that they can easily be transformed into standard normal distributions as the property below shows.

Standardization

If X has normal distribution with mean μ and standard deviation σ then the random variable

$$\frac{X - \mu}{\sigma}$$

is a standard normal random variable.

What is remarkable here is not that $\frac{X-\mu}{\sigma}$ has mean 0 and standard deviation 1. This is true for any random variable that has a mean and a standard deviation as will be shown below. What is remarkable is that after shifting and scaling a normal random variable we still get a normal random variable.

We now compute the expected value and standard deviation of $\frac{X-\mu}{\sigma}$.

$$E\left(\frac{X - \mu}{\sigma}\right) = \frac{1}{\sigma}(E(X) - \mu) = 0,$$

where the last equality comes from the fact that $E(X) = \mu$. For the variance we have

$$\text{Var}\left(\frac{X-\mu}{\sigma}\right) = \frac{1}{\sigma^2}\text{Var}(X-\mu) = \frac{1}{\sigma^2}\text{Var}(X) = 1.$$

We now give a few examples on how to use the property above.

Example 6. Assume that heights of 6 years old are normally distributed with mean 100 cm and standard deviation 2 cm. What is the probability that a 6 years old taken at random is at least 105 cm tall?

Let X be height of the child picked at random. We want $P(X > 105)$. We standardize X to get

$$P(X > 105) = P\left(\frac{X-100}{2} > \frac{105-100}{2}\right) = P(Z > 2.5) \sim 0.01.$$

So there is only a 1% probability that a child taken at random be at least 105 cm tall.

Example 7. What is the height above which 90% of the 6 years old are?

We want h such that $P(X > h)$. We standardize X again to get

$$P(X > h) = P\left(\frac{X-100}{2} > \frac{h-100}{2}\right) = P\left(Z > \frac{h-100}{2}\right) = 0.9.$$

Note that $\frac{h-100}{2}$ must be negative. By symmetry of the distribution of Z we have that

$$P\left(Z > \frac{h-100}{2}\right) = P\left(Z < \frac{-h+100}{2}\right) = 0.9.$$

So according to the Normal table we have

$$\frac{-h+100}{2} = 1.28.$$

We solve for h and get that h is approximately 97.44 cm.

Example 8. Let X be normally distributed with mean μ and standard deviation σ. What is the probability that X is 2σ or more away from its mean?

We want

$$P(\{X > \mu + 2\sigma\} \cup \{X < \mu - 2\sigma\}) = P(X > \mu + 2\sigma) + P(X < \mu - 2\sigma),$$

where the last equality comes from the fact that the two events are disjoint. We standardize X to get

$$P(\{X > \mu + 2\sigma\} \cup \{X < \mu - 2\sigma\}) = P\left(\frac{X-\mu}{\sigma} > 2\right) + P\left(\frac{X-\mu}{\sigma} < -2\right)$$
$$= P(Z > 2) + P(Z < -2) = 0.05.$$

2.5.1 Extreme Observations

As we have just seen the normal distribution is concentrated around its mean and it is unlikely that an observation taken at random is more than 2σ away from its mean (see Example 8). However, if we make several independent observations what is the probability that the largest or the smallest of the observations is far away from the mean? We look next at a particular example.

Example 9. Assume that heights of 6 years old are normally distributed with mean 100 cm and standard deviation 2 cm. In a group of 25 children what is the probability that the tallest of the group is at least 105 cm tall?

Let $X_1, \ldots, X_{25}$ be the heights of the 25 children in the group. We are interested in the probability that the maximum of these random variables be at least 105. It is easier to deal with the complement of the preceding event. Note that the maximum of the 25 observations is less than 105 cm if and only if each one of the observations is less than 105 cm. Thus,

$$P(\max(X_1, \ldots, X_{25}) < 105) = P(\{X_1 < 105\} \cap \{X_2 < 105\} \cap \cdots \cap \{X_{25} < 105\})$$
$$= P(X_1 < 105)P(X_2 < 105)\ldots P(X_{25} < 105),$$

where the last equality comes from the independence of the X_i. According to Example 6, we have that $P(X_1 < 105)$ is $P(Z > 2.5) = 0.9876$ and this probability is the same for each X_i since they all have the same distribution. Thus,

$$P(\max(X_1, \ldots, X_{25}) < 105) = (0.9876)^{25} \sim 0.73.$$

That is, the probability that the tallest child in a group of 25 is at least 105 is 0.27. As the group increases this probability increases as well. For a group of 50 this probability becomes about 0.5. For a group of 100 this probability becomes about 0.7.

The important conclusion of this example is the following. Extreme observations (especially if there are many observations) are likely to be far from a typical observation.

Exercises 2.5

1. Let Z be a standard normal random variable. Compute the following.

(a) $P(Z > 1.52)$.
(b) $P(Z > -1.15)$.
(c) $P(-1 < Z < 2)$.
(d) $P(-2 < Z < -1)$.

2. Let Z be a standard normal random variable. What is the value above which Z is with 99% of probability?

3. Assume that X is normally distributed with mean 3 and standard deviation 2.

(a) $P(X > 3) =?$
(b) $P(X > -1) =?$
(c) $P(-1 < X < 3) =?$
(d) $P(|X - 2| < 1) =?$

4. Assume that the diameter of a ball bearing is normally distributed with mean 1 cm and standard deviation 0.05 cm. A ball bearing is considered defective if its diameter is larger than 1.1 cm or smaller than 0.9 cm.

(a) What is the proportion of defective ball bearings?
(b) Find the diameter above which 99% of the diameters are.

5. Assume that X is normally distributed with mean 5 and standard deviation σ. Find σ so that $P(X > 4) = 0.95$.

6. Assume that the annual snow fall at some place is normally distributed with mean 20 in. and standard deviation 8 in.

(a) What is the probability that the snow fall be less than 5 in. on a given year?
(b) What is the probability that the smallest annual snow fall in the next 20 years will be less than 5 in.?

7. Let Z be a standard normal random variable with density

$$f(z) = \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-z^2/2}.$$

In this exercise we will check that f is actually a density.

(a) Change the variables from Cartesian to polar to show that

$$\int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \mathrm{e}^{-(x^2+y^2)/2} \mathrm{d}x\mathrm{d}y = \int_0^{\infty} \int_0^{2\pi} \mathrm{e}^{-\rho^2/2} \rho \mathrm{d}\rho\mathrm{d}\theta.$$

(b) Show that the r.h.s. of (a) is 2π.

(c) Show that the l.h.s. of (a) is

$$\left(\int_{-\infty}^{+\infty} \mathrm{e}^{-x^2/2}\mathrm{d}x\right)^2.$$

(d) Conclude that f is a density.

8. Let Z be a standard normal random variable.

(a) Compute $E(Z)$.
(b) Compute $\mathrm{Var}(Z)$.

9. Let

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma}\mathrm{e}^{-(x-\mu)^2/2}.$$

Show that f has inflection points at $\mu + \sigma$ and $\mu - \sigma$.

Review Exercises for Chap. 2

1. Three people toss one fair coin each. The winner is the one whose coin shows a face different from the two others. If the three coins show the same face then there is a new round of tosses, until someone wins.

(a) What is the probability of exactly one round of tosses?
(b) What is the probability that at least three rounds of tosses are necessary?

2. A and B take turns rolling a die. A starts. The winner is the first one that rolls a 6. What is the probability that A wins?

3. Two people play the following game. They toss two fair coins. If the two coins land on heads then A wins. If one coin lands on heads and the other on tails then B wins. If the two coins land on tails then the coins are tossed again until someone wins. What is the probability that B wins?

4. The probability of finding someone in favor of a certain initiative is 0.01. We interview people at random until we find a person in favor of the initiative. What is the probability that we need to conduct 50 or more interviews?

5. Draw five cards from a 52 cards deck.

(a) Explain why the probability that the second card is red is the same as the probability that the second card is black.
(b) What is the expected number of red cards among the five cards that have been drawn.

(c) What is the expected number of hearts in five cards dealt from a deck of 52 cards?

6. Assume that car batteries lifetimes follow an exponential distribution with mean 3 years.

(a) What is the probability that a battery lasts 10 years or more?
(b) In a group of ten batteries what is the probability that at least one will last 10 years or more?
(c) How many batteries do we need in order to have at least one last 10 years or more with probability 0.9?

7. Let X a random variable with density $f(x) = c\mathrm{e}^{-|x|}$.

(a) Find c.
(b) What is the $P(X > 1)$?

8. Let X have density $g(x) = c(x-1)^2$ for x in [0,2].

(a) Find c.
(b) Find $E(X)$.
(c) Find $\mathrm{Var}(X)$.

9. Let Y be a random variable with density $f(y) = c(-(y-1)^2 + 2)$ for y in [0,2].

(a) Sketch the graphs of g in Exercise 8 and of f.
(b) Which random variable X or Y do you expect to have the highest variance?
(c) Confirm your prediction by doing a computation.

10. Roll two dice. I win \$1 if the sum is 7 or more. I lose \$$b$ if the sum is 6 or less. Find b so that this is a fair game.

11. Toss five fair coins.

(a) What is the expected number of heads?
(b) What is the variance of the number of heads?

12. Suppose atoms of a given kind have an exponential distributed lifetime with mean 30 years. What is the expected number of atoms still present after 30 years if we start with 10^{23} atoms?

13. Ball bearings are manufactured with diameters that are normally distributed with mean 1 cm and standard deviation 0.05 cm. Assume that 1,000 ball bearings are manufactured. What is the expected number of ball bearings whose diameter is at least 1.1 cm?

14. Assume that the random variable T is such that $E(T) = 1$ and $E(T(T-1) = 2$. What is the standard deviation of T?

15. It is believed that in the 1700s in Europe life expectancy at birth was only around 40 years. That is, a newborn baby could expect on average to live 40 years.

It is also known that child mortality was extremely high. Maybe, as many as 50% of all babies did not make it to their fifth birthday.

(a) Compare the median life span to the expected life span.
(b) Were people old at 35?

16. I have 100 balls in an urn numbered from 1 to 100. I draw at random one ball at a time and then I put it back in the urn.

(a) What is the expected number of draws to get ten different numbers?
(b) What is the expected number of draws to get all the 100 different numbers?

Chapter 3
Binomial and Poisson Random Variables

3.1 Counting Principles

Before stating the fundamental principle of counting we give an example.

Example 1. Assume that a restaurant offers five different specials and for each one of them you can pick either a salad or a soup. How many choices do you have?

In this simple example we can just enumerate all the possibilities. Number the specials from 1 to 5 and let S denote the salad and O denote the soup. There are ten possibilities:

$$(1,S)\ (2,S)\ (3,S)\ (4,S)\ (5,S)$$
$$(1,O)\ (2,O)\ (3,O)\ (4,O)\ (5,O)$$

This is so because we have two selections to make, one with two choices and the other one with five choices. Thus, in all there are $2 \times 5 = 10$ choices.

The Multiplication Rule

If we have r successive selections with n_k choices at the kth step, for $k = 1, \ldots, r$, then in all we have $n_1 \times n_2 \times \cdots \times n_r$ possibilities.

Example 2. Consider an answer sheet with five categories for age, two categories for sex, three categories for education. How many possible answer sheets are there?

In this example we have $r = 3$, $n_1 = 5$, $n_2 = 2$, and $n_3 = 3$. Thus, in all there are $5 \times 2 \times 3 = 30$ possibilities.

Example 3. In a true/false test there are ten questions. How many different ways can this test be answered?

This time we have ten successive selections to be made and for each selection we have two choices. In all there are $2 \times 2 \times \cdots \times 2 = 2^{10}$ possibilities.

R.B. Schinazi, *Probability with Statistical Applications*,
DOI 10.1007/978-0-8176-8250-7_3, © Springer Science+Business Media, LLC 2012

Example 4. How many arrival orders are there for three runners?

We call the three runners A, B, and C. A has three possible arrival positions. Once the arrival of A is fixed then B has only two possible arrival positions. Once the arrivals of A and B are fixed there is only one possible arrival position for C. Thus, we may use the multiplication rule to get that in all there are $3 \times 2 \times 1 = 6$ possibilities.

The preceding example illustrates a consequence of the multiplication rule which is particularly important.

Permutations

For any positive integer n define n factorial as

$$n! = n \times (n-1) \times (n-2) \times \cdots \times 1 \text{ for } n \geq 1$$

and

$$0! = 1.$$

A particular labeling of n distinct objects is called a permutation of these n objects. The number of possible permutations of n objects is $n!$.

Note that in Example 4 we are counting the number of permutations of three runners. The number of permutations is $3! = 6$.

Example 5. How many ways are there to put seven different books on a shelf?

Again we need to count the number of permutations of seven distinct objects. We get $7! = 5,040$ possibilities.

Note that the factorials can be computed inductively by using the formula

$$n! = n \times (n-1)!.$$

Factorials grow very fast (see Exercise 10).

In many situations we want to pick a set of (nonordered) k objects among n objects where $k \leq n$. How many ways are there to do that?

Let $\binom{n}{k}$ (it is read "n choose k") be the number of ways to pick a subset of k objects among n objects. For the first object we pick we have n choices, for the second one we have $n-1$ choices, for the third one $n-2$ choices and so on. For the kth object we have $(n-k+1)$ choices. So according to the multiplication rule we have $n \times (n-1) \times (n-2) \times \cdots \times (n-k+1)$ ways to pick an *ordered* set of k objects. We know that a set of k objects has $k!$ permutations. That is, for every set of k objects there is $k!$ ways to order it. Thus, we have that

$$\text{The number of ways to pick an ordered set of } k \text{ objects} = k! \binom{n}{k}.$$

So

$$n \times (n-1) \times (n-2) \times \cdots \times (n-k+1) = k! \binom{n}{k}.$$

Observe that

$$n \times (n-1) \times (n-2) \times \cdots \times (n-k+1) = \frac{n!}{(n-k)!}.$$

Thus,

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}.$$

Ordered and Nonordered Sets

The number of ways to pick an ordered set of k elements out of n elements is

$$n \times (n-1) \times (n-2) \times \cdots \times (n-k+1).$$

A particular way to pick a nonordered set of k elements out of n is called a combination. The number of combinations of k elements out of n is given by the binomial coefficient

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}.$$

Example 6. Three awards will be given to three distinct students in a group of ten students. How many ways are there to give these three awards?

We want to know how many subsets of three students can be picked out of a set of ten students. This is exactly

$$\binom{10}{3} = \frac{10!}{3!7!} = \frac{10 \times 9 \times 8}{3 \times 2} = 120.$$

Example 7. In a contest, ten students will be ranked and the top 3 will get gold, silver, and bronze medals, respectively. How many ways are there to give these three medals?

This is different from Example 6 because the order of the three students picked is important. There are ten possible choices for the gold medal, there are nine choices for the silver medal, and there are eight choices for the bronze. So according to the multiplication rule there are $10 \times 9 \times 8$ ways to give these medals. That is 720 ways. Note that this is 6 (i.e., 3!) times more ways than in Example 6.

Example 8. In a business meeting seven people shake hands. How many handshakes are there in all?

There are as many handshakes as there are sets of 2 people among 7. So the number is

$$\binom{7}{2} = \frac{7!}{2!5!} = 21.$$

Example 9. How many distinct strings of letters can be made out of the word CARE?

Every permutation of theses four distinct letters will give a distinct string of letters. Thus, there are $4! = 24$ distinct strings of letters.

Example 10. How many distinct strings of letters can be made out of the word PEPPER?

There are 6! possible permutations of these six letters. However, there are only four distinct letters in this word. For instance, if we permute the P's only (there are 3! such permutations) we get the same string of letters. If we permute the E's only (there are 2! such permutations) we also get the same string. Thus, there are

$$\frac{6!}{2!3!} = 60$$

distinct strings of letters.

Example 11. How many distinct strings can we make with three 1's and two 0's?

This is exactly the same problem as Example 10. Note that there are 5! permutations but since there are three 1's and two 0's the total number of distinct strings is:

$$\frac{5!}{3!2!} = \binom{5}{2} = 10.$$

Example 12. You are dealt five cards from a 52 cards deck. What is the probability of getting a full house (three of a kind and a pair of another kind)?

We first observe that there are $\binom{52}{5}$ equally likely hands. Next we use the multiplication rule. There are 13×12 ways to pick two distinct kinds (one for the pair, another one for the triplet). Once we have picked the pair kind there are $\binom{4}{2}$ choices to make a pair. For the triplet there are $\binom{4}{3}$ choices. So there are

$$13 \times 12 \times \binom{4}{2} \times \binom{4}{3}$$

ways to pick a full house. Assuming that all hands are equally likely we get that the probability of a full house is

$$\frac{13 \times 12 \times \binom{4}{2} \times \binom{4}{3}}{\binom{52}{5}} \sim 0.001$$

Example 13. You are dealt five cards from a 52 cards deck. What is the probability of getting three of a kind ?

There are $\binom{13}{1}$ ways to pick the kind for the triplet. Once the kind of the triplet is picked there are $\binom{4}{3}$ ways to pick three cards to make a triplet. There are $\binom{12}{2}$ ways to pick the two remaining kinds. Note that this is NOT 12×11, this is so because the two remaining cards are exchangeable: a queen and a king is the same as a king and a queen for the two remaining cards. Once the kind of each remaining card has been picked then there are $\binom{4}{1}$ to pick a card for each kind. Thus, the number of ways to pick three of a kind is

$$\binom{13}{1}\binom{4}{3}\binom{12}{2}\binom{4}{1}\binom{4}{1}.$$

By dividing the formula above by $\binom{52}{5}$ we get a probability of 0.02.

3.1.1 Properties of the Binomial Coefficients

The $\binom{n}{k}$ are also called binomial coefficients because of their role in the binomial theorem that we will see below. We start by listing a few useful properties of these coefficients.

P1. Recall that $0! = 1$ so

$$\binom{n}{0} = 1 \text{ for every integer } n \geq 0.$$

P2. For all integers $n \geq 1$ and $k \geq 1$ we have that

$$\binom{n}{k} = \binom{n-1}{k-1} + \binom{n-1}{k}.$$

In order to see the preceding identity fix a particular element out of the n elements we have and call it O. We have two possible types of subsets of k elements. The ones that contain O and the ones that do not contain O. There are $\binom{n-1}{k-1}$ subsets of k elements that contain O. This is so because if we pick

O then we need to pick $k-1$ elements out of $n-1$. There are $\binom{n-1}{k}$ subsets of k elements that do not contain O. By adding the two preceding binomial coefficients we get all the subsets of k elements out of n. This proves P2.

P3. For all integers $n \geq 0$ and $k \geq 0$ we have that

$$\binom{n}{k} = \binom{n}{n-k}.$$

Each time we pick k out of n elements, we do not pick $n-k$ out of n elements. So there are as many subsets with k elements as there are with $n-k$ elements. This proves P3.

P4. Pascal's triangle. This is a convenient way to compute the binomial coefficients by using the preceding properties.

	k	0	1	2	3	4	5
n							
0		1					
1		1	1				
2		1	2	1			
3		1	3	3	1		
4		1	4	6	4	1	
5		1	5	10	10	5	1

One reads $\binom{n}{k}$ at the intersection of row n and column k. For instance, $\binom{4}{2} = 6$. The triangle is constructed by using Property P2. For instance,

$$\binom{4}{2} = \binom{3}{1} + \binom{3}{2}.$$

That is, we get 6 by adding the 3 immediately above and the 3 above and to the left. Note that Pascal's triangle is symmetric and that is a consequence of P3.

We now turn to the binomial theorem.

Binomial Theorem

For any integer $n \geq 0$ and any real numbers a and b we have that

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

We will see why the Theorem holds on a particular example. Take $n = 4$ then

$$(a+b)^4 = (a+b) \times (a+b) \times (a+b) \times (a+b).$$

All the terms in the expansion come from these four products. So all the terms must have degree 4. That is all the terms are of the type $a^i b^j$ where $i + j = 4$. To get a^4 we must pick a in each one of the four terms in the product and there is only one way to do that. In the final expansion there is only one a^4. To get $a^3 b$ we need to pick a's from three of the four terms in the product and there are $\binom{4}{3} = 4$ ways to do that. In the final expansion there are 4 $a^3 b$. To get $a^2 b^2$ we need to pick 2 a's and there are $\binom{4}{2} = 6$ ways to do that. Using the symmetry property P3 we get

$$(a+b)^4 = a^4 + 4a^3 b + 6a^2 b^2 + 4ab^3 + b^4.$$

Exercises 3.1

1. Someone has three pairs of shoes, two pairs of pants and four shirts. In how many ways can he get dressed?

2. A test is composed of 12 questions. Each question can be true, false, or blank. How many ways are there to answer this test?

3. In how many ways can seven persons stand in line?

4. How many five cards hands are there out of a deck of 52?

5. Two balls are red and three are blue. How many ways are there to line the balls?

6. License plates have three letters and four numbers. How many different license plates can be made?

7. In a class of 21 in how many ways can a professor give out three A's?

8. In a class of 21 in how many ways can a professor give out three A's and three B's?

9. Assume that eight horses are running and that three will win.

(a) How many ways are there to pick the unordered three winners?
(b) How many ways are there to pick the ordered three winners?

10. According to Stirling's formula we have that

$$n! \sim \sqrt{2\pi} n^{n+1/2} \mathrm{e}^{-n}.$$

That is, the ratio of the two sides tends to 1 as n goes to infinity. Use Stirling's formula to approximate 10!, 20!, and 50!. How good are these approximations?

11. Use Pascal's triangle to compute $\binom{10}{k}$ for $k = 0, \ldots, 10$.

12. You are dealt five cards from a 52 cards deck. What is the probability that

(a) You get exactly one pair?
(b) You get two pairs?
(c) You get a straight flush (five consecutive cards of the same suit)?
(d) A flush (five of the same suit but not a straight flush)?
(e) A straight (five consecutive cards but not a straight flush)?

13. (a) Show that

$$\sum_{k=0}^{n} \binom{n}{k} = 2^n.$$

(b) Use (a) to show that a set of n elements has 2^n subsets.

14. Compute

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

15. Expand $(x + y)^7$.

3.2 Binomial Random Variables

Recall that Bernoulli random variable X is a random variable with two possible outcomes, usually denoted by 0 and 1. Think of 0 as being a failure and 1 as being a success. Assume that $P(X = 1) = p$ and $P(X = 0) = 1 - p$. Consider n independent and identically distributed Bernoulli random variables $X_1, X_2, \ldots, X_n$ and let B be the number of successes among these n experiments. In other words, we have that

$$B = X_1 + X_2 + \cdots + X_n.$$

The random variable B is said to have a binomial distribution with parameters n and p.

We are now going to derive the distribution of B. One of the ways B may be equal to k is if the first k Bernoulli random variables are successes and the last $n - k$ are failures. This happens with probability $p^k(1 - p)^{n-k}$. However, there are as many ways for $B = k$ as there are ways to distribute k 1's and $n - k$ 0's among n places. This is the same problem as the one we solved in Example 11 in 3.1.

We want the number of distinct strings that have length n and k 1's. There are $\frac{n!}{k!(n-k)!}$ distinct strings. Thus,

$$P(B = k) = \binom{n}{k} p^k (1-p)^{n-k}.$$

We now summarize these facts about the binomial distribution in the box below.

Binomial Random Variables

Consider n independent and identically distributed Bernoulli random variables $X_1, X_2, \ldots, X_n$. Let $P(\text{success in the } i\text{th trial}) = P(X_i = 1) = p$, for $i = 1, \ldots, n$. Let B be the number of successes among these n experiments. That is,

$$B = X_1 + X_2 + \cdots + X_n.$$

The random variable B is said to have a binomial distribution with parameters n and p. We have that,

$$P(B = k) = \binom{n}{k} p^k (1-p)^{n-k}.$$

Note that for a binomial B with parameters n and p the formula simplifies for the extreme values

$$P(B = 0) = (1-p)^n \text{ and } P(B = n) = p^n$$

and that by the binomial Theorem

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

Example 1. Roll a fair die 5 times. What is the probability of getting exactly two 6's?

In this case we are doing $n = 5$ identical experiments. The probability of success is $p = 1/6$ and B is the number of 6's (or successes) we get in 5 trials. Thus,

$$P(B = 2) = \binom{5}{2} \left(\frac{1}{6}\right)^2 \left(\frac{5}{6}\right)^3 = 10\frac{5^3}{6^5} \sim 0.16.$$

Example 2. What is the probability of getting at least one 6 in five rolls?

We want the probability of $\{B \geq 1\}$. It is quicker to compute the probability of the complement of $\{B \geq 1\}$ which is $\{B = 0\}$.

$$P(B = 0) = (1 - p)^n = \left(\frac{5}{6}\right)^5 \sim 0.4.$$

Thus, the probability of getting at least one 6 in 5 rolls is approximately 0.6.

Example 3. Assume that births of boys and girls are equally likely. What is the probability that a family with three children have three girls?

This time we have $n = 3$ trials and each has a probability of success (having a girl) equal to $p = 1/2$. We want

$$P(B = 3) = \left(\frac{1}{2}\right)^3 = \frac{1}{8}.$$

Example 4. Consider four families, each with three children. What is the probability that exactly one family has three girls?

We have $n = 4$ trials and a trial is success if the corresponding family has exactly three girls. According to Example 3 the probability of success is 1/8. Thus,

$$P(B = 1) = \binom{4}{1}\left(\frac{1}{8}\right)^1\left(\frac{7}{8}\right)^3 \sim 0.33.$$

Binomial coefficients grow very fast. Next we give an algorithm that allows the computation of a binomial distribution while avoiding the computation of the binomial coefficients.

Computational Formula for the Binomial Distribution

Let B be a binomial random variable with parameters n and p. We have that

$$P(B = 0) = (1 - p)^n$$

and

$$P(B = k) = \frac{p}{1 - p}\frac{n - k + 1}{k}P(B = k - 1) \text{ for } k = 1, 2, \ldots, n.$$

We derive the preceding formula. Let $k \geq 1$,

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

We now apply the preceding formula to an example.

Example 5. Find the distribution of a binomial random variable with $n = 8$ and $p = 0.2$.

We have that

$$P(B = 0) = (1-p)^n = (0.8)^8 \sim 0.17.$$

We use the recursion for $k \geq 1$

$$P(B = k) = \frac{p}{1-p} \frac{n-k+1}{k} P(B = k-1) = \frac{1}{4} \frac{8-k+1}{k} P(B = k-1).$$

For instance,

$$P(B = 1) = \frac{1}{4} 8 P(B = 0) \sim 0.34.$$

We summarize the distribution in the table below.

k	0	1	2	3	4	5	6	7	8
$P(B = k)$	0.17	0.34	0.29	0.15	0.05	0.01	0.001	0	0

Note that $P(B = 7)$ and $P(B = 8)$ are small but strictly positive. In the table above we are keeping only the first three decimals and this is why they appear to be 0.

We now turn to the mean and variance of the binomial distribution.

Mean and Variance of a Binomial Distribution

Assume that B is a binomial random variable with parameters n and p. We have that

$$E(B) = np$$

and

$$\text{Var}(B) = np(1-p) = npq.$$

Recall that a binomial random variable B is a sum of independent identically distributed Bernoulli random variables. That is,

$$B = X_1 + X_2 + \cdots + X_n.$$

Thus,

$$E(B) = E(X_1) + E(X_2) + \cdots + E(X_n)$$

and using that $E(X_i) = p$ for $i = 1, \ldots, n$ we get

$$E(B) = np.$$

Recall that $\mathrm{Var}(X_i) = p(1-p) = pq$ and that the variance of the sum of *independent* random variables is the sum of the variances. Thus,

$$\mathrm{Var}(B) = \mathrm{Var}(X_1) + \cdots + \mathrm{Var}(X_n) = npq.$$

Example 6. Roll a die 30 times, what is the expected number of 5's?

The number of 5's is a binomial random variable with parameters $n = 30$ and $p = 1/6$. So the expected number of 5's is $np = 5$.

Example 7. Assume that 100 components have exponential lifetimes with mean 1 year. Assume that the components fail independently one of the other. What is the expected number of components that have not failed after 2 years?

Let B be the number of components that have not failed after 2 years. We may write

$$B = X_1 + \cdots + X_{100},$$

where $X_i = 1$ if the ith component has not failed after 2 years and $X_i = 0$ otherwise, for $i = 1, \ldots, 100$. The X_i are independent identically distributed Bernoulli random variables with probability of success

$$p = \int_2^\infty \mathrm{e}^{-t}\mathrm{d}t = \mathrm{e}^{-2}.$$

So B is a binomial random variable with parameters 100 and p and the expected value of B is

$$E(B) = np = 100\mathrm{e}^{-2} \sim 13.53.$$

Another measure of location is the mode. Recall that a mode M (not necessarily unique) of a discrete random variable B is such that $P(B = M)$ is the maximum of all the $P(X = k)$.

Mode of a Binomial Random Variable

Let B be a binomial random variable with parameters n and p. If $np + p$ is not an integer then there is a unique mode M which is the greatest integer less than $np + p$. If $np + p$ is an integer then there are two modes $np + p$ and $np + p - 1$.

See Exercise 10 for a proof of the formula above.

Example 8. Roll a die 30 times. What is the most likely number of 6's we will get?

The number of 6's is a binomial random variable with parameters $n = 30$ and $p = 1/6$. We first examine $np + p = 5 + 1/6$. This is not integer, therefore we have a unique mode: the largest integer below $5 + 1/6$, that is 5. The most likely number of 6's is 5.

Note that if we roll the die 35 times then $np + p = 6$ and we have two modes $np + p = 6$ and $np + p - 1 = 5$. So if we roll the die 35 times there are two most likely numbers of 6's: 5 and 6.

3.2.1 Normal Approximation to the Binomial Distribution

As n gets big the computation of something like $P(B \geq a)$ may involve the computation of many binomial probabilities. The most important technique around this problem is the following normal approximation.

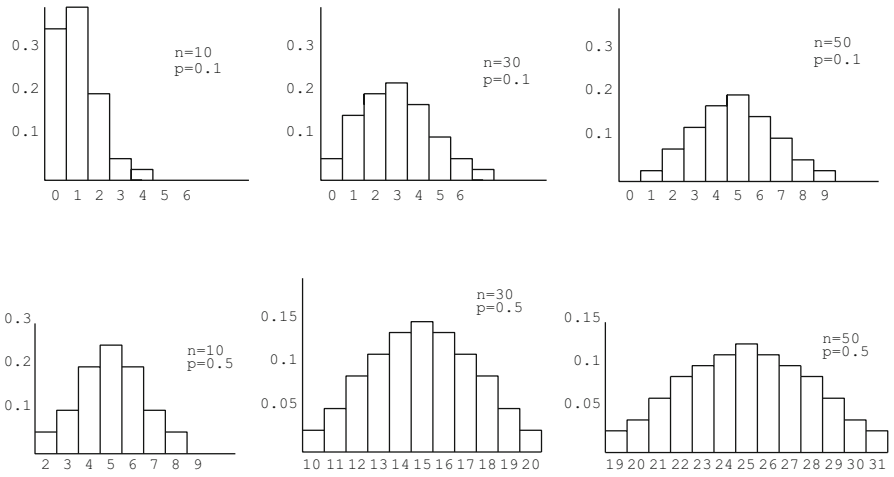
Normal Approximation

Let B be a binomial distribution with parameters n and p. As n increases the distribution of $\frac{B-np}{\sqrt{npq}}$ approaches the distribution of a standard normal random variable Z in the sense that for any $a \leq b$ we have

$$P(a \leq B \leq b) \sim P\left(\frac{a - np - 1/2}{\sqrt{npq}} \leq Z \leq \frac{b - np + 1/2}{\sqrt{npq}}\right) \text{ as } n \to \infty.$$

We are using a continuous random variable Z to approximate a discrete random variable B. This is why we enlarge the interval by 1/2 on both sides. This is especially important if $a = b$ or if $\sqrt{npq}$ is small.

As the figures below illustrate, the larger the n is the closer a binomial histogram is from a normal curve. Another fact that can be seen below is that the convergence toward the normal curve is quicker when p is closer to 1/2.



Example 9. Roll a fair die 36 times, what is the probability that we get exactly six 6's?

Let B be the number of 6's we get in 36 rolls. Then B is a binomial distribution with parameters 36 and 1/6. We first compute the exact probability.

$$P(B = 6) = \binom{36}{6}\left(\frac{1}{6}\right)^6\left(\frac{5}{6}\right)^{30} \sim 0.176$$

We now use the normal approximation. Note that $np = 6$ and $npq = 5$.

$$P(B = 6) \sim P\left(\frac{6 - 6 - 1/2}{\sqrt{5}} \le Z \le \frac{6 - 6 + 1/2}{\sqrt{5}}\right) \sim 0.174.$$

So even in this example with n not so large and p not close to 1/2 the approximation is good.

Example 10. A hotel has accepted 210 reservations but it has only 200 rooms. It is assumed that guests will actually show up independently of each other with probability 0.9. What is the probability that the hotel will not have enough rooms?

Let B be the number of guests that will actually show up. This is a binomial random variable with parameters 210 and 0.9. The mean number of guests showing up is $np = 189$ and the variance is $npq = 18.9$. The normal approximation yields

$$P(201 \le B) \sim P\left(\frac{201 - 189 - 1/2}{\sqrt{18.9}} \le Z\right) = P(2.64 \le Z) \sim 0.004.$$

It is rather unlikely that not enough rooms will be available.

Example 11. Assume that a fair coin is tossed 10,000 times. Let B be the number of heads. What is the probability of getting exactly 5,000 heads?

We use the normal approximation to get

$$P(B = 5000) \sim P\left(\frac{5000 - np - 1/2}{\sqrt{npq}} \leq Z \leq \frac{5000 - np + 1/2}{\sqrt{npq}}\right).$$

The mean is $np = 5,000$ and the standard deviation $\sqrt{npq} = 50$. Thus,

$$P(B = 5000) \sim P(-0.01 \leq Z \leq 0.01) \sim 0.008.$$

So the probability of getting exactly 5,000 heads is rather slim: less than 1%.

Note that $np + p = 5{,}000 + 1/2$ and so the most likely number of heads is 5,000. However, there are so many possible values that any fixed number of heads is rather unlikely.

Example 12. Assume that a fair coin is tossed 10,000 times. Let B be the number of heads. Find a so that B is between $E(B) - a$ and $E(B) + a$ with probability 0.99.

The expected value for B is $np = 10{,}000 \times 1/2 = 5{,}000$. We want a so that

$$P(E(B) - a \leq B \leq E(B) + a) \sim P\left(\frac{-a - 1/2}{\sqrt{npq}} \leq Z \leq \frac{a + 1/2}{\sqrt{npq}}\right) = 0.99.$$

Using the normal table we get

$$\frac{a + 1/2}{\sqrt{npq}} = 2.57.$$

Thus,

$$a = 2.57\sqrt{npq} - \frac{1}{2}.$$

In this particular case we get $a = 128$. So with 99% of confidence the number of heads will be in the interval $[5{,}000 - 128;\ 5{,}000 + 128]$, which is rather narrow considering that we are performing 10,000 tosses. The important lesson of this example is that the number of successes of a binomial with parameters n and p is in the interval $(np - (2.57\sqrt{npq} - 1/2), np + (2.57\sqrt{npq} - 1/2))$ with probability 0.99 when n is large. In particular, typical deviations from the mean are of order $\sqrt{n}$.

3.2.2 The Negative Binomial

Example 13. Roll a fair die. What is the probability that the second 6 appears at the 10th roll?

Note that the event $A = \{$the second 6 appears at the 10th roll$\}$ is the intersection of the two events $B = \{$there is exactly one 6 in the first 9 rolls$\}$ and $C = \{$the 10th roll is a 6$\}$. Moreover, B and C are independent since B depends on the first 9 rolls and C depends on the 10th roll. Note that the number of 6's in 9 rolls is a binomial with parameters 9 and 1/6. Moreover, $P(C) = 1/6$. Thus,

$$P(A) = P(B)P(C) = \binom{9}{1}\left(\frac{1}{6}\right)\left(\frac{5}{6}\right)^8\left(\frac{1}{6}\right) = 9\frac{5^8}{6^{10}} \sim 0.06.$$

More generally, we have the following

Negative Binomial

Assume that we perform identical and independent trials, each trial having a probability of success equal to p. Let r be an integer larger than or equal to 1. Let B_r be the number of trials until the rth success. Then B_r is called a negative binomial random variable with parameters r and p. Moreover, we have

$$P(B_r = k) = \binom{k-1}{r-1} p^{r-1}(1-p)^{k-r} p = \binom{k-1}{r-1} \times p^r (1-p)^{k-r} \text{ for } k = r, r+1, \ldots.$$

Note that in the case $r = 1$, B_1 is the number of trials until the first success. This is exactly a geometric random variable and the formula above simplifies to

$$P(B_1 = k) = p(1-p)^{k-1} \text{ for } k = 1, 2, \ldots.$$

Example 14. What is the probability that the fifth child of a couple is their second girl?

This is a negative binomial with $r = 2$ and $p = 1/2$. Thus,

$$P(B_2 = 5) = \binom{4}{1}\left(\frac{1}{2}\right)\left(\frac{1}{2}\right)^3\left(\frac{1}{2}\right) = \frac{4}{2^5} = \frac{1}{8}.$$

We now turn to the mean and variance of a negative binomial random variable.

Mean and Variance of a Negative Binomial

Let B_r be a negative binomial random variable with parameters r and p. Then,

$$E(B_r) = \frac{r}{p}$$

$$\text{Var}(B_r) = \frac{r(1-p)}{p}.$$

We derive the formulas above by breaking B_r into a sum of simpler random variables. Let G_1 the number trials until the first success, G_1 is a geometric random variable with parameter p. Let G_2 be the number of trials from the first success until the second success. The random variable G_2 is also geometric and G_1 and G_2 are independent. More generally, we define G_i to be the number of trials between the $(i-1)$th success and the ith success for $i = 1, 2, \ldots, r$. It is easy to see that

$$B_r = G_1 + G_2 + \cdots + G_r.$$

All the G_i are independent and identically distributed according to a geometric. Recall that $E(G_1) = 1/p$. Thus,

$$E(B_r) = rE(G_1) = \frac{r}{p}.$$

By using the independence of the G_i and the fact that $\text{Var}(G_1) = (1-p)/p$ we get

$$\text{Var}(B_r) = r\,\text{Var}(G_1) = \frac{r(1-p)}{p}.$$

Example 15. Roll a fair die. How many rolls are expected to get the third 6?

The number of trials to get the third 6 is a negative binomial with parameters $r = 3$ and $p = 1/6$. So

$$E(B_3) = \frac{3}{1/6} = 18.$$

Exercises 3.2

1. Toss a fair coin 4 times.

(a) What is the probability of getting at least 1 heads?
(b) What is the probability of getting at least 3 heads?
(c) What is the probability of getting exactly 2 heads?

2. Roll two fair dice 5 times.

(a) What is the probability of getting at least one sum equal to 7?
(b) What is the probability of getting at least two sums larger than or equal to 7?

3. Toss a fair coin 7 times. Let B be the number of heads.

(a) Draw the histogram of the distribution of B.
(b) What is the mean of B?
(c) What is the mode of B?

4. Toss a fair coin 11 times. What is the most likely number of heads?

5. Given that there were 5 heads in 12 tosses of a fair coin.

(a) What is the probability that the first toss was head?
(b) What is the probability that the last two tosses were heads?
(c) What is the probability that at least two of the first five tosses were heads?

6. Assume that 100 components have normal lifetimes with mean 1 year and standard deviation 6 months. Assume that the components fail independently one of the other.

(a) What is the probability that at least 2 components have not failed after 2 years?
(b) What is the expected number of components that have not failed after 2 years?

7. Assume that 500 invitations have been sent out for a given event. Assume that each person shows up independently of the others with probability 0.6.

(a) What is the probability that 250 or less people show up?
(b) Find b so that the number of people that show up is b or larger with probability 0.9.

8. Roll a fair die 360 times.

(a) What is the probability to get exactly 60 1's?
(b) Find a so that the number of 1's is in the interval $[60-a, 60+a]$ with probability 95%.

9. Toss a fair coin 100 times.

(a) What is the probability of getting exactly 50 heads?
(b) Assume that 25 probability students toss a fair coin 100 times each. What is the probability that at least one student gets exactly 50 heads?

10. In this exercise we derive the formulas for the mode. Let B be a binomial with parameters n and p.

(a) By using the computational formula for the binomial distribution show that $P(B = k-1) \leq P(B = k)$ if and only if $k \leq np + p$.
(b) By definition of the mode M we must have simultaneously $P(B = M-1) \leq P(B = M)$ and $P(B = M+1) \leq P(B = M)$. Use (a) to show that

$$np + p - 1 \leq M \leq np + p.$$

(c) Show that there is only one integer M solution of the double inequality in (b) when $np + p$ is not an integer.
(d) Show that there are two solutions to the double inequality in (b) when $np + p$ is an integer.

11. In 1975, in Columbus Ohio there were 12 cases of childhood leukemia. The expected number is 6 per year (Morbidity and mortality weekly report, July 25 1997, p 671–674). Assume that there are 200,000 children under 15 in that area and that each one has the same probability 3×10^{-5} of being hit by leukemia in a given year.

(a) Use the computational formula for the binomial distribution to compute the probability of having 12 or more cases of leukemia in a given year.
(b) Assume that there are 200 regions in the United States with the same number of children and the same probability for each child to be struck by leukemia. What is the probability that at least one region will get 12 cases or more?
(c) Considering (a) and (b), would you attribute the cluster in Columbus to chance alone?

12. Toss a fair coin.

(a) What is the probability that the third head occurs at the eighth toss?
(b) What is the expected number of tosses to get the tenth head?

13. Items are examined sequentially at a manufacturing plant. The probability that an item is defective is 0.05.

(a) What is the probability that the first 20 items examined are not defective?
(b) What is the expected number of examined items until we get the fifth defective?

14. What is the probability that the fifth child of a couple is their third girl?

3.3 Poisson Random Variables

We start with the definition.

Poisson Random Variables

The random variable N is said to have a Poisson distribution with mean λ if

$$P(N = k) = \mathrm{e}^{-\lambda} \frac{\lambda^k}{k!} \text{ for } k = 0, 1, \ldots.$$

Later in this section we will show that a random variable with the distribution above has indeed mean λ. Typically the Poisson distribution appears when we count the number of occurrences of events that have small probabilities and are independent.

Example 1. Consider a fire station that serves a given neighborhood. Each resident has a small probability of needing help on a given day and most of the time people need help independently of each other. The number of calls a fire station gets on a given day may be approximated by a Poisson random variable with mean λ. The parameter λ may be taken to be the observed average. Assume that $\lambda = 6$. What is the probability that a fire station get 2 or more calls in a given day?

$$P(N \geq 2) = 1 - P(N = 0) - P(N = 1) = 1 - \mathrm{e}^{-\lambda} - \lambda \mathrm{e}^{-\lambda} = 1 - 7\mathrm{e}^{-6} \sim 0.98.$$

Example 2. Assume that a book has an average of one misprint every ten pages. What is the probability that a given page has no misprint?

Consider all the words in a given page, we may assume that each one of them has a small probability of being misprinted. We may also assume that each word is misprinted independently of the other words. With these assumptions the Poisson distribution is adequate. The mean number of misprints per page is $\lambda = 0.1$. Thus,

$$P(N = 0) = \mathrm{e}^{-\lambda} = \mathrm{e}^{-0.1} \sim 0.9.$$

The next property shows that a binomial distribution with parameters n and p may be approximated by a Poisson distribution with mean $\lambda = np$.

Poisson Approximation of the Binomial

Let B be a binomial random variable with parameters n and p. Let N be a Poisson random variable with mean $\lambda = np$ then for every $k \geq 0$

$$P(B = k) \sim P(N = k) \text{ for small } p.$$

The smaller the p is the better the approximation above is, for more details see Hodges and Le Cam, Annals of Mathematical Statistics (1960), pp 737–740. Thanks to the Poisson approximation we replace a two parameters distribution by a one parameter distribution and we avoid the computation of binomial coefficients. Note that if B is a binomial random variable with parameters n and p then

$$P(B = 0) = (1 - p)^n.$$

Recall from Calculus that

$$\lim_{p\to 0} \frac{\ln(1-p)}{-p} = 1.$$

Therefore,

$$P(B=0) = (1-p)^n = \exp(n\ln(1-p)) \sim \exp(-np) = P(N=0),$$

where the approximation holds for p small enough. In order to prove that a binomial with small p may be approximated by a Poisson we need to show that for every $k \geq 0$ it is true that $P(B=k) \sim P(N=k)$. For a proof see the reference above.

Example 3. During a recent meteor shower it was estimated that the probability of a given satellite to be hit by a meteor is 1/1,000. Assuming that there are 500 satellites around the Earth and that they get hit independently one of the other, what is the probability that no satellite will be hit?

Let B be the number of satellites hit. Under these assumptions B has a binomial distribution with parameters 500 and 1/1,000. We have

$$P(B=0) = \left(1 - \frac{1}{1,000}\right)^{500} \sim 0.6064.$$

We now use the Poisson approximation with $np = 1/2$. We have

$$P(N=0) = 1 - \mathrm{e}^{-\lambda} = \mathrm{e}^{-1/2} \sim 0.6065.$$

One can see that the approximation is excellent in this case.

What is the probability that two or more satellites are hit?

This time we want

$$P(N \geq 2) = 1 - P(N=0) - P(N=1) = 1 - \mathrm{e}^{-\lambda} - \lambda\mathrm{e}^{-\lambda} = 1 - \frac{3}{2}\mathrm{e}^{-1/2} \sim 0.09.$$

The next example will use the following algorithm to compute the distribution of a Poisson random variable.

Computational Formula for the Poisson Distribution

Let N be a Poisson random variable with mean λ then its distribution may be computed inductively by using the following algorithm.

$$P(N=0) = \mathrm{e}^{-\lambda}$$

$$P(N=k) = \frac{\lambda}{k} P(N=k-1) \text{ for all } k \geq 1.$$

The formula above is easy to derive. Assume $k \geq 1$, then

$$P(N = k) = \mathrm{e}^{-\lambda}\frac{\lambda^k}{k!} = \mathrm{e}^{-\lambda}\frac{\lambda}{k}\frac{\lambda^{k-1}}{(k-1)!} = \frac{\lambda}{k}P(N = k-1).$$

Example 4. Assume that a hospital serves 100,000 people and that each person may get hit by a certain disease with probability 3×10^{-5} per year, independently one of the other. What is the probability that the hospital will see six or more cases of the disease in a given year?

Under the assumptions the number of cases of the disease follows a binomial with parameters $n = 100{,}000$ and $p = 3 \times 10^{-5}$. We use the Poisson approximation with mean $\lambda = np = 3$. We want

$$P(N \geq 6) = 1 - \sum_{k=0}^{5} P(N = k).$$

We use the computational formula to get

k	0	1	2	3	4	5
$P(N = k)$	0.05	0.15	0.22	0.22	0.17	0.1

Thus,

$$P(N \geq 6) \sim 1 - 0.91 = 0.09.$$

In many situations we may need more involved models than the simple binomial in Example 4. For instance, in the case of cancer the probability of getting hit increases significantly with age. So a more realistic model should separate people in age classes. The total number of cancer cases is then a sum of binomial random variables with different p's. This is not a binomial random variable. However, the next result shows that we may still use the Poisson approximation when all the p's are small.

Poisson Approximation of a Sum of Binomial Random Variables

Let B_i, for $i = 1, \ldots, r$, be independent binomial random variables with parameters n_i and p_i. Let

$$\lambda = n_1 p_1 + \cdots + n_r p_r$$

and N be a Poisson random variable with mean λ. Then for every $k \geq 0$ we have

$$P(B_1 + B_2 + \cdots + B_r = k) \sim P(N = k) \text{ when all the } p_i's \text{ are small.}$$

Example 5. Assume that a hospital serves 100,000 people that are in three different class ages. Assume that an individual in class i has a probability p_i (independently of all the other individuals) of getting a certain disease. Class 1 has $n_1 = 50,000$ individuals and $p_1 = 2 \times 10^{-5}$, class 2 has $n_2 = 30,000$ individuals and $p_2 = 5 \times 10^{-5}$ and class 3 has $n_3 = 20,000$ individuals and $p_3 = 10^{-4}$. What is the probability that on a given year this hospital sees three or more cases of the disease?

For each class i the number of cases B_i follows a binomial with parameters n_i and p_i. We are interested in the event $B_1 + B_2 + B_3 \geq 3$. Since the B_i are independent and the p_i's are small we may use the Poisson approximation. Let

$$\lambda = n_1 p_1 + n_2 p_2 + n_3 p_3 = 4.5$$

and let N be a Poisson random variable with mean λ. We have

$$\begin{aligned} P(B_1 + B_2 + B_3 \geq 3) \sim P(N \geq 3) &= 1 - (P(N = 0) + P(N = 1) + P(N = 2)) \\ &= 1 - \mathrm{e}^{-\lambda} - \lambda \mathrm{e}^{-\lambda} - \lambda^2 \mathrm{e}^{-\lambda}/2 \sim 0.83. \end{aligned}$$

We now turn to a property that shows that the Poisson distribution is bound to appear in many situations. Consider a finite interval I that gets random hits (the interval may represent a time interval and the hits may represent incoming telephone calls). Assume the following two hypotheses:

(1) A given point of I may get hit at most once
and
(2) Divide I in equal subintervals then each subinterval gets hit with the same probability and independently of the other subintervals.

Poisson Scatter Theorem

Under hypotheses (1) and (2) there is a number $\lambda > 0$ such that the total number of hits on I follows a Poisson distribution with mean λ. Let L be the length of I then any subinterval of I with length ℓ has a Poisson distribution with mean $\lambda \ell / L$.

For a proof of this theorem, see *Probability* by Pitman.

Example 6. Consider a telephone exchange on a Monday from 2:00 to 3:00 p.m. Assume that there is an average of 120 calls during this time period. What is the probability of getting at least four calls in a 3-min interval?

It may be reasonable to assume that hypotheses (1) and (2) hold (the only question about this is whether each subinterval of time is equally likely to get calls). Then according to the Poisson scatter Theorem the number of calls during a 3-min interval follows a Poisson distribution with mean $120 \times 3/60 = 6$.

$$P(N \geq 4) = 1 - (P(N = 0) + P(N = 1) + P(N = 2) + P(N = 3))$$
$$= 1 - \mathrm{e}^{-6} - 6\mathrm{e}^{-6} - \frac{6^2}{2}\mathrm{e}^{-6} - \frac{6^3}{3!}\mathrm{e}^{-6} \sim 0.85.$$

The Poisson scatter Theorem holds in any dimension. For instance, it may be used to count the number of stars that appear on a photographic plate or the number of raisins in a cookie. In the first case we replace length by area and in the second one we replace length by volume.

Example 7. Assume that rain drops are hitting a square with side 10 in. Assume that the average is 30 drops per minute. What is the probability that a subsquare with side 2 in. does not get hit in a given minute?

Again it seems reasonable to assume that hypotheses (1) and (2) hold. The number of rain drops in the subsquare follows a Poisson distribution with mean $30 \times 2^2/10^2 = 1.2$. Thus,

$$P(N = 0) = \mathrm{e}^{-1.2} \sim 0.3.$$

Example 8. Assume that a given document has in average two misprints per page. Given that there are no misprints in the first half of a page, what is the probability that there will be two or more misprints in the second half of this page?

It is reasonable to assume that hypotheses (1) and (2) hold and therefore the number of misprints in the two half pages are independent. Let A be the event "there are no misprints in the first half of the page" and let B be the event "there are at least two misprints in the second half of the page." Let N be the number of misprints in the second half page. According to the Poisson scatter Theorem N follows a Poisson distribution with mean $2 \times 1/2 = 1$. Thus, we have

$$P(B|A) = P(B) = P(N \geq 2) = 1 - (P(N = 0) + P(N = 1)) = 1 - 2\mathrm{e}^{-1}.$$

For some of the computations below it will be useful to recall the following from Calculus.

Taylor Series for the Exponential Function

We have

$$\mathrm{e}^x = \sum_{k=0}^{\infty} \frac{x^k}{k!} \text{ for every } x.$$

In particular we see that if N is a Poisson random variable with mean λ we have

$$\sum_{k=0}^{\infty} P(N = k) = \sum_{k=0}^{\infty} \mathrm{e}^{-\lambda} \frac{\lambda^k}{k!} = \mathrm{e}^{-\lambda}\mathrm{e}^{\lambda} = 1.$$

We are now going to compute the mean and variance of a Poisson random variable N with mean λ. We have

$$E(N) = \sum_{k=0}^{\infty} kP(N = k) = \sum_{k=1}^{\infty} k\mathrm{e}^{-\lambda}\frac{\lambda^k}{k!} = \mathrm{e}^{-\lambda}\lambda\sum_{k=1}^{\infty}\frac{\lambda^{k-1}}{(k-1)!}.$$

By shifting the summation index we get

$$\sum_{k=1}^{\infty}\frac{\lambda^{k-1}}{(k-1)!} = \sum_{k=0}^{\infty}\frac{\lambda^k}{k!} = \mathrm{e}^{\lambda}.$$

Thus,

$$E(N) = \mathrm{e}^{-\lambda}\lambda\sum_{k=1}^{\infty}\frac{\lambda^{k-1}}{(k-1)!} = \mathrm{e}^{-\lambda}\lambda\mathrm{e}^{\lambda} = \lambda.$$

We now turn to the computation of the variance of N. It turns out that it is easier to compute $E(N(N-1))$ than $E(N^2)$. We have

$$E(N(N-1)) = \sum_{k=0}^{\infty} k(k-1)P(N = k) = \sum_{k=2}^{\infty} k(k-1)\mathrm{e}^{-\lambda}\frac{\lambda^k}{k!} = \mathrm{e}^{-\lambda}\lambda^2\sum_{k=2}^{\infty}\frac{\lambda^{k-2}}{(k-2)!}.$$

We shift again the summation index to get

$$E(N(N-1)) = \mathrm{e}^{-\lambda}\lambda^2\sum_{k=2}^{\infty}\frac{\lambda^{k-2}}{(k-2)!} = \mathrm{e}^{-\lambda}\lambda^2\mathrm{e}^{\lambda} = \lambda^2.$$

So $E(N(N-1)) = \lambda^2$ and therefore

$$E(N^2) = E(N(N-1) + E(N) = \lambda^2 + \lambda.$$

By definition of the variance we have

$$\mathrm{Var}(N) = E(N^2) - E(N)^2 = \lambda.$$

We now summarize the computations above.

Mean and Variance of a Poisson Random Variable

Let N be a Poisson random variable with mean λ then

$$E(N) = \mathrm{Var}(N) = \lambda.$$

Example 9. We have seen that the distribution of binomial random variable B with parameters n and p may be approximated by a Poisson distribution with mean $\lambda = np$ if p is small. We also know that B may be approximated by a Normal distribution if n is large. This implies that if n is large and p small then a Poisson random variable with mean λ may be approximated by a normal distribution. In fact, the larger λ the better the approximation. In this example, we will compute $P(N = 5)$ exactly and by using a normal approximation for a Poisson random variable N with mean 7. The exact computation is

$$P(N = 5) = \mathrm{e}^{-7}\frac{7^5}{5!} \sim 0.13.$$

We now use the normal approximation.

$$P(N = 5) = P(4.5 \le N \le 5.5) = P\left(\frac{4.5 - E(N)}{SD(N)} \le \frac{N - E(N)}{SD(N)} \le \frac{5.5 - E(N)}{SD(N)}\right).$$

We now approximate the distribution of $\frac{N-E(N)}{SD(N)}$ by the distribution of a standard normal distribution Z. Thus,

$$P(N = 5) \sim P\left(\frac{4.5 - 7}{\sqrt{7}} \le Z \le \frac{5.5 - 7}{\sqrt{7}}\right) = P(-0.94 \le Z \le -0.57) \sim 0.11.$$

Mode of a Poisson Random Variable

Let N be a Poisson random variable with mean λ. If λ is an integer then N has two modes: λ and $\lambda - 1$. If λ is not integer then N has a unique mode: the largest integer smaller than λ.

For a proof see Exercise 10 below.

Exercises 3.3

1. Assume that books from a certain publisher have an average of one misprint every 20 pages.

(a) What is the probability that a given page has two or more misprints?
(b) What is the probability that a book of 200 pages has at least one page with two or more misprints?

2. Suppose that cranberries muffins have an average of six cranberries.

(a) What is the probability that half a muffin has at least four cranberries?

(b) What is the probability that half a muffin has two or less cranberries?
(c) Given that the first half of my muffin had two cranberries or less, what is the probability that the second half has four or more cranberries?

3. Assume that you bet 200 times on 7 at the roulette (there are 38 possible slots). What is the probability that you win at least 3 times?

4. (a) Use the computational formula to compute $P(N = k)$ for $k = 0, 1, \ldots, 10$ for a Poisson random variable with mean $\lambda = 5$.
(b) What are the modes of the distribution in (a)?

5. Assume that 1,000 individuals are screened for a condition that affect 1% of the general population. What is the probability that exactly ten individuals have the condition?

6. Assume that an elementary school has 500 children.

(a) What is the probability that at least one child was born on April 15?
(b) What is the probability that at least three children were born on April 15?

7. The number of incoming phone calls at a telephone exchange is modeled using a Poisson distribution with mean $\lambda = 2$ per minute.

(a) What is the probability of having five or less calls in a 3-min interval?
(b) Given that there were seven calls in the first 3 min, what is the probability that there were no calls during the first minute?
(c) Show that given that there were n calls during the first t minutes the number of calls during the first $s < t$ minutes follows a binomial with parameters s/t and n.

8. Suppose that the probability of a genetic disorder is 0.05 for men and 0.01 for women. Assume that 50 men and 100 women are screened.

(a) Compute the exact probability that exactly two individuals among the 150 that have been screened have the disorder.
(b) Use the Poisson approximation for a sum of binomial random variables to compute the approximate probability of the event in (a).

9. Assume that 1% of men under 20 experience hair loss and that 10% of men over 30 experience hair loss. A sample of 20 men under 20 and 30 men over 30 are examined. What is the probability that four or more men experience hair loss?

10. In this problem we are going to find a formula for the mode of a Poisson distribution.

(a) Use that

$$P(N = k) = \frac{\lambda}{k} P(N = k - 1) \text{ for } k \geq 1$$

to show that $P(N = k) \geq P(N = k - 1)$ if and only if $\lambda \geq k$.

(b) Show that $P(N = k) \geq P(N = k + 1)$ if and only if $\lambda \leq k + 1$.
(c) Show that the mode M of N must satisfy the double inequality

$$\lambda - 1 \leq k \leq \lambda.$$

(d) Show that if λ is an integer then there are two modes λ and $\lambda - 1$. Show that if λ is not an integer then there is a unique mode which is the largest integer smaller than λ.

11. Let N be a Poisson random variable with mean 10.

(a) What is the exact probability that $N = 10$?
(b) Use the normal approximation of Example 9 to compute the probability in (a).

Review Exercises for Chap. 3

1. How many distinct string of letters can we get from the word TOUGH?

2. How many distinct string of letters can we get from the word PROBABILITY?

3. Roll three dice.

(a) What is the probability of getting a sum equal to 10?
(b) What is the probability of getting a sum equal to 9?

4. Assume you toss a coin 100 times and you get 32 heads. Do you think this is a fair coin? (hint: assume it is a fair coin and compute the probability of getting 32 or less heads).

5. Roll a pair of dice 10 times.

(a) What is the probability to get at least once a pair of 6's?
(b) What is the probability of getting twice a pair of 6's?
(c) What is the probability of getting the first pair of 6's at the tenth roll?

6. Roll a die.

(a) What is the probability of getting the first 6 at or before the fifth roll?
(b) What is the probability of getting the third 6 at the tenth roll?
(c) What is the expected number of rolls to get the fifth 6?
(d) Given that the second 6 occurred at the tenth roll, what is the probability that the first 6 occurred at the fifth roll?

7. A gambler bets repeatedly \$1 on red at the roulette (there are 18 red slots and 38 slots in all). He wins \$1 if red comes up loses \$1 otherwise. What is the probability that he will be ahead

(a) After 100 bets?
(b) After 1,000 bets?

8. Assume that each passenger shows up independently of the others with probability 0.95. How many tickets should the airline sell for a flight on an airplane with 200 seats so that, with probability 0.99, each passenger that shows up gets a seat on the flight?

9. A company has three factories A, B, and C. A has manufactured 1,000 items, B has manufactured 1,500 items, and C has manufactured 2,000 items. Assume that the probability that an item be defective is 0.003 for A, 0.002 for B, and 0.001 for C. What is the probability that the total number of defective items is 7 or larger?

10. Assume that lamp bulbs have exponential life times with mean 2 years. What is the probability that in a box of ten

(a) Exactly two will last at least 2 years?
(b) None will last more than 1 year?
(c) What is the expected number of lamp bulbs that will last at least 2 years?

11. Assuming that boys and girls are equally likely, how many children should a couple plan to have in order to have at least one boy and one girl with probability 0.99?

12. In average there is one defect per 100 m of magnetic tape.

(a) What is the probability that 150 m of tape have no defect?
(b) Given that the first 100 m of tape had no defect what is the probability that the whole 150 m have no defect?
(c) Given that the first 100 m of tape had at least one defect what is the probability that the whole 150 m have exactly two defects?

13. Assume you bet \$1 100 times on 7 (there are 38 equally likely slots). If 7 comes up you win \$35, otherwise you lose your \$1.

(a) What are your expected winnings?
(b) What is the probability that you are ahead after 100 bets?
(c) What is the probability that you have lost \$100?

14. Assume you bet \$1 100 times on red (there are 38 equally likely slots and 18 are red). If red comes up you win \$1, otherwise you lose your \$1.

(a) What are your expected winnings?
(b) What is the probability that you are ahead after 100 bets?
(c) What is the probability that you have lost \$100?

15. Assume that 49 students each toss a fair coin 100 times.

(a) What is the probability that at least one student gets 60 or more heads?
(b) What is the probability that at least three students get at least 60 heads?

16. Assume that on average there are five raisins per cookie.

(a) What is the probability that in a package of ten cookies all the cookies have at least one raisin?

(b) How many raisins should each cookie have in average so that the probability in (a) is 0.99?

17. Assume that 10% of the population are left-handers. What is the probability that in a class of 40 there are at least three left-handers?

18. Roll a die four times. What is the probability of

(a) Getting a pair?
(b) Getting three of a kind?
(c) Getting four of a kind?
(d) Getting two pairs?
(e) Four distinct faces?

Chapter 4
Limit Theorems

4.1 The Law of Large Numbers

Assume that we want to know the mean lifetime of a certain type of battery. A natural way to do that is to pick at random a sample of 100 identical batteries, measure the lifetime for each battery and then compute the average lifetime in our sample. The law of large numbers will show that if the sample is large enough then the sample average should be close to the true mean with high probability. We now formalize these ideas.

Let $X_1, \ldots, X_n$ be n independent identically distributed (i.i.d. in short) random variables. These may represent, for instance, the lifetimes of a sample of n batteries. Typically, the distributions of the X_i will not be known. However, we will assume that the mean and the variance exist (but are not known). We denote the variance and the mean of X_i by μ and σ, respectively.

$$E(X_1) = E(X_2) = \cdots = E(X_n) = \mu \text{ and } \mathrm{Var}(X_1) = \mathrm{Var}(X_2) = \cdots = \mathrm{Var}(X_n) = \sigma^2.$$

We would like to estimate μ. A natural estimator for μ is

$$\bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}.$$

That is, we estimate the true mean μ by using the average over the sample $\bar{X}$. Note that $\bar{X}$ is a random variable whose value varies with the sample over which we are averaging. We start by computing the mean and the variance of $\bar{X}$ in function of μ and σ. Recall that the expectation is a linear operator.

R.B. Schinazi, *Probability with Statistical Applications*,
DOI 10.1007/978-0-8176-8250-7_4, © Springer Science+Business Media, LLC 2012

The Expectation is Linear

Let a_i, $1 \leq i \leq n$, be a sequence of real numbers. Let X_i, $1 \leq i \leq n$, be a sequence of random variables defined on the same sample space. We have

$$E(a_1X_1 + a_2X_2 + \cdots + a_nX_n) = a_1E(X_1) + a_2E(X_2) + \cdots + a_nE(X_n).$$

Note that by the linearity of the expectation we have

$$E(X_1+X_2+\cdots+X_n) = E(X_1)+E(X_2)+\cdots+E(X_n) = \mu+\mu+\cdots+\mu = n\mu.$$

Again by the linearity of the expectation we have

$$E(\bar{X}) = E\left(\frac{X_1 + X_2 + \cdots + X_n}{n}\right) = \frac{1}{n}E(X_1 + X_2 + \cdots + X_n) = \frac{1}{n}n\mu = \mu.$$

That is, the expected value of $\bar{X}$ is the same as the expected value of each random variable X_i. We now compute the variance of $\bar{X}$ in order to investigate the dispersion of $\bar{X}$. We first recall an important property of the variance.

Variance of a Sum of INDEPENDENT Random Variables

Let a_i, $1 \leq i \leq n$, be a sequence of real numbers. Let X_i, $1 \leq i \leq n$, be a sequence of independent random variables defined on the same sample space. We have

$$\mathrm{Var}(a_1X_1+a_2X_2+\cdots+a_nX_n) = a_1^2\mathrm{Var}(X_1)+a_2^2\mathrm{Var}(X_2)+\cdots+a_n^2\mathrm{Var}(X_n).$$

Let $X_1, X_2, \ldots, X_n$ be i.i.d. random variables. We start by computing

$$\begin{aligned}\mathrm{Var}(X_1 + X_2 + \cdots + X_n) &= \mathrm{Var}(X_1) + \mathrm{Var}(X_2) + \cdots + \mathrm{Var}(X_n)\\ &= \sigma^2 + \sigma^2 + \cdots + \sigma^2 = n\sigma^2.\end{aligned}$$

$$\begin{aligned}\mathrm{Var}(\bar{X}) &= \mathrm{Var}\left(\frac{X_1 + X_2 + \cdots + X_n}{n}\right) = \frac{1}{n^2}\mathrm{Var}(X_1 + X_2 + \cdots + X_n)\\ &= \frac{1}{n^2}n\sigma^2 = \frac{\sigma^2}{n}.\end{aligned}$$

We summarize these results below.

Expected Value and Variance of the Sample Average

Let $X_1, X_2, \ldots, X_n$ be independent and identically distributed random variables with mean μ and variance σ^2. Then,

$$E(\bar{X}) = E\left(\frac{X_1 + X_2 + \cdots + X_n}{n}\right) = \mu$$

and

$$\text{Var}(\bar{X}) = \text{Var}\left(\frac{X_1 + X_2 + \cdots + X_n}{n}\right) = \frac{\sigma^2}{n}.$$

That is, the expected value of $\bar{X}$ is μ and its distribution is more and more concentrated around μ as the sample size n increases.

The variance of $\bar{X}$ is one measure of the distance between $\bar{X}$ and its mean μ. Observe that the variance of $\bar{X}$ converges to 0 as n goes to infinity. In this sense this means that $\bar{X}$ converges to μ as n goes to infinity. In other words, we have justified mathematically the natural idea of taking the sample average to estimate the mean of the distribution.

Example 1. Assume that we use a sample of 100 identical batteries to estimate the lifetime of a battery. Denote the mean and standard deviation of the lifetime distribution by μ and σ, respectively. What are the mean and standard deviation of the sample average $\bar{X}$?

According to the formula above

$$E(\bar{X}) = \mu \text{ and } \text{Var}(\bar{X}) = \frac{\sigma^2}{100}.$$

Thus, $SD(\bar{X}) = \sigma/10$. That is, the distribution of $\bar{X}$ is 10 times more concentrated that the distribution of X_1.

Since the variance of $\bar{X}$ goes to 0 as the sample size n goes to infinity we know that $\bar{X}$ approaches μ. But this is not very precise. For instance, it would be more useful to be able to say that $\bar{X}$ is within 0.1 of μ with probability 0.95. We are now going to work toward this goal.

Markov's Inequality

Let $X \geq 0$ be a positive random variable with mean μ. Then, for any $b > 0$ we have

$$P(X \geq b) \leq \frac{\mu}{b}.$$

Example 2. Find a bound on the probability that a positive random variable be larger than 10 times its mean.

We want a bound on $P(X > 10\mu)$. We use Markov's inequality with $b = 10\mu$ to get

$$P(X > 10\mu) \leq \frac{\mu}{10\mu} = \frac{1}{10}.$$

What is interesting here is that for ANY positive random variable X (that has a mean) this probability is bound by 0.1.

We now prove Markov's inequality for a continuous random variable. The proof for a discrete random variable is very similar. Let f be the density of X. We have

$$E(X) = \int_0^\infty xf(x)\mathrm{d}x = \int_0^b xf(x)\mathrm{d}x + \int_b^\infty xf(x)\mathrm{d}x \geq \int_b^\infty xf(x)\mathrm{d}x.$$

The preceding inequality holds since X is assumed to be positive. Note that

$$\int_b^\infty xf(x)\mathrm{d}x \geq b\int_b^\infty f(x)\mathrm{d}x = bP(X \geq b).$$

Thus,

$$P(X \geq b) \leq \frac{E(X)}{b}$$

and this proves Markov's inequality.

A consequence of Markov's inequality is Chebyshev's inequality. The latter gives a bound on the dispersion of a random variable.

Chebyshev's Inequality

Let X be a random variable with mean μ and variance σ^2. Then, for any $b > 0$ we have

$$P(|X - \mu| \geq b) \leq \frac{\sigma^2}{b^2}.$$

Example 3. Let X be a random variable with mean μ and variance σ^2. Give an upper bound on the probability that X is more than 2σ away from its mean μ.

We want $P(|X - \mu| \geq 2\sigma)$. We use Chebyshev's inequality with $b = 2\sigma$ to get

$$P(|X - \mu| \geq 2\sigma) \leq \frac{\sigma^2}{(2\sigma)^2} = \frac{1}{4}.$$

So an upper bound is 1/4. Again, what is remarkable here is that this bound holds for ANY random variable with a variance. The bound given by Chebyshev may be quite crude. For instance, note that if X is a normal random variable then the probability of being at least 2σ away from μ is about 0.05 while the bound given by Chebyshev's inequality for the same probability is 0.25.

We now prove Chebyshev's inequality. Let X be a random variable with mean μ and standard deviation σ. Define the random variable

$$Y = (X - \mu)^2.$$

Note that $E(Y) = \text{Var}(X) = \sigma^2$. We apply Markov's inequality to Y (which is a positive random variable)

$$P(Y > b^2) \leq \frac{E(Y)}{b^2}.$$

The event $\{Y > b^2\}$ may also be written as the event $\{|X - \mu| > b\}$. Thus,

$$P(Y > b^2) = P(|X - \mu| > b) \leq \frac{E(Y)}{b^2} = \frac{\sigma^2}{b^2}$$

and the proof of Chebyshev's inequality is complete.

We are now ready to state the Law of large numbers.

Law of Large Numbers

Let $X_1, X_2, \ldots, X_n$ be a sequence of independent identically distributed random variables with mean μ and variance σ^2. Then, for any $b > 0$ we have

$$\lim_{n\to\infty} P\left(|\frac{X_1 + X_2 + \cdots + X_n}{n} - \mu| \geq b\right) = 0.$$

Moreover, we have that

$$P\left(|\frac{X_1 + X_2 + \cdots + X_n}{n} - \mu| \geq b\right) \leq \frac{\sigma^2}{b^2 n}.$$

What the Law of large numbers tells us is that the probability that $\bar{X}$ deviates from μ by an arbitrarily small $b > 0$ goes to 0 as the sample size n goes to infinity. Note that there are different versions of the Law of large numbers. The version we just stated is the *weak* law of large numbers as opposed to the *strong* law of large numbers:

$$P\left(\lim_{n\to\infty} \frac{X_1 + X_2 + \cdots + X_n}{n} = \mu\right) = 1.$$

The strong law of large numbers is proved in more advanced probability courses because it involves more advanced mathematics.

We now prove the weak version. We apply Chebyshev's inequality to the random variable $\bar{X}$. By using that $E(\bar{X}) = \mu$ and $\text{Var}(\bar{X}) = \sigma^2/n$ we get for any $b > 0$

$$P(|\bar{X} - \mu| \geq b) \leq \frac{\sigma^2}{nb^2}.$$

This yields the inequality above. Moreover, as n goes to infinity the right-hand side converges to 0 and this proves the Law of large numbers.

Example 4. Consider a fair die whose probability to show a 6 is p. Roll the die n times, for each roll let $X_i = 1$ if the die shows a 6 and $X_i = 0$ otherwise, for $i = 1, \ldots, n$. The random variables $X_1, X_2, \ldots, X_n$ are i.i.d. and have a Bernoulli distribution with probability of success p. Recall that for each i we have

$$E(X_i) = 0 \times (1 - p) + 1 \times p = p.$$

Thus, according to the Law of large numbers we have that for any $b > 0$

$$\lim_{n \to \infty} P\left(|\frac{X_1 + X_2 + \cdots + X_n}{n} - p| \geq b\right) = 0.$$

This gives a physical interpretation to the notion of probability. When we say that the probability of a 6 is 1/6 (i.e., $p = 1/6$), it means that if we roll the die n times the ratio of the number of 6's that appear over n will approach 1/6 as n goes to infinity.

Example 5. Assume we roll a die 3,600 times and we get 557 6's. Let p be the probability of getting a 6. Find an interval that contains p with probability 0.95.

We use the Bernoulli random variables defined in Example 4. According to the inequality above, we have

$$P\left(|\frac{X_1 + X_2 + \cdots + X_n}{n} - p| \geq b\right) \leq \frac{\sigma^2}{b^2 n},$$

where σ^2 is the variance of each X_i. Since the X_i are Bernoulli random variables $\sigma^2 = p(1 - p)$. We do not know p (this is what we are estimating) so we do not know σ. However, $p(1 - p)$ is always less than 1/4 for p in [0,1] (graph $p(1 - p)$ as a function of p and you will see why). Hence,

$$\frac{\sigma^2}{b^2 n} \leq \frac{1}{4b^2 n}.$$

Therefore,

$$P\left(|\frac{X_1 + X_2 + \cdots + X_n}{n} - p| \geq b\right) \leq \frac{1}{4b^2 n}.$$

We want

$$\frac{1}{4b^2n} = 0.05.$$

Using that $n = 3{,}600$ we get $b = 0.04$. For this sample we have $\bar{X} = \frac{557}{3600}$. So the confidence interval with confidence at least 0.95 is

$$\begin{aligned}&\left(\frac{X_1 + X_2 + \cdots + X_n}{n} - b, \frac{X_1 + X_2 + \cdots + X_n}{n} + b\right)\\ &= (0.15 - 0.04, 0.15 + 0.04)\\ &= (0.11, 0.19).\end{aligned}$$

Such an interval (with a probability attached to it) is called a *confidence* interval.

The confidence interval above is obtained by using Chebyshev's inequality. This inequality holds for all random variables (that have a variance) and in particular for the most dispersed ones. As a consequence the confidence interval we got above is larger than it could be (its confidence is also larger than 0.95). The Central Limit Theorem that we will see in the next section will give us a narrower interval for 0.95 confidence.

Example 6. This example will introduce a numerical integration method called Monte Carlo integration. Let g be a continuous function on [0,1]. Let U_1, $U_2, \ldots, U_n$ be a sequence of i.i.d. uniform random variables on [0,1]. Observe that $g(U_1), g(U_2), \ldots, g(U_n)$ is also a sequence of i.i.d.random variables and the Law of large numbers can be applied to get

$$\lim_{n\to\infty} \frac{g(U_1) + \cdots + g(U_n)}{n} = E(g(U_1)).$$

Recall that if U_1 has density f then

$$E(g(U_1)) = \int g(x) f(x) \mathrm{d}x.$$

In this case $f(x) = 1$ for x in [0,1] and $f(x) = 0$ otherwise. Thus,

$$E(g(U_1)) = \int_0^1 g(x) \mathrm{d}x$$

and

$$\lim_{n\to\infty} \frac{g(U_1) + \cdots + g(U_n)}{n} = \int_0^1 g(x) \mathrm{d}x.$$

In other words, the average $\frac{g(U_1)+\cdots+g(U_n)}{n}$ approaches $\int_0^1 g(x)\mathrm{d}x$ as n goes to infinity. For instance, we take $g(x) = x$, $n = 10$ and use the random numbers: 0.382, 0.101, 0.596, 0.885, 0.899, 0.958, 0.014, 0.407, 0.863, 0.139. We get

$$\frac{g(U_1) + \cdots + g(U_n)}{n} = \frac{U_1 + \cdots + U_n}{n} = 0.52$$

Thus, 0.52 is the approximation we get for the integral

$$\int_0^1 g(x)\mathrm{d}x = \int_0^1 x\mathrm{d}x = 0.5.$$

Exercises 4.1

1. Assume that $X_1, \ldots, X_n$ is a sequence of i.i.d. random variables with mean 3 and standard deviation 2.

(a) What is the mean of $X_1 + X_2 + \cdots + X_n$?
(b) What is the mean of $\bar{X}$?
(c) What is the standard deviation of $X_1 + X_2 + \cdots + X_n$?
(d) What is the standard deviation of $\bar{X}$?

2. Find an upper bound on the probability that a positive random variable be 100 times larger than its mean.

3. Assume that the random variable X has mean 3 and standard deviation 2.

(a) Find an upper bound on the probability that X is at least 3σ away from its mean.
(b) Find an upper bound on the probability that X is larger than 11.

4. Let U be uniform in [0,1].

(a) Compute the probability that U be at least σ away from its mean.
(b) Use Chebyshev's inequality to give an upper bound on the probability that U be at least σ away from its mean.

5. Let S be a binomial random variable with $n = 10$ and $p = 0.2$.

(a) Compute the probability that S be at least 2σ away from its mean.
(b) Use Chebyshev's inequality to give an upper bound on the probability that S be at least 2σ away from its mean.

6. (a) Use Monte Carlo integration to estimate

$$\int_0^1 \mathrm{e}^{-\frac{x^2}{2}}\mathrm{d}x.$$

(b) Use the normal table to check the accuracy of the estimate in (a).

7. It is assumed that each line of a given document has a mean of 15 words.

(a) Find an upper bound on the probability that a given line has 30 words or more.
(b) Assume that the standard deviation is $\sigma = 3$. Find a better upper bound for the event in (a).

8. Consider a random variable X such that $P(X = -1) = P(X = 1) = 1/2$.

(a) Compute $E(X)$.
(b) Compute $\text{Var}(X)$.
(c) Compute $P(|X - \mu| \geq 1)$.
(d) Show that Chebyshev's inequality is an equality for $P(|X - \mu| \geq 1)$. (This shows that Chebyshev's inequality may not be improved if it is to hold for all random variables with a variance).

9. Show that if p belongs to [0,1] then $p(1 - p) \leq 1/4$.

10. This is concerned with Monte Carlo integration. Let g be a continuous function on [0,1] and let U be an uniform random variable. Show that $E(g(U))$ and $\text{Var}(g(U))$ exist. (Hence, the Law of Large Numbers applies to the sequence $g(U_1), g(U_2), \ldots, g(U_n) \ldots$).

11. Let $U_1, U_2, \ldots$ be a sequence of i.i.d. uniform random variables on $[0, 1]$.

(a) Show that

$$\lim_{n\to\infty} \frac{1}{n}\Big(\frac{1}{1+U_1} + \frac{1}{1+U_2} + \cdots + \frac{1}{1+U_n}\Big) = \ln 2.$$

(b) Do a simulation to check the result in (a).

12. Do a simulation to estimate the number π.

4.2 Central Limit Theorem

Consider a sequence $X_1, X_2, \ldots, X_n$ of i.i.d. random variables with mean μ and standard deviation σ. We have seen that $\bar{X}$ the sample average has mean μ and standard deviation $\sigma/\sqrt{n}$. This shows that $\bar{X}$ has a distribution that is more and more concentrated around the mean μ. Actually a lot more is true: the distribution of $\bar{X}$ approaches a normal distribution with mean μ and standard deviation $\sigma/\sqrt{n}$.

Central Limit Theorem

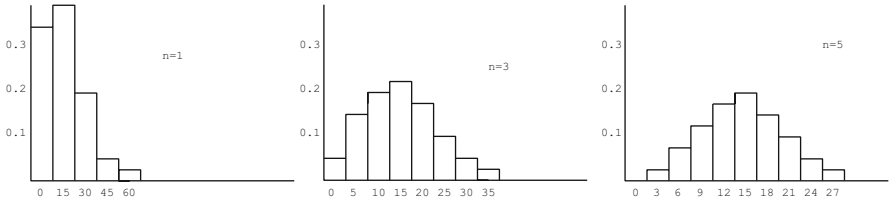
Let $X_1, X_2, \ldots, X_n$ be a sequence of independent identically distributed random variables with mean μ and variance σ^2. Then the distribution of $\bar{X} = \frac{X_1+X_2+\cdots+X_n}{n}$ approaches a normal distribution in the following sense. For any $a < b$ we have that

$$\lim_{n\to\infty} P\left(a < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < b\right) = P(a < Z < b),$$

where Z has a standard normal distribution. Equivalently, the sum $S = X_1 + X_2 + \cdots + X_n$ also approaches a normal distribution. That is,

$$\lim_{n\to\infty} P\left(a < \frac{S - n\mu}{\sigma\sqrt{n}} < b\right) = P(a < Z < b)$$

We will sketch a proof of the Central Limit Theorem in 7.2. The remarkable fact of this result is that it does not matter what the distribution of the X_i is (provided it has a variance) when we average or sum many i.i.d. random variables we get a normal distribution. The Central Limit Theorem (CLT in short) shows why the normal distribution is so crucial in Probability. We illustrate this Theorem with the histograms of $\bar{X}$ for $n = 1$, $n = 3$, and $n = 5$.



One can see above that when we average even a few random variables there is a departure from the original shape and there is a tendency toward the bell-like shape.

Note that in order to apply the CLT to a random variable Y (we have two choices for Y: $\bar{X}$ and S) we need to standardize it. The CLT states that the distribution of

$$\frac{Y - E(Y)}{SD(Y)} \text{ approaches the distribution of } Z.$$

In particular, if $Y = S = X_1 + X_2 + \cdots + X_n$ we have that $E(S) = n\mu$ and $\text{Var}(S) = n\sigma^2$. Thus, the distribution of

$$\frac{S - n\mu}{\sigma\sqrt{n}} \text{ approaches the distribution of } Z.$$

On the other hand if $Y = \bar{X}$ then we use that $E(\bar{X}) = \mu$ and that $SD(\bar{X}) = \sigma/\sqrt{n}$ to show that the distribution of

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \text{ approaches the distribution of } Z.$$

Example 1. We will show that the normal approximation to the binomial is a particular case of the CLT. Consider a binomial random variable S with parameters

n and p. Then, S can be written as a sum of n Bernoulli random variables X_i with probability of success p. We have $E(X_i) = p$ and $\text{Var}(X_i) = p(1-p)$. Since the X_i are i.i.d. we may apply the CLT to the sum of X_i to get that the distribution of

$$\frac{S - np}{\sqrt{np(1-p)}} \text{ approaches the distribution of } Z.$$

This is what the normal approximation to the binomial says. Recall that because we are using a continuous random variable to approach a discrete one we also enlarge the interval by 1/2 on both sides. That is, we replace $P(a \leq S \leq b)$ by $P(a-1/2 \leq S \leq b+1/2)$ before applying the Central Limit Theorem.

Example 2. Toss a fair coin. Each time it lands on heads you win \$1, each time it lands on tails I win \$1. What is the probability that after 100 tosses you will be winning at least 10\$?

Let T be your winnings after 100 tosses. We may write T as the sum of the winnings for each toss:

$$T = X_1 + \cdots + X_{100},$$

where X_i (for $1 \leq i \leq 100$) is 1 with probability 1/2 or -1 with probability 1/2. For each i we have that

$$E(X_i) = \frac{1}{2} \times 1 + \frac{1}{2} \times (-1) = 0.$$

We also have that

$$E(X_i^2) = \frac{1}{2} \times (1)^2 + \frac{1}{2} \times (-1)^2 = 1.$$

Since $E(X_i) = 0$, $\text{Var}(X_i) = E(X_i^2) = 1$. The random variables X_i are i.i.d. so we may use the CLT to get that the distribution of

$$\frac{T - n \times 0}{1\sqrt{100}} \text{ approaches the distribution of } Z.$$

Thus,

$$P(T \geq 10) = P\left(\frac{T}{10} \geq 1\right) \sim P(Z \geq 1) = 0.16.$$

So the probability that you are ahead by at least \$10 is about 0.16. Note that with the same reasoning we get that the probability that you are ahead by at least \$20 after 100 bets is about $P(Z \geq 2) = 0.02$. This is rather unlikely and if this happens one may start to get suspicious about the fairness of the coin.

Example 3. Assume we roll a die 3,600 times and we get 557 6's. Let p be the probability of getting a 6. Use the CLT to find an interval that contains p with probability 0.95.

This is the same question as in Example 5 in Sect. 4.1. We now have the CLT at our disposal so we will use it. Let $X_i = 1$ if the die shows a 6 and $X_i = 0$ otherwise, for $i = 1, \ldots, n$. The random variables $X_1, X_2, \ldots, X_n$ are i.i.d. and have a Bernoulli distribution with probability of success p. Recall that for each i we have

$$E(X_i) = p \text{ and } \operatorname{Var}(X_i) = p(1-p).$$

We want c so that

$$P(|\bar{X} - p| < c) = 0.95.$$

According to the CLT the distribution of

$$\frac{\bar{X} - p}{\sqrt{p(1-p)}/\sqrt{n}} \text{ approaches the distribution of } Z.$$

Thus,

$$P\left(\frac{|\bar{X} - p|}{\sqrt{p(1-p)}/\sqrt{n}} < \frac{c}{\sqrt{p(1-p)}/\sqrt{n}}\right) \sim P\left(|Z| < \frac{c}{\sqrt{p(1-p)}/\sqrt{n}}\right).$$

We use the normal table to get

$$\frac{c}{\sqrt{p(1-p)}/\sqrt{n}} = 1.96.$$

Thus,

$$c = 1.96\frac{\sqrt{p(1-p)}}{\sqrt{n}}.$$

Since we do not know p (this is what we are estimating) we use again that $\sqrt{p(1-p)} \leq 1/2$ to get that

$$c \leq \frac{1.96}{2\sqrt{3600}} = 0.0016.$$

The confidence interval is

$$\left(\frac{557}{3600} - 0.0016; \frac{557}{3600} + 0.0016\right).$$

Note that this interval is much narrower (and therefore better) than the interval we got in Example 5 in Sect. 4.1. This is so because CLT is a much stronger result than Chebyshev's inequality. The price to pay is that CLT is much more difficult to prove than Chebyshev's inequality.

Example 4. Assume that we have 25 batteries whose lifetime are exponentially distributed with mean 2 h. If the batteries are used one at the time, with a failed battery replaced immediately by a new one, what is the probability that after 50 h there is still a working battery?

Let X_i be the lifetime of the ith battery for $i = 1, \ldots, 25$. We want to compute

$$P(X_1 + \cdots + X_{25} > 50).$$

The distribution of a sum of exponentially distributed random variables is not exponentially distributed. To solve this question it is easier to use the CLT rather than use the exact distribution of the sum. The CLT applies since we have an i.i.d. sequence of random variables. Recall that for an exponential random variable the mean and the standard deviation are equal. Thus, in this case we have $\mu = \sigma = 2$. According to the CLT the distribution of

$$\frac{X_1 + \cdots + X_{25} - 25\mu}{\sigma\sqrt{25}} \text{ approaches the distribution of } Z.$$

We have that

$$\begin{aligned} P(X_1 + \cdots + X_{25} > 50) &= P\left(\frac{X_1 + \cdots + X_{25} - 25 \times 2}{2\sqrt{25}} > \frac{50 - 25 \times 2}{2\sqrt{25}}\right) \\ &\sim P(Z > 0) = 0.5. \end{aligned}$$

So there is a probability of around 50% that the batteries will last at least 50 h.

Example 5. We continue Example 4. How many batteries should we have so that there is still a working battery after 50 h with probability 0.9?

This time we are looking for n such that

$$P(X_1 + \cdots + X_n > 50) = 0.9,$$

where X_i is the lifetime of the ith battery. Again we use the CLT to get

$$\begin{aligned} P(X_1 + X_2 + \cdots + X_n > 50) &= P\left(\frac{X_1 + X_2 + \cdots + X_n - n\mu}{\sigma\sqrt{n}} > \frac{50 - n\mu}{\sigma\sqrt{n}}\right) \\ &\sim P\left(Z > \frac{50 - n\mu}{\sigma\sqrt{n}}\right) \end{aligned}$$

Since we want the probability above to be 0.9, we use the normal table to get

$$\frac{50 - n\mu}{\sigma\sqrt{n}} = -1.28.$$

We have that $\mu = \sigma = 2$. Set $x = \sqrt{n}$ and we get the quadratic equation

$$x^2 - 1.28x - 25 = 0.$$

The only positive solution is $x = \sqrt{n} = 5.68$, thus the smallest corresponding integer n is 33. That is, we need at least 33 batteries if we want that there is still a working battery after 50 h, with probability 0.9.

The CLT tells us that the approximate distribution of a sum (or average) of n i.i.d. random variables is normal when n is large. However, it does NOT say that every distribution is normal! If the sample size n is not large enough we need to have more information about the specific distribution we are dealing with and we cannot use the CLT.

Exercises 4.2

1. Assume that you bet 100 times \$1 on red at the roulette (probability of winning \$1 is 18/38).

(a) What are your expected winnings (or losses) after 100 bets?
(b) What is the probability that you are at least \$10 ahead after 100 bets?
(c) Compare (b) to Example 2.

2. Assume that we toss a coin 400 times and we get 260 heads.

(a) Give a confidence interval for p the probability of getting heads with confidence 0.99.
(b) Is this a fair coin?

3. A small airplane can take off with a maximum of 2,000 kg (no luggage!). Assume that passengers have a mean weight of 70 kg with a SD of 15 kg.

(a) What is the probability that 25 passengers will overload the plane?
(b) Find the maximum number of passengers that will not overload the plane with probability 0.99.

4. Assume that first graders have a mean height of 100 cm with SD of 8 cm.

(a) What is the probability that the average height in a class of 30 is over 105 cm?
(b) What is the probability that at least one child is more than 105 cm high?
(c) What assumption did you make to answer (b)?

5. How many times should you toss a fair coin in order to get at least 100 heads with probability 0.9?

6. A bank teller takes a mean of 2 min with a standard deviation of 30 s to serve a client. Assuming that there is at least one client waiting at all times, what is the probability that the teller will serve at least 25 clients in 1 h?

7. The average grade a professor hands out is 80 with SD of 10.

(a) What is the probability that in a class of 50 the average grade is below 75?
(b) How large should the class be so that the average grade is in the interval [75,85] with probability 0.95?

8. Roll a fair die. What is the probability that the sum of the first 100 rolls will be over 300?

9. Let $X_1, \ldots, X_n$ be a sequence of i.i.d. random variables with mean 0 and SD 5. Let S be the sum of the X_i.

(a) What is the probability that S exceeds 10 for $n = 100$?
(b) How large should n be so that at least one of the X_i is larger than 10?
(c) What assumption did you make to answer (b).

10. Example 3 shows that a confidence interval with confidence a for a proportion p has length $2c$ where

$$c = \frac{z_a}{2\sqrt{n}}$$

and z_a is such that

$$P(|Z| < z_a) = a.$$

(a) Does c increase or decrease as the sample size n increases?
(b) What is z_a for $a = 0.9$?
(c) Does c increase or decrease as the confidence a increases?

Chapter 5
Estimation and Hypothesis Testing

5.1 Large Sample Estimation

5.1.1 Confidence Interval for a Proportion

Example 1. In a poll of 100 randomly selected voters, 35 expressed support for initiative A. How does one estimate the proportion of voters in the whole population that supports initiative A based on the sample of 100? How much confidence do we have on our estimate?

Let p be the population proportion of voters in favor of A. A natural estimator for p is $\hat{p}$: the sample proportion of voters in favor of A. Note that p is an unknown constant while $\hat{p}$ is a random variable: every sample we pick gives a different $\hat{p}$.

Let $X_i = 0$ if the ith voter is against A and let $X_i = 1$ otherwise, for $i = 1, 2, \ldots, 100$. Then,

$$\hat{p} = \frac{X_1 + X_2 + \cdots + X_n}{n}.$$

Note that $X_1, X_2, \ldots, X_n$ are Bernoulli random variables and that $P(X_i = 1) = p$, the parameter we want to estimate. We assume that the sample is picked at random. That is, the X_i are independent and identically distributed. Recall that $E(X_i) = p$ and so by the linearity of the expectation we get

$$\begin{aligned} E(\hat{p}) &= E\left(\frac{X_1 + X_2 + \cdots + X_n}{n}\right) = \frac{1}{n} E(X_1 + X_2 + \cdots + X_n) \\ &= \frac{1}{n} n E(X_1) = E(X_1) = p. \end{aligned}$$

That is, the expected value of $\hat{p}$ is p, $\hat{p}$ is said to be an *unbiased* estimator of p. We also know, by the Law of large numbers, that if the observations X_i are i.i.d. then

R.B. Schinazi, *Probability with Statistical Applications*,
DOI 10.1007/978-0-8176-8250-7_5, © Springer Science+Business Media, LLC 2012

$\hat{p}$ converges to p as the sample size n increases. So at this point we may say that $\hat{p} = \frac{35}{100}$ is an estimate of p. But what confidence do we have in this estimate? In order to answer this question, we will now compute a confidence interval for p based on $\hat{p}$. That is, we would like to find c such that p is in the interval $(\hat{p}-c, \hat{p}+c)$ with probability 0.95. We need the variance of $\hat{p}$. We use that the X_i are independent, that the variance is a quadratic operator, and that the variance of a Bernoulli random variable is $p(1-p)$ to get

$$\begin{aligned}\text{Var}(\hat{p}) &= \text{Var}\left(\frac{X_1 + X_2 + \cdots + X_n}{n}\right) = \frac{1}{n^2}\text{Var}(X_1 + X_2 + \cdots + X_n) \\ &= \frac{1}{n^2} n \text{Var}(X_1) = \frac{p(1-p)}{n}.\end{aligned}$$

We start by writing that c should be such that

$$P(|\hat{p} - p| < c) = 0.95.$$

If the sample size n is large enough we know, by the Central Limit Theorem, that

$$\frac{\hat{p} - p}{\sqrt{p(1-p)/n}} \text{ has an approximately standard normal distribution.}$$

Hence,

$$P(|\hat{p} - p| < c) = P\left(\left|\frac{\hat{p} - p}{\sqrt{p(1-p)/n}}\right| < \frac{c}{\sqrt{p(1-p)/n}}\right) = 0.95.$$

Let Z be a standard normal random variable, we get

$$P\left(|Z| < \frac{c}{\sqrt{p(1-p)/n}}\right) = 0.95.$$

Using the normal table we have that

$$\frac{c}{\sqrt{p(1-p)/n}} = 1.96.$$

The above equation has two unknowns: n and p. If n is large enough it is reasonable to estimate p by $\hat{p}$. We get that c is approximately

$$c = 1.96\frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}}.$$

Numerically, we get that $c = 0.09$. Therefore, we may say that with confidence 0.95 the population proportion is in the interval

$$(\hat{p} - c, \hat{p} + c) = (0.35 - 0.09, 0.35 + 0.09) = (0.26, 0.44).$$

One interpretation for the confidence interval above is the following. If we take many samples of 100 voters then 95% of the confidence intervals we get contain p. Of course, we may be unlucky and draw a sample that will yield an interval that does not contain p. This will happen 5% of the time.

Note also that the computations above work only for RANDOM samples. Asking the opinion of your 100 best friends does not work! One way to draw a random sample from a population is to label all the population and then pick labels at random to get a sample. This is more or less what is done with political polls: phone numbers are selected at random to make up a sample. However, more and more people have only cellular phones and they will not be selected if only land lines are picked in the sample. Since cellular phones only households tend to be younger this could skew the sample toward older people. There are many other things to be cautious about when designing a statistical experiment, see for instance "Introduction to the practice of Statistics" by Moore and McCabe, Freeman.

We now give the general form of a confidence interval for a proportion.

Confidence Interval for a Proportion

Draw a random sample of size n from a large population with unknown proportion p of successes. Let $\hat{p}$ be the sample proportion of successes. Then, for large n

$$(\hat{p} - c, \hat{p} + c)$$

is a confidence interval with confidence a where

$$c = z_a \frac{\sqrt{\hat{p}(1 - \hat{p})}}{\sqrt{n}},$$

where z_a is such that

$$P(|Z| < z_a) = a$$

and Z is a standard normal distribution.

Example 2. Find a confidence interval with confidence 0.99 for the proportion in Example 1.

The only difference with Example 1 is the level of confidence. This time $a = 0.99$. According to the normal table

$$P(|Z| < 2.57) = 0.99$$

so

$$c = 2.57\frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}}.$$

Numerically, we get $c = 0.12$. Therefore, we may say that with confidence 0.99 the population proportion is in the interval

$$(\hat{p} - c, \hat{p} + c) = (0.35 - 0.12, 0.35 + 0.12) = (0.23, 0.47).$$

Note that at the level 0.99 we get a larger confidence interval. So we increased the confidence (from 95 to 99%) but we decreased the precision (we got a larger interval). The only way to increase the confidence without decreasing the precision is to increase the sample size.

Example 3. How large should a random sample be to get an estimate of the population proportion within 0.01 with confidence 0.95?

We want to know how large n should be in order to get $c = 0.01$. Since the confidence is $a = 0.95$ we get that $z_a = 1.96$. We need to solve in n the following equation

$$c = z_a\frac{\sqrt{p(1-p)}}{\sqrt{n}}.$$

Here we do not know $\hat{p}$ so we use the original p in our formula. A little algebra yields

$$n = \left(\frac{z_a}{c}\right)^2 p(1-p).$$

However, we do not know p. Note that p is in [0,1] and that the function $g(p) = p(1-p)$ has a maximum for $p = 1/2$. Thus,

$$p(1-p) \leq \frac{1}{4} \text{ for all } p \text{ in}[0,1].$$

We get that

$$n \leq \left(\frac{z_a}{c}\right)^2 \frac{1}{4}.$$

Numerically, we get

$$n \leq 9{,}604.$$

That is, in order to get a precision of 0.01 with confidence 0.95 we need a sample of the order of 10,000. Note that this estimate is based on a worst case scenario: we did the computation assuming $p = 0.5$. If we assume $p = 0.1$ for instance then $n = 3{,}457$, about three times smaller! In practice we use $p = 0.5$ when we have no idea what to expect for p.

5.1.2 *Confidence Interval for a Mean*

Example 4. Assume that 500 lamp bulbs have been tested and the average lifetime for this sample has been 562 days. Give a confidence interval at the level 90% for the mean lifetime of this brand of lamp bulb.

We assume that we have a random sample. That is, if we denote by $X_1, X_2, \ldots, X_{500}$ the 500 lifetimes observed in the sample then we assume that this is an i.i.d. sequence of random variables. Denote the mean lifetime by μ and the corresponding standard deviation by σ. We want to estimate μ. A natural estimator for μ is the sample average

$$\bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}.$$

Recall from Chap. 4 that

$$E(\bar{X}) = \mu \text{ and } \operatorname{Var}(\bar{X}) = \frac{\sigma^2}{n}.$$

That is, $\bar{X}$ is an unbiased estimator of μ. One way to measure the precision of our estimator is to compute $E((\bar{X} - \mu)^2)$. This is one way to measure the distance between the random variable $\bar{X}$ and the constant μ. Since the expected values of $\bar{X}$ is μ we get by the definition of the variance that

$$E((\bar{X} - \mu)^2) = \operatorname{Var}(\bar{X}) = \frac{\sigma^2}{n}.$$

That is, the precision of our estimator $\bar{X}$ increases as the sample size n increases. In order to compute a confidence interval for μ we need c such that

$$P(|\bar{X} - \mu| < c) = 0.9.$$

We standardize the left-hand side

$$P\left(\frac{|\bar{X} - \mu|}{\sigma/\sqrt{n}} < \frac{c}{\sigma/\sqrt{n}}\right) = 0.9.$$

Since we are assuming that the X_i are i.i.d. and n is large we may use the Central Limit Theorem to get that

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \text{ has an approximately standard normal distribution.}$$

Let Z be a standard normal random variable, we have

$$P\left(|Z| < \frac{c}{\sigma/\sqrt{n}}\right) = 0.9.$$

From the normal table we get

$$\frac{c}{\sigma/\sqrt{n}} = 1.64.$$

However, we usually do not know σ. We will show later on that

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

is an unbiased estimator for σ^2.

We use S to estimate σ and get that

$$c = \frac{1.64S}{\sqrt{n}}.$$

Going back to the observations $X_1, X_2, \ldots, X_n$ one may compute S. Assume that in this example $S = 112$ days. Then, $c = 8$. Thus, a confidence interval for the mean lifetime μ of a lamp bulb, at the 90% level, is

$$(562 - 8, 562 + 8) = (556, 570).$$

Note that this is rather precise thanks to the large size of the sample.

Confidence Interval for a Mean

Draw a random sample of size n, $X_1, X_2, \ldots, X_n$, from a large population with unknown mean μ. Let $\bar{X}$ be the sample mean. Let

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

be the sample variance. Then, for large n

$$(\bar{X} - c, \bar{X} + c)$$

is a confidence interval for μ with confidence a where

$$c = z_a \frac{S}{\sqrt{n}}$$

and z_a is such that

$$P(|Z| < z_a) = a$$

and Z is a standard normal distribution.

Example 5. Find a confidence interval for the mean in Example 4 with confidence 0.95

We have that

$$c = z_a \frac{S}{\sqrt{n}} = 1.96 \frac{112}{\sqrt{500}} \sim 10.$$

So at the level 0.95 we have the confidence interval

$$(562 - 10{,}562 + 10) = (552{,}572)$$

for the mean lifetime of a lamp bulb.

We now go back to the sample standard deviation S. We first establish a computational formula for S. We expand the square below to get

$$\sum_{i=1}^{n} (X_i - \bar{X})^2 = \sum_{i=1}^{n} X_i^2 - 2 \sum_{i=1}^{n} X_i \bar{X} + \sum_{i=1}^{n} (\bar{X})^2.$$

Using that $\sum_{i=1}^{n} X_i = n\bar{X}$ we get that

$$\sum_{i=1}^{n} (X_i - \bar{X})^2 = \sum_{i=1}^{n} X_i^2 - 2n(\bar{X})^2 + n(\bar{X})^2 = \sum_{i=1}^{n} X_i^2 - n(\bar{X})^2.$$

Thus, we get the following computational formula for S^2:

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} X_i^2 - \frac{n}{n-1} (\bar{X})^2.$$

We use the preceding formula to compute the expected value of S^2. First, recall that for any random variable X (that has a variance)

$$\text{Var}(X) = E(X^2) - E(X)^2.$$

Hence,

$$E(X^2) = \text{Var}(X) + E(X)^2. \tag{5.1}$$

Going back to

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} X_i^2 - \frac{n}{n-1} (\bar{X})^2$$

and taking expectations we have

$$E(S^2) = \frac{1}{n-1} \sum_{i=1}^{n} E(X_i^2) - \frac{n}{n-1} E((\bar{X})^2).$$

Using formula (5.1) we get

$$E(X_i^2) = \text{Var}(X_i) + E(X_i)^2 = \sigma^2 + \mu^2,$$

and

$$E((\bar{X})^2) = \text{Var}(\bar{X}) + E(\bar{X})^2 = \frac{\sigma^2}{n} + \mu^2.$$

Hence,

$$E(S^2) = \frac{1}{n-1} n(\sigma^2 + \mu^2) - \frac{n}{n-1}\left(\frac{\sigma^2}{n} + \mu^2\right) = \sigma^2.$$

This shows that S^2 is an unbiased estimator of σ^2.

Sample Variance

Let $X_1, X_2, \ldots, X_n$ be i.i.d. random variables with mean μ and variance σ^2 from a large population with unknown mean μ. Then, the sample variance

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$$

is an unbiased estimator of σ^2. Moreover, we have the following computational formula for S^2:

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} X_i^2 - \frac{n}{n-1} (\bar{X})^2.$$

5.1.3 Confidence Interval for a Difference of Proportions

Example 6. In a political poll of 100 randomly selected voters, 35 expressed support for initiative A in Boulder. In Colorado Springs in a poll of 200 randomly selected voters, 50 expressed support for initiative A. Find a confidence interval, with confidence 0.9, for the difference between the proportions of supporters of initiative A in Boulder and in Colorado Springs.

Let p_1 and p_2 be the proportions of the population in Boulder and in Colorado Springs that support A, respectively. Let n_1 and n_2 be the sample sizes taken in Boulder and Colorado Springs, respectively. We would like a confidence interval

for $p_1 - p_2$ with confidence 0.9. A natural estimator for $p_1 - p_2$ is $\hat{p}_1 - \hat{p}_2$. We would like to find c so that

$$P(|(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)| < c) = 0.9.$$

In order to find c we need some information regarding the distribution of $\hat{p}_1 - \hat{p}_2$. We know that if the sample size n_1 is large enough then by the Central Limit Theorem $\hat{p}_1$ is approximately normally distributed. The same holds for $\hat{p}_2$. Since $\hat{p}_1$ and $\hat{p}_2$ are independent one can show that $\hat{p}_1 - \hat{p}_2$ is approximately normally distributed as well. Since $\hat{p}_1$ and $\hat{p}_2$ are unbiased estimators of p_1 and p_2, respectively, we have that

$$E(\hat{p}_1 - \hat{p}_2) = p_1 - p_2.$$

Since $\hat{p}_1$ is the average of n_1 i.i.d. Bernoulli random variables we have that

$$\text{Var}(\hat{p}_1) = \frac{1}{n_1^2} n_1 p_1(1 - p_1) = \frac{p_1(1 - p_1)}{n_1}.$$

Using that $\hat{p}_1$ and $\hat{p}_2$ are independent we get that

$$\text{Var}(\hat{p}_1 - \hat{p}_2) = \text{Var}(\hat{p}_1) + \text{Var}(\hat{p}_2) = \frac{p_1(1 - p_1)}{n_1} + \frac{p_2(1 - p_2)}{n_2}.$$

We are now ready to normalize to get

$$P\left(|\frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}}| < \frac{c}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}} \right) = 0.9.$$

We use that

$$\frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}} \text{ is approximately a standard normal distribution}$$

to get that

$$P\left(|Z| < \frac{c}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}} \right) = 0.9,$$

where Z is a standard normal random variable. Using the normal table we get

$$\frac{c}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}} = 1.64.$$

For n_1 and n_2 large enough we may use $\hat{p}_1$ and $\hat{p}_2$ to approximate p_1 and p_2, respectively. Thus,

$$c \sim 1.64\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}.$$

In this example we have $n_1 = 100$, $n_2 = 200$, $\hat{p}_1 = 0.35$ and $\hat{p}_2 = 0.25$. Thus, $c = 0.09$. At the level 90% the confidence interval for $p_1 - p_2$ is

$$(\hat{p}_1 - \hat{p}_2 - c, \hat{p}_1 - \hat{p}_2 + c) = (0.01, 0.19).$$

We now summarize the preceding technique.

Confidence Interval for the Difference Between Two Proportions

Draw a random sample of size n_1 from a large population with unknown proportion p_1 of successes and an independent random sample of size n_2 from another large population having a proportion p_2 of successes. For large n_1 and large n_2

$$(\hat{p}_1 - \hat{p}_2 - c, \hat{p}_1 - \hat{p}_2 + c)$$

is a confidence interval with confidence a where

$$c = z_a\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

and z_a is such that

$$P(|Z| < z_a) = a$$

and Z is a standard normal distribution.

5.1.4 Confidence Interval for a Difference of Two Means

Example 7. Assume that $n_1 = 500$ lamp bulbs from brand 1 have been tested. The average lifetime for this sample is $\bar{X}_1 = 562$ days the standard deviation for the sample is $S_1 = 112$. Similarly, $n_2 = 300$ lamp bulbs from brand 2 have been tested. The average lifetime for this sample is $\bar{X}_2 = 551$ days and the standard deviation for the sample is $S_2 = 121$. Give a confidence interval at the level 95% for the difference of mean lifetimes of the two brands of lamp bulb. Is there evidence that brand 1 lasts longer than brand 2?

Let μ_1 and μ_2 be the unknown mean lifetimes of brands 1 and 2, respectively. We would like a confidence interval for $\mu_1 - \mu_2$. We use $\bar{X}_1 - \bar{X}_2$ as an estimator of $\mu_1 - \mu_2$. We want c such that

$$P(|(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)| < c) = 0.95.$$

In order to find c we need to know the distribution of the expression above. By the Central Limit Theorem $\bar{X}_1$ and $\bar{X}_2$ are approximately normally distributed if the samples sizes are large enough. If the two samples are independent then one can show that $\bar{X}_1 - \bar{X}_2$ is also approximately normally distributed. Since $\bar{X}_1$ and $\bar{X}_2$ are unbiased estimators of μ_1 and μ_2 we have that

$$E(\bar{X}_1 - \bar{X}_2) = \mu_1 - \mu_2.$$

Since $\bar{X}_1$ and $\bar{X}_2$ are independent we get

$$\mathrm{Var}(\bar{X}_1 - \bar{X}_2) = \mathrm{Var}(\bar{X}_1) + \mathrm{Var}(\bar{X}_2) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}.$$

We are now ready to normalize to get

$$P\left(\frac{|(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)|}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}} < \frac{c}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}\right) = 0.95.$$

By the normal approximation,

$$P\left(|Z| < \frac{c}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}\right) = 0.95.$$

According to the normal table,

$$\frac{c}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}} = 1.96.$$

However, the variances σ_1^2 and σ_2^2 are not known. If n_1 and n_2 are large it is reasonable to use the sample variances S_1^2 and S_2^2 in order to estimate them.

$$c \sim 1.96\sqrt{S_1^2/n_1 + S_2^2/n_2}.$$

Numerically, we get $c \sim 17$. Thus, the confidence interval for $\mu_1 - \mu_2$ at the level 0.95 is

$$(\bar{X}_1 - \bar{X}_2 - c, \bar{X}_1 - \bar{X}_2 + c) = (-6, 28).$$

Note that 0 is in the above interval. This shows that μ_1 and μ_2 could be equal. So there is no evidence that brand 1 lasts longer than brand 2.

Confidence Interval for the Difference Between Two Means

Draw a random sample of size n_1 from a large population with unknown mean μ_1 and an independent random sample of size n_2 from another population having mean μ_2. We denote the sample average by $\bar{X}_i$ and the sample standard deviation by S_i, for $i = 1, 2$. For large n_1 and large n_2

$$(\bar{X}_1 - \bar{X}_2 - c, \bar{X}_1 - \bar{X}_2 + c)$$

is a confidence interval with confidence a where

$$c = z_a \sqrt{S_1^2/n_1 + S_2^2/n_2}$$

and z_a is such that

$$P(|Z| < z_a) = a$$

and Z is a standard normal distribution.

Exercises 5.1

1. Consider the following scores: 87, 92, 58, 64, 72, 43, 75. Compute the average score $\bar{X}$ and the standard deviation S.

2. The English statistician Karl Pearson once tossed a coin 24,000 times and obtained 12,012 heads. Find a confidence at the level 0.99 for the probability of heads.

3. A poll institute claims that its estimate of a proportion is within 0.02 of the true value with confidence 0.95. How large must the sample be?

4. A poll institute has interviewed 1,000 people and gives an estimate of a proportion within 0.01. What is the confidence of this estimate?

5. Of 250 Ponderosa pines attacked with a certain type of beetle, 34 died. Find a confidence interval at the level 0.9 for the proportion of trees that die when attacked by this type of beetle.

6. The heights of 25 6-year-old boys average 85 cm with a standard deviation of 5 cm. Find a confidence interval at the level 0.95 for the mean height of a 6-year-old boy of that population.

7. A researcher has measured the yields of 40 tomato plants and found that the sample average yield per plant to be 5 pounds with a sample standard deviation of 1.7 pound. Find a confidence interval for the mean yield at the level 0.9.

8. The English statistician Karl Pearson once tossed a coin 24,000 times and obtained 12,012 heads. The English mathematician John Kerrich tossed a coin 10,000 times and obtained 5,067. Find a confidence at the level 0.99 for the difference of the probabilities of heads for the two coins.

9. The same final exam is given to several sections of calculus. Each professor gets to grade 50 papers taken at random from the pile. Professor A has an average of 75 with a standard deviation of 12. Professor B has an average of 79 with a standard deviation of 8.

(a) Find a confidence interval for the difference in mean scores between Professors A and B.
(b) Is there evidence that Professor A is harsher than Professor B?

10. In March a poll indicates that 104 out of 250 voters are in favor of initiative A. In October another (independent of the first one) indicates that 140 out of 300 voters are in favor of initiative A.

(a) Find a confidence interval for the difference of proportions of voters in favor of initiative A in March and October.
(b) Based on (a) would you say that there is statistical evidence that support has increased for initiative A?

11. A researcher wants to compare the yield of two varieties of tomatoes. The first variety of 40 tomato plants has a sample average yield per plant of 5 pounds with a sample standard deviation of 1.7 pound. The second variety of 50 tomato plants has a sample average yield per plant of 4.5 pounds with a sample standard deviation of 1.2 pound.

(a) Find a confidence interval for the difference in mean yield at the level 0.9.
(b) Based on (a), would you say that variety 1 yields more than variety 2?

12. Consider a random sample $X_1, \ldots, X_n$ of size of a uniform random variable on $[0, a]$. Recall that $E(X_1) = a/2$.

(a) Find an unbiased estimator $\hat{a}$ for a.
(b) Compute $E((\hat{a} - a)^2)$ (this indicates how close $\hat{a}$ is to a).

13. Compute the expected value of

$$\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

(You can use that $E(S^2) = \sigma^2$).

5.2 Hypothesis Testing

5.2.1 Testing a Proportion

Example 1. A manufacturer claims that he produces strictly less than 5% defective items. A sample of 100 items is taken at random and 4 are found to be defective. Test the claim of the manufacturer.

We denote the true proportion of defective items by p. The claim of the manufacturer is that $p < 0.05$. We want to test whether this claim holds based on the observations. There are two possible errors. We may reject the claim of the manufacturer although it is true or we may accept the claim of the manufacturer although it is not true. The test we will perform is not symmetric and the errors cannot be both small. We set up the test so that the error we minimize is the one that accepts the claim of the manufacturer although it is not true. This is so because we are testing the manufacturer's claim and he should have the burden of proof. The manufacturer's claim is called the *alternative hypothesis* and it is denoted by H_a. The negation of this claim is called the *null hypothesis* and is denoted by H_0. So the test we would like to perform is

$$H_0 : p \geq 0.05.$$

$$H_a : p < 0.05.$$

It is convenient to have an equality for the null hypothesis. It turns out that it is always possible to replace an inequality by an equality in the null hypothesis without changing the test. The reason is a little involved so we will omit it but we will actually test

$$H_0 : p = 0.05.$$

$$H_a : p < 0.05.$$

Hypothesis Testing

The claim you want to test should be your alternative hypothesis and is denoted by H_a. The negation of that claim is called the null hypothesis and is denoted by H_0. The test is determined by the level of significance. This is the probability of making the so-called error I: rejecting H_0 although H_0 is true.

We need to make a decision: reject H_0 (the manufacturer's claim is accepted) or do not reject H_0 (the manufacturer's claim is not accepted). We make this decision based on the observations. Given that the alternative hypothesis is $p < 0.05$, we will reject H_0 is the sample proportion is low. We define the *rejection region* (the region where the null hypothesis is rejected) to be

$$R = \{\hat{p} < 0.04\},$$

where 0.04 is the sample proportion for this particular example. We now compute the probability of error I (reject the null hypothesis when in fact the null hypothesis is true). This is the so-called P value.

$$P = P(\text{reject } H_0 | H_0 \text{ is true}) = P(\hat{p} < 0.04 | p = 0.05).$$

Since the sample is large enough we may use the CLT to get

$$P = P\left(\frac{\hat{p} - 0.05}{\sqrt{0.05(0.95)/n}} < \frac{0.04 - 0.05}{\sqrt{0.05(0.95)/n}}\right) = P(Z < -0.46) = 0.32.$$

This computation shows that the probability of making error I is pretty high: 32%. What this test is telling us is that although we observe 4% of defective items in the sample there is a high probability (32%) that this was due to chance and that the actual proportion of defective items is equal to or larger than 5%. Therefore, we do not reject the null hypothesis. We conclude that there is not enough evidence to support the claim of the manufacturer.

P Value and Significance Level

The P value of a test is the probability that we make error I: reject H_0 although H_0 is true. We give ourselves a significance level a (usually 5%). To perform a test at the significance level a we do the following. If $P < a$ we reject H_0. If the $P > a$ we do not reject H_0.

We summarize the P value method for testing a proportion.

P Value for Testing a Proportion

Assume we have a large random sample with a proportion $\hat{p}$ of successes. We use this sample to test the true proportion of successes p. Let p_0 be a fixed number in [0,1]. For the test

$$H_0 : p = p_0$$
$$H_a : p < p_0$$

and a sample size n large enough the P value is

$$P = P\left(Z < \frac{\hat{p} - p_0}{\sqrt{p_0(1 - p_0)/n}}\right),$$

where $\hat{p}$ is the sample proportion and n is the size of the random sample. For the test

$$H_0 : p = p_0$$
$$H_a : p > p_0$$

and a sample size n large enough the P value is

$$P = P\left(Z > \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}}\right).$$

Example 2. Candidate A claims that more than 10% of the voters are in his favor. In a poll candidate A got 121 votes in a random sample of 1,000. Test the claim of A.

Let p be the proportion of voters in favor of candidate A. The alternative hypothesis should be $p > 0.1$ since this is the claim we want to test. Therefore, the test is

$$H_0 : p = 0.1$$
$$H_a : p > 0.1.$$

Since the alternative hypothesis is $p > 0.1$ we will reject the null hypothesis if the sample proportion $\hat{p}$ is too large. Hence, the rejection region is

$$\left\{\hat{p} > \frac{121}{1{,}000}\right\}.$$

The P value is

$$P = P\left(Z > \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}}\right)$$
$$= P\left(Z > \frac{0.121 - 0.1}{\sqrt{0.1(0.9)/(1{,}000)}}\right) = P(Z > 2.21) = 0.01.$$

Since $P < 0.05$, at the level 5% we reject H_0: there is statistical evidence supporting the claim of candidate A.

5.2.2 Testing a Mean

Example 3. A manufacturer of lamp bulbs claims that the mean lifetime of his lamp bulbs is 1,000 h. The average lifetime in a sample of 200 bulbs is 1,016 h with a standard deviation of 102 h. Test the claim of the manufacturer.

Let μ be the true lifetime mean of a lamp bulb. The claim of the manufacturer is that $\mu > 1{,}000$. So this should be our alternative hypothesis. Therefore, the test is going to be

$$H_0 : \mu = 1{,}000$$

$$H_a : \mu > 1{,}000.$$

To take our decision on μ we use $\bar{X}$: the average lifetime in the sample. Given that the alternative hypothesis is $\mu > 1,000$ the rejection region is

$$R = \{\bar{X} > 1{,}016\}.$$

We compute the P value.

$$P = P(\bar{X} > 1{,}016 | \mu = 1{,}000).$$

If the sample size is large enough we may use the CLT to get

$$P = P\left(\frac{\bar{X} - 1{,}000}{S/\sqrt{n}} > \frac{1{,}016 - 1{,}000}{S/\sqrt{n}}\right) \sim P(Z > 2.21) = 0.01.$$

So at any level larger than 0.01 we reject H_0. In particular at the standard level 0.05 we reject H_0. There is statistical evidence supporting the manufacturer's claim.

As we mentioned before the test is determined by the error I level. The other possible error is the so-called error II: reject H_a when H_a is true. We now compute the probability of error II. The probability of error II can be computed for any $\mu > 1{,}000$ (since this is the alternate hypothesis). For this example we pick $\mu = 1{,}020$. In order to make an error II the sample needs to be outside the rejection region (so that we do not reject H_0). Therefore,

$$P(\text{error II}) = P(\bar{X} \leq 1{,}016 | \mu = 1{,}020) = P\left(\frac{\bar{X} - 1{,}020}{S/\sqrt{n}} \leq \frac{1{,}016 - 1{,}020}{S/\sqrt{n}}\right).$$

That is,

$$P(\text{error II}) = P(Z < -1.94) = 0.03.$$

P Value for Testing a Mean

Assume we have a large random sample with average $\bar{X}$ and standard deviation S. We would like to use this sample to test the true mean of the population μ. Let μ_0 be a fixed number. For the test

$$H_0 : \mu = \mu_0$$

$$H_a : \mu < \mu_0$$

the P value is

$$P = P\left(Z < \frac{\bar{X} - \mu_0}{S/\sqrt{n}}\right).$$

For the test

$$H_0 : \mu = \mu_0$$

$$H_a : \mu > \mu_0$$

the P value is

$$P = P\left(Z > \frac{\bar{X} - \mu_0}{S/\sqrt{n}}\right).$$

Example 4. A farmer is supposed to deliver chickens to a grocery store that weigh 3 pounds in average. The grocery store claims that the chickens are in average under 3 pounds. A random sample of 100 chicken has an average of 46 ounces and a standard deviation of 5 ounces. Test the claim of the store.

The claim we want to test is $\mu < 48$. This should be our alternative hypothesis. So we perform the test

$$H_0 : \mu = 48$$

$$H_a : \mu < 48.$$

We compute the P value.

$$P = P\left(Z < \frac{\bar{X} - \mu_0}{S/\sqrt{n}}\right) = P\left(Z < \frac{46 - 48}{5/\sqrt{100}}\right) = P(Z < -4).$$

This P value is practically 0. So at any reasonable level (1, 5, or 10%) we should reject the null hypothesis. There is strong statistical evidence to support the claim of the grocery store.

All the tests we have performed so far were one-sided. Next we give an example of a two-sided test.

Example 5. We want to test whether a certain medication has an effect on the blood pressure. A group of 100 male patients 35–44 years is given the medication. The average blood pressure in this group is 120 with a standard deviation of 10. For this age range the blood pressure in the population is known to be 128.

Since we are told to just test any effect we should test

$$H_0 : \mu = 128$$
$$H_a : \mu \neq 128$$

The rejection region is two-sided:

$$R = \{\bar{X} < 120 \text{ or } \bar{X} > 136\}.$$

The value 120 is as always what was observed in the sample and since $128-120 = 8$ we get the value 136 by doing $128 + 8$. We now compute the P value

$$P = P(\bar{X} < 120 \text{or} \bar{X} > 136|\mu = 128) = 2P(\bar{X} > 136|\mu = 128)$$

since the rejection region is symmetric with respect 128. Therefore,

$$P = 2P\left(Z > \frac{136-128}{S/\sqrt{n}}) = 2P(Z > 8\right).$$

This P value is extremely small. We reject the null hypothesis, there is strong evidence that the medication has an effect on blood pressure.

5.2.3 Testing Two Proportions

Example 6. Test whether candidate A has more support in Colorado Springs than in Boulder. In a poll of 1,000 voters candidate A got 42% of the votes in Colorado Springs. In Boulder he got 39% of the votes in a poll of 500 voters.

Let p_1 and p_2 be respectively the true proportions of voters in favor of A in Colorado Springs and in Boulder. We want to test whether $p_1 > p_2$. So we want to perform the test

$$H_0 : p_1 = p_2$$
$$H_a : p_1 > p_2.$$

It is convenient to observe that this test can be expressed as a one parameter test with parameter $p_1 - p_2$ by writing it as

$$H_0 : p_1 - p_2 = 0$$
$$H_a : p_1 - p_2 > 0$$

The parameter $p_1 - p_2$ is estimated by $\hat{p}_1 - \hat{p}_2$ and for these samples we have $\hat{p}_1 - \hat{p}_2 = 0.42 - 0.39 = 0.03$. Therefore the rejection region is

$$R = \{\hat{p}_1 - \hat{p}_2 > 0.03\}.$$

We compute the P value for this test.

$$P = P(\hat{p}_1 - \hat{p}_2 > 0.03 | p_1 - p_2 = 0).$$

We use that

$$E(\hat{p}_1 - \hat{p}_2) = p_1 - p_2 \text{ and } \text{Var}(\hat{p}_1 - \hat{p}_2) = p_1(1 - p_1)/n_1 + p_2(1 - p_2)/n_2.$$

For n_1 and n_2 large and if the two random samples are independent we may use the CLT to get

$$P = P\left(Z > \frac{0.03}{\sqrt{p_1(1 - p_1)/n_1 + p_2(1 - p_2)/n_2}} | p_1 - p_2 = 0\right).$$

We need to estimate p_1 and p_2 in the expression. Given that we are assuming that $p_1 = p_2$ we use the pooled estimate

$$\hat{p} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2}.$$

Thus,

$$P = P\left(Z > \frac{0.03}{\sqrt{\hat{p}(1 - \hat{p})(1/n_1 + 1/n_2)}}\right).$$

Numerically, we get $\hat{p} = 0.41$ and that $P = P(Z > 1.11) = 0.13$. At the 5% (or 10%) level we do not reject H_0. There is no evidence that the support of candidate A is larger in Colorado Springs than in Boulder.

P Value for Testing Two Proportions

We have two independent random samples of size n_1 and n_2 respectively from two distinct populations. Let p_1 and p_2 be respectively the true proportions of

successes in populations 1 and 2. Let $\hat{p}_1$ and $\hat{p}_2$ be the corresponding sample proportions. For the test

$$H_0 : p_1 = p_2$$
$$H_a : p_1 < p_2$$

the P value is

$$P = P\left(Z < \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})(1/n_1 + 1/n_2)}}\right),$$

where

$$\hat{p} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2}.$$

For the test

$$H_0 : p_1 = p_2$$
$$H_a : p_1 > p_2$$

the P value is

$$P = P\left(Z > \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})(1/n_1 + 1/n_2)}}\right).$$

5.2.4 Testing Two Means

Example 7. Test the claim that lamp bulbs from A last longer than lamp bulbs from B. A sample of 200 lamp bulbs from A gave a sample average of 1,052 h and a standard deviation of 151. A sample of 100 lamps from B gave a sample average of 980 h and a standard deviation of 102.

Let μ_1 and μ_2 be respectively the mean lifetimes of the lamp bulbs from manufacturers A and B. We want to test whether $\mu_1 > \mu_2$. This is our alternative hypothesis. We perform the test

$$H_0 : \mu_1 = \mu_2$$
$$H_a : \mu_1 > \mu_2$$

We rewrite the test as a one parameter test

$$H_0 : \mu_1 - \mu_2 = 0$$
$$H_a : \mu_1 - \mu_2 > 0$$

Let n_1 and n_2 be the sample sizes from A and B, respectively. We denote the sample averages from A and B by $\bar{X}_1$ and $\bar{X}_2$, respectively, and the sample standard deviations by S_1 and S_2. The rejection region is of the type

$$R = \{\bar{X}_1 - \bar{X}_2 > c\}.$$

We compute the P value so we take $c = 1,052 - 980 = 72$. We have that

$$P = P(\bar{X}_1 - \bar{X}_2 > 72 | \mu_1 - \mu_2 = 0).$$

Assuming the sample sizes are large enough and that the two random samples are independent we get by the Central Limit Theorem that

$$P \sim P\left(Z > \frac{72}{\sqrt{S_1^2/n_1 + S_2^2/n_2}}\right) = P(Z > 4.87).$$

This is an extremely small P value. At any reasonable level we reject H_0. There is strong statistical evidence supporting the claim that lamp bulbs from A last longer than lamp bulbs from B. In order to estimate how much longer lamp bulbs from brand A last we may compute a confidence interval. For instance with 95% confidence we get the following confidence interval for $\mu_1 - \mu_2$

$$(\bar{X}_1 - \bar{X}_2 - c, \bar{X}_1 - \bar{X}_2 + c),$$

where

$$c = z_a \sqrt{S_1^2/n_1 + S_2^2/n_2}.$$

We have $z_a = 1.96$ and $c \sim 29$. So the confidence interval for $\mu_1 - \mu_2$ with 0.95 confidence is $(43, 101)$.

P Value for Testing Two Means

Draw a random sample of size n_1 from a large population with unknown mean μ_1 and an independent random sample of size n_2 from another population having mean μ_2. We denote the sample average by $\bar{X}_i$ and the sample standard deviation by S_i, for $i = 1, 2$. For large n_1 and large n_2 to test

$$H_0 : \mu_1 = \mu_2$$
$$H_a : \mu_1 > \mu_2$$

the P value is

$$P = P\left(Z > \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{S_1^2/n_1 + S_2^2/n_2}}\right).$$

To test

$$H_0 : \mu_1 = \mu_2$$
$$H_a : \mu_1 < \mu_2$$

the P value is

$$P = P\left(Z < \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{S_1^2/n_1 + S_2^2/n_2}}\right).$$

5.2.5 A Few Remarks

The confidence intervals and hypothesis testing we have performed all assume that the samples are RANDOM and LARGE. However, it is possible to analyze small samples (this might be necessary in areas like medicine for which one does not always control the sample sizes) by using different techniques. We will give two such examples in the next section.

For hypothesis testing we have concentrated on one error: rejecting the null hypothesis when it is true. This is called a type I error. There is another possible error which is not rejecting the null hypothesis when it is not true. This is called a type II error. As we have seen it is the type I error that determines a test. However, two different tests with the same type I error may be compared by computing the type II errors. The test with the lower type II error (given a type I error) is the better one.

Exercises 5.2

1. The manufacturer claims that less than 5% of the items it manufactured are defective. Assume that in a random sample of 1,000 items 40 are defective.

(a) Test the claim of the manufacturer at the level 10%.
(b) Compare the conclusion of (a) to the conclusion of Example 1.

2. A pharmaceutical company claims that its new drug is more efficient than the existing one that cures about 70% of the cases treated. In a random sample of 96 patients 81 were cured by the new drug.

(a) What test should the company perform to prove its point?
(b) Perform the test stated in (a) and draw a conclusion.

3. Test whether drug A is more effective than drug B. Drug A was given to 31 patients and 25 recovered. Drug B was given to 42 patients and 32 recovered.

4. Pesticide A killed 15 of 35 cockroaches and pesticide B killed 20 of 35 cockroaches. Compare the two pesticides.

5. Test whether a certain brand of radon detectors are undermeasuring radon levels. Each detector is exposed to 100 standard units of radon. For a sample of 25 detectors, the average reading was 97 with a standard deviation of 8.

6. Is there evidence that children from vegetarian families are not as tall as children in the general population? The heights of 25 6-year-old boys from vegetarian families average 85 cm with a standard deviation of 5 cm. The average height for the general population of 6-year-old boys is 88 cm.

7. With pesticide A in a sample of 40 plants the average yield per plant is 5 pounds with a sample standard deviation of 1.7 pound. Using pesticide B in a sample of 30 plants the average yield is 4.5 pounds with a standard deviation of 1.5 pound. Compare the two pesticides.

8. To study the effect of a drug on pulse rate the available subjects were divided at random in two groups of 30 persons each. The first group was given the drug. The second group was given a placebo. The treatment group had an average pulse rate of 67 with a standard deviation of 8. The placebo group had an average pulse rate of 71 with a standard deviation of 10. Test the effectiveness of the drug.

9. An aptitude test is given in 6th grade. The 150 boys average 75 with a standard deviation of 10 and the 160 girls average 87 with a standard deviation of 8. Test the claim that girls are in average 10 points above boys.

10. A coin is tossed 12 times and nine tails are observed. Is there evidence that the coin is biased? (The sample is too small to use the CLT but you may use the binomial distribution for the number of heads in 12 tosses).

11. I want to test whether the random number generator on my computer is consistent with an uniform distribution on $[0, 1]$. It generated 100 random numbers with average 0.53 and standard deviation 0.09.

(a) Perform a test (recall that an uniform has mean 0.5).
(b) Compute the probability of error II if the mean of the random number generator is actually 0.51.

12. A roulette has 38 slots and 18 red slots.

(a) What is the probability of a red slot for a well-balanced roulette wheel?
(b) Explain how you would proceed to test whether this roulette wheel is well balanced.

5.3 Small Samples

In the previous two sections we have used the Central Limit Theorem to get confidence intervals and perform hypothesis testing. Usually the CLT may be safely applied for random samples of size 25 or larger. In this section, we will see two alternatives for smaller sample sizes.

5.3.1 If the Population is Normal

Assuming we have a random sample of size n from a normal population then it is possible to compute the exact distribution of the normalized sample mean (recall that the CLT only gives an approximate distribution). We now state this result precisely.

Distribution of the Sample Mean

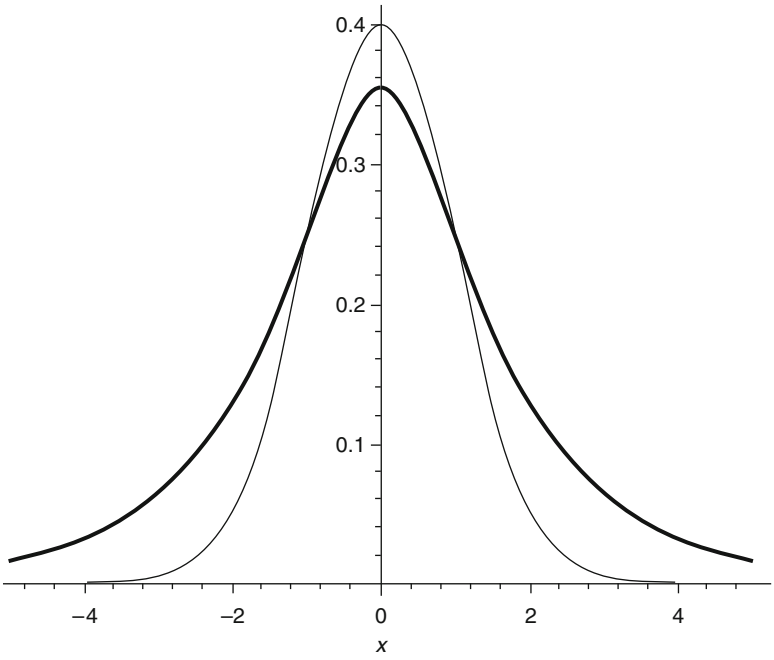
Assume that $X_1, X_2, \ldots, X_n$ are observations from a random sample taken in a NORMAL population. Let μ be the true mean. Let $\bar{X}$ and S be respectively the sample average and the sample standard deviation. Then

$$\frac{\bar{X} - \mu}{S/\sqrt{n}}$$

follows a student distribution with $n-1$ degrees of freedom. A Student random variable with r degrees of freedom will be denoted by $t(r)$.

Student distributions are very similar to the normal standard distributions: they are bell shaped and symmetric around the y axis. The only difference is that the tails of the Student distribution are larger than the tails of the standard normal distribution. That is, the probability of an extreme value is higher for a Student distribution than for a standard normal distribution. However, as the number of degrees of freedom increases Student distributions are closer and closer to the standard normal distribution (as it should be according to the CLT).

The graph below compares a Student distribution with 2 degrees of freedom with a standard normal distribution. One sees for instance that it is a lot more likely for $t(2)$ to be larger than 3 than it is for Z to be larger than 3.



Example 1. Assume that the weights of five 9-year-old boys are 25, 28, 24, 26, and 24 kg in a certain population. Find a confidence interval for the mean weight of 9-year-old boys in that population.

We first need to compute the sample average and standard deviation.

$$\bar{X} = \frac{2 \times 24 + 25 + 26 + 28}{5} = 25.4.$$

We have the following computational formula for S^2

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} X_i^2 - \frac{n}{n-1} (\bar{X})^2.$$

We compute the sum of the squares first

$$\sum_{i=1}^{n} X_i^2 = 3,237$$

and we get

$$S^2 = \frac{1}{4} 3,237 - \frac{5}{4} 25.4^2 = 2.8.$$

As always we use the sample average to estimate the true mean μ. Thus, we look for c such that

$$P(|\bar{X} - \mu| < c) = 0.9.$$

We normalize to get

$$P\left(\frac{|\bar{X} - \mu|}{S/\sqrt{n}} < \frac{c}{S/\sqrt{n}}\right) = 0.9.$$

At this point we need the distribution of $\frac{\bar{X}-\mu}{S/\sqrt{n}}$. The sample is much too small to invoke the CLT. However, it may be reasonable to ASSUME that the weight is normally distributed. In that case $\frac{\bar{X}-\mu}{S/\sqrt{n}}$ follows a Student distribution with 4 degrees of freedom $t(4)$. Thus,

$$P\left(|t(4)| < \frac{c}{S/\sqrt{n}}\right) = 0.9.$$

We now use the t table to get that

$$\frac{c}{S/\sqrt{n}} = 2.13.$$

Solving for c we get $c = 1.6$. Thus, the confidence interval for the true mean weight of 9 years old is

$$(\bar{X} - c, \bar{X} + c) = (25.4 - 1.6, 25.4 + 1.6) = (23.8, 27).$$

If we were using the standard normal distribution then we would have had

$$\frac{c}{S/\sqrt{n}} = 1.64$$

instead of 2.13 and the confidence interval would have been narrower. However, the sample size is too small to invoke the CLT in this case. Next we summarize the method to find a confidence interval for the mean.

Confidence Interval for a Mean

Draw a random sample of size n, $X_1, X_2, \ldots, X_n$, from a NORMAL population with unknown mean μ. Let $\bar{X}$ be the sample mean. Let S^2 be the sample variance. Then,

$$(\bar{X} - c, \bar{X} + c)$$

is a confidence interval for μ with confidence a where

$$c = t_a \frac{S}{\sqrt{n}}$$

and t_a is such that

$$P(|t(n-1)| < t_a) = a$$

and $t(n-1)$ is a Student distribution with $n-1$ degrees of freedom.

Example 2. A certain pain medication is said to provide more than 3 h of relief. We would like to test this claim. The medication is given to 6 patients. The average relief time is 200 min and the standard deviation is 40 min. The test is

$$H_0 : \mu = 180$$
$$H_a : \mu > 180$$

We compute the P value for this test.

$$P = P(\bar{X} > 200|\mu = 180) = P\left(\frac{\bar{X} - 180}{S/\sqrt{n}} > \frac{200 - 180}{S/\sqrt{n}}\right).$$

Assuming that the time of relief is normally distributed we get that $\frac{\bar{X}-180}{S/\sqrt{n}}$ follows a Student distribution $t(5)$. Thus,

$$P = P(t(5) > 1.22) \sim 0.15$$

At the 5% level the null hypothesis is not rejected. There is not enough evidence to support the claim that the medication provides more than 3 h of relief.

P Value for Testing a Mean

Assume we have a random sample of size n from a NORMAL population with average $\bar{X}$ and standard deviation S. The true mean of the population is μ. Let μ_0 be a fixed number. For the test

$$H_0 : \mu = \mu_0$$
$$H_a : \mu < \mu_0$$

the P value is

$$P = P\left(t(n-1) < \frac{\bar{X} - \mu_0}{S/\sqrt{n}}\right),$$

where $t(n-1)$ is a Student distribution with $n-1$ degrees of freedom. For the test

$$H_0 : \mu = \mu_0$$
$$H_a : \mu > \mu_0$$

the P value is
$$P = P\left(t(n-1) > \frac{\bar{X} - \mu_0}{S/\sqrt{n}}\right).$$

5.3.2 Comparing Two Means with Two Small Samples

A company claims that its new fertilizer works better than the old one. To test the claim ten identical small plots are fertilized, five with the new fertilizer and five with the old fertilizer. The average yield with the new fertilizer is 123 pounds of tomatoes and the standard deviation is 6 pounds. For the old fertilizer the average is 116 pounds and the standard deviation is 7 pounds. We perform the test

$$H_0 : \mu_1 = \mu_2$$
$$H_a : \mu_1 > \mu_2$$

The rejection region is

$$\{\bar{X}_1 - \bar{X}_2 > 123 - 116\}.$$

The samples are too small to use the CLT. However, assuming that the yields are normally distributed and that the true variances are equal we have that under $\mu_1 = \mu_2$

$$\frac{\bar{X}_1 - \bar{X}_2}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

follows a Student distribution with $n_1 + n_2 - 2$ degrees of freedom where S^2 is defined by

$$S^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)^2 S_2^2}{n_1 + n_2 - 2}.$$

Going back to our example we have $\bar{X}_1 = 123$, $S_1 = 6$, $\bar{X}_2 = 116$, $S_1 = 7$, $n_1 = n_2 = 5$. Hence,

$$S^2 = \frac{4 \times 6^2 + 4 \times 7^2}{8} = 42.5.$$

Hence, the P value is

$$P = P\left(\frac{\bar{X}_1 - \bar{X}_2}{S\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} > \frac{7}{S\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}\right) = P(t(8) > 1.7).$$

According to the Student table the P value is between 5 and 10%. So at the 5% level we would not reject the null hypothesis. There is not enough evidence to claim that the new fertilizer yields more tomatoes.

Note that we assumed that the variances were the same to compare the two means. When the variances are not the same a different Student test must be performed. See for instance 6.5 in "Statistical Methods in Biology" by N.T.J. Bailey, Cambridge University Press, Third Edition.

5.3.3 Matched Pairs

Assume we want to test the effect of a course on students. We test the students before and after a course to assess the effectiveness of the course. We should not analyze such data as two independent samples. Our two samples techniques work for two INDEPENDENT samples. We do not have independence here since we are testing the same individuals in the two samples. We have matched pairs instead. In such a case we should analyze the difference between the two tests for each individual. We may apply a one sample technique to the differences. Next we treat such an example.

Example 3. Does a certain course help the students? Ten students are given two similar tests before and after a course. Here are their grades

Before	71	72	85	90	55	61	76	78	79	85
After	73	75	89	92	50	68	82	81	86	80

We start by computing the gain per student. We get 2, 3, 4, 2, −5, 7, 6, 3, 7, and −5. We get an average gain of 2.4 and a sample standard deviation of 4.3. Let μ be the true gain after the course. We would like to test

$$H_0 : \mu = 0$$
$$H_a : \mu > 0$$

Assuming that the gains are normally distributed we use a Student test. The P value is

$$P = P\left(t(9) > \frac{\bar{X}}{S/\sqrt{10}}\right) = P(t(9) > 1.75).$$

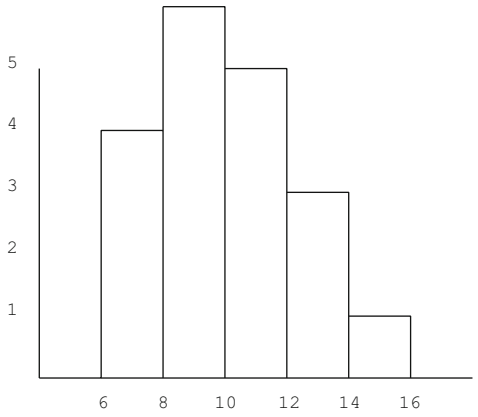
A Student table yields that the P value is strictly between 0.05 and 0.1. At the 5% level we cannot reject the null hypothesis. There is not enough evidence to claim that the course increases the test scores.

Note that we can use this matched pair technique for large samples as well. If the sample size is large enough we do not need the normality assumption we just use the CLT.

5.3.4 Checking Normality

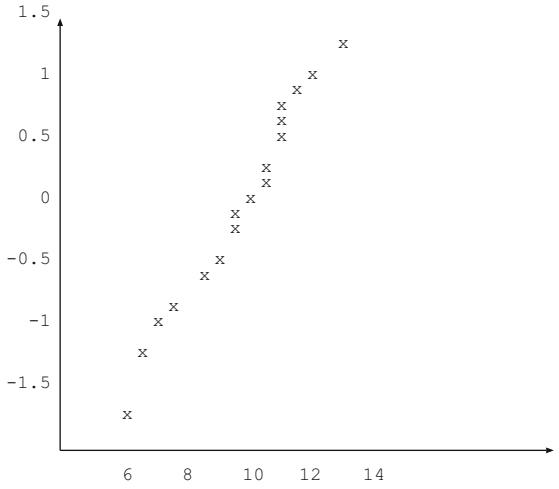
How do we decide whether the assumption of normality is reasonable? This is what the next example is about.

Example 4. Consider the following data: 10.8, 9.6, 10.2, 9.8, 6.9, 8.7, 12.2, 10.4, 11.7, 9.0, 7.4, 13.2, 10.9, 9.5, 11.0, 6.9, 12.9, 6.2, 9.2, 16.9. Could these numbers come from a normal population? We first draw an histogram.



This histogram is kind of bell shaped and symmetric. However, interpreting an histogram is subjective. We may quantify this interpretation by using the fact that 68% of normal observations should be within one standard deviation of the mean and 95% of the observations should be within two standard deviations of the mean. See Exercise 5.2.

However, there is a more precise way to assess normality which is provided by a normal quantile plot. We have a sample of 20 and the smallest observation is 6.2. So 6.2 is the 1/20 or 0.05 quantile of the data. The 0.05 quantile for a standard normal distribution is the number with an area of 0.05 to its left. So the 0.05 quantile for a standard normal is -1.64. The first point on the normal quantile plot is (6.2,-1.64). The second smallest observation is 6.9. So this is the 2/20 or 0.1 quantile of the data. The 0.1 quantile of a standard normal distribution is -1.28. The second point of our normal quantile plot is (6.9,-1.28) and so on. Below is the normal quantile plot for our data.



If the distribution of the observations is normally distributed then the normal quantile plot is close to a line. This is so because if X is normally distributed with mean μ and standard deviation σ then $\frac{X-\mu}{\sigma}$ is a standard normal distribution. So there is a linear relation between the quantiles of any two normally distributed random variables.

In this particular example one sees that the points are reasonably aligned and we may conclude that our observations come from an approximately normal population.

5.3.5 The Sign Test

If the population is clearly not normal and the sample is too small to use the CLT we still have the following sign test at our disposal. This is a test that may be performed without assuming that the random variables follow a normal distribution. We still need to have a sample of n i.i.d. random variables but nothing will be assumed about the distribution of these random variables. We will explain the test on an example.

Example 5. We would like to test the following claim: the median height of a 6-year-old boy in this population is at least 84 cm. Assume that the heights in centimeter of 11 6-year-old boys are the following: 80, 93, 85, 87, 79, 85, 85, 86, 89, 90, and 91. So our test is

$$H_0 : m = 84$$
$$H_a : m > 84,$$

where m is the true median of the population. If the median of the distribution of the continuous random variable X is m then by definition of m we have

$$P(X > m) = P(X \leq m) = 1/2.$$

Let B be the number of observations above 84. Under the null hypothesis $m = 84$, B is a binomial random variable with parameters $n = 11$ and $p = 1/2$ (since there is the same chance for an observation to fall below or above 84). The sign test is based on the random variable B. We should reject the null hypothesis if B is too large. In this sample we note $\binom{0}{1}$ that $B = 9$. We compute the P value for the sign test

$$P = P(B \geq 9 | m = 84) = \binom{11}{9}\left(\frac{1}{2}\right)^{11} + \binom{11}{10}\left(\frac{1}{2}\right)^{11} + \binom{11}{11}\left(\frac{1}{2}\right)^{11}.$$

We get a P value of 0.03. Thus, at the 5% level we reject the null hypothesis: there is statistical evidence that the median height in this population is at least 84 cm.

Example 6. Test the claim that the median weight loss for a certain diet is larger than 5 pounds. The diet is tested on eight people. Here are the weights before and after the diet

Before	181	178	205	195	202	176	180	177
After	175	171	196	192	190	168	176	171

Let m be the true median weight loss. We want to test

$$H_0 : m = 5$$
$$H_a : m > 5$$

We first compute the weight losses: 6, 7, 9, 3, 12, 8, 4, 6. Let B be the number of weight losses larger than 5. Under $m = 5$, B follows a binomial with parameters $n = 8$ and $p = 1/2$. Thus, the P value is

$$P = P(B \geq 6 | m = 5) = \binom{8}{6}\left(\frac{1}{2}\right)^{8} + \binom{8}{7}\left(\frac{1}{2}\right)^{8} + \binom{8}{8}\left(\frac{1}{2}\right)^{8} = 0.14.$$

At the level 5 or 10% there is not enough evidence to reject the null hypothesis. That is, there is not enough evidence to claim that the median weight loss of the diet is at least 5 pounds.

Example 7. Farm worker X claims that he picks more apples than farm worker Y. Here are the quantities picked by both workers over 7 days:

X	201	179	159	192	177	170	182
Y	172	165	161	184	174	142	190

We test

$$H_0 : \text{X and Y pick the same quantity}$$

$$H_a : \text{X picks more than Y}$$

In order to perform the test we use the random variable B: the number of days that X outperforms Y. In this sample $B = 5$. Under the null hypothesis B is a binomial with parameters $n = 7$ and $p = 1/2$. The P value is

$$P(B \geq 5|H_0) = \binom{7}{5}\left(\frac{1}{2}\right)^7 + \binom{7}{6}\left(\frac{1}{2}\right)^7 + \binom{7}{7}\left(\frac{1}{2}\right)^7 = \frac{29}{2^7} = 0.23$$

We do not reject the null hypothesis. There is not enough evidence to claim that X picks more apples than Y.

Exercises 5.3

1. (a) What is $P(t(3) > 2)$?
(b) Compare (a) with $P(Z > 2)$.

2. Consider the observations of Example 4.

(a) What percentage of the observations are within 1 standard deviation of the mean?
(b) What percentage of the observations are within 2 standard deviations of the mean?
(c) Is it reasonable to assume that these observations come from a normal population?

3. Some components in the blood tend to vary normally over time for each individual. Assume that the following levels for a given component were measured on a single patient: 5.5, 5.2, 4.5, 4.9, 5.6, and 6.3.

(a) Test the claim that the mean level for this patient is above 4.7.
(b) Find a confidence interval with 0.95 confidence for the mean level of this patient.

4. Assume that a group of ten eighth graders taken at random averaged 85 on a test with a standard deviation of 7.

(a) Is there evidence that the true mean grade for this population is above 80?
(b) Find a confidence interval for the true mean grade.
(c) What assumptions did you make to answer (a) and (b)?

5. Eight students were given a placement test and after a week of classes were given again a placement test at the same level. Here are their scores.

Before	71	78	80	90	55	65	76	77
After	75	71	89	92	61	68	80	81

(a) Test whether the scores improved after 1 week by performing a student test.
(b) Test whether the scores improved after 1 week by performing a sign test.

6. In an agricultural field trial, researchers tested two varieties of tomatoes in ten plots. In eight of the plots variety A yielded more than variety B. Is this enough evidence to say that variety A yields more than variety B?

7. A diet was tested on nine people. Here are their weights before and after the diet.

Before	171	178	180	190	165	165	176	177	182
After	175	171	182	161	168	156	165	171	175

(a) Test whether the diet makes lose at least 5 pounds by performing a Student test.
(b) Check whether it is reasonable to assume normality in (a).
(c) Perform a sign test for the hypothesis in (a).

8. A test given to 12 male students has an average of 75 and standard deviation of 11. The same test given to ten female students has an average of 81 with a standard deviation of 8.

(a) Is there evidence that the female students outperform the male students?
(b) Find a confidence interval for the difference of the true means.

9. Does Calculus improve the algebra skills of the students? At the beginning of the semester 100 Calculus students were given an algebra test and got an average of 79 and a standard deviation of 15. At the end of the semester the same 100 students were given another algebra test for which the average was 85 and the standard deviation was 12. The standard deviation for the differences between the two tests is 5. Perform a test.

10. We would like to assess the effect of a medication on blood pressure. The medication is given to 12 patients. This group has an average blood pressure of 131 and a standard deviation of 5. Another group of ten patients is given a placebo and their average is 127 and the standard deviation is 4.

11. Consider the data in Example 7.

(a) Perform a student test.
(b) What assumptions do you need to perform the test in (a)?

5.4 Chi-Square Tests

In this section we will see two Chi-Square tests. We will first test whether two variables are independent. Our second test will check whether given observations fit a theoretical model. We start by introducing the Chi-Square distributions.

Chi-Square Distributions

The random variable X is said to have a chi-square distribution with n degrees of freedom if it is a continuous random variable with density

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

where

$$\Gamma(a) = \int_0^\infty x^{a-1} \mathrm{e}^{-x} \mathrm{d}x \text{ for } a > 0.$$

The notation for a chi-square random variable with n degrees of freedom is $\chi^2(n)$.

The function Γ appears in a number of different branches of mathematics but cannot be defined more explicitly. However, it is possible to compute some values of the function explicitly. For instance,

$$\Gamma(1) = \int_0^\infty x^{1-1} \mathrm{e}^{-x} \mathrm{d}x = -\mathrm{e}^{-x} \Big]_0^\infty = 1.$$

An integration by parts gives

$$\Gamma(2) = \int_0^\infty x^{2-1} \mathrm{e}^{-x} \mathrm{d}x = -x\mathrm{e}^{-x} \Big]_0^\infty + \int_0^\infty \mathrm{e}^{-x} \mathrm{d}x = \Gamma(1) = 1.$$

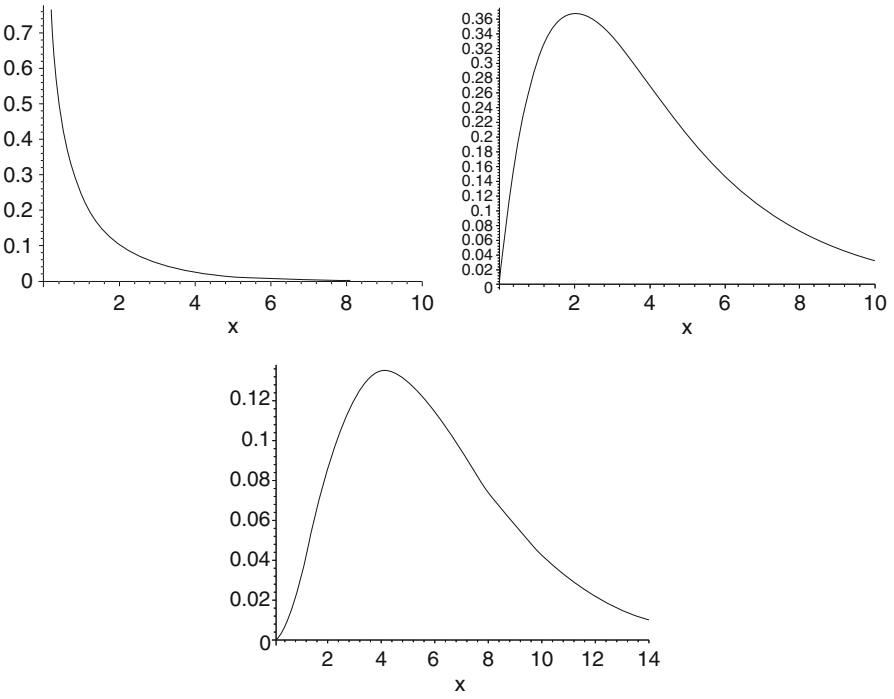
More generally, an integration by parts shows that

$$\Gamma(a+1) = a\Gamma(a).$$

In particular, for a positive integer n using the identity above repeatedly gives

$$\Gamma(n) = (n-1)!$$

We now sketch the graphs of three densities of chi-square random variables.



From left to right we have the chi distributions with 1, 4, and 6 degrees of freedom, respectively.

5.4.1 Testing Independence

Example 1. Is there a relation between the level of education and smoking? Assume that a random sample of 200 was taken with the following results.

	Smoker	Non smoker
Education		
8 years or less	9	38
12 years	21	80
16 years	5	47

In the test we are going to perform the null hypothesis will be that there is no association between the two variables. That is, H_0 will be that education level and smoking are independent. The alternative hypothesis is that there is an association between the two variables. In order to take a decision we will compare the counts in our sample to the expected counts under the null hypothesis. We now explain how

to compute the expected counts under the null hypothesis. There are 9 people with 8 years of education or less that smoke. The probability that someone in the sample has 8 years or less of education is

$$\frac{9+38}{200} = \frac{47}{200}.$$

The probability that someone in the sample be a smoker is

$$\frac{9+21+5}{200} = \frac{35}{200}.$$

Thus, under the assumption that level of education and smoking are independent we get that the probability that someone taken at random in the sample has 8 years or less of education and smoke is

$$\frac{47}{200} \times \frac{35}{200}.$$

The expected number of people who have 8 years or less of education and smoke is therefore

$$200 \times \frac{47}{200} \times \frac{35}{200} = \frac{47 \times 35}{200}.$$

More generally we have the following,

Expected Count Under the Independence Assumption

$$\text{Expected count} = \frac{\text{row total} \times \text{column total}}{\text{sample size}}$$

We now go back to the data of Example 1 and compute the expected counts for all the cells.

	Expected Counts	
	Smoker	Non smoker
Education		
8 years or less	8.225	38.775
12 years	17.675	83.325
16 years	9.1	42.9

Testing Independence

Assume that we want to test whether two variables are related. The null hypothesis is H_0 : the two variables are independent. We use the statistic

$$X^2 = \sum \frac{(\text{observed}-\text{expected})^2}{\text{expected}}$$

The random variable X^2 follows approximately a $\chi^2((r-1)(c-1))$ distribution where r and c are the number of rows and columns, respectively. Therefore the P value for this test is given by

$$P(\chi^2((r-1)(c-1)) > X^2).$$

The approximation of X^2 by a chi-square distribution gets better as the sample size increases and is more reliable if every expected cell has a count of five or more.

We now go back to the data of Example 1 to perform the test. We compute X^2.

$$X^2 = \frac{(9-8.225)^2}{8.225} + \frac{(38-38.775)^2}{38.775} + \frac{(21-17.675)^2}{17.675}$$
$$+\frac{(80-83.325)^2}{83.325} + \frac{(5-9.1)^2}{9.1} + \frac{(47-42.9)^2}{42.9} = 3.09.$$

We have three categories for education so $r = 3$ and two categories for smoking so $c = 2$. Thus, $(r-1)(c-1) = 2$ and X^2 follows approximately a $\chi^2(2)$. The P value for Example 1 is

$$P = P(\chi^2(2) > 3.09).$$

According to the chi-square table the P value is larger than 0.1. At the 5% level we do not reject H_0. It does not appear that there is an association between education level and smoking.

5.4.2 *Goodness-of-Fit Test*

We now turn to another important test: goodness-of-fit. We start with an example.

Example 2. Consider the following 25 observations: 0, 3, 1, 0, 1, 1, 1, 3, 4, 3, 2, 0, 2, 0, 0, 0, 4, 2, 3, 4, 1, 6, 1, 4, 1. Could these observations come from a Poisson distribution?

Recall that a Poisson distribution depends only on one parameter: its mean. We use the sample average to estimate the mean. We get

$$\bar{X} = \frac{47}{25} = 1.88.$$

Let N be a Poisson random variable with mean 1.88. We have that

$$P(N = 0) = \mathrm{e}^{-1.88} = 0.15$$

and therefore the expected number of 0's in 25 observations is $25 \times \mathrm{e}^{-1.88} = 3.81$. Likewise we have that

$$P(N = 1) = 1.88\mathrm{e}^{-1.88} = 0.29$$

and the expected number of 1's in 25 observations is 7.17. We also get that the expected number of 2's is 6.74 and the expected number of 3's is 4.22. The probability that N is 4 or more is

$$P(N \geq 4) = 0.12.$$

Thus, the expected number of observations larger than 4 is 3. We summarize these computations in the table below.

	0	1	2	3	4 or more
Observed	6	7	3	4	5
Expected	3.81	7.17	6.74	4.22	3

The test we are going to perform compares the expected and observed counts in the following way.

Goodness-of-Fit

We want to test whether some observations are consistent with a certain distribution F (F may be for instance a Poisson distribution or a normal distribution). The null hypothesis is H_0 : The observations follow a distribution F. We use the statistic

$$X^2 = \sum \frac{(\text{observed}-\text{expected})^2}{\text{expected}}.$$

The random variable X^2 follows approximately a $\chi^2(r-1-d)$ distribution where r is the number of categories of observations and d is the number of parameters that must be estimated for the distribution F. Therefore the P value for this test is given by

$$P(\chi^2(r-1-d) > X^2).$$

We use the preceding test on Example 2. In that case the categories are 0, 1, 2, 3, and 4 or more. So $r = 5$. The Poisson distribution depends on one parameter (its mean) therefore $d = 1$. We now compute X^2.

$$X^2 = \frac{(6-3.81)^2}{3.81} + \frac{(7-7.17)^2}{7.17} + \frac{(3-6.74)^2}{6.74}$$
$$+\frac{(4-4.22)^2}{4.22} + \frac{(5-3)^2}{3} = 4.68.$$

We know that X^2 follows approximately a chi-square distribution with $r-1-d = 5-1-1 = 3$ degrees of freedom so the P value is.

$$P = P(\chi^2(3) > 4.68).$$

Since the P-value is larger than 0.1 we do not reject the null hypothesis. That is, these observations are consistent with a Poisson distribution.

Example 3. The observations of Example 2 were in fact generated as Poisson observations with mean 2 by a computer random generator. We now test whether these observations are consistent with a Poisson distribution with mean 2. That is our null hypothesis is now H_0: the observations follow a Poisson distribution with mean 2. The only difference with Example 1 is that now we do not need to estimate the mean of the Poisson distribution. In particular, for this example $d = 0$. We compute the expected counts.

	0	1	2	3	4 or more
Observed	6	7	3	4	5
Expected	3.38	6.77	6.77	4.51	3.75

This time $X^2 = 4.62$. We have that $r-d-1 = 5-0-1 = 4$.

$$P = P(\chi^2(4) > 4.62) > 0.1.$$

We do not reject the null hypothesis. These observations are consistent with a mean 2 Poisson distribution.

The following example deals with continuous distributions.

Example 4. Are the following observations consistent with a normal distribution?
66, 64, 59, 65, 81, 82, 64, 60, 78, 62
65, 67, 67, 80, 63, 61, 62, 83, 78, 65
66, 58, 74, 65, 80

The sample average is 69 and the sample standard deviation is 8 (we are rounding to the closest 1 to simplify the subsequent computations). We will now try to fit the observations to a normal distribution with mean 69 and standard deviation 8.

We first pick the number of categories, keeping in mind that the chi-square approximation is best when there are at least five expected counts per cell. We pick

five categories. Using the standard normal table we find the 20th, 40th, 60th, and 80th percentiles. For instance, we read in the standard normal table that

$$P(Z < -0.84) = 0.2$$

and so the 20th percentile of a standard normal distribution is -0.84. Likewise we find the four percentiles in increasing order

$$-0.84, -0.25, 0.25, 0.84$$

Recall that if X is a normal random variable with mean 69 and standard deviation 8, then

$$\frac{X - 69}{8}$$

is a standard normal random variable. So, for instance, the 20th percentile of a normal random variable with mean 69 and standard deviation 8 is

$$69 + 8(-0.84) = 62.28.$$

Likewise the 40th, 60th, and 80th percentiles of a normal random variable with mean 69 and standard deviation 8 are 67, 71, 75.72. We round these percentiles to the nearest one. We now compare the observed and expected counts.

Category	$(-\infty, 62]$	$(62, 67]$	$(67,71)$	$[71, 76)$	$[76, \infty)$
Observed	6	11	0	1	7
Expected	5	5	5	5	5

We compute the statistic

$$nX^2 = \sum \frac{(\text{observed}-\text{expected})^2}{\text{expected}} = 16.4.$$

We had to estimate two parameters (μ and σ) so $d = 2$ and X^2 is approximately a chi-square random variable with $r - d - 1 = 5 - 2 - 1 = 2$ degrees of freedom. We get the P value

$$P(\chi^2(2) > 16.4) < 0.01.$$

So we reject the null hypothesis. These observations are not consistent with a normal distribution.

Observe that the goodness of fit test is especially useful when we reject the null hypothesis. In that case we conclude that the observations are unlikely to come from the distribution we are testing. On the other hand when we do not reject the

null hypothesis we are simply saying that the observations are consistent with the distribution we are testing. There may be a number of other distributions for which this is true as well.

Exercises 5.4

Problems 1–5 use data from the American Mathematical Society regarding new doctorates in mathematics (Notices of the AMS January 1998). Types I–III are groups of mathematics departments as ranked by the AMS.

1. The following table gives the number of new PhD's in mathematics according to their area of concentration and the type of department that granted their degree.

Area	Algebra	Geometry	Probability and stat.
Institution			
I	21	28	9
II	10	7	4
III	3	1	3

Is it true that certain types of institutions graduate more students in one area than in others?

2. The table below breaks down the number of graduates in 1997 according to their gender and area of concentration.

Area	Algebra	Geometry	Probability and stat.
Gender			
Male	123	118	194
Female	37	23	98

Is the distribution of area of concentrations for female students different from the distribution for male students?

3. The following table gives the numbers of employed new graduates according to the type of their granting institution and the type of employer. Does the type of employer depend on the type of granting institution?

Granting inst.	I public	I private	II
Employer			
I public	35	19	6
I private	13	25	2
II	14	7	16

4. The next table breaks down the number of new graduates according to gender and granting institution.

Granting inst.	I public	I private	II	III
Gender				
Male	239	157	175	96
Female	58	30	63	36

Is the distribution of granting institutions for female students different from the distribution for male students?

5. The table below breaks down the number of employed new graduates per type of employer and citizenship. Does the type of employer depend on citizenship?

Citizenship	US	Non-US
Employer		
PhD dept.	100	111
Non-PhD dept.	177	59
Nonacademic	104	160

6. Test whether the following observations are consistent with a Poisson distribution: 1, 4, 2, 7, 4, 3, 0, 2, 5, 2, 3, 2, 1, 5, 5, 0, 3, 2, 2, 2, 2, 1, 4, 1, 2, 4.

7. Test whether the following observations are consistent with a standard normal distribution:
1.70, 0.11, 0.14, 0.81, 2.19
−1.56, −0.67, 0.89, −1.24, 0.26
−0.05, 0.72, 0.29, −1.09, −0.43
−2.23, −1.68, 0.23, 1.17, −0.87
−0.28, 1.11, −0.43, −0.16, −0.07

8. Test whether the following observations are consistent with a uniform distribution on [0,100].
99, 53, 18, 21, 20, 53, 58, 4, 32, 51
24, 51, 62, 98, 2, 48, 97, 64, 61, 18
25, 57, 92, 72, 95

9. Let T be an exponential random variable with mean 2. That is, the density of T is $f(t) = \frac{1}{2}\mathrm{e}^{\frac{-t}{2}}$ for $t \geq 0$. Find the 20th percentile of T.

10. Test whether the following observations are consistent with an exponential distribution.
13, 7, 14, 10, 12, 8, 8, 8, 10, 9
8, 10, 5, 14, 13, 7, 11, 11, 10, 8
10, 10, 13, 9, 10

11. (a) Show that the function Γ:

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

is defined on $(0, \infty)$.

(b) Show that $\Gamma(a+1) = a\Gamma(a)$.

(c) Show that $\Gamma(n) = (n-1)!$ for any positive integer n.

Chapter 6
Linear Regression

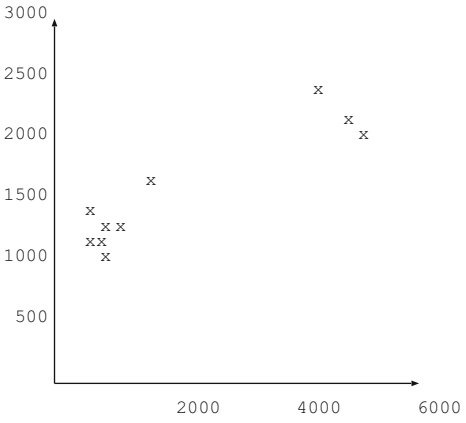
6.1 Fitting a Line to Data

We start by an example.

Example 1. We consider state taxes and state debts, per capita, for 10 American states (data from The World Almanac and Book of Facts 1998).

x = Debt	884	720	798	1,526	899	4,719	4,916	1,085	781	4,377
y = Taxes	1,194	1,475	1,365	1,686	1,209	2,282	2,224	1,311	1,317	2,422

We would like to see whether there is a linear relationship between per capita taxes and per capita debt. We start by plotting the points. Let x denote the state debt per capita and y denote the tax per capita. We get



R.B. Schinazi, *Probability with Statistical Applications*,
DOI 10.1007/978-0-8176-8250-7_6, © Springer Science+Business Media, LLC 2012

It looks like there is an approximate linear relation between x and y. How can we find a line that fits the data? The most used criterion is the least-squares criterion.

The Least Squares Regression Line

Assume that we have n observations (x_i, y_i) and we want the line that best fits these observations. More precisely, we would like to predict y from x by using a line. The line $\hat{b}x + \hat{a}$ is said to be the least squares regression line of y on x if

$$\sum_{i=1}^{n}(y_i - (bx_i + a))^2$$

is minimum for $a = \hat{a}$ and $b = \hat{b}$. The values of $\hat{a}$ and $\hat{b}$ are given by

$$\hat{b} = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}$$

and

$$\hat{a} = \bar{y} - \hat{b}\bar{x}.$$

Remarks

1. Note that $\sum_{i=1}^{n}(y_i - (bx_i + a))^2$ represents the total error we make when we replace y_i by $bx_i + a$. This is why we want a and b to minimize this sum. Other choices are possible for the error such as $\sum_{i=1}^{n}|y_i - (bx_i + a)|$. However, the advantage of the sum of squares is that we can get explicit expressions for $\hat{a}$ and $\hat{b}$. Explicit expressions are not available if the error is taken to be the sum of absolute values.
2. Note also that the variables x and y do not play symmetric roles here. We are trying to predict y from x. If we want to predict x from y then we should minimize $\sum_{i=1}^{n}(x_i - (by_i + a))^2$. We would get the regression line of x on y and the reader may check that this is a different line from the regression line of y on x.
3. Finally, note that the relation $\hat{a} = \bar{y} - \hat{b}\bar{x}$ shows that the regression line passes through the point of averages $(\bar{x}, \bar{y})$.

We now go back to the data of Example 1 and compute $\hat{a}$ and $\hat{b}$. We write the intermediate computations in a table.

	x_i	y_i	x_i^2	y_i^2	$x_i\, y_i$
	884	1,194	781,456	1,425,636	1,055,496
	720	1,475	518,400	2,175,625	1,062,000
	798	1,365	636,804	1,863,225	1,089,270
	1,526	1,686	2,328,676	2,842,596	2,572,836
	899	1,209	808,201	1,461,681	1,086,891
	4,719	2,282	22,268,961	5,207,524	10,768,758
	4,916	2,224	24,167,056	4,946,176	10,933,184
	1,085	1,311	1,177,225	1,718,721	1,422,435
	781	1,317	609,961	1,734,489	1,028,577
	4,377	2,422	19,158,129	5,866,084	10,601,094
Sums:	20,705	16,485	72,454,869	29,241,757	41,620,541

We get that

$$\hat{b} = \frac{10(41{,}620{,}541) - (20{,}705)(16{,}485)}{10(72{,}454{,}869) - (20{,}705)^2} \sim 0.25$$

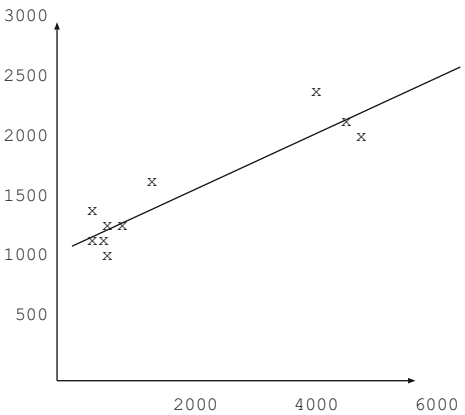
and that

$$\hat{a} = \bar{y} - \hat{a}\bar{x} = 1{,}648.5 - (0.25)2{,}070.5 \sim 1{,}131.$$

So that the equation of the regression line is

$$y = 0.25x + 1{,}131.$$

One can see below that the regression line fits well the scatter plot.



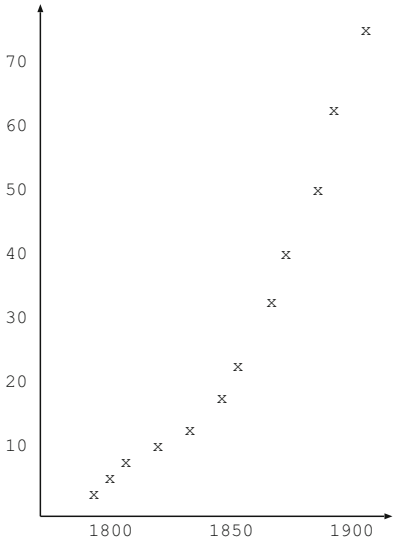
The regression line main use is to make predictions. For instance, if the debt per capita is \$6,000 in a state then we plug $x = 6{,}000$ in the regression line and get that the predicted tax y is

$$y = 0.25 \times 6{,}000 + 1{,}131 = 2{,}631.$$

Example 2. We consider the population, in millions, of the United States from 1790 to 1900.

Year	Population	Year	Population
1790	3.9	1850	23.2
1800	5.3	1860	31.4
1810	7.2	1870	39.8
1820	9.6	1880	50.2
1830	12.9	1890	62.9
1840	17.1	1900	76.0

The scatter plot below shows that there is a relation between population and year but that this relation is not linear.



It looks like the population has grown exponentially over the period of time we consider. Let x be the year and y the population, if y is an exponential function of x as

$$y = c\mathrm{e}^{\mathrm{d}x}$$

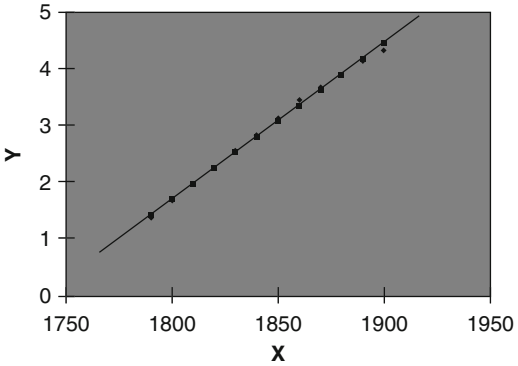
then by taking logarithms on both sides we get

$$\ln(y) = \mathrm{d}x + \ln(c).$$

Therefore, we may test our hypothesis that the population grew exponentially fast during the period 1790–1900 by trying to find a linear relation between the logarithm of the population and the year. Our transformed data is

Year	ln(Population)	Year	ln(Population)
1790	1.36	1850	3.14
1800	1.67	1860	3.45
1810	1.97	1870	3.68
1820	2.26	1880	3.92
1830	2.56	1890	4.14
1840	2.84	1900	4.33

The regression line fits the data remarkably well. See below.



The equation of the regression line is

$$\ln(y) = 0.0275x - 47.77.$$

We plug $x = 1,872$ in the preceding equation to predict the population in 1872. We get $\ln(y) = 3.71$ and therefore $y = 40.8$. Thus, the model predicts that the population in 1872 was 40.8 millions in the United States. This figure is in good agreement with the actual figure. We plug $x = 2,000$ and we get $y = 1,380.2$. The prediction of the model is that the population of the United States in 2000 will be 1 billion and 380 millions people. This is in gross disagreement with the actual figure (around 260 millions). We used the data from 1790 to 1900 to get this regression line. The year 2000 is well off this range. This example illustrates the fact that as we get away from the range of x for which the regression was made the predictions may become very unreliable.

6.1.1 Sample Correlation

When one looks for a relation between x and y the starting point should always be a scatter plot. However, the relation between x and y is rarely obvious and it is interesting to have a measure of the strength of the linear relation between x and y. This is where correlation comes into play.

Sample Correlation

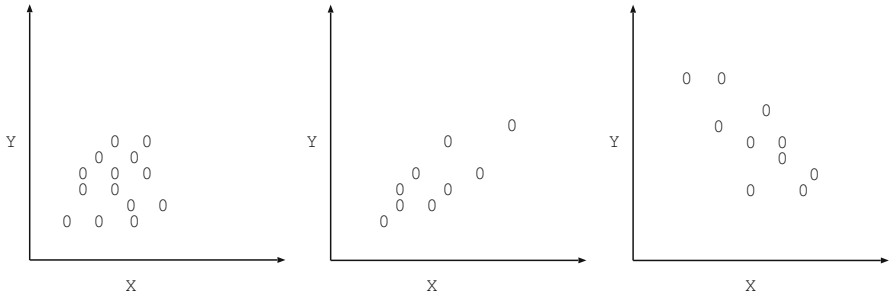
Assume that we have n observations of the variables x and y that we denote by (x_i, y_i). The sample correlation r for this data is

$$r = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right),$$

where $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$ and $s_x^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$. The coefficient r is always in the interval $[-1,1]$ and measures the strength of the LINEAR relation between x and y. If $|r|$ is close to 1 there is a strong linear relation between x and y. If $|r|$ is close to 0 then there is no linear relation between x and y. A computational formula for r is

$$r = \frac{\sum_{i=1}^{n} x_i y_i - \frac{1}{n} \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{(n-1) s_x s_y}.$$

Next we examine three typical situations.



For the scatter plot on the left there is no linear relation between x and y. In this case the correlation coefficient r will be close to 0. For the scatter plot in the middle there is a positive linear relation between x and y and so r will be close to 1. Finally, for the scatter plot on the right there is a negative relation between x and y and so r will be close to -1. It is possible to show that $r = 1$ or -1 if and only if all the points are aligned.

Example 3. We compute the sample correlation for the data of Example 1. We use the computational formula for s_x and s_y,

$$s_x^2 = \frac{1}{n-1} \left(\sum_{i=1}^{n} x_i^2 - n(\bar{x})^2 \right) = \frac{1}{9}(72{,}454{,}869 - 10 \times (2{,}070.5)^2) = 3{,}287{,}241.$$

Similarly,

$$s_y^2 = \frac{1}{9}(29{,}241{,}757 - 10 \times (1{,}648.5)^2) = 229{,}582.$$

Therefore we get that $s_x = 1,813$ and $s_y = 479$. Thus, the correlation coefficient

$$r = \frac{\sum_{i=1}^n x_i y_i - \frac{1}{n}\sum_{i=1}^n x_i \sum_{i=1}^n y_i}{(n-1)s_x s_y} = \frac{41{,}620{,}541 - \frac{1}{10}20{,}705 \times 16{,}485}{9(1{,}813)(479)} \sim 0.96.$$

This computation shows a very strong positive linear relation between tax and debt.

We now state an interesting relation between the correlation coefficient and the equation of the regression line.

Correlation and Regression

Let r be the sample correlation for variables x and y. Let $\hat{b}x + \hat{a}$ be the equation of the regression line of y on x then

$$\hat{b} = r\frac{s_y}{s_x}$$

and

$$\hat{a} = \bar{y} - \hat{b}\bar{x}.$$

Example 4. We compute the equation of the regression line of y on x for the data in Example 1 by using the formulas above. First,

$$\hat{b} = r\frac{s_y}{s_x} = 0.96\frac{479}{1,813} = 0.25.$$

For $\hat{a}$ we use

$$\hat{a} = \bar{y} - \hat{b}\bar{x} = 1,648.5 - (0.25)(2,070.5) = 1,131.$$

Exercises 6.1

1. Here is the population data for the United States from 1900 to 1990.

Year	Population	Year	Population
1900	76.0	1950	151.3
1910	92.0	1960	179.3
1920	105.7	1970	203.3
1930	122.8	1980	226.5
1940	131.7	1990	248.7

(a) Make a scatter plot.
(b) Did the population increase exponentially fast during this period?
(c) Do a regression explaining the log of the population in function of the year.
(d) Does the model in (c) look adequate?
(e) Predict the population in 2000 by using this model.

2. Use the data of Example 1 to do a regression where the roles of x and y are inverted. That is, take y to be the debt per capita and x the tax per capita.

(a) Find the equation of the regression line.
(b) Is this the same line as the one we found in Example 1?

3. As the scatter plot from Example 1 shows the data we have analyzed so far has two clusters. One with small tax debt and the other one with high tax debt. We now add 4 points with intermediate tax debt, the first coordinate represents the debt, the second one the corresponding tax: (3066,1713), (3775,1891), (2282,952), and (2851,1370). So the modified data of Example 1 has now 14 points.

(a) Compute the correlation coefficient debt/tax for the modified data.
(b) Compute the equation of the new regression line.
(c) Compare the fit of the regression line in this problem to the fit of the regression line in Example 1.

4. In recent years there has been a significant increase of tuberculosis in the US. We will examine whether there seems to be a linear relation between the number of cases of AIDS and the number of cases of tuberculosis. In the following table we write the rate per 100,000 population for nine states in 1998 (the data are from the Center for Disease Control).

State	Tuberculosis	AIDS
California	11.8	367
Florida	8.7	538
Georgia	8.3	307
Illinois	7.1	189
Maryland	6.3	390
New Jersey	7.9	496
New York	11.0	715
Pennsylvania	3.7	175
Texas	9.2	284

(a) Draw a scatter plot.
(b) Compute the correlation coefficient between AIDS and tuberculosis.

5. Here are the death rates (per 100,000 persons) per age in the US in 1978 (data from US Department of Health).

Age	Death rate	Age	Death rate
42	296.1	67	2,463.0
47	471.6	72	3,787.4
52	742.4	77	6,024.2
57	1,115.9	82	8,954.0
62	1,774.2		

(a) Draw a scatter plot.
(b) Based on (a) should you transform the data in order to have a linear relation?
(c) Compute an adequate regression line.

6. The following table gives the per capita health expenditures in the US.

Year	Expenditure	Year	Expenditure
1940	30	1970	367
1950	82	1971	394
1955	105	1972	438
1960	146	1973	478
1965	211	1974	534
		1975	604

(a) Draw a scatter plot.
(b) Do a regression of the log of expenditures on the year.
(c) Compute the correlation coefficient for log of expenditures and year.
(d) Does the model in (b) seem adequate?

7. The following table gives the death rates (100,000 persons) for cancer of the respiratory system in the US.

Year	Rate	Year	Rate
1950	14.5	1970	47.6
1955	20.6	1975	56.7
1960	30.5	1977	61.5
1965	36	1978	62.5

(a) Draw a scatter plot.
(b) Find the regression line for the death rate on the year.
(c) What is the correlation coefficient for the rate and year?

6.2 Inference for Regression

The regression line introduced in the preceding section is based on a sample. If we change the sample we will get another regression line. Therefore, it is important to know how much confidence we may have on our regression line. In particular, in this section we will get confidence intervals for the coefficients of the regression line. In order to do so we need an underlying probability model that we now formulate. We will assume that the variable y we wish to explain is random and that the explanatory variable x is deterministic (i.e., nonrandom). We have n observations (x_i, y_i) for $i = 1, 2, \ldots, n$. We assume that

$$y_i = bx_i + a + e_i,$$

where the e_i are random variables. We also assume that the e_i are independent and normally distributed with mean 0 and standard deviation σ. Observe that our model has three parameters a, b, and σ. Note also that since $bx_i + a$ is a constant and $E(e_i) = 0$ we get

$$E(y_i) = bx_i + a + E(e_i) = bx_i + a.$$

That is, the model assumes that the mean of a variable y_i corresponding to x_i is $bx_i + a$. However, the random variable y_i varies around its mean. The model assumes that e_i, the variation of y_i around its mean, is normally distributed with standard deviation σ. Therefore, the model makes four major assumptions: the mean of y_i is a linear function of x_i, the variables e_i are normally distributed, the standard deviation is the same for all the e_i and the e_i are independent of each other.

To estimate a and b we use the estimators $\hat{a}$ and $\hat{b}$ that we have given in Sect. 6.1. To predict the y corresponding to x_i we use

$$\hat{y}_i = \hat{b}x_i + \hat{a}.$$

For $i = 1, 2, \ldots, n$ let

$$\hat{e}_i = y_i - \hat{y}_i.$$

They are the *residuals* and they represent the error between the observation y_i and the prediction $\hat{y}_i$.

Estimating the Regression Parameters

Assume that

$$y_i = bx_i + a + e_i \text{ for } i = 1, \ldots, n$$

and that the e_i are independent and normally distributed with mean 0 and standard deviation σ. Then a and b are estimated by

$$\hat{b} = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}$$

and

$$\hat{a} = \bar{y} - \hat{b}\bar{x}.$$

The parameter σ^2 is estimated by

$$s^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - (\hat{b}x_i + \hat{a}))^2 = \frac{1}{n-2}\sum_{i=1}^{n}\hat{e}_i^2.$$

Note that to get s^2 we divide by $n-2$ even though there are n terms in the sum. This is similar to what we do to compute the sample variance of a random variable (there, we divide by $n-1$). By dividing by $n-2$ we get an unbiased estimator of σ^2. That is,

$$E(s^2) = \sigma^2.$$

Example 1. We use the data of Example 1 in Sect. 7.1 to get estimates of the regression parameters. We have already computed the regression line

$$y = 0.25x + 1,131.$$

So to get the predicted $\hat{y}_i$ below we compute

$$\hat{y}_i = 0.25x_i + 1{,}131.$$

	x_i	y_i	$\hat{y}_i$	$\hat{e}_i$	$\hat{e}_i^2$
	884	1,194	1,352	−158	24,964
	720	1,475	1,311	164	26,896
	798	1,365	1,331	−34	1,156
	1,526	1,686	1,513	−173	29,929
	899	1,209	1,356	−147	21,609
	4,719	2,282	2,311	−29	841
	4,916	2,224	2,360	−136	18,496
	1,085	1,311	1,402	−91	8,281
	781	1,317	1,326	−9	81
	4,377	2,422	2,225	197	38,809
Sum:					171,062

Thus, we get

$$s^2 = \frac{1}{n-2}\sum_{i=1}^{n}\hat{e}_i^2 = \frac{1}{8}171{,}062 = 21{,}383.$$

Therefore, $s = 146$ is an estimate of σ. We will now use the estimators $\hat{a}$, $\hat{b}$, and s to get confidence intervals for a and b. We first need information about the distributions of $\hat{a}$ and $\hat{b}$.

Distribution of $\hat{a}$ and $\hat{b}$

Consider the model

$$y_i = bx_i + a + e_i$$

with the assumptions that the e_i are independent, normally distributed with mean 0 and variance σ^2. Assume that we have n observations. Then,

$$\frac{\hat{a} - a}{s_{\hat{a}}} \text{ and } \frac{\hat{b} - b}{s_{\hat{b}}}$$

follow a Student distribution with $n - 2$ degrees of freedom where,

$$s_{\hat{a}} = s\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

and

$$s_{\hat{b}} = \frac{s}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}.$$

Note that the above holds only if we assume that the e_i are independent, NORMALLY distributed with mean 0 and the SAME σ. The statistical analysis we will do below is only valid under these assumptions. We will discuss below how to check these assumptions.

The crucial step in any regression analysis is to test whether or not $b = 0$. If we cannot reject the null hypothesis $b = 0$ then the conclusion should be that there is no statistical evidence that there is a linear relation between the variables y and x. In other words our model is not adequate for the problem and we have to look for another model.

Example 2. We are going to test whether $b = 0$ for the data of Example 1. In order to perform the test we need to compute

$$\sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n}x_i^2 - 2\bar{x}\sum_{i=1}^{n}x_i + n\bar{x}^2 = \sum_{i=1}^{n}x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n}x_i\right)^2.$$

We now may use the table in Example 1, Sect. 6.1 to get

$$\sum_{i=1}^{n}(x_i - \bar{x})^2 = 72{,}454{,}869 - \frac{1}{10}(20{,}705)^2 = 29{,}585{,}166.5.$$

From Example 1, we know that $s = 146$. Thus,

$$s_{\hat{b}} = \frac{s}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}} = \frac{146}{\sqrt{29{,}585{,}166.5}} = 0.03.$$

We have already computed $\hat{b} = 0.25$. The test we want to perform is

$$H_0 : b = 0$$

$$H_a : b \neq 0$$

This is a two-sided test. The rejection region is

$$\{\hat{b} > 0.25 \text{or} \hat{b} < -0.25\}.$$

We now compute the P value for this test. By the symmetry of the rejection region we get

$$P = 2P(\hat{b} > 0.25 | b = 0) = 2P\left(t(n-2) > \frac{0.25}{0.03}\right) = 2P(t(8) > 8).$$

According to the Student table this P value is extremely small and we reject the null hypothesis with very high confidence. That is, there is strong statistical evidence that there is a positive relation between x (state debt) and y (state tax).

Example 3. Since we have rejected the hypothesis $b = 0$ we should find a confidence interval for b. For a confidence interval with 0.95 confidence we want c such that

$$P(|\hat{b} - b| < c) = 0.95.$$

Therefore,

$$P(|\hat{b} - b| < c) = P\left(\frac{|\hat{b} - b|}{s_{\hat{b}}} < \frac{c}{s_{\hat{b}}}\right) = P\left(|t(n-2)| < \frac{c}{s_{\hat{b}}}\right) = 0.95.$$

We use the Student table to get

$$\frac{c}{s_{\hat{b}}} = 2.3.$$

So

$$c = 2.3 \times 0.03 = 0.07.$$

Therefore, a confidence interval at the 95% level for b is (0.18;0.32).

We now turn our attention to confidence intervals for the predicted and mean values of y. For a given value x_0 of the variable x there are two possible interpretations for $\hat{b}x_0 + \hat{a}$. It could be an estimate of the mean value of y corresponding to x_0 or it could be a prediction for a y corresponding to x_0. As we are going to see below there is more variation in predicting an individual y than in predicting its mean.

Predicting $E(y)$

The ratio

$$\frac{E(y) - (bx_0 + a)}{s_y}$$

follows a Student distribution with $n - 2$ degrees of freedom where $E(y) = \hat{b}x_0 + \hat{a}$ and

$$s_y = s\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}.$$

Example 4. We will now compute a confidence interval for the mean tax per capita corresponding to debt per capita of $3,000. We use the data of Example 1. We have $x_0 = 3{,}000$. In Example 2 we have already computed

$$\sum_{i=1}^{n}(x_i - \bar{x})^2 = 29{,}585{,}166.5.$$

We also have that $\bar{x} = 2,070.5$ and that $s = 146$. Therefore,

$$s_y = s\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}} = 146\sqrt{\frac{1}{10} + \frac{(3{,}000 - 2{,}070.5)^2}{29{,}585{,}166.5}} = 52.$$

In this example we have that

$$\hat{b}x_0 + \hat{a} = 1{,}881.$$

Since $\frac{(\hat{b}x_0+\hat{a})-(bx_0+a)}{s_y}$ follows a Student with $n-2=8$ degrees of freedom we get that a confidence interval with 95% confidence for the mean tax corresponding to a debt of 3,000 per capita is

$$(1{,}881-2.3s_y, 1{,}881+2.3s_y) = (1{,}761; 2{,}001).$$

Predicting *y*

The ratio

$$\frac{y-(bx_0+a)}{s_y}$$

follows a Student distribution with $n-2$ degrees of freedom where $y=\hat{b}x_0+\hat{a}$ and

$$s_y = s\sqrt{1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}.$$

Note that the standard error for predicting an individual y is larger than the standard error for estimating the mean $E(y)$.

Example 5. We now compute a confidence interval for a predicted tax based on a debt of \$3,000. The only difference with Example 4 is that the standard deviation s_y is now

$$s\sqrt{1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}} = 155.$$

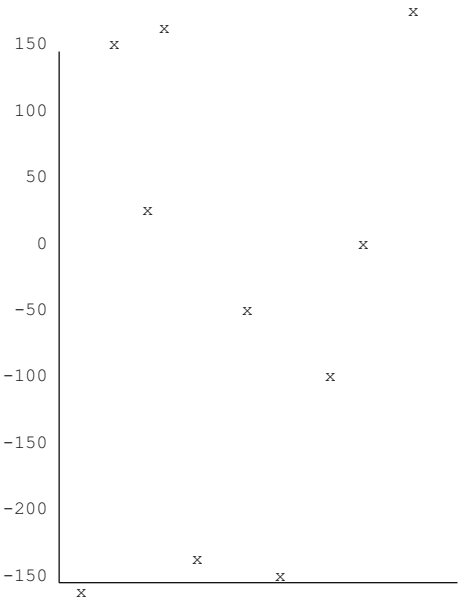
Note that the standard deviation has tripled compared to Example 4. With 95% confidence we get that the predicted tax corresponding to 3,000 debt is in the interval

$$(1,881-2.3s_y; 1,881+2.3s_y) = (1,515; 2,238).$$

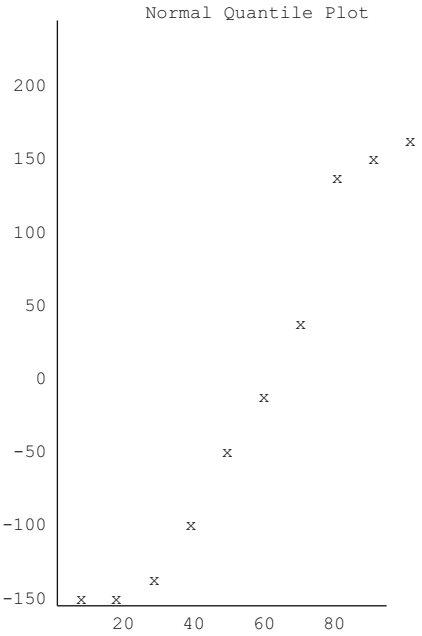
6.2.1 Checking the Assumptions of the Model

We will now check some of the assumptions we made for the model. We start by plotting the residuals $\hat{e}_i$ for the data of Example 1.

The important thing to check here is that there is no special pattern. For instance, it could be that the residuals increase, decrease in a regular way or have a special clustering. In this case there seems that there is no special pattern emerging.



The second important plot is the normal quantile plot for the residuals (see Sect. 5.3 for more details). We now plot the normal quantile plot for the residuals in Example 1.



Recall that when the variable is normally distributed the points in this plot are aligned. The pattern here is not too far from a straight line so the assumption of normality seems reasonable in this case. In summary, based on the two plots above we may conclude that the assumptions of the model (normality and independence of the residuals, same σ) are not violated in this example.

Exercises 6.2

1. (a) Test whether a is 0 for the data of Example 1.
(b) Find a confidence interval for a with confidence 0.99.

2. Consider the data about the US population in Exercise 1 in Sect. 6.1.

(a) Do a regression of the log of the population on the year.
(b) Test whether $b = 0$.
(c) Give a confidence interval for the US population in 2000.

3. Consider the data on death rates from Exercise 5 in Sect. 6.1.

(a) Compute the regression line of log of death rate on age.
(b) Test whether $b = 0$ for the model in (a).
(c) Plot the residuals of the model in (a).
(d) Plot the normal probability quantiles for the residuals.
(e) Based on (c) and (d) would you say that the assumptions of the model hold in this case?

4. Consider the data in Exercise 6 in Sect. 6.1.

(a) Do a regression of the log of expenditures on the year.
(b) Test the adequacy of the model.
(c) Test the assumptions of the model.

5. Consider the data of Exercise 4 in Sect. 6.1.

(a) Do a regression of the tuberculosis rate on the AIDS rate.
(b) Test whether there is a linear relation between the two rates.

6. Consider the data of Exercise 6.7 in Sect. 6.1 about the death rate for cancer of the respiratory system.

(a) Test whether there is a linear relation between death rate and year.
(b) Give a confidence interval for the death rate of year 1985.

7. The table below gives the number of years a person alive at 65 in a given year is expected to live.

Year	Life expectancy	Year	Life expectancy
1900	11.9	1975	16.0
1950	13.9	1977	16.3
1960	14.3	1979	16.3
1970	15.2		

(a) Test whether there is a linear relation between life expectancy and year.
(b) Give a confidence interval for the life expectancy at age 65 in the year 2000.
(c) Does the answer in (b) look accurate?

Chapter 7
Moment Generating Functions and Sums of Independent Random Variables

7.1 Moment Generating Functions

The purpose of this chapter is to introduce moment generating functions (mgf). We have two applications in mind that will be covered in the next section. We will compute the distribution of some sums of independent random variables and we will indicate how moment generating functions may be used to prove the Central Limit Theorem. We start by defining moment generating functions.

Moment Generating Functions

The moment generating function of a random variable X is defined by

$$M_X(t) = E(\mathrm{e}^{tX}).$$

In particular, if X is a discrete random variable then

$$M_X(t) = \sum_k \mathrm{e}^{tk} P(X = k).$$

If X is a continuous random variable and has a density f then

$$M_X(t) = \int \mathrm{e}^{tx} f(x)\mathrm{d}x.$$

Note that the mgf is not necessarily defined for all t (because of convergence problems of the series or of the generalized integral). It is useful even if it is defined on a small interval. We start by computing some mgf.

R.B. Schinazi, *Probability with Statistical Applications*,
DOI 10.1007/978-0-8176-8250-7_7,

Example 1. Consider a binomial random variable S with parameters n and p. Compute its mgf.

We have that

$$M_S(t) = E(\mathrm{e}^{tS}) = \sum_{k=0}^{n} \mathrm{e}^{tk} P(S = k) = \sum_{k=0}^{n} \mathrm{e}^{tk} \binom{n}{k} p^k (1-p)^{n-k}$$

$$= \sum_{k=0}^{n} \binom{n}{k} (\mathrm{e}^t p)^k (1-p)^{n-k}.$$

We now use the binomial Theorem

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

with $x = \mathrm{e}^t p$ and $y = (1-p)$ to get

$$M_S(t) = (p\mathrm{e}^t + 1 - p)^n \text{ for all } t.$$

Example 2. Let N be a Poisson random variable with mean λ. We have

$$M_N(t) = E(\mathrm{e}^{tN}) = \sum_{k=0}^{\infty} \mathrm{e}^{tk} P(N = k) = \sum_{k=0}^{\infty} \mathrm{e}^{tk} \mathrm{e}^{-\lambda} \frac{\lambda^k}{k!} = \sum_{k=0}^{\infty} \mathrm{e}^{-\lambda} \frac{(\mathrm{e}^t \lambda)^k}{k!}.$$

Recall that

$$\mathrm{e}^x = \exp(x) = \sum_{k=0}^{\infty} \frac{x^k}{k!}.$$

We use this power series expansion with $x = \mathrm{e}^t \lambda$ to get

$$M_N(t) = \mathrm{e}^{-\lambda} \exp(\mathrm{e}^t \lambda) = \exp(\lambda(-1 + \mathrm{e}^t)) \text{ for all } t.$$

We now give an example of computation of mgf for a continuous random variable.

Example 3. Assume X is exponentially distributed with rate λ. Its mgf is

$$M_X(t) = E(\mathrm{e}^{tX}) = \int_0^{\infty} \mathrm{e}^{tx} \lambda \mathrm{e}^{-\lambda x} \mathrm{d}x = \int_0^{\infty} \lambda \mathrm{e}^{(t-\lambda)x} \mathrm{d}x.$$

Note that the preceding improper integral is convergent only if $t - \lambda < 0$. In that case we get

$$M_X(t) = \frac{\lambda}{\lambda - t} \text{ for } t < \lambda.$$

The moment generating functions get their name from the following property.

Moments of a Random Variable

Let X be a random variable and $k \geq 1$ be an integer. The expectation $E(X^k)$ is called the kth moment of X. If X has a moment generating function M_X defined on some interval $(-r, r)$ for $r > 0$ then all the moments of X exist and

$$E(X^k) = M_X^{(k)}(0),$$

where $M_X^{(k)}$ designates the kth derivative of M_X.

Example 4. We will use the formula above to compute the moments of the Poisson distribution. Let N be a Poisson random variable with mean λ. Then M_N is defined everywhere and

$$M_N(t) = \exp(\lambda(-1 + \mathrm{e}^t)).$$

Note that the first derivative is

$$M_N'(t) = \lambda \mathrm{e}^t \exp(\lambda(-1 + \mathrm{e}^t)).$$

Letting $t = 0$ in the formula above yields

$$E(X) = M_N'(0) = \lambda.$$

We now compute the second derivative

$$M_N''(t) = \lambda \mathrm{e}^t \exp(\lambda(-1 + \mathrm{e}^t)) + \lambda^2 \mathrm{e}^{2t} \exp(\lambda(-1 + \mathrm{e}^t)).$$

Thus,

$$E(X^2) = M_N''(0) = \lambda + \lambda^2.$$

Note that

$$\mathrm{Var}(X) = E(X^2) - E(X)^2 = \lambda.$$

Example 5. We now compute the mgf of a standard normal distribution. Let Z be a standard normal distribution. We have

$$M_Z(t) = E(\mathrm{e}^{Zt}) = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{zt} \mathrm{e}^{-z^2/2} \mathrm{d}z = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{zt - z^2/2} \mathrm{d}z.$$

Note that we may "complete the square" to get

$$zt - z^2/2 = -(z-t)^2/2 + t^2/2.$$

Thus,

$$M_Z(t) = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-(z-t)^2/2+t^2/2} \mathrm{d}z = \mathrm{e}^{t^2/2} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-(z-t)^2/2} \mathrm{d}z.$$

Note that $g(z) = \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-(z-t)^2/2}$ is the density of a normal distribution with mean t and standard deviation 1. Thus,

$$\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-(z-t)^2/2} \mathrm{d}z = 1$$

and

$$M_Z(t) = \mathrm{e}^{t^2/2}.$$

Example 6. We may use Example 5 to compute the moments of a standard normal distribution.

$$M_Z'(t) = t\mathrm{e}^{t^2/2}.$$

Letting $t = 0$ above we get

$$E(Z) = 0.$$

We have

$$M_Z''(t) = \mathrm{e}^{t^2/2} + t^2\mathrm{e}^{t^2/2}.$$

So

$$E(Z^2) = M_Z''(0) = 1.$$

We also compute the third moment

$$M_Z^{(3)}(t) = t\mathrm{e}^{t^2/2} + 2t\mathrm{e}^{t^2/2} + t^3\mathrm{e}^{t^2/2}.$$

We get

$$E(Z^3) = M_Z^{(3)}(0) = 0.$$

Example 7. We now use the computation in Example 5 to compute the mgf of a normal random variable X with mean μ and standard deviation σ. We have used already many times the fact that the random variable Z defined as

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

is a standard normal distribution. We will provide a proof in Sect. 8.1. Assuming this fact we have

$$M_X(t) = M_{\sigma Z+\mu}(t) = E(\mathrm{e}^{t(\sigma Z+\mu)}).$$

Observe that $\mathrm{e}^{t\mu}$ is a constant so

$$M_X(t) = \mathrm{e}^{t\mu} E(\mathrm{e}^{t\sigma Z}) = M_Z(t\sigma).$$

We now use that $M_Z(t) = \mathrm{e}^{t^2/2}$ to get

$$M_X(t) = \exp(t\mu)\exp(t^2\sigma^2/2) = \exp(t\mu + t^2\sigma^2/2).$$

Our next example deals with the Gamma distribution.

Example 8. A random variable X is said to have a Gamma distribution with parameters $r > 0$ and $\lambda > 0$ if its density is

$$f(x) = \frac{\lambda^r}{\Gamma(r)}\mathrm{e}^{-\lambda x}x^{r-1} \text{ for } x > 0,$$

where

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

The improper integral above is convergent for all $r > 0$. Moreover, an easy induction proof shows that

$$\Gamma(n) = (n-1)! \text{ for all integers } n \geq 1.$$

Observe that a Gamma random variable with parameters $r = 1$ and λ is an exponential random variable with parameter λ.

We now compute the mgf of a Gamma random variable with parameters r and λ.

$$M_X(t) = E(\mathrm{e}^{tX}) = \int_0^\infty \mathrm{e}^{tx}\frac{\lambda^r}{\Gamma(r)}\mathrm{e}^{-\lambda x}x^{r-1}\mathrm{d}x.$$

The preceding improper integral converges only for $t < \lambda$. Note that

$$g(x) = \frac{(\lambda - t)^r}{\Gamma(r)} x^{r-1} \mathrm{e}^{-(\lambda - t)x}$$

is the density of a Gamma random variable with parameters r and $\lambda - t$. Thus,

$$\int_0^\infty g(x)\mathrm{d}x = \frac{(\lambda - t)^r}{\Gamma(r)} \int_0^\infty x^{r-1} \mathrm{e}^{-(\lambda - t)x} \mathrm{d}x = 1.$$

Hence,

$$\int_0^\infty x^{r-1} \mathrm{e}^{-(\lambda - t)x} \mathrm{d}x = \frac{\Gamma(r)}{(\lambda - t)^r}$$

and

$$M_X(t) = \int_0^\infty \mathrm{e}^{tx} \frac{\lambda^r}{\Gamma(r)} \mathrm{e}^{-\lambda x} x^{r-1} \mathrm{d}x = \frac{\lambda^r}{\Gamma(r)} \frac{\Gamma(r)}{(\lambda - t)^r} = \frac{\lambda^r}{(\lambda - t)^r} \text{ for } t < \lambda.$$

Example 9. We now compute the expected value and the variance of a Gamma distribution with parameters $r > 0$ and $\lambda > 0$. According to Example 8 we have

$$M_X(t) = \frac{\lambda^r}{(\lambda - t)^r} \text{ for } t < \lambda.$$

Hence,

$$M_X'(t) = (-r)(-1) \frac{\lambda^r}{(\lambda - t)^{r+1}}$$

and

$$M_X'(0) = r \frac{\lambda^r}{\lambda^{r+1}} = \frac{r}{\lambda}.$$

Therefore,

$$E(X) = \frac{r}{\lambda}.$$

We now turn to the second moment

$$M_X''(t) = r(-r - 1)(-1) \frac{\lambda^r}{(\lambda - t)^{r+2}}.$$

In particular,

$$M_X''(0) = r(r + 1) \frac{\lambda^r}{\lambda^{r+2}} = \frac{r(r + 1)}{\lambda^2}.$$

Therefore,

$$\text{Var}(X) = E(X^2) - E(X)^2 = \frac{r(r+1)}{\lambda^2} - \left(\frac{r}{\lambda}\right)^2 = \frac{r}{\lambda^2}.$$

We summarize our findings about the Gamma distribution below.

Gamma Distribution

A random variable X is said to have a Gamma distribution with parameters $r > 0$ and $\lambda > 0$ if its density is

$$f(x) = \frac{\lambda^r}{\Gamma(r)} e^{-\lambda x} x^{r-1} \text{ for } x > 0$$

where

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

Its moment generating function is

$$M_X(t) = \frac{\lambda^r}{(\lambda - t)^r} \text{ for } t < \lambda.$$

We also have

$$E(X) = \frac{r}{\lambda} \text{ and } \text{Var}(X) = \frac{r}{\lambda^2}.$$

Exercises 7.1

1. Compute the moment generating function of a geometric random variable with parameter p.

2. Compute the mgf of an uniform random variable on [0,1].

3. Compute the first three moments of a binomial random variable by taking derivatives of its mgf.

4. Compute the first moment of a geometric random variable by using Exercise 7.1.

5. Compute the first two moments of an uniform random variable on [0,1] by using Exercise 7.2.

6. Compute the fourth moment of a standard normal distribution Z.

7. What is the mgf of a normal distribution with mean 1 and standard deviation 2?

8. Use the mgf in Example 7 to compute the first two moments of a normal distribution with mean μ and standard deviation σ.

9. We defined the function Γ for all $r > 0$ by

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

(a) Make a change of variables to show that for all $\lambda > 0$ and $r > 0$

$$\int_0^\infty e^{-\lambda x} \lambda^r x^{r-1} dx = \Gamma(r).$$

(b) Show that for all $r > 0$ and $\lambda > 0$

$$\frac{\lambda^r}{\Gamma(r)} \int_0^\infty e^{-\lambda x} x^{r-1} dx = 1$$

(c) Show that

$$\Gamma(n) = (n-1)! \text{ for all integers } n \geq 1.$$

10. A random variable with density

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

is said to be a Chi-square random variable with n degrees of freedom ($n \geq 1$ is an integer).

(a) Show that a Chi-square random variable is also a Gamma random variable.
(b) Find the expected value and the variance of a Chi-square random variable with n degrees of freedom.

11. Show that the improper integral

$$\int_0^\infty e^{-\lambda x} \lambda^r x^{r-1} dx$$

converges for all $r > 0$. (You need to show convergence near 0 when $0 < r < 1$ and near ∞ for all $r > 0$).

12. Let Z be a standard normal distribution.

(a) Show that for any integer $k \geq 2$ we have

$$E(Z^k) = (k-1)E(Z^{k-2}).$$

(Use integration by parts.)

(b) Use (a) to compute $E(Z^4)$ and $E(Z^5)$.
(c) More generally, show that for all integer $n \geq 2$ we have

$$E(Z^{2n-1}) = 0 \text{ and } E(Z^{2n}) = 1 \times 3 \times 5 \times \ldots (2n-1).$$

7.2 Sums of Independent Random Variables

We first summarize the mgf we have computed in Sect. 7.1.

Random variable	Moment generating function
Binomial (n, p)	$(pe^t + 1 - p)^n$
Poisson (λ)	$\exp(\lambda(-1 + e^t))$
Exponential (λ)	$\frac{\lambda}{\lambda - t}$ for $t < \lambda$
Normal (μ, σ^2)	$\exp(t\mu + t^2\sigma^2/2)$
Gamma (r, λ)	$\frac{\lambda^r}{(\lambda - t)^r}$ for $t < \lambda$

We will use moment generating functions to show the following important property of normal random variables.

Linear Combination of Independent Normal Random Variables

Assume that $X_1, X_2, \ldots, X_n$ are independent normal random variables with mean μ_i and variance σ_i^2. Let $a_1, a_2, \ldots, a_n$ be a sequence of real numbers. Then

$$a_1X_1 + a_2X_2 + \cdots + a_nX_n$$

is also a normal variable with mean

$$a_1\mu_1 + a_2\mu_2 + \cdots + a_n\mu_n$$

and variance

$$a_1^2\sigma_1^2 + a_2^2\sigma_2^2 + \cdots + a_n^2\sigma_n^2.$$

In words, a linear combination of independent normal random variables is normal. We apply this property in the following two examples.

Example 1. Assume that in a population heights are normally distributed. The mean height for men is 172 cm with SD 5 cm and for women the mean is 165 cm with SD 3 cm. What is the probability that a woman taken at random be taller than a man taken at random?

Let X be the man's height and let Y be the woman's height. We want $P(X < Y) = P(Y - X > 0)$. According to the preceding property $Y - X$ is normally distributed with

$$E(Y - X) = E(Y) - E(X) = 165 - 172 = -7$$

and

$$\text{Var}(Y - X) = \text{Var}(Y) + \text{Var}(X) = 3^2 + 5^2 = 34.$$

We normalize $Y - X$ to get

$$P(X < Y) = P(Y - X > 0) = p\left(\frac{Y - X - (-7)}{\sqrt{34}} > \frac{0 - (-7)}{\sqrt{34}}\right)$$
$$= P\left(Z > \frac{7}{\sqrt{34}}\right) = 0.12.$$

Example 2. Assume that at a University salaries of junior faculty are normally distributed with mean 40,000 and SD 5,000. Assume also that salaries of senior faculty are normally distributed with mean 60,000 and SD 10,000. What is the probability that the salary of a senior faculty taken at random is at least twice the salary of a junior faculty taken at random?

Let X be the salary of the junior faculty and Y be the salary of the senior faculty. We want $P(Y > 2X)$. We know that $Y - 2X$ is normally distributed. We express all the figures in thousands of dollars to get

$$E(Y - 2X) = -20 \text{ and } \text{Var}(Y - 2X) = \text{Var}(Y) + 4\text{Var}(X) = 10^2 + 4 \times 5^2 = 200.$$

We normalize to get

$$(Y - 2X > 0) = P\left(\frac{Y - 2X - (-20)}{\sqrt{200}} > \frac{0 - (-20)}{\sqrt{200}}\right)$$
$$= P\left(Z > \frac{20}{\sqrt{200}}\right) = 0.08.$$

Before proving that a linear combination of independent normally distributed random variables is normally distributed we need two properties of moment generating functions that we now state.

P1. The moment generating function of a random variable characterizes its distribution. That is, if two random variables X and Y are such that

$$M_X(t) = M_Y(t) \text{ for all } t \text{ in } (-r, r)$$

for some $r > 0$ then X and Y have the same distribution.

P1 is a crucial property. It tells us that if we recognize a moment generating function then we know what the underlying distribution is.

P2. Assume that the random variables $X_1, X_2, \ldots, X_n$ are independent and have moment generating functions. Let $S = X_1 + X_2 + \cdots + X_n$, then

$$M_S(t) = M_{X_1}(t)M_{X_2}(t)\ldots M_{X_n}(t).$$

The proof of P1 involves mathematics that are beyond the scope of this book. For a proof of P2 see P2 in Sect. 8.3. We now prove that a linear combination of independent normally distributed random variables is normally distributed. Assume that $X_1, X_2, \ldots, X_n$ are independent normal random variables with mean μ_i and variance σ_i^2. Let $a_1, a_2, \ldots, a_n$ be a sequence of real numbers. We compute the mgf of $a_1X_1 + a_2X_2 + \cdots + a_nX_n$. The random variables a_iX_i are independent so by P2 we have

$$M_{a_1X_1+a_2X_2+\cdots+a_nX_n}(t) = M_{a_1X_1}(t)M_{a_2X_2}(t)\ldots M_{a_nX_n}(t).$$

Note that by definition

$$M_{a_iX_i}(t) = E(\mathrm{e}^{ta_iX_i}) = M_{X_i}(a_it).$$

We now use the mgf corresponding to the normal distribution to get

$$M_{a_iX_i}(t) = \exp(a_it\mu_i + a_i^2t^2\sigma_i^2/2).$$

Thus,

$$M_{a_1X_1+a_2X_2+\cdots+a_nX_n}(t) = \exp(a_1t\mu_1 + a_1^2t^2\sigma_1^2/2) \times \cdots \times \exp(a_nt\mu_n + a_n^2t^2\sigma_2^2/2).$$

Therefore,

$$M_{a_1X_1+a_2X_2+\cdots+a_nX_n}(t) = \exp((a_1\mu_1 + \cdots + a_n\mu_n)t + (a_1^2\sigma_1^2 + \cdots + a_n^2\sigma_n^2)t^2/2).$$

This is exactly the mgf of a normal random variable with mean

$$a_1\mu_1 + \cdots + a_n\mu_n$$

and variance

$$a_1^2\sigma_1^2 + \cdots + a_n^2\sigma_n^2.$$

So according to property P1 this shows that $a_1X_1 + \cdots + a_nX_n$ follows a normal distribution with mean and variance given above.

Example 3. Let $T_1, \ldots, T_n$ be i.i.d. exponentially distributed random variables with rate λ. What is the distribution of $T_1 + T_2 + \cdots + T_n$?

We compute the mgf of the sum by using Property P2.

$$M_{T_1+T_2+\cdots+T_n}(t) = M_{T_1}(t)M_{T_2}(t)\ldots M_{T_n}(t) = M_{T_1}^n(t)$$

since all the T_i have all the same distribution they also have the same mgf

$$M_{T_1}(t) = \frac{\lambda}{\lambda - t}.$$

Hence,

$$M_{T_1+T_2+\cdots+T_n}(t) = \left(\frac{\lambda}{\lambda - t}\right)^n.$$

This is not the mgf of an exponential distribution. However, it is the mgf of a Gamma distribution with parameters n and λ. That is, we have the following.

Sum of i.i.d. Exponential Random Variables

Let $T_1, \ldots, T_n$ be i.i.d. exponentially distributed random variables with rate λ. Then $T_1 + T_2 + \cdots + T_n$ has a Gamma distribution with parameters n and λ.

Example 4. Assume that you have two batteries that have an exponential lifetime with mean 2 h. As soon as the first battery fails you replace it with a second battery. What is the probability that the batteries will last at least 4 h?

The total time, T, the batteries will last is a sum of two exponential i.i.d. random variables. Therefore, T follows a Gamma distribution with parameters $n = 2$ and $\lambda = 1/2$. We use the density of a Gamma distribution (see Example 8 in 1 and note that $\Gamma(2) = 1$) to get

$$P(T > 4) = \int_4^\infty \lambda^2 t \mathrm{e}^{-\lambda t}\mathrm{d}t = 3\mathrm{e}^{-2} = 0.41,$$

where we use an integration by parts to get the second equality.

Example 5. Let X and Y be two independent Poisson random variables with means λ and μ, respectively. What is the distribution of $X + Y$?

We compute the mgf of $X + Y$. By property P2 we have that

$$M_{X+Y}(t) = M_X(t)M_Y(t).$$

Thus,

$$M_{X+Y}(t) = \exp(\lambda(-1 + \mathrm{e}^t)) \times \exp(\mu(-1 + \mathrm{e}^t)) = \exp((\lambda + \mu)(-1 + \mathrm{e}^t)).$$

This is the moment generating function of a Poisson random variable with mean $\lambda + \mu$. Thus, by property P1, $X + Y$ is a Poisson random variable with mean $\lambda + \mu$.

We state the general result.

Sum of Independent Poisson Random Variables

Let $N_1, \ldots, N_n$ be independent Poisson random variables with means $\lambda_1, \ldots, \lambda_n$, respectively. Then,

$$N_1 + N_2 + \cdots + N_n$$

is a Poisson random variable with mean

$$\lambda_1 + \lambda_2 + \cdots + \lambda_n.$$

Only a few distributions are stable under addition. Normal and Poisson distributions are two of them.

Example 6. Assume that at a given hospital there is on average two births of twins per month and one birth of triplets per year. Assume that both are Poisson random variables. What is the probability that on a given month there are four or more multiple births?

Let N_1 and N_2 be the number of births of twins and of triplets on a given month, respectively. Then $N = N_1 + N_2$ is a Poisson random variable with mean $\lambda = 2 + 1/12 = 25/12$. We have that

$$\begin{aligned} P(N \geq 4) &= 1 - P(N = 0) - P(N = 1) - P(N = 2) - P(N = 3) \\ &= 1 - \mathrm{e}^{-\lambda} - \lambda \mathrm{e}^{-\lambda} - \lambda^2 \mathrm{e}^{-\lambda}/2 - \lambda^3 \mathrm{e}^{-\lambda}/3! = 0.16. \end{aligned}$$

As noted before, when we sum two random variables with the same type of distribution we do not, in general, get the same distribution. Next, we will look at such an example.

Example 7. Roll two fair dice. What is the distribution of the sum?

Let X and Y be the faces shown by the two dice. The random variables X and Y are discrete uniform random variables on $\{1, 2 \ldots, 6\}$. Let $S = X + Y$. Note that S must be an integer between 2 and 12. We have that

$$P(S = 2) = P(X = 1; Y = 1) = P(X = 1)P(Y = 1) = \frac{1}{36},$$

where we use the independence of X and Y to get the second equality. Likewise, we have that

$$P(S = 3) = P(X = 1; Y = 2) + P(X = 2; Y = 1) = \frac{2}{36}.$$

The method above can be applied to get $P(S = n)$ for any n. We get

$$P(S = n) = \sum_{k=1}^{n-1} P(X = k)P(Y = n - k) \text{ for } n = 2, 3 \ldots, 12.$$

The computations yield

k	2	3	4	5	6	7	8	9	10	11	12
$P(X = k)$	1/36	2/36	3/36	4/36	5/36	6/36	5/36	4/36	3/36	2/36	1/36

Note that S is not an uniform random variable. In this case using the moment generating function does not help. We could compute the mgf of S but it would not correspond to any distribution we know.

We now state the general form of the distribution of the sum of two independent random variables.

Sum of Two Independent Random Variables

Let X and Y be two discrete independent random variables. The distribution of $X + Y$ may be computed by using the following formula.

$$P(X + Y = n) = \sum_{k} P(X = k)P(Y = n - k).$$

If X and Y are independent continuous random variables with densities f and g then $X + Y$ has density h that may be computed by using the formula

$$h(x) = \int_{-\infty}^{+\infty} f(y)g(x - y)\mathrm{d}y = \int_{-\infty}^{+\infty} g(y)f(x - y)\mathrm{d}y.$$

The operation $\int_{-\infty}^{+\infty} f(y)g(x - y)\mathrm{d}y$ is called the convolution of f and g. The convolution formula for continuous random variables will be proved in Chap. 8. Next, we apply the preceding formula to uniform random variables.

Example 8. Let U and V be two independent uniform random variables on [0,1]. The density for both of them is $f(x) = 1$ for x in [0,1]. Let $S = U + V$ and let h be the density of S. We have that

$$h(x) = \int_{-\infty}^{+\infty} f(y)f(x - y)\mathrm{d}y.$$

Note that $f(y) > 0$ if and only if y is in [0,1]. Note also that $f(x-y) > 0$ if and only if $x - y$ is in [0,1], that is y is in $[-1+x, x]$. Thus, $f(y)f(y-x) > 0$ if and only if y is simultaneously in [0,1] and in $[-1+x, x]$. So

$$h(x) = \int_0^x \mathrm{d}y = x \text{ if } x \text{ is in } [0,1]$$

and

$$h(x) = \int_{-1+x}^1 \mathrm{d}y = 2 - x \text{ if } x \text{ is in } [1,2].$$

Observe that the sum of two uniform random variables is not uniform, the density has a triangular shape instead.

Example 9. Let X and Y be two independent exponentially distributed random variables with rates 1 and 2, respectively. What is the density of $X + Y$?

The densities of X and Y are $f(x) = \mathrm{e}^{-x}$ for $x > 0$ and $g(x) = 2\mathrm{e}^{-2x}$ for $x > 0$, respectively. The density h of the sum $X + Y$ is

$$h(x) = \int_{-\infty}^{+\infty} f(y)g(x-y).$$

In order for $f(y)g(x-y) > 0$ we need $y > 0$ and $x - y > 0$. Thus,

$$h(x) = \int_0^x \mathrm{e}^{-y} 2\mathrm{e}^{-2(x-y)} \mathrm{d}y \text{ for } x > 0.$$

We get

$$h(x) = 2(\mathrm{e}^{-x} - \mathrm{e}^{-2x}) \text{ for } x > 0.$$

Note that this is not the density of an exponential distribution. If the two rates were the same we would have obtained a Gamma distribution for the sum but with different rates we get yet another distribution.

Our final application of moment generating functions regards the convergence of sequences of random variables. Our main tool will be the following Theorem.

Convergence in Distribution

Consider a sequence of random variables $X_1, X_2, \ldots$ and their corresponding moment generating functions $M_1, M_2, \ldots$. Assume that there is $r > 0$ such that for every t in $(-r, r)$

$$\lim_{n\to\infty} M_n(t) = M(t).$$

Then M is the moment generating function of some random variable X and the distribution of X_n approaches the distribution of X as n goes to infinity.

The result above is a particular case of a deep probability result called Levy's Continuity Theorem.

Example 10. Consider a sequence of binomial random variables X_n for $n \geq 1$. Each X_n is a binomial with parameters (n, p_n) where p_n is a sequence of strictly positive numbers. Assume also that np_n converges to some $\lambda > 0$ as n goes to infinity. We will show that X_n converges in distribution to a mean λ Poisson random variable.

For $n \geq 1$ the mgf of X_n is

$$M_n(t) = (1 - p_n + p_n e^t)^n$$

and so

$$\ln M_n(t) = n \ln(1 - p_n + p_n e^t).$$

Observe that

$$p_n = \frac{np_n}{n}$$

and since np_n converges to λ we see that p_n converges to 0. Hence, for fixed t, $-p_n + p_n e^t$ converges to 0 as well. We multiply and divide by $-p_n + p_n e^t$ to get

$$\ln M_n(t) = \frac{\ln(1 - p_n + p_n e^t)}{-p_n + p_n e^t} n(-p_n + p_n e^t).$$

Now, recall from Calculus that

$$\lim_{x \to 0} \frac{\ln(1 + x)}{x} = 1.$$

In particular if x_n is a nonzero sequence converging to 0 we have

$$\lim_{n \to \infty} \frac{\ln(1 + x_n)}{x_n} = 1.$$

We apply this result to $x_n = -p_n + p_n e^t$ to get

$$\lim_{n \to \infty} \frac{\ln(1 - p_n + p_n e^t)}{-p_n + p_n e^t} = 1.$$

Observe also that

$$n(-p_n + p_n e^t) = np_n(-1 + e^t)$$

converges to $\lambda(-1 + e^t)$. Hence,

$$\ln M_n(t) \text{ converges to } \lambda(-1 + e^t)$$

and

$$M_n(t) \text{ converges to } \exp(\lambda(-1 + e^t)).$$

Since $\exp(\lambda(-1 + e^t))$ is the mgf of a mean λ Poisson random variable we have proved that the sequence X_n converges in distribution to a mean λ Poisson distribution.

Next we prove the Central Limit Theorem using the same technique.

7.2.1 Proof of the Central Limit Theorem

We now sketch the proof of the Central Limit Theorem in the particular case where the random variables have moment generating functions (in the general case it is only assumed that the random variables have finite second moments). Let $X_1, X_2, \ldots, X_n$ be a sequence of independent identically distributed random variables with mean μ and variance σ^2. Assume that there is $r > 0$ such that each X_i has a mgf defined on $(-r, r)$. Let

$$\bar{X}_n = \frac{X_1 + X_2 + \cdots + X_n}{n}.$$

We want to show that the distribution of

$$T_n = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}$$

approaches the distribution of a standard normal distribution. We start by computing the moment generating function of T_n. For every n we denote the mgf of T_n by M_n. By definition of the mgf

$$M_n(t) = E(e^{tT_n}) = E\left(\exp\left(t\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}\right)\right) = E\left(\exp\left(t\sqrt{n}\frac{\bar{X}_n - \mu}{\sigma}\right)\right).$$

Observe now that

$$\frac{\bar{X}_n - \mu}{\sigma} = \frac{1}{n}\sum_{i=1}^{n}\frac{X_i - \mu}{\sigma}.$$

Let $Y_i = \frac{X_i - \mu}{\sigma}$. We have that

$$M_n(t) = E\left(\exp\left(t\frac{\sqrt{n}}{n}\sum_{i=1}^{n} Y_i\right)\right).$$

Since the Y_i are independent we get by P2 that

$$M_n(t) = M_Y\left(\frac{t}{\sqrt{n}}\right)^n.$$

We now write a third degree Taylor expansion for M_Y.

$$M_Y\left(\frac{t}{\sqrt{n}}\right) = M_Y(0) + \frac{t}{\sqrt{n}}M_Y'(0) + \frac{t^2}{2n}M_Y''(0) + \frac{t^3}{6n^{3/2}}M_Y'''(s_n)$$

for some s_n in $(0, \frac{t}{\sqrt{n}})$. Since the Y_i are standardized we have that

$$M_Y'(0) = E(Y) = 0 \text{ and } M_Y''(0) = E(Y^2) = \text{Var}(Y) = 1.$$

We also have (for any random variable) that $M_Y(0) = 1$. Thus,

$$M_Y\left(\frac{t}{\sqrt{n}}\right) = 1 + \frac{t^2}{2n} + \frac{t^3}{6n^{3/2}}M_Y'''(s_n),$$

and

$$\begin{aligned}\ln(M_n(t)) &= \ln\left(M_Y\left(\frac{t}{\sqrt{n}}\right)^n\right) = n\ln\left(M_Y\left(\frac{t}{\sqrt{n}}\right)\right)\\ &= n\ln\left(1 + \frac{t^2}{2n} + \frac{t^3}{6n^{3/2}}M_Y'''(s_n)\right).\end{aligned}$$

Let $x_n = \frac{t^2}{2n} + \frac{t^3}{6n^{3/2}}M_Y'''(s_n)$. Then

$$n\ln\left(M_Y\left(\frac{t}{\sqrt{n}}\right)\right) = n\ln(1 + x_n) = nx_n\frac{\ln(1 + x_n)}{x_n}.$$

Since s_n converges to 0 $M_Y'''(s_n)$ converges to $M_Y'''(0)$ as n goes to infinity and x_n converges to 0. Thus,

$$\lim_{n\to\infty} nx_n = \frac{t^2}{2} \text{ and } \lim_{n\to\infty}\frac{\ln(1 + x_n)}{x_n} = 1.$$

Therefore,

$$\lim_{n\to\infty}\ln(M_n(t)) = \frac{t^2}{2} \text{ and } \lim_{n\to\infty} M_n(t) = \mathrm{e}^{t^2/2}.$$

That is, the sequence of moment generating functions of the sequence $\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}$ converges (as n goes to infinity) to the moment generating function of a standard normal distribution. This is enough to prove that the distribution of $\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}$ converges to the distribution of a standard normal random variable and concludes the sketch of the proof of the CLT.

Exercises 7.2

1. The weight of a manufactured product is normally distributed with mean 5 kg and SD 0.1 kg.

(a) Take two items at random what is the probability that they have a weight difference of at least 0.3 kg?
(b) Take three items at random, what is the probability that the sum of the three weights is less than 14.8 kg?

2. Consider X a binomial random variable with parameter $n = 10$ and p. Let Y be independent of X and be a binomial random variable with $n = 15$ and p. Let $S = X + Y$.

(a) Find the mgf of S.
(b) What is the distribution of S?

3. Let X be normally distributed with mean 10 and SD 1. Let $Y = 2X - 30$.

(a) Compute the mgf of Y.
(b) Use (a) to show that Y is normally distributed and to find the mean and SD of Y.

4. Let X be the number of students from University A that get into Medical School at University B. Let Y be the number of students from University A that get into Law School at University B. Assume that X and Y are two independent Poisson random variables with means 2 and 3, respectively. What is the probability that $X + Y$ is larger than 5?

5. Assume that 6 years old weights are normally distributed with mean 20 kg and SD 3 kg. Assume that male adults weights are normally distributed with mean 70 kg and SD 6 kg. What is the probability that the sum of the weights of three children is larger than an adult's weight?

6. Assume you roll a die 3 times, you win each time you get a 6. Assume you toss a coin twice, you win each time heads come up. Compute the distribution of your total number of wins.

7. Find the density of a sum of three independent uniform random variables on [0,1]. You may use the result for the sum of two uniform random variables in Example 8.

8. Let X and Y be two independent random variables with density $f(x) = 2x$ for $0 < x < 1$. Find the density of $X + Y$.

9. (a) Use moment generating functions to show that if X and Y are independent binomial random variables with parameters n and p, and m and p, respectively, then $X + Y$ is also a binomial random variable.
(b) If the probability of success are distinct for X and Y, is $X + Y$ a binomial random variable?

10. Let X and Y be two geometric random variables with the same parameter p. That is,

$$P(X = k) = P(Y = k) = p(1-p)^{k-1} \text{ for } k = 1, 2, \ldots$$

Show that

$$P(X + Y = n) = (n-1)p^2(1-p)^{n-2} \text{ for } n = 2, 3, \ldots.$$

11. Let X and Y be two independent exponential random variables with parameters a and b, respectively. Assume that $a \neq b$. Find the density of $X + Y$.

12. Let $X_1, X_2, \ldots, X_n$ be a sequence of independent Gamma random variables with parameters $(r_1, \lambda), (r_2, \lambda), \ldots, (r_n, \lambda)$. That is, they all have the same parameter λ but have possibly different parameters r_i. Show that $X_1 + X_2 + \cdots + X_n$ is also a Gamma random variable. With what parameter?

13. Let X and Y be two independent Gamma random variables with parameters $(r, 1)$ and $(s, 1)$, respectively.

(a) We know that $X + Y$ has density h such that

$$h(x) = \int f(y)g(x-y)\mathrm{d}y,$$

where f and g are the densities of Y and X, respectively. Show that

$$h(x) = \frac{1}{\Gamma(s)\Gamma(r)} \exp(-x) \int_0^x (x-y)^{s-1} y^{r-1} \mathrm{d}y.$$

(b) By Exercise 12 we know that $X + Y$ is a Gamma random variable with parameters $(1, r + s)$. Hence, we have that h must also be

$$h(x) = \frac{1}{\Gamma(r+s)} \exp(-x) x^{r+s-1}.$$

(c) Let $x = 1$ in the formulas in (a) and (b) to get

$$\int_0^1 (1-y)^{s-1} y^{r-1} \mathrm{d}y = \frac{\Gamma(r)\Gamma(s)}{\Gamma(r+s)}.$$

(d) Use that $\Gamma(n) = (n-1)!$ for all integers $n \geq 1$ to compute

$$\int_0^1 (1-y)^5 y^6 \mathrm{d}y.$$

14. Let X be a Gamma random variable with parameters (r, λ).

(a) Let $Y = \lambda X$. Show that

$$M_Y(t) = M_X(\lambda t).$$

(b) Show that Y is also a Gamma random variable. With what parameter?

15. Let X_n be a sequence of geometric random variables. Each X_n has a parameter p_n. Assume that p_n is a strictly positive sequence converging to 0. Let M_n be the mgf of $p_n X_n$.

(a) Show that

$$M_n(t) = \frac{p_n e^{t p_n}}{1 - (1 - p_n)e^{t p_n}}.$$

(b) Show that for every t

$$\lim_{n\to\infty} M_n(t) = \frac{1}{1-t}.$$

(Recall that $\lim_{x\to 0} \frac{e^x - 1}{x} = 1$.)

(c) What is the limiting distribution of $n X_n$ as n goes to infinity?

16. Let X_n be a sequence of Poisson random variables. For each n, X_n has mean n. We are going to show that

$$Y_n = \frac{X_n - n}{\sqrt{n}}$$

converges in distribution to a standard normal. Let M_n be the mgf of Y_n.

(a) Show that

$$\ln(M_n(t)) = -t\sqrt{n} + n\left(-1 + e^{\frac{t}{\sqrt{n}}}\right).$$

(b) Show that for every t

$$\lim_{n\to\infty} \ln(M_n(t)) = \frac{t^2}{2},$$

and conclude. To compute the limit you may use that

$$\lim_{x\to 0} \frac{e^x - (1 + x)}{x^2/2} = 1.$$

Chapter 8
Transformations of Random Variables and Random Vectors

8.1 Distribution Functions and Transformations of Random Variables Distribution Functions

The notion of distribution function is especially useful when dealing with continuous random variables.

Distribution Function

Let X be a random variable. The function

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

is called the distribution function of X. If X is a continuous random variable with density f then

$$F(x) = \int_{\infty}^{x} f(t)\mathrm{d}t.$$

Recall that if X is a continuous random variable with density f then for any a and b such that $-\infty \leq a \leq b \leq +\infty$ we have

$$P(a \leq X \leq b) = \int_{a}^{b} f(x)\mathrm{d}x.$$

Therefore, if X has density f and distribution function F then

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

R.B. Schinazi, *Probability with Statistical Applications*,
DOI 10.1007/978-0-8176-8250-7_8, © Springer Science+Business Media, LLC 2012

By the Fundamental Theorem of Calculus if f is continuous at x then F is differentiable at x and

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

The preceding equality shows that if we know the distribution function then we know the density and therefore the distribution of a continuous random variable. This is true for any random variable: a distribution function determines the distribution of a random variable.

Next we compute a few distribution functions.

Example 1. Let U be a uniform random on [0,1]. That is, the density of U is $f(u) = 1$ for u in [0,1] and $f(u) = 0$ elsewhere. The distribution function F of U is

$$F(u) = \int_{-\infty}^{u} f(x)\mathrm{d}x.$$

In order to compute F explicitly we need to consider three cases. If $u \leq 0$ then f is 0 on $(-\infty, u)$ and $F(u) = 0$. If $0 < u < 1$ then

$$F(u) = \int_{-\infty}^{u} f(x)\mathrm{d}x = \int_{0}^{u} f(x)\mathrm{d}x = \int_{0}^{u} \mathrm{d}x = u.$$

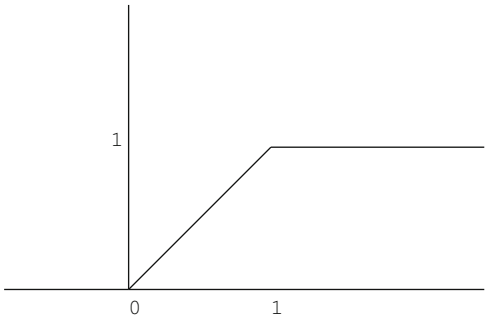
Finally, if $u \geq 1$ then

$$F(u) = \int_{-\infty}^{u} f(x)\mathrm{d}x = \int_{-\infty}^{+\infty} f(x)\mathrm{d}x = 1.$$

This is so because if $u \geq 1$ then f is 0 on $(u, +\infty)$ and the integral of a density function on the whole line is always 1. Summarizing the computations above we get

$$F(u) = 0 \text{ if } u \leq 0$$
$$F(u) = u \text{ if } 0 < u < 1$$
$$F(u) = 1 \text{ if } u \geq 1$$

Below we sketch the graph of F.



There are three features of the graph above that are typical of all distribution functions and that we now state without proof (the proofs are not especially difficult but require more mathematics than we need at this level).

Properties of Distribution Functions

Let F be the distribution function of a random variable X. Then, we have the following three properties.

(*i*) $\lim_{x\to-\infty} F(x) = 0$.
(*ii*) F is an increasing function. That is, if $x_1 < x_2$ then $F(x_1) \leq F(x_2)$.
(*iii*) $\lim_{x\to+\infty} F(x) = 1$.

Example 2. Let T be an exponential random variable with rate λ. What is its distribution function?

The density of T is $f(t) = \lambda e^{-\lambda t}$ for $t \geq 0$. Thus, $F(t) = 0$ for $t \leq 0$ and for $t > 0$ we have that

$$F(t) = \int_{-\infty}^{t} f(x)\mathrm{d}x = \int_{0}^{t} f(x)\mathrm{d}x = -e^{-\lambda x}]_0^t = 1 - e^{-\lambda t}.$$

We have that

$$F(t) = 0 \text{ if } t \leq 0$$
$$F(t) = 1 - e^{-\lambda t} \text{ if } t > 0$$

One can easily check that the three properties of distribution functions hold here as well.

In the next example we give a first application of the notion of distribution function.

Example 3. Assume that T_1 and T_2 are two independent exponentially distributed random variables with rates λ_1 and λ_2, respectively. Let T be the minimum of T_1 and T_2, what is the distribution of T?

Let F be the distribution function of T. We have that

$$F(t) = P(T \leq t) = P(\min(T_1, T_2) \leq t) = 1 - P(\min(T_1, T_2) > t).$$

Observe that $\min(T_1, T_2) > t$ if and only if $T_1 > t$ and $T_2 > t$. Thus, since we are assuming that T_1 and T_2 are independent we get

$$F(t) = 1 - P(T_1 > t)P(T_2 > t) = 1 - (1 - F_1(t))(1 - F_2(t)),$$

where F_1 and F_2 are the distribution functions of T_1 and T_2, respectively. Using the form of the distribution function given in Example 2 we get

$$F(t) = 1 - e^{-\lambda_1 t} e^{-\lambda_2 t} = 1 - e^{-(\lambda_1+\lambda_2)t} \text{ for } t \geq 0.$$

Therefore, the computation above shows that the minimum of two independent exponential random variables is also exponentially distributed and its rate is the sum of the two rates.

Next we look at the maximum of three random variables.

Example 4. Let U_1, U_2, and U_3 be three independent random variables uniformly distributed on [0,1]. Let M be the maximum of U_1, U_2, U_3. What is the density of M?

We are going to compute the distribution function of M and then differentiate to get the density. Let F be the distribution function of M. We have that

$$F(x) = P(M \leq x) = P(\max(U_1, U_2, U_3) \leq x).$$

We have that $\max(U_1, U_2, U_3) \leq x$ if and only if $U_i \leq x$ for $i = 1, 2, 3$. Thus, due to the independence of the U_i we get

$$F(x) = P(U_1 \leq x)P(U_2 \leq x)P(U_3 \leq x) = F_1(x)F_2(x)F_3(x).$$

According to Example 1 we have

$$F(x) = 0 \text{ if } x \leq 0$$
$$F(x) = x^3 \text{ if } 0 < x < 1$$
$$F(x) = 1 \text{ if} x \geq 1$$

Thus, the density of M that we denote by f is

$$f(x) = F'(x) = 3x^2 \text{ for } x \text{ in } [0, 1].$$

Observe that the maximum of uniform random variables is not uniform!

In Examples 3 and 4 in order to compute the distribution of a minimum and a maximum we have used distribution functions. This is a general method that we may summarize below.

Maximum and Minimum of Independent Random Variables

Let $X_1, X_2, \ldots, X_n$ be independent random variables with distribution functions $F_1, F_2, \ldots, F_n$, respectively. Let $F_{\max}$ and $F_{\min}$ be the distribution functions of the random variables $\max(X_1, X_2, \ldots, X_n)$ and $\min(X_1, X_2, \ldots, X_n)$,

respectively. Then,

$$F_{\max} = F_1 F_2 \dots F_n$$

and

$$F_{\min} = 1 - (1 - F_1)(1 - F_2) \dots (1 - F_n).$$

We now give an example of distribution function for a discrete random variable.

Example 5. Flip two fair coins and let X be the number of tails. The distribution of X is given by

x	0	1	2
$P(X = x)$	1/4	1/2	1/4

Observe that if $0 \leq x < 1$ then

$$F(x) = P(X \leq x) = P(X = 0) = \frac{1}{4}$$

while if $1 \leq x < 2$ then

$$F(x) = P(X \leq x) = P(X = 0) + P(X = 1) = \frac{3}{4}.$$

Therefore the distribution function of this discrete random variable is then given by

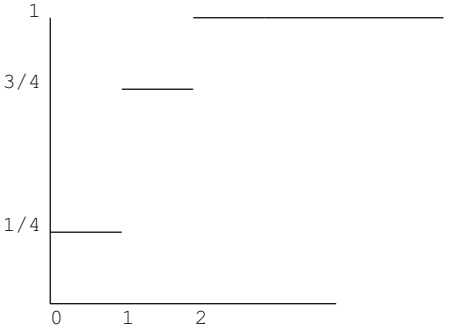
$$F(x) = 0 \text{ if } x < 0$$

$$F(x) = \frac{1}{4} \text{ if } 0 \leq x < 1$$

$$F(x) = \frac{3}{4} \text{ if } 1 \leq x < 2$$

$$F(x) = 1 \text{ if } x \geq 2$$

The graph is given below.



Note that distribution functions of discrete random variables are always discontinuous.

8.1.1 Simulations

Simulations are another type of application of distribution functions. We will see how one can simulate a random variable with a given distribution by using a simulation of an uniform random variable. Many computer random simulators create numbers that behave approximately as observations of independent uniform random variable on [0,1]. So our problem is to go from an uniform distribution to another distribution. We start by considering a continuous random variable X. Assume that the distribution function F of X is strictly increasing and continuous so that the inverse function F^{-1} is well defined. Let U be an uniform random variable on [0,1]. We have that

$$P(F^{-1}(U) \leq x) = P(U \leq F(x)) = F(x)$$

since $F(x)$ is always in [0,1] and $P(U \leq x) = x$ for x in [0,1]. This shows the following.

Simulation of a Continuous Random Variable

Let X be a continuous random variable with a strictly increasing distribution function F. Let U be an uniform random variable on [0,1]. Then $F^{-1}(U)$ has the same distribution as X. That is, to simulate X it is enough to simulate an uniform random variable U and then compute $F^{-1}(U)$.

Example 6. A computer random simulator gives us the following ten random numbers: 0.38, 0.1, 0.6, 0.89, 0.96, 0.89, 0.01, 0.41, 0.86, 0.13. Simulate ten independent exponential random variables with rate 1.

By Example 2 we know that the distribution function F of an exponential random variable with rate 1 is

$$F(x) = 1 - \mathrm{e}^{-x}.$$

We compute F^{-1}. If

$$y = 1 - \mathrm{e}^{-x}$$

then

$$x = -\ln(1 - y).$$

Thus,

$$F^{-1}(x) = -\ln(1 - x).$$

We now compute $F^{-1}(x)$ for $x = 0.38, 0.1, 0.6, 0.89, 0.96, 0.89, 0.01, 0.41, 0.86, 0.13$. We get the following ten observations for ten independent exponential rate 1 random variables: 4.78, 1.05, 0.92, 2.21, 3.22, 2.21,4.6, 5.28, 1.97, 1.39.

Example 7. How do we simulate a standard normal distribution? In this case the distribution function is

$$F(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \mathrm{e}^{-t^2/2} \mathrm{d}t.$$

This is not an expression which is easy to use. Instead we use the normal table. For instance, if we want $F^{-1}(0.38)$ we are looking for z such that $P(Z \leq z) = 0.38$. Hence, $P(0 \leq Z \leq -z) = 0.12$. We read in the table $-z = 0.31$, that is, $z = -0.31$. Using the 10 random numbers from Example 6 we get we following ten observations for a standard normal distribution: -0.31, -1.28, 0.25, 1.22, 1.75, 1.22, -2.33, -0.23, 1.08, -1.13.

Example 8. Simulate a normal distribution X with mean $\mu = 5$ and variance $\sigma^2 = 4$. We know that if Z is a standard normal distribution then $\mu + \sigma Z$ is a normal distribution with mean μ and variance σ^2. We can use the simulation of Z in Example 7 to get simulations of X. For instance, if $Z = -0.31$ then $X = 5 + (-0.31) \times 2 = 4.38$. Here are the ten observations for a normal distribution X with mean $\mu = 5$ and variance $\sigma^2 = 4$. We have: 4.38, 2.44, 5.5, 7.44, 8.5, 7.44, 0.34, 4.54, 7.16, 2.74.

We now turn to the simulation of discrete random variables. Consider a discrete random variable X with k values: $0, 1, 2, \ldots, k$. Denote $P(X = i) = p_i$ for $i = 1, 2, \ldots, k$. Let U be an uniform random variable. The following algorithm uses a simulation of U to give a simulation of X.

$$\text{If } U < p_0 \text{ set } X = 0.$$

$$\text{If } p_0 \leq U < p_0 + p_1 \text{ set } X = 1.$$

More generally, for $i = 1, 2, \ldots, k$.

$$\text{If } p_0 + p_1 + \cdots + p_{i-1} \leq U < p_0 + p_1 + \cdots + p_{i-1} + p_i \text{ set } X = i.$$

Recall that for $0 \leq a \leq b \leq 1$ we have

$$P(a \leq U \leq b) = b - a.$$

Thus,

$$P(X = 0) = P(U < p_0) = p_0 \text{ and } P(X = 1) = P(p_0 \leq U < p_0 + p_1) = p_1.$$

More generally, this algorithm yields $P(X = i) = p_i$ for $i = 1, 2, \ldots, k$. That is, we are able to simulate X from a simulation of U. We now use this algorithm on an example.

Example 9. Let X be a binomial with parameters $n = 2$ and $p = 1/2$. The distribution of X is given by $p_0 = 1/4$, $p_1 = 1/2$ and $p_2 = 1/4$. A random generator gives us the following random numbers: 0.38, 0.1, 0.6, 0.89, 0.96, 0.89, 0.01, 0.41, 0.86, 0.13. Note that

$$p_0 \leq 0.38 \leq p_0 + p_1.$$

Thus, the first simulation for the random variable X is $X = 1$. The second random number is $0.1 < p_0$. This corresponds to $X = 0$ and so on. We get the following simulation of 10 independent random variables with the same distribution as X: 1, 0, 1, 2, 2, 2, 0, 1, 2, 0.

8.1.2 Transformations of Random Variables

At this point we know relatively few different continuous distributions: uniform, exponential, and normal are the main distributions we have seen. In this section we will see a general method to obtain many more distributions from the known ones. We start with an example.

Example 10. Let U be an uniform random variable on [0,1]. Define $X = U^2$. What is the distribution of X?

We use the distribution function F of X.

$$F(x) = P(X \leq x) = P(U^2 \leq x) = P(U \leq \sqrt{x}).$$

Recall that $P(U \leq y) = y$ for y in [0,1]. Thus,

$$F(x) = \sqrt{x} \text{ for } 0 \leq x \leq 1.$$

Note that F is differentiable on $(0, 1]$ and let

$$f(x) = F'(x) = \frac{1}{2\sqrt{x}} \text{ for } 0 < x \leq 1.$$

Let a be in $(0, 1)$. By the Fundamental Theorem of Calculus we have for x in $(0, 1)$,

$$\int_a^x f(t)\mathrm{d}t = F(x) - F(a).$$

Now let a go to 0 to get

$$\int_0^x f(t)\mathrm{d}t = F(x).$$

That is, f is the density of X.

In fact the preceding example gives a general method to compute the density of the transformed random variable. We first compute the distribution function of the transformed random variable. Assuming the distribution function is regular enough (which will always be the case for us) the density of the transformed variable is the derivative of the distribution function.

Example 11. Let X be a continuous random variable with density f, let a be a constant and $Y = X - a$, what is the density of Y?

We first compute the distribution function of Y.

$$F_Y(y) = P(Y \leq y) = P(X - a \leq y) = P(X \leq y + a) = F_X(y + a),$$

where F_X is the distribution function of X. By taking the derivative of F_Y with respect to y we get

$$f_Y(y) = \frac{\mathrm{d}}{\mathrm{d}y} F_X(y + a) = f_X(y + a),$$

where f_X and f_Y are the densities of X and Y, respectively.

Example 12. The Chi-Square distribution. Let Z be a standard normal random variable. What is the density of $Y = Z^2$?

We have for $y \geq 0$

$$F_Y(y) = P(Y \leq y) = P(Z^2 \leq y) = P(-\sqrt{y} \leq Z \leq \sqrt{y}) = F_Z(\sqrt{y}) - F_Z(-\sqrt{y}).$$

Recall that the density of Z is

$$f_Z(z) = \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-z^2/2}.$$

Using the chain rule we get

$$\frac{\mathrm{d}}{\mathrm{d}y} F_Y(y) = f_Y(y) = f_Z(\sqrt{y}) \times \frac{1}{2\sqrt{y}} - f_Z(-\sqrt{y}) \times \frac{-1}{2\sqrt{y}}.$$

Hence, the density of Y is

$$f_Y(y) = \frac{1}{\sqrt{2\pi}} y^{-1/2} \mathrm{e}^{-y/2} \text{ for } y > 0.$$

This is the density of the so-called Chi-Square distribution with one degree of freedom.

Example 13. Let T be exponentially distributed with mean 1. What is the distribution of $X = \sqrt{T}$?

Take $x \geq 0$,

$$F_X(x) = P(X \leq x) = P(\sqrt{T} \leq x) = P(T \leq x^2) = \int_0^{x^2} \mathrm{e}^{-t}\mathrm{d}t = 1 - \mathrm{e}^{-x^2}.$$

Thus, the density of X is

$$\frac{\mathrm{d}}{\mathrm{d}x} F_X(x) = f_X(x) = 2x\mathrm{e}^{-x^2} \text{ for } x \geq 0.$$

Next we finally prove a property of normal random variables that we have already used many times.

Example 14. Let X be normal random variable with mean μ and standard deviation σ. Show that $Y = \frac{X-\mu}{\sigma}$ is a standard normal random variable.

We compute the distribution function of Y:

$$F_Y(y) = P(Y \leq y) = P\left(\frac{X-\mu}{\sigma} \leq y\right) = P(X \leq \mu + \sigma y) = F_X(\mu + \sigma y).$$

We take derivatives to get

$$f_Y(y) = f_X(\mu + \sigma y) \times \sigma.$$

Recall that the density of X is

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \mathrm{e}^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

Thus,

$$f_Y(y) = \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{y^2}{2}}.$$

This proves that Y is a standard normal distribution.

Exercises 8.1

1. Compute the distribution function of a uniform random variable on $[-1,2]$.

2. Assume that waiting times for buses from lines 5 and 8 are exponentially distributed with means 10 and 20 min, respectively. I can take either line so I will take the first bus that comes.

(a) Compute the probability that I will have to wait at least 15 min?
(b) What is the mean time I will have to wait?

3. Consider a circuit with two components in parallel. Assume that both components have independent exponential lifetimes with means 1 and 2 years, respectively.

(a) What is the probability that the circuit lasts more than 3 years?
(b) What is the expected lifetime of the circuit?

4. Assume that T_1 and T_2 are two independent exponentially distributed random variables with rates λ_1 and λ_2, respectively. Let M be the maximum of T_1 and T_2, what is the density of M?

5. Roll a fair die. Let X be the face shown. Graph the distribution function of X.

6. Consider a standard normal random variable Z. Use a normal table to sketch the graph of the distribution function of Z.

7. Simulate 10 observations of a normal distribution with mean 3 and standard deviation 2.

8. Simulate 10 observations of a Poisson distribution with mean 1.

9. Simulate 20 observations of a Bernoulli distribution with parameter $p = 1/4$.

10. Simulate 10 observations of an exponential distribution with mean 2.

11. Simulate 10 observations of a geometric distribution with parameter $p = 1/3$.

12. Let X be a random variable with distribution function $F(x) = x^2$ for x in [0,1].

(a) What is $P(X < 1/3) =$?
(b) What is the expected value of X?

13. Let $U_1, U_2, \ldots, U_n$ be n i.i.d. uniform random variables on [0,1].

(a) Find the density of the maximum of the U_i.
(b) Find the density of the minimum of the U_i.

14. Let U be an uniform random variable on [0,1]. Define $X = \sqrt{U}$. What is the density of X?

15. Let T be exponentially distributed with mean 1. What is the expected value of $T^{1/3}$?

16. Let Z be a standard normal distribution. Find the density of $X = e^Z$. (X is called a lognormal random variable).

17. Let U be uniform on [0,1]. Find the density of $Y = \ln(1 - U)$.

18. Let T be exponentially distributed with rate λ. Find the density of $T^{1/a}$ where $a > 0$. ($T^{1/a}$ is called a Weibull random variable with parameters a and λ).

19. Let X be a continuous random variable, let $a > 0$ and b be two real numbers and let $Y = aX + b$.

(a) Show that

$$f_Y(y) = \frac{1}{a} f_X\left(\frac{y-b}{a}\right).$$

(b) Show that if X is normally distributed then so is $Y = aX + b$.
(c) If X is exponentially distributed, is $Y = aX + b$ also exponentially distributed?

20. Consider the discrete random variable X with the following distribution.

x	-2	-1	2
$P(X = x)$	1/4	1/2	1/4

Find the distribution $Y = X^2$.

8.2 Random Vectors

In this section we introduce the notion of random vectors and joint distributions.

Density of a Continuous Random Vector

Let X and Y be two continuous random variables. The density of the vector (X, Y) is a positive function f such that

$$\int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} f(x, y)\mathrm{d}x\mathrm{d}y = 1.$$

For $a < b$ and $c < d$ we have

$$P(a < X < b; c < Y < d) = \int_a^b \int_c^d f(x, y)\mathrm{d}x\mathrm{d}y.$$

More generally, for a function g we have

$$E(g(X, Y)) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} g(x, y) f(x, y)\mathrm{d}x\mathrm{d}y$$

provided the expectation of $g(X, Y)$ exists.

Example 1. Assume that (X, Y) is uniformly distributed on the disc $\mathcal{C} = \{(x, y) : x^2 + y^2 \leq 1\}$. What is the density of (X, Y)?

Since we want a uniform distribution, we let $f(x, y) = c$ for (x, y) in $\mathcal{C}$, $f(x, y) = 0$ elsewhere. We want

$$\int\int_{\mathcal{C}} f(x, y) = 1 = c \times \text{area}(\mathcal{C}).$$

Thus, $c = 1/\pi$.

At this point the reader may want to review Fubini's Theorem from Calculus. It gives sufficient conditions to integrate multiple integrals one variable at the time.

Note that if the random vector (X, Y) has a density f then for any $a < b$ we have

$$P(a < X < b) = P(a < X < b; -\infty < Y < +\infty) = \int_a^b \int_{-\infty}^{+\infty} f(x, y)\mathrm{d}x\mathrm{d}y.$$

Let

$$f_X(x) = \int_{-\infty}^{+\infty} f(x, y)\mathrm{d}y$$

then

$$P(a < X < b) = \int_a^b f_X(x)\mathrm{d}x.$$

That is, f_X is the density of X. We now state this result.

Marginal Densities

Let (X, Y) be a random vector with density f. Then the densities of X and Y are denoted respectively by f_X and f_Y and are called the marginal densities. They are given by

$$f_X(x) = \int_{-\infty}^{+\infty} f(x, y)\mathrm{d}y \text{ and } f_Y(y) = \int_{-\infty}^{+\infty} f(x, y)\mathrm{d}x.$$

Example 2. We consider again the uniform random vector on the unit disc from Example 1. What are the marginals of X and Y?

Since $x^2 + y^2 \leq 1$, if we fix x in $[-1,1]$ then y varies between $-\sqrt{1 - x^2}$ and $+\sqrt{1 - x^2}$. Thus,

$$f_X(x) = \int_{-\infty}^{+\infty} f(x, y)\mathrm{d}y = f_X(x) = \int_{-\sqrt{1-x^2}}^{+\sqrt{1-x^2}} 1/\pi\mathrm{d}y.$$

Therefore,

$$f_X(x) = \frac{2}{\pi}\sqrt{1 - x^2} \text{ for } x \text{ in } [-1, 1].$$

By symmetry we get that

$$f_Y(y) = \frac{2}{\pi}\sqrt{1-y^2} \text{ for } y \text{ in } [-1,1].$$

Note that although the vector (X, Y) is uniform X and Y are not uniform random variables.

Recall that two random variables X and Y are said to be independent if for any $a < b$ and $c < d$ we have

$$P(a < X < b; c < Y < d) = P(a < X < b)P(c < Y < d).$$

This definition translates nicely into a property of densities that we now state without proof (the proof is beyond the mathematical level of this text).

Independence

Let (X, Y) be a random vector with density f and marginal densities f_X and f_Y. The random variables X and Y are independent if and only if

$$f(x, y) = f_X(x) f_Y(y).$$

Example 3. We continue to analyze the uniform distribution on a disc from Example 1. Are X and Y independent?

Recall that in this case we have $f(x, y) = 1/\pi$ on $\mathcal{C} = \{(x, y) : x^2 + y^2 \leq 1\}$ and 0 elsewhere. We computed f_X and f_Y in Example 2 and clearly $f(x, y) \neq f_X(x) f_Y(y)$. We conclude that X and Y are not independent.

Example 4. Consider two electronic components that have independent exponential lifetimes with means 1 and 2 years, respectively. What is the probability that component 1 outlasts component 2?

Let T and S be respectively the lifetimes of components 1 and 2. We want $P(T > S)$. In order to compute this type of probability we need the joint distribution of (T, S). Since the two random variables are assumed to be independent we have that the joint density is

$$f(t, s) = f_T(t) f_S(s) = \mathrm{e}^{-t}\mathrm{e}^{-s/2}/2 \text{ for } t \geq 0, s \geq 0.$$

We now compute

$$P(T > S) = \int_{s=0}^{\infty}\int_{t=s}^{\infty} \mathrm{e}^{-t}\mathrm{e}^{-s/2}/2 \mathrm{d}t\,\mathrm{d}s.$$

We first integrate in t and then in s to get

$$P(T > S) = \int_{s=0}^{\infty} \mathrm{e}^{-s}\mathrm{e}^{-s/2}/2\mathrm{d}s = \frac{1}{3}.$$

Example 5. Assume that my arrival time at the bus stop is uniformly distributed between 7:00 and 7:05. Assume that the arrival time of the bus I want to take is uniformly distributed between 7:02 and 7:04. What is the probability that I catch the bus?

To simplify the notation we do a translation of 7 h. Let U be my arrival time, it is uniformly distributed on [0,5]. Let V be the arrival time of the bus, it is uniformly distributed on [2,4]. We want the probability $P(U < V)$. It is natural to assume that U and V are independent. So we get

$$P(U < V) = \int_{v=2}^{4}\int_{u=0}^{v} \frac{1}{2} \times \frac{1}{5}\mathrm{d}u\mathrm{d}v = \int_{v=2}^{4} \frac{v}{10}\mathrm{d}v = \frac{3}{5}.$$

We now turn to an example of a discrete joint distribution.

Example 6. Let X and Y be two random variables with the following joint distribution.

X	0	1	2
Y			
1	1/8	1/8	1/4
2	1/8	0	1/8
3	1/8	1/8	0

By replacing integrals by sums we get the marginals of X and Y in a way which is analogous to the continuous case. To get the distribution of X we sum the joint probabilities from top to bottom.

X	0	1	2
$P(X = x)$	3/8	1/4	3/8

To get the distribution of Y we sum the joint probabilities from left to right.

Y	1	2	3
$P(Y = y)$	1/2	1/4	1/4

Two discrete random variables X and Y are independent if and only if

$$P(X = x; Y = y) = P(X = x)P(Y = y) \text{ for all } x, y.$$

In this example we see that X and Y are not independent since

$$P(X = 1; Y = 2) = 0 \text{ and } P(X = 1)P(Y = 2) = 3/16.$$

8.2.1 Proof That the Expectation is Linear

We start by proving the addition formula for expectation that we have already used many times. Assume that X and Y are continuous random variables with joint density f and that their expectations exist. Then, by using the linearity of the integral we get

$$\begin{aligned} E(X+Y) &= \int_{x=-\infty}^{\infty}\int_{y=-\infty}^{\infty}(x+y)f(x,y)\mathrm{d}x\mathrm{d}y \\ &= \int_{x=-\infty}^{\infty}\int_{y=-\infty}^{\infty}xf(x,y)\mathrm{d}x\mathrm{d}y + \int_{x=-\infty}^{\infty}\int_{y=-\infty}^{\infty}yf(x,y)\mathrm{d}x\mathrm{d}y. \end{aligned}$$

Note that

$$\begin{aligned} \int_{x=-\infty}^{\infty}\int_{y=-\infty}^{\infty}xf(x,y)\mathrm{d}x\mathrm{d}y &= \int_{x=-\infty}^{\infty}x\left(\int_{y=-\infty}^{\infty}f(x,y)\mathrm{d}y\right) \\ &= \int_{x=-\infty}^{\infty}xf_X(x)\mathrm{d}x = E(X). \end{aligned}$$

Similarly, we have that

$$\int_{x=-\infty}^{\infty}\int_{y=-\infty}^{\infty}yf(x,y)\mathrm{d}x\mathrm{d}y = E(Y).$$

Hence,

$$E(X+Y) = E(X) + E(Y).$$

The preceding computation holds provided the integrals are finite. A sufficient condition for that is to assume that $E|X|$ and $E|Y|$ are finite.

For any constant a and random variable X we have

$$E(aX) = \int axf_X(x)\mathrm{d}x = a\int xf_X(x)\mathrm{d}x = aE(X).$$

Therefore, we have proved the following (the proof is analogous for discrete random variables):

The Expectation is a Linear Operator

Let X and Y be random variables such that $E|X|$ and $E|Y|$ are finite then

$$E(X+Y) = E(X) + E(Y).$$

For any constant a we have

$$E(aX) = aE(X).$$

8.2.2 Covariance

As we will see covariance and correlation are measures of the joint variations of X and Y.

Covariance

Assume that X and Y are two random variables such that $E(X^2)$ and $E(Y^2)$ exist. The covariance of X and Y is defined by

$$\text{Cov}(X, Y) = E[(X - E(X))(Y - E(Y))].$$

A computational formula for the covariance is

$$\text{Cov}(X, Y) = E(XY) - E(X)E(Y).$$

We prove the computational formula. Note first that

$$(X - E(X))(Y - E(Y)) = XY - XE(Y) - E(X)Y + E(X)E(Y).$$

By taking the expectation on both sides we get

$$\text{Cov}(X, Y) = E[XY - XE(Y) - E(X)Y + E(X)E(Y)].$$

Recalling that $E(X)$ and $E(Y)$ are constants and that the expectation is linear we have

$$\begin{aligned}\text{Cov}(X, Y) &= E[XY] - E[XE(Y)] - E[E(X)Y] + E[E(X)E(Y)] \\ &= E(XY) - E(X)E(Y) - E(X)E(Y) + E(X)E(Y) \\ &= E(XY) - E(X)E(Y).\end{aligned}$$

The formula is proved.

Properties of the Covariance

The covariance is symmetric. That is,

$$\text{Cov}(X, Y) = \text{Cov}(Y, X).$$

The covariance is bilinear. That is, for any constants a and b

$$\text{Cov}(aX, bY) = a\text{Cov}(X, bY) = ab\text{Cov}(X, Y),$$

and for any random variables U, V, and W

$$\text{Cov}(U, V + W) = \text{Cov}(U, V) + \text{Cov}(U, W).$$

Note that

$$\text{Cov}(Y, X) = E(YX) - E(Y)E(X) = E(XY) - E(X)E(Y) = \text{Cov}(X, Y).$$

Hence, the covariance is symmetric. That the covariance is bilinear (i.e., linear in each coordinate) is a direct consequence of the linearity of the expectation. We now prove bilinearity. Let a be a constant then

$$\text{Cov}(aX, Y) = E(aXY) - E(aX)E(Y) = aE(XY) - aE(X)E(Y) = a\text{Cov}(X, Y).$$

Using the fact just proved and symmetry we get

$$\text{Cov}(aX, bY) = a\text{Cov}(X, bY) = a\text{Cov}(bY, X) = ab\text{Cov}(Y, X) = ab\text{Cov}(X, Y).$$

Finally, let U, V, and W be random variables we have

$$\begin{aligned}\text{Cov}(U, V + W) &= E[U(V + W)] - E(U)E(V + W) \\ &= E(UV + UW) - E(U)(E(V) + E(W)).\end{aligned}$$

Using the linearity of expectation we get

$$\begin{aligned}\text{Cov}(U, V + W) &= E(UV) - E(U)E(V) + E(UW) - E(U)E(W) \\ &= \text{Cov}(U, V) + \text{Cov}(U, W).\end{aligned}$$

This completes the proof that covariance is bilinear.

Example 7. Let (X, Y) be uniformly distributed on the triangle $\mathcal{T} = \{(x, y) : 0 < y < x < 2\}$. It is easy to see that the density of (X, Y) is $f(x, y) = 1/2$ for (x, y) in $\mathcal{T}$ and 0 elsewhere. We start by computing the marginal densities of X and Y.

$$f_X(x) = \int_0^x 1/2 \mathrm{d}y = \frac{1}{2}x \text{ for } x \text{ in } [0, 2].$$

$$f_Y(y) = \int_y^2 1/2 \mathrm{d}x = \frac{1}{2}(2 - y) \text{ for } y \text{ in } [0, 2].$$

We note that X and Y are not uniformly distributed and are not independent. We now compute the expectations and standard deviations of X and Y.

$$E(X) = \int_0^2 x\frac{1}{2}x \mathrm{d}x = \frac{4}{3}.$$

We have that

$$E(X^2) = \int_0^2 x^2 \frac{1}{2} x \mathrm{d}x = 2.$$

Thus,

$$\mathrm{Var}(X) = E(X^2) - E(X)^2 = \frac{2}{9}.$$

We have

$$E(Y) = \int_0^2 y \frac{1}{2}(2-y)\mathrm{d}y = \frac{2}{3}$$

and

$$E(Y^2) = \int_0^2 y^2 \frac{1}{2}(2-y)\mathrm{d}y = \frac{2}{3}.$$

Thus,

$$\mathrm{Var}(Y) = \frac{2}{9}.$$

We still need to compute

$$E(XY) = \int_{x=0}^2 \int_{y=0}^x xyf(x,y)\mathrm{d}x\mathrm{d}y = \frac{1}{2}\int_{x=0}^2 x\left(\frac{x^2}{2}\right)\mathrm{d}x = 1.$$

We now may compute the covariance of X and Y.

$$\mathrm{Cov}(X,Y) = E(XY) - E(X)E(Y) = 1 - \left(\frac{4}{3}\right) \times \left(\frac{2}{3}\right) = \frac{1}{9}.$$

Our next goal is a formula for the variance of the sum of random variables. Recall that

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

Therefore, for any two random variables X and Y we have

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

By using the linearity of the expectation and the definition of the covariance we get

$$\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\mathrm{Cov}(X,Y).$$

Variance of a Sum

For any random variables X and Y we have

$$\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\mathrm{Cov}(X,Y)$$

provided $\text{Var}(X)$ and $\text{Var}(Y)$ exist. In particular

$$\text{Var}(X+Y) = \text{Var}(X) + \text{Var}(Y)$$

if and only if $\text{Cov}(X,Y) = 0$.

We now introduce the notion of correlation.

Correlation

Assume that $E(X^2)$ and $E(Y^2)$ exist. The correlation of X and Y is defined by

$$\text{Corr}(X,Y) = \frac{\text{Cov}(X,Y)}{SD(X)SD(Y)}.$$

For any random variables X and Y the correlation between X and Y is always in $[-1,1]$. The correlation between X and Y is -1 or 1 if and only if there are constants a and b such that $Y = aX + b$.

Correlations are standardized covariances: Correlations are always in $[-1,1]$. So it is easier to interpret a correlation than a covariance. We now prove that correlations are always in $[-1,1]$. Let X and Y be two random variables such that $E(X^2)$ and $E(Y^2)$ exist. Assume that $SD(X)$ and $SD(Y)$ are strictly positive. If $SD(X)$ or $SD(Y)$ is 0, see the exercises. Let

$$U = \frac{X}{SD(X)} + \frac{Y}{SD(Y)}.$$

Using the formula for the variance of a sum we get

$$\text{Var}(U) = \text{Var}\left(\frac{X}{SD(X)}\right) + \text{Var}\left(\frac{Y}{SD(Y)}\right) + 2\text{Cov}\left(\frac{X}{SD(X)}, \frac{Y}{SD(y)}\right).$$

Recall that Var is quadratic. That is, for any constant a, $\text{Var}(aX) = a^2\text{Var}(X)$. In particular,

$$\text{Var}\left(\frac{X}{SD(X)}\right) = \frac{1}{SD(X)^2}\text{Var}(X) = 1$$

and similarly $\text{Var}(\frac{Y}{SD(Y)}) = 1$. Using the bilinearity of covariance we get

$$\text{Cov}\left(\frac{X}{SD(X)}, \frac{Y}{SD(y)}\right) = \frac{1}{SD(X)SD(Y)}\text{Cov}(X,Y) = \text{Corr}(X,Y).$$

Hence, going back to the variance of U we have

$$\text{Var}(U) = 1 + 1 + 2\text{Corr}(X, Y) = 2(1 + \text{Corr}(X, Y)).$$

Since $\text{Var}(U) \geq 0$ (a variance is always positive) this yields $1 + \text{Corr}(X, Y) \geq 0$. That is, a correlation is always larger than or equal to -1.

Note also that $\text{Corr}(X, Y) = -1$ only if $\text{Var}(U) = 0$. The variance can be 0 only if the random variable is a constant. That is,

$$U = \frac{X}{SD(X)} + \frac{Y}{SD(Y)} = c$$

for some constant c. Since $SD(X)$, $SD(Y)$ and c are constants there is a linear relation between X and Y when $\text{Corr}(X, Y) = -1$.

We now turn to the inequality $\text{Corr}(X, Y) \leq 1$. It is very similar to what we just did. Let

$$V = \frac{X}{SD(X)} - \frac{Y}{SD(Y)}.$$

Doing computations very similar to the ones we just did we get

$$\text{Var}(V) = 2(1 - \text{Corr}(X, Y)).$$

Since $\text{Var}(V) \geq 0$ we get $\text{Corr}(X, Y) \leq 1$. Moreover, $\text{Corr}(X, Y) = 1$ only if V is a constant and therefore there is a linear relation between X and Y.

Remark 1. A positive correlation indicates that when one random variable is large the other one tends to be large too. Conversely, a negative correlation indicates that when one random variable is large the other one tends to be small.

Correlation and Independence

If $\text{Corr}(X, Y) = 0$ then X and Y are said be uncorrelated. If X and Y are independent then they are uncorrelated. However, uncorrelated random variables need not be independent.

We now show that independent random variables are uncorrelated. We do the proof in the continuous case, the discrete case is similar. Let X and Y be independent. Thus,

$$f(x, y) = f_X(x) f_Y(y).$$

By using the independence property above we get

$$E(XY) = \int_{x=-\infty}^{\infty} \int_{y=-\infty}^{\infty} xyf(x,y)\mathrm{d}x\mathrm{d}y$$
$$= \int_{x=-\infty}^{\infty} f_X(x)\mathrm{d}x \int_{y=-\infty}^{\infty} f_Y(y)\mathrm{d}y = E(X)E(Y).$$

Therefore $\mathrm{Cov}(X,Y) = 0$ and $\mathrm{Corr}(X,Y) = 0$. As the next example shows uncorrelated random variables do not need to be independent.

Example 8. We go back to Example 7 for which we have already computed the variances and covariance. The correlation between X and Y is

$$\mathrm{Corr}(X,Y) = \frac{\mathrm{Cov}(X,Y)}{SD(X)SD(Y)} = \frac{1/9}{\sqrt{2/9}\sqrt{2/9}} = \frac{1}{2}.$$

This positive correlation indicates that when one variable is large it is likely that the other variable will be large as well.

Example 9. We go back to the uniform random vector on the disc $\mathcal{C} = \{(x,y) : x^2 + y^2 \leq 1\}$. We have shown already in Example 3 that X and Y are not independent. However, we will show now that they are uncorrelated.

$$E(XY) = \int_{x=-1}^{1} \int_{y=-\sqrt{1-x^2}}^{\sqrt{1-x^2}} xy\frac{1}{\pi}\mathrm{d}y\mathrm{d}x.$$

Note that when we integrate in y we get

$$\int_{y=-\sqrt{1-x^2}}^{\sqrt{1-x^2}} y\mathrm{d}y = 0.$$

Therefore $E(XY) = 0$. On the other hand from Example 2 we have the following density for X

$$f_X(x) = \frac{2}{\pi}\sqrt{1-x^2} \text{ for } x \text{ in } [-1,1].$$

We compute

$$E(X) = \int_{-1}^{1} xf_X(x)\mathrm{d}x = \frac{2}{\pi}\left(\frac{-1}{2}\right)(1-x^2)^{3/2}\Big]_{x=-1}^{1} = 0.$$

By symmetry we also have that $E(Y) = 0$. Therefore,

$$\mathrm{Cov}(X,Y) = \mathrm{Corr}(X,Y) = 0$$

although X and Y are not independent. This is so because correlation measures the strength of the LINEAR relation between X and Y. Here there is no linear relation but the random variables are related in some other way.

8.2.3 Transformations of Random Vectors

A consequence of multivariate calculus is the following formula for the density of a transformed random vector.

Density of a Transformed Random Vector

Let (X, Y) be a random vector with density f. Let (U, V) be such that

$$U = g_1(X, Y) \text{ and } V = g_2(X, Y).$$

Assume that the transformation $(x, y) \longrightarrow (g_1(x, y), g_2(x, y))$ is one to one with inverse

$$X = h_1(U, V) \text{ and } Y = h_2(U, V).$$

Then the density of the transformed random vector (U, V) is

$$f(h_1(u, v), h_2(u, v))|J(u, v)|,$$

where $J(u, v)$ is the following determinant

$$\begin{vmatrix} \partial h_1/\partial u & \partial h_1/\partial v \\ \partial h_2/\partial u & \partial h_2/\partial v \end{vmatrix}$$

We now use the preceding formula on an example.

Example 10. Let X and Y be two independent standard normal distributions. Let $U = X/Y$ and $V = X$. What is the density of (U, V)?

We see that $(x, y) \longrightarrow (u, v)$ is a one to one transformation from $\mathbf{R}^* \times \mathbf{R}^*$ on to itself where $\mathbf{R}^*$ is the set of all reals different from 0. We invert the transformation to get

$$X = V \text{ and } Y = \frac{V}{U}.$$

We now compute the Jacobian

$$J(u, v) = \begin{vmatrix} 0 & 1 \\ -v/u^2 & 1/u \end{vmatrix} = \frac{v}{u^2}.$$

Since we assume that X and Y are independent standard normal distributions we have

$$f(x,y) = \frac{1}{\sqrt{2\pi}}\mathrm{e}^{-x^2/2}\frac{1}{\sqrt{2\pi}}\mathrm{e}^{-y^2/2}.$$

Therefore, the density of (U,V) is

$$\frac{1}{2\pi}\mathrm{e}^{-v^2/2}\mathrm{e}^{-v^2/(2u^2)}|J(u,v)| = \frac{1}{2\pi}\exp\left(\frac{-v^2}{2}\left(1+\frac{1}{u^2}\right)\right)\frac{|v|}{u^2}.$$

We may now use this joint density to get the marginal density of U. We integrate the density above in v to get

$$f_U(u) = \int_{-\infty}^{\infty}\frac{1}{2\pi}\exp\left(\frac{-v^2}{2}\left(1+\frac{1}{u^2}\right)\right)\frac{|v|}{u^2}\mathrm{d}v.$$

Observe that the integrand above is an even function of v. Thus,

$$\begin{aligned} f_U(u) &= 2\int_0^{\infty}\frac{1}{2\pi}\exp\left(\frac{-v^2}{2}\left(1+\frac{1}{u^2}\right)\right)\frac{v}{u^2}\mathrm{d}v \\ &= -\frac{1}{\pi}\frac{1}{1+1/u^2}\frac{1}{u^2}\exp\left(\frac{-v^2}{2}\left(1+\frac{1}{u^2}\right)\right)\Bigg]_{v=0}^{\infty}. \end{aligned}$$

Hence,

$$f_U(u) = \frac{1}{\pi}\frac{1}{1+u^2}.$$

Therefore, the ratio of two standard normal random variables follows the density above which is called the Cauchy density. Note that $E(U)$ does not exist (see Exercise 10).

Example 11. Let X and Y be two exponential and independent random variables with rates a and b, respectively. Let $U = \min(X,Y)$ and $V = \max(X,Y)$. What is the joint density of (U,V)?

Note that if $X < Y$ then $U = X$ and $V = Y$. The Jacobian is then 1. The portion of the density of (U,V) corresponding to the domain $X < Y$ is then

$$a\mathrm{e}^{-au}b\mathrm{e}^{-bv} \text{ for } 0 < u < v.$$

If $X > Y$ then $U = Y$ and $V = X$. Again the Jacobian is 1. The portion of the density of (U,V) corresponding to the domain $X > Y$ is

$$a\mathrm{e}^{-av}b\mathrm{e}^{-bu} \text{ for } 0 < u < v.$$

We add the two parts to get the joint density of (U, V):

$$ae^{-au}be^{-bv} + ae^{-av}be^{-bu} \text{ for } 0 < u < v.$$

Are U and V independent?

We compute

$$\begin{aligned} f_U(u) &= \int_{v=u}^{\infty} (ae^{-au}be^{-bv} + ae^{-av}be^{-bu})dv \\ &= ae^{-(au+bu)} + be^{-(au+bu)} = (a+b)e^{-(a+b)u}. \end{aligned}$$

That is, the minimum of two independent exponential random variable is exponentially distributed and its rate is the sum of the rates. Using distribution functions in 8.1 we had already seen this result. We now compute the density of V.

$$f_V(v) = \int_{u=0}^{v} (ae^{-au}be^{-bv} + ae^{-av}be^{-bu})du = b(1-e^{-av})e^{-bv} + a(1-e^{-bv})e^{-av}.$$

It is easy to see that the joint distribution of (U, V) is not the product of f_U and f_V. Therefore, U and V are not independent.

Example 12. Assume that X and Y are independent exponential random variables with rates a and b, respectively. Find the density of X/Y.

We could set $U = X/Y$ and $V = X$, find the density of (U, V) and then find the density of U. However, in this case since exponential functions are easy to integrate we may use the distribution function technique of 8.1. Let $U = X/Y$. We have that

$$F_U(u) = P(U \le u) = P(X/Y \le u) = P(X \le uY).$$

Since X and Y are independent we know the joint density of (X, Y). Thus, by integrating in y we get

$$F_U(u) = \int_{x=0}^{\infty} \int_{y=x/u}^{\infty} ae^{-ax}be^{-by}dydx = \int_{x=0}^{\infty} ae^{-ax}e^{-bx/u}dx.$$

Hence, by integrating in x we have

$$F_U(u) = \frac{a}{a + b/u} \text{ for } u > 0.$$

We now differentiate the distribution function to get the density of U:

$$f_U(u) = \frac{ab}{(au+b)^2} \text{ for } u > 0.$$

We can now prove a formula that we used in Chap. 7.

Convolution Formula

Assume that X and Y are independent continuous random variables with densities f_X and f_Y, respectively. Let $U = X + Y$ then the density f_U is given by

$$f_U(u) = \int_{-\infty}^{+\infty} f_X(v)\, f_Y(u-v)\mathrm{d}v.$$

To prove the formula set $U = X + Y$ and $V = X$. The Jacobian of this transformation is -1 and the joint density of (U, V) is

$$f(u, v) = f_X(v)\, f_Y(u-v).$$

Then

$$f_U(u) = \int_{-\infty}^{+\infty} f_X(v)\, f_Y(u-v)\mathrm{d}v$$

and the formula is proved.

Example 13. Assume that X and Y are independent Gamma random variables with parameters (r, λ) and (s, λ), respectively. What is the distribution of X/Y?

Let $U = X/Y$ and $V = X$. We see that $(x, y) \longrightarrow (u, v)$ is a one to one transformation from $(0, \infty) \times (0, \infty)$ to itself and the Jacobian already computed in Example 10 is v/u^2. The density of (X, Y) is

$$\frac{\lambda^r}{\Gamma(r)} x^{r-1} \exp(-\lambda x) \frac{\lambda^s}{\Gamma(s)} y^{s-1} \exp(-\lambda y).$$

Hence, the density of (U, V) is

$$f(u, v) = \frac{\lambda^{r+s}}{\Gamma(r)\Gamma(s)} v^{r-1} \exp(-\lambda v) \left(\frac{v}{u}\right)^{s-1} \exp\left(-\lambda \frac{v}{u}\right) \frac{v}{u^2}.$$

Our goal is to compute the marginal density of U and hence to integrate the preceding joint density with respect to v. This is why we rearrange the joint density as follows:

$$f(u, v) = \frac{\lambda^{r+s}}{u^{s+1}\Gamma(r)\Gamma(s)} v^{r+s-1} \exp\left(-\lambda \left(1 + \frac{1}{u}\right) v\right).$$

Now, for any $\mu > 0$ and $a > 0$ we have

$$\int_0^{\infty} \frac{1}{\Gamma(a)} \mu^a x^{a-1} \exp(-\mu x)\mathrm{d}x = 1.$$

This is so because the integrand is a Gamma density with parameters (a, μ). Hence,

$$\int_0^\infty x^{a-1} \exp(-\mu x) \mathrm{d}x = \frac{\Gamma(a)}{\mu^a}.$$

We use this formula with $a = r + s$ and $\mu = \lambda(1 + 1/u)$ to get

$$\int_0^\infty v^{r+s-1} \exp\left(-\lambda\left(1 + \frac{1}{u}\right)v\right) \mathrm{d}v = \frac{\Gamma(r+s)}{\lambda^{r+s}(1+\frac{1}{u})^{r+s}}.$$

Therefore,

$$\int_0^\infty f(u, v) \mathrm{d}v = \frac{\lambda^{r+s}}{u^{s+1}\Gamma(r)\Gamma(s)} \frac{\Gamma(r+s)}{\lambda^{r+s}(1+\frac{1}{u})^{r+s}}.$$

Hence, the density of $U = X/Y$ is

$$\frac{\Gamma(r+s)}{\Gamma(r)\Gamma(s)} \frac{u^{r-1}}{(u+1)^{r+s}} \text{ for } u > 0.$$

We will use this computation to introduce in the next example the beta distribution which plays an important role in statistics.

Example 14. Assume that X and Y are independent Gamma random variables with parameters (r, λ) and (s, λ), respectively. What is the distribution of $B = X/(X + Y)$?

Since X and Y are positive random variables the random variable B takes values in $(0, 1)$. We will get the density of B through its distribution function F_B. Let t be in $(0, 1)$. Note that B can be written as

$$B = \frac{U}{U+1},$$

where $U = X/Y$. Thus, $B \leq t$ is equivalent to

$$U \leq \frac{t}{1-t}.$$

Using this observation we get

$$F_B(t) = P(B \leq t) = P\left(U \leq \frac{t}{1-t}\right).$$

By the chain rule we have

$$\frac{\mathrm{d}}{\mathrm{d}t} P\left(U \leq \frac{t}{1-t}\right) = f_U\left(\frac{t}{1-t}\right) \times \frac{1}{(1-t)^2},$$

where f_U is the density of U which was computed in Example 13. Hence, the density f_B of B is

$$f_B(t) = \frac{\Gamma(r+s)}{\Gamma(r)\Gamma(s)} \frac{u^{r-1}}{(u+1)^{r+s}} \times \frac{1}{(1-t)^2},$$

where

$$u = \frac{t}{1-t}.$$

After a little algebra we get

$$f_B(t) = \frac{\Gamma(r+s)}{\Gamma(r)\Gamma(s)} t^{r-1}(1-t)^{s-1} \text{ for } t \text{ in } (0,1).$$

This is the density of a so-called Beta distribution with parameters (r, s).

The Beta Distribution

Let $r > 0$ and $s > 0$ then the Beta distribution with parameters (r, s) has density

$$\frac{\Gamma(r+s)}{\Gamma(r)\Gamma(s)} t^{r-1}(1-t)^{s-1} \text{ for } t \text{ in } (0,1).$$

The expected value is $\frac{r}{r+s}$ and the variance is $\frac{rs}{(r+s)^2(r+s+1)}$.

We first compute the expected value. Let B be a Beta distribution with parameters (r, s). We have

$$E(B) = \int_0^1 t \frac{\Gamma(r+s)}{\Gamma(r)\Gamma(s)} t^{r-1}(1-t)^{s-1} \mathrm{d}t = \int_0^1 \frac{\Gamma(r+s)}{\Gamma(r)\Gamma(s)} t^{r}(1-t)^{s-1} \mathrm{d}t.$$

Note that for any $a > 0$ and $b > 0$ we have

$$\int_0^1 \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} t^{a-1}(1-t)^{b-1} \mathrm{d}t = 1,$$

since the integrand is the density of a Beta distribution with parameters (a, b). Hence,

$$\int_0^1 t^{a-1}(1-t)^{b-1} \mathrm{d}t = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)} \tag{8.1}$$

We apply this formula with $a = r + 1$ and $b = s$ to get

$$\int_0^1 t^r (1-t)^{s-1} = \frac{\Gamma(r+1)\Gamma(s)}{\Gamma(r+s+1)}.$$

Thus,

$$E(B) = \frac{\Gamma(r+s)}{\Gamma(r)\Gamma(s)} \int_0^1 t^r (1-t)^{s-1} \mathrm{d}t = \frac{\Gamma(r+s)}{\Gamma(r)\Gamma(s)} \frac{\Gamma(r+1)\Gamma(s)}{\Gamma(r+s+1)}.$$

The function Γ is defined on $(0, +\infty)$ by

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

Hence, by integration by parts

$$\Gamma(a+1) = \int_0^\infty x^a \mathrm{e}^{-x} \mathrm{d}x = -\mathrm{e}^{-x} x^a \Big]_0^\infty + \int_0^\infty a x^{a-1} \mathrm{e}^{-x} \mathrm{d}x = a\Gamma(a),$$

where we use that $a > 0$. This yields the useful formula:

$$\Gamma(a+1) = a\Gamma(a) \tag{8.2}$$

We apply this to get

$$\frac{\Gamma(r+1)}{\Gamma(r)} = r \text{ and } \frac{\Gamma(r+s)}{\Gamma(r+s+1)} = \frac{1}{r+s}.$$

Therefore,

$$E(B) = \frac{r}{r+s}.$$

We now turn to the second moment of B

$$E(B^2) = \int_0^1 t^2 \frac{\Gamma(r+s)}{\Gamma(r)\Gamma(s)} t^{r-1} (1-t)^{s-1} \mathrm{d}t = \int_0^1 \frac{\Gamma(r+s)}{\Gamma(r)\Gamma(s)} t^{r+1} (1-t)^{s-1} \mathrm{d}t.$$

We apply formula (8.1) with $a = r + 2$ and $b = s$ to get

$$\int_0^1 t^{r+1} (1-t)^{s-1} = \frac{\Gamma(r+2)\Gamma(s)}{\Gamma(r+s+2)}.$$

Hence,

$$E(B^2) = \frac{\Gamma(r+s)}{\Gamma(r)\Gamma(s)} \frac{\Gamma(r+2)\Gamma(s)}{\Gamma(r+s+2)}.$$

By (8.2)

$$\Gamma(r+s+2) = (r+s+1)\Gamma(r+s+1) = (r+s+1)(r+s)\Gamma(r+s)$$

and

$$\Gamma(r+2) = (r+1)r\Gamma(r).$$

Therefore,

$$E(B^2) = \frac{r(r+1)}{(r+s+1)(r+s)}.$$

The variance is

$$\begin{aligned}\mathrm{Var}(B) = E(B^2) - E(B)^2 &= \frac{r(r+1)}{(r+s+1)(r+s)} - \left(\frac{r}{r+s}\right)^2 \\ &= \frac{rs}{(r+s+1)(r+s)^2}.\end{aligned}$$

This completes the computation.

Exercises 8.2

1. Consider an uniform random vector on the triangle $\{(x, y) : 0 \le x \le y \le 1\}$.

(a) Find the density of the vector (X, Y).
(b) Find the marginal densities f_X and f_Y.
(c) Are the two random variables X and Y independent?

2. Consider an uniform random vector on the square $\{(x, y) : 0 \le x \le 1; 0 \le y \le 1\}$.

(a) Find the density of the vector (X, Y).
(b) Find the marginal densities f_X and f_Y.
(c) Are the two random variables X and Y independent?

3. Redo Example 5 assuming that the bus leaves at 7:03 precisely. What is the probability that I catch the bus?

4. Two friends have set an appointment between 8:00 and 8:30. Assume that the arrival times of the two friends are independent and uniformly distributed between 8:00 and 8:30. Assume also that the first that arrives waits for 15 min and then leaves. What is the probability that the friends miss each other?

5. Roll two dice. Let X be the sum and Y the minimum of the two dice.

(a) Find the joint distribution of (X, Y).
(b) Are X and Y independent?

6. Consider an uniform random vector on the triangle $\{(x, y) : 0 \leq x \leq y \leq 1\}$. Find the correlation between X and Y.

7. Roll two dice. Let X be the sum and Y the minimum of the two dice. Find the correlation between X and Y.

8. Compute the correlation for the random variables of Example 6.

9. Let X and Y be two independent exponential random variables with rates λ and μ, respectively. What is the probability that X is less than Y?

10. Show that if U has a Cauchy density $f_U(u) = \frac{1}{\pi}\frac{1}{1+u^2}$ then $E(U)$ does not exist.

11. Let X and Y be two independent exponential random variables with rate λ.

(a) Find the joint density of $(X + Y, X/Y)$.
(b) Find the density of X/Y.
(c) Show that $X + Y$ and X/Y are independent.

12. Let X and Y be two exponential and independent random variables with rate a. Let $U = \min(X, Y)$, $V = \max(X, Y)$ and $D = V - U$.

(a) Find the joint density of (U, D).
(b) Are U and D independent?

13. Let X and Y be two exponential independent random variables with rate 1. Let $U = X/(X + Y)$.

(a) Find the distribution function of U.
(b) Find the density of U.

14. Let X and Y be two independent uniform random variables.

(a) Find the density of XY.
(b) Find the density of X/Y.

15. Let X and Y be two exponential and independent random variables with rate a. Let $U = X$ and $V = X + Y$. Find the joint density of (U, V).

16. Let T_1, $T_2 \ldots T_n$ be independent exponential random variables with rates a_1, $a_2 \ldots, a_n$. Let $S = \min(T_1, T_2, \ldots, T_n)$.

(a) Show that S is exponentially distributed with rate $a_1 + a_2 + \cdots + a_n$.
(b) Fix k in $\{1, \ldots, n\}$, let $S_k = \min_{i \neq k} T_i$. That is S_k is the minimum of the T_i for $i \neq k$. Find the density of the vector (T_k, S_k).
(c) Show that the event $\{S = T_k\}$ has the same probability as the event $\{T_k < S_k\}$.
(d) Prove that the probability that the minimum of the T_i for $1 \leq i \leq n$ is T_k is

$$\frac{a_k}{a_1 + a_2 + \cdots + a_n}.$$

17. Let X and Y be two random variables that have a variance. Find a formula for $\text{Var}(X - Y)$ that uses $\text{Var}(X)$, $\text{Var}(Y)$ and $\text{Cov}(X, Y)$.

18. (a) Using that $\text{Corr}(X, Y)$ is in $[-1, 1]$ show that

$$|E(XY)| \leq SD(X)SD(Y).$$

(b) In which cases is the inequality in (a) an equality?

19. Let Z be a standard normal random variable. Let $Y = Z^2$.

(a) Show that Y and Z are uncorrelated.
(b) Are Y and Z independent?

20. Assume that $SD(X) = 0$.

(a) Show that X is a constant.
(b) Show that $\text{Cov}(X, Y) = 0$ for all Y.

21. Consider a Beta distribution with parameters (r, s).

(a) Sketch the graph of the density for $r = 1$ and $s = 1$.
(b) Sketch the graph of the density for $r = 10$ and $s = 30$.

22. Consider a Beta random variable B with parameters (r, s).

(a) Compute the third moment of B.
(b) Compute the fourth moment of B.
(c) Can you guess what the formula is for the rth moment?

8.3 Transformations of Normal Vectors

We start this section by tying some lose ends concerning normal random variables.

Two Important Integrals

$$\int_{-\infty}^{\infty} \mathrm{e}^{-\frac{x^2}{2}} \mathrm{d}x = \sqrt{2\pi}$$

and

$$\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}.$$

Note that the first result proves that $\frac{1}{\sqrt{2\pi}}\mathrm{e}^{-\frac{x^2}{2}}$ is indeed a probability density! To prove it let

$$I = \int_{-\infty}^{\infty} \mathrm{e}^{-\frac{x^2}{2}} \mathrm{d}x.$$

Consider the double integral

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \mathrm{e}^{-\frac{x^2}{2}}\mathrm{e}^{-\frac{y^2}{2}}\mathrm{d}x\mathrm{d}y.$$

By integrating first in x and then in y we get

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \mathrm{e}^{-\frac{x^2}{2}}\mathrm{e}^{-\frac{y^2}{2}}\mathrm{d}x\mathrm{d}y = I^2.$$

We now compute the double integral by changing to polar coordinates. Let $x = r\cos\theta$ and $y = r\sin\theta$. The Jacobian of this transformation is r and the domain $(-\infty, +\infty) \times (-\infty, +\infty)$ for (x, y) corresponds to the domain $(0, \infty) \times (0, 2\pi)$ for (r, θ). Hence,

$$I^2 = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \mathrm{e}^{-\frac{x^2}{2}}\mathrm{e}^{-\frac{y^2}{2}}\mathrm{d}x\mathrm{d}y = \int_{\theta=0}^{2\pi}\int_{r=0}^{\infty} \mathrm{e}^{-\frac{r^2}{2}} r\mathrm{d}r\mathrm{d}\theta.$$

Note that

$$\int_{r=0}^{\infty} \mathrm{e}^{-\frac{r^2}{2}} r\mathrm{d}r = -\mathrm{e}^{-\frac{r^2}{2}}\Big]_0^{\infty} = 1$$

and so

$$I^2 = 2\pi \text{ and } I = \sqrt{2\pi}$$

This completes the computation of the first integral.

To compute the second integral we will use probability densities. Recall from 8.1 that if Z is a standard normal distribution then $Y = Z^2$ has the following density

$$f_Y(y) = \frac{1}{\sqrt{2\pi}} y^{-1/2}\mathrm{e}^{-y/2} \text{ for } y > 0.$$

On the other hand a Gamma random variable X with parameters $(1/2, 1/2)$ has density

$$f_X(x) = \frac{(1/2)^{1/2}}{\Gamma(1/2)} x^{-1/2}\mathrm{e}^{-x/2} \text{ for } x > 0.$$

Let $g(x) = x^{-1/2}\mathrm{e}^{-x/2}$. We see that

$$f_Y(x) = C_1 g(x) \text{ and } f_X(x) = C_2 g(x),$$

where C_1 and C_2 are constants. Since

$$\int_0^{\infty} f_Y(x)\mathrm{d}x = \int_0^{\infty} f_X(x)\mathrm{d}x = 1$$

we have

$$C_1 \int_0^\infty g(x)\mathrm{d}x = C_2 \int_0^\infty g(x)\mathrm{d}x.$$

Since $\int_0^\infty g(x)\mathrm{d}x > 0$ we conclude that $C_1 = C_2$. That is,

$$\frac{1}{\sqrt{2\pi}} = \frac{(1/2)^{1/2}}{\Gamma(1/2)}$$

and we get

$$\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}.$$

Next we construct Chi-square and Student distributions by using normal random variables. First, recall that a Gamma random variable with parameters $r > 0$ and $\lambda > 0$ has density

$$f(x) = \frac{\lambda^r}{\Gamma(r)} x^{r-1} \mathrm{e}^{-\lambda x} \text{ for all } x > 0,$$

where

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

The moment generating function of the Gamma random variable with parameters r and λ has been computed in 7.1 and we found that

$$M_X(t) = \left(\frac{\lambda}{\lambda - t}\right)^r \text{ for } t < \lambda.$$

Assume that $X_1, X_2, \ldots, X_n$ are independent Gamma random variables with parameters $(r_1, \lambda), (r_2, \lambda), \ldots, (r_n, \lambda)$, respectively. Then,

$$M_{X_1+X_2+\cdots+X_n}(t) = E(\mathrm{e}^{t(X_1+X_2+\cdots+X_n)}) = E(\mathrm{e}^{tX_1})E(\mathrm{e}^{tX_2})\ldots E(\mathrm{e}^{tX_n}),$$

where the last equality comes from the independence of the X_i. We now use the formula for the moment generating function of a Gamma to get

$$M_{X_1+X_2+\cdots+X_n}(t) = \left(\frac{\lambda}{\lambda - t}\right)^{r_1}\left(\frac{\lambda}{\lambda - t}\right)^{r_2}\cdots\left(\frac{\lambda}{\lambda - t}\right)^{r_n} = \left(\frac{\lambda}{\lambda - t}\right)^{r_1+r_2+\cdots+r_n}.$$

Recall that a moment generating function determines the distribution of a random variable. Therefore, the computation above shows that the sum of independent Gamma random variables with the same λ is a Gamma distribution.

Sum of Gamma Random Variables

Assume that $X_1, X_2, \ldots, X_n$ are independent Gamma random variables with parameters $(r_1, \lambda), (r_2, \lambda), \ldots, (r_n, \lambda)$, respectively (note that they all have the same λ). Then, $X_1 + X_2 + \cdots + X_n$ is a Gamma random variable with parameters $(r_1 + r_2 + \cdots + r_n, \lambda)$.

We use the preceding fact about Gamma distributions to show the following property of standard normal variables.

Chi-Square Distribution

Assume that $Z_1, Z_2, \ldots, Z_n$ are independent standard normal distributions. Then,

$$X = Z_1^2 + Z_2^2 + \cdots + Z_n^2$$

is called a Chi-square random variable with n degrees of freedom. Its density is given by

$$f(x) = \frac{1}{2^{n/2}\Gamma(n/2)} x^{n/2-1} \mathrm{e}^{-x/2} \text{ for } x > 0.$$

We note that $\frac{1}{\sqrt{2\pi}} y^{-1/2} \mathrm{e}^{-y/2}$ is the density of a Gamma with parameters 1/2 and 1/2. Hence,

$$X = Z_1^2 + Z_2^2 + \cdots + Z_n^2$$

is a sum of n independent Gamma random variables with parameters 1/2 and 1/2. Therefore, X is a Gamma random variable with parameters $r = n/2$ and $\lambda = 1/2$. Thus, the density of X is

$$f(x) = \frac{(1/2)^{n/2}}{\Gamma(n/2)} x^{n/2-1} \mathrm{e}^{-x/2} \text{ for all } x > 0.$$

We have seen in Chap. 5 that Chi-square distributions play a pivotal role in statistics. See Sect. 5.4 for sketches of the graph of Chi-square densities and for statistical applications. Another important distribution in statistics is the Student distribution.

Student Distribution

Let Z be a standard normal random variable and X be a Chi-square random variable with r degrees of freedom. Assume that X and Z are independent. Then,

$$T = \frac{Z}{\sqrt{X/r}}$$

is a Student random variable with r degrees of freedom. Its density is

$$f(t) = \frac{\Gamma(\frac{r+1}{2})}{\Gamma(r/2)\sqrt{\pi r}}(1 + t^2/r)^{-(r+1)/2}.$$

As r increases to infinity the density of a Student with r degrees of freedom approaches the density of a standard normal distribution (see Sect. 5.3). We now find the density of a Student with r degrees of freedom. Let

$$T = \frac{Z}{\sqrt{X/r}} \text{ and } U = X.$$

We invert the relations above to get

$$Z = T\sqrt{U/r} \text{ and } X = U.$$

The Jacobian of the transformation above is

$$\begin{vmatrix} \sqrt{u/r} & t/(2\sqrt{ur}) \\ 0 & 1 \end{vmatrix} = \sqrt{u/r}.$$

Since Z and X are independent the joint density of (Z, X) is

$$\frac{1}{\sqrt{2\pi}2^{r/2}\Gamma(r/2)}\mathrm{e}^{-z^2/2}x^{r/2-1}\mathrm{e}^{-x/2} \text{ for any } z, x > 0$$

The joint density of (T, U) is then

$$\frac{1}{\sqrt{2\pi}2^{r/2}\Gamma(r/2)}\mathrm{e}^{-t^2u/(2r)}u^{r/2-1}\mathrm{e}^{-u/2}\sqrt{u/r} \text{ for any } t, u > 0.$$

In order to get the density of T we integrate the joint density of (T, U) in u. We get

$$f_T(t) = \frac{1}{\sqrt{2\pi r}2^{r/2}\Gamma(r/2)}\int_0^\infty u^{\frac{r+1}{2}-1}\mathrm{e}^{-u(\frac{t^2}{2r}+1/2)}\mathrm{d}u.$$

To compute the preceding integral, we use a Gamma density in the following way. We know that

$$\int_0^\infty \frac{\lambda^s}{\Gamma(s)}x^{s-1}\mathrm{e}^{-\lambda x}\mathrm{d}x = 1$$

for all $s > 0$ and $\lambda > 0$. This is so because the integrand above is the density of a Gamma random variable with parameters s and λ. Therefore,

$$\int_0^\infty x^{s-1} e^{-\lambda x} dx = \frac{\Gamma(s)}{\lambda^s}.$$

Now let $s = \frac{r+1}{2}$ and $\lambda = \frac{t^2}{2r} + 1/2$ we get

$$\int_0^\infty u^{\frac{r+1}{2}-1} e^{-u(\frac{t^2}{2r}+1/2)} du = \frac{\Gamma(\frac{r+1}{2})}{\left(\frac{t^2}{2r} + 1/2\right)^{\frac{r+1}{2}}}.$$

Thus,

$$f_T(t) = \frac{1}{\sqrt{2\pi r} 2^{r/2} \Gamma(r/2)} \frac{\Gamma(\frac{r+1}{2})}{(\frac{t^2}{2r} + 1/2)^{\frac{r+1}{2}}} = \frac{\Gamma(\frac{r+1}{2})}{\Gamma(r/2)\sqrt{\pi r}} (1 + t^2/r)^{-(r+1)/2},$$

and the computation of the Student density is complete. We now introduce the Fisher distribution.

F Distribution

Assume that X and Y are independent Chi-square random variables with m and n degrees of freedom, respectively. Then the distribution of

$$\frac{X/m}{Y/n}$$

is called the F distribution (after R.A. Fisher) with (m, n) degrees of freedom. Its density is

$$f(v) = \frac{\Gamma((m+n)/2)}{\Gamma(m/2)\Gamma(n/2)} \left(\frac{m}{n}\right)^{m/2} v^{m/2-1} \left(1 + \frac{m}{n} v\right)^{-(n+m)/2} \quad \text{for } v > 0.$$

In order to compute the density of a F distribution we go back to Gamma distributions. Recall from 8.2 that if X and Y are independent Gamma random variables with parameters (r, λ) and (s, λ), respectively then the density of $U = X/Y$ is

$$\frac{\Gamma(r+s)}{\Gamma(r)\Gamma(s)} \frac{u^{r-1}}{(u+1)^{r+s}} \text{ for } u > 0.$$

Since a Chi-square random variable with m degrees of freedom is a Gamma random variable with parameters $r = m/2$ and $\lambda = 1/2$ we get that X/Y has density

$$\frac{\Gamma((m+n)/2)}{\Gamma(m/2)\Gamma(n/2)} \frac{u^{m/2-1}}{(u+1)^{(m+n)/2}} \text{ for } u > 0.$$

Let

$$V = \frac{X/m}{Y/n},$$

then

$$V = \frac{n}{m}\frac{X}{Y} = \frac{n}{m}U = cU,$$

where $c = n/m$. Let F_U and F_V be the distribution functions of U and V. We have

$$F_V(v) = P(V \leq v) = P(U \leq v/c) = F_U(v/c).$$

By taking derivatives and using the chain rule we get

$$f_V(v) = \frac{1}{c} f_U(v/c) = \frac{m}{n} f_U\left(\frac{m}{n}v\right),$$

where f_U and f_V are the densities of U and V. Hence, for $v > 0$

$$\begin{aligned} f_V(v) &= \frac{\Gamma((m+n)/2)}{\Gamma(m/2)\Gamma(n/2)} \frac{(\frac{m}{n}v)^{m/2-1}}{(\frac{m}{n}v+1)^{(m+n)/2}} \\ &= \frac{\Gamma((m+n)/2)}{\Gamma(m/2)\Gamma(n/2)} \left(\frac{m}{n}\right)^{m/2} v^{m/2-1} \left(1+\frac{m}{n}v\right)^{-(n+m)/2} \end{aligned}$$

This completes our computation.

8.3.1 Variance of a Vector

We first need to define the expected value of a random vector or a random matrix.

Expected Value of a Random Matrix

Let $\mathbf{X}$ be a random matrix with i, j component $X_{i,j}$. We define the expected value of $\mathbf{X}$ as the matrix with i, j components $E(X_{i,j})$. If A and B are constant matrices then

$$E(A\mathbf{X}) = AE(\mathbf{X})$$

and

$$E(\mathbf{X}B) = E(\mathbf{X})B.$$

In order for the formulas above to make sense the dimensions of the matrices X, A and B must be adequate to allow for matrix multiplication. We prove the first property. The i, j component of $A\mathbf{X}$ is

$$(A\mathbf{X})_{i,j} = \sum_k A_{i,k} X_{k,j}.$$

Therefore, the i, j component of $E(A\mathbf{X})$ is

$$E(A\mathbf{X})_{i,j} = E\left(\sum_k A_{i,k} X_{k,j}\right) = \sum_k A_{i,k} E(X_{k,j}),$$

where we use that the expectation is linear and the $A_{i,k}$ are constant (i.e., nonrandom). On the other hand the i, j component of $AE(\mathbf{X})$ is

$$(AE(\mathbf{X}))_{i,j} = \sum_k A_{i,k} E(X_{k,j}).$$

This proves that

$$E(A\mathbf{X}) = AE(\mathbf{X}).$$

The proof of the right multiplication formula is done in a similar way and we omit it.

The transpose of a matrix A is denoted by A'. It is obtained by switching the rows and columns of A. For instance, if

$$A = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{pmatrix},$$

then

$$A' = \begin{pmatrix} 1 & 4 \\ 2 & 5 \\ 3 & 6 \end{pmatrix}.$$

We now define the variance of a random vector $\mathbf{X}$.

Variance of a Random Vector

The variance of a random vector $\mathbf{X}$ is the matrix

$$\text{Var}(\mathbf{X}) = E(\mathbf{X}\mathbf{X}') - E(\mathbf{X})E(\mathbf{X}').$$

The (i, j) component of the matrix $\text{Var}(\mathbf{X})$ is $\text{Cov}(X_i, X_j)$. If A is a constant matrix then

$$\text{Var}(A\mathbf{X}) = A\text{Var}(\mathbf{X})A'.$$

Recall that $\mathrm{Cov}(X_i, X_i) = \mathrm{Var}(X_i)$ so that the diagonal of the variance matrix is made up of variances.

Next we check that the variance matrix is composed of covariances in the case of a vector with two components.

Assume that

$$\mathbf{X} = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}.$$

Then,

$$\mathbf{X}' = (X_1 X_2),$$

and

$$\mathbf{X}\mathbf{X}' = \begin{pmatrix} X_1^2 & X_1 X_2 \\ X_1 X_2 & X_2^2 \end{pmatrix}.$$

Hence,

$$E(\mathbf{X}\mathbf{X}') = \begin{pmatrix} E(X_1^2) & E(X_1 X_2) \\ E(X_1 X_2) & E(X_2^2) \end{pmatrix}.$$

On the other hand

$$E(\mathbf{X})E(\mathbf{X}') = \begin{pmatrix} E(X_1) \\ E(X_2) \end{pmatrix} E(X_1)E(X_2) = \begin{pmatrix} E(X_1)^2 & E(X_1)E(X_2) \\ E(X_1)E(X_2) & E(X_2)^2 \end{pmatrix}.$$

Thus,

$$\mathrm{Var}(\mathbf{X}) = E(\mathbf{X}\mathbf{X}')) - E(\mathbf{X})E(\mathbf{X}') = \begin{pmatrix} \mathrm{Var}(X_1) & \mathrm{Cov}(X_1, X_2) \\ \mathrm{Cov}(X_1, X_2) & \mathrm{Var}(X_2) \end{pmatrix}.$$

Note also that if $\mathbf{X}$ has only one component ($\mathbf{X}$ is then a random variable) then $\mathbf{X}' = \mathbf{X}$ and the variance matrix has just one component which is the variance of the random variable $\mathbf{X}$.

We now prove that $\mathrm{Var}(A\mathbf{X}) = A\mathrm{Var}(\mathbf{X})A'$. Let A and B be two matrices. We have from linear algebra that

$$(AB)' = B'A'.$$

Hence,

$$A\mathbf{X}(A\mathbf{X})' = A\mathbf{X}\mathbf{X}'A'.$$

Taking the expectation on both sides and using the linearity of the expectation we get

$$E[A\mathbf{X}(A\mathbf{X})'] = AE[\mathbf{X}\mathbf{X}']A'.$$

We also have

$$E(A\mathbf{X}) = AE(\mathbf{X}) \text{ and } E[(A\mathbf{X})'] = E(\mathbf{X}'A') = E(\mathbf{X}')A'.$$

Thus,

$$\begin{aligned} \text{Var}(AX) &= E[A\mathbf{X}(A\mathbf{X})'] - E(A\mathbf{X})E[(A\mathbf{X})'] \\ &= AE[\mathbf{X}\mathbf{X}']A' - AE(\mathbf{X})E(\mathbf{X}')A' \\ &= A\text{Var}(\mathbf{X})A'. \end{aligned}$$

8.3.2 Normal Random Vectors

We now define normal vectors.

Normal Random Vectors

The vector $\mathbf{X} = \begin{pmatrix} X_1 \\ X_2 \\ . \\ . \\ X_n \end{pmatrix}$ is said to be a normal random vector if there is $n \times n$ matrix A, n independent standard normal variables $Z_1, Z_2, \ldots, Z_n$ and a constant vector $\mathbf{b}$ such that

$$\mathbf{X} = \mathbf{A}\mathbf{Z} + \mathbf{b},$$

where $\mathbf{Z} = \begin{pmatrix} Z_1 \\ Z_2 \\ . \\ . \\ Z_n \end{pmatrix}$.

In order to analyze some of the properties of normal random vectors our main tool will be multivariate moment generating functions. Let $\mathbf{X}$ be a random vector with n components. The moment generating function of $\mathbf{X}$ is defined by

$$M_{\mathbf{X}}(\mathbf{t}) = E(\exp(\mathbf{t}'\mathbf{X})),$$

where $\mathbf{t}$ is a column vector with n components.

Multivariate moment generating functions have the two following important properties.

P1. Assume that

$$M_{\mathbf{X}}(\mathbf{t}) = M_{\mathbf{Y}}(\mathbf{t}) \text{ for all } \mathbf{t} \text{ in } [-r, r]^n$$

for some $r > 0$ then the random vectors $\mathbf{X}$ and $\mathbf{Y}$ have the same distribution.

Property (P1) tells us the moment generating function completely characterizes the distribution of a random vector.

P2. Let $X_1, X_2, \ldots, X_n$ be random variables and $\mathbf{X}$ be the random vector whose components are $X_1, X_2, \ldots, X_n$. Then, $X_1, X_2, \ldots, X_n$ are independent if and only if

$$M_{\mathbf{X}}(\mathbf{t}) = M_{X_1}(t_1) M_{X_2}(t_2) \ldots M_{X_n}(t_n)$$

for all column vectors $\mathbf{t}$ in $[-r, r]^n$ for some $r > 0$ and where $t_1, t_2, \ldots, t_n$ are the components of $\mathbf{t}$.

That is, in order for $X_1, X_2, \ldots, X_n$ to be independent it is necessary and sufficient that the moment generating function of the vector $\mathbf{X}$ be the product of the moment generating functions of the random variables $X_1, X_2, \ldots, X_n$.

The proof of (P1) involves some advanced mathematics and we will skip it. However, assuming (P1) it is easy to prove (P2) and we will now do that. First, assume that the random variables $X_1, X_2, \ldots, X_n$ are independent. The density of the vector $\mathbf{X}$ is $f_{X_1}(x_1) f_{X_2}(x_2) \ldots f_{X_n}(x_n)$. Therefore,

$$M_{\mathbf{X}}(\mathbf{t}) = \int_{x_1=-\infty}^{\infty} \int_{x_2=-\infty}^{\infty} \cdots \int_{x_n=-\infty}^{\infty} \exp(t_1 x_1 + \cdots + t_n x_n) f_{X_1}(x_1) \ldots f_{X_n}(x_n) \mathrm{d}x_1 \ldots \mathrm{d}x_n.$$

Since,

$$\exp(t_1 x_1 + t_2 x_2 + \cdots + t_n x_n) = \exp(t_1 x_1) \exp(t_2 x_2) \ldots \exp(t_n x_n)$$

we can integrate with respect to each x_i separately to get

$$M_{\mathbf{X}}(\mathbf{t}) = M_{X_1}(t_1) M_{X_2}(t_2) \ldots M_{X_n}(t_n).$$

We now prove the converse. Assume that the moment generating function of $\mathbf{X}$ is the product of the generating functions of the variables $X_1, X_2, \ldots, X_n$. Then $M_{\mathbf{X}}$ is equal to the moment generating function of a vector whose components $X_1, X_2, \ldots, X_n$ are independent. By using (P1) we see that the random variables $X_1, X_2, \ldots, X_n$ are independent. This completes the proof of (P2).

We will now compute the joint moment generating function of a normal vector. First recall two important properties of normal random variables from Chap. 7.

Properties of Normal Random Variables

Let X be a normal random variable with mean μ and standard deviation σ then the moment generating function of X is

$$M_X(t) = E(\mathrm{e}^{tX}) = \exp\left(\mu t + \sigma^2 \frac{t^2}{2}\right).$$

Let $X_1, X_2, \ldots, X_n$ be a sequence of independent normal random variables and let $t_1, t_2, \ldots, t_n$ be a sequence of real numbers. Let $\mathbf{t}$ be the column vector with components $t_1, t_2, \ldots, t_n$ then the linear combination

$$\mathbf{t}'\mathbf{X} = \sum_{i=1}^{n} t_i X_i$$

is also a normal random variable.

We use the properties of normal random variables to get the following properties of normal vectors.

Properties of Normal Random Vectors

Assume that $\mathbf{X}$ is a normal vector. Then, each component X_i is a normal random variable and a linear transformation of $\mathbf{X}$ is also a normal vector.

We now prove the properties. We know that

$$\mathbf{X} = A\mathbf{Z} + \mathbf{b},$$

where $\mathbf{Z}$ is a random vector whose components are i.i.d. standard normal random variables, A is a constant matrix and $\mathbf{b}$ is a constant vector. Therefore, the i component of $\mathbf{X}$ is

$$X_i = \sum_j A_{i,j} Z_j + b_i.$$

That is, X_i is the sum a linear combination of independent normal random variables and a constant b_i. Thus, X_i is a normal random variable.

Assume now that $\mathbf{Y} = C\mathbf{X}$ for some matrix C. Then,

$$\mathbf{Y} = CA\mathbf{Z} + C\mathbf{b}.$$

This shows that $\mathbf{Y}$ is a normal vector where in the definition CA plays the role of A and $C\mathbf{b}$ plays the role of $\mathbf{b}$. This completes the proof.

Let

$$L = \mathbf{t}'\mathbf{X},$$

where $\mathbf{X}$ is a normal vector. Then L is also a normal vector since it is a linear transformation of $\mathbf{X}$. Actually, L has only one component and is therefore a normal random variable.

We compute the moment generating function of L.

$$M_L(s) = E(\mathrm{e}^{sL}) = \exp(E(L)s + \mathrm{Var}(L)s^2/2).$$

Recall that if A is a matrix then

$$E(A\mathbf{X}) = AE(\mathbf{X})$$

and

$$\mathrm{Var}(A\mathbf{X}) = A\mathrm{Var}(\mathbf{X})A'.$$

Hence, with $\mathbf{t}'$ playing the role of A

$$E(L) = E(\mathbf{t}'\mathbf{X}) = \mathbf{t}'E(\mathbf{X})$$

and

$$\mathrm{Var}(L) = \mathbf{t}'\mathrm{Var}(\mathbf{X})(\mathbf{t}')' = \mathbf{t}'\mathrm{Var}(\mathbf{X})\mathbf{t},$$

where we used that $(\mathbf{t}')' = \mathbf{t}$.

We substitute $E(L)$ and $\mathrm{Var}(L)$ in M_L to get

$$M_L(s) = \exp\left(s\mathbf{t}'E(\mathbf{X}) + \frac{s^2}{2}\mathbf{t}'\mathrm{Var}(\mathbf{X})\mathbf{t}\right).$$

There is a simple relation between the moment generating function of the vector $\mathbf{X}$ at $\mathbf{t}$ and the moment generating function of L.

$$M_{\mathbf{X}}(\mathbf{t}) = E(\exp(\mathbf{t}'\mathbf{X})) = E(\exp(L)) = M_L(1).$$

We plug $s = 1$ in the formula for $M_L(s)$ to get

$$M_{\mathbf{X}}(\mathbf{t}) = \exp\left(\mathbf{t}'E(\mathbf{X}) + \frac{1}{2}\mathbf{t}'\mathrm{Var}(\mathbf{X})\mathbf{t}\right).$$

The Moment Generating of a Normal Vector

Let $\mathbf{X} = \begin{pmatrix} X_1 \\ X_2 \\ . \\ . \\ X_n \end{pmatrix}$ is a normal vector if and only if the moment generating function of $\mathbf{X}$ is

$$M_{\mathbf{X}}(\mathbf{t}) = \exp(\mathbf{t}' E(\mathbf{X}) + \frac{1}{2}\mathbf{t}' \text{Var}(\mathbf{X})\mathbf{t}).$$

The following is useful to prove that a vector is normal.

Independent Components

If the random variables $X_1, X_2, \ldots, X_n$ are independent and normally distributed then the vector $\mathbf{X} = \begin{pmatrix} X_1 \\ X_2 \\ . \\ . \\ X_n \end{pmatrix}$ is a normal vector.

We have at least two ways to show this. We may find a matrix A and a vector $\mathbf{b}$ such that $\mathbf{X} = A\mathbf{Z} + \mathbf{b}$ where $\mathbf{Z}$ is a vector whose components are independent standard normal random variables (see the exercises) or we may use moment generating functions. We use moment generating functions. Since the X_i are assumed to be independent we have by property (P2)

$$M_{\mathbf{X}}(\mathbf{t}) = M_{X_1}(t_1) M_{X_2}(t_2) \ldots M_{X_n}(t_n).$$

Since the X_i are normally distributed, we have for each $i = 1, 2, \ldots, n$

$$M_{X_i}(s) = \exp(\mu_i s + s^2 \sigma_i^2/2).$$

Thus,

$$\begin{aligned} M_{\mathbf{X}}(\mathbf{t}) &= \exp(\mu_1 t_1 + t_1^2 \sigma_1^2/2) \exp(\mu_2 t_2 + t_2^2 \sigma_2^2/2) \ldots \exp(t_n \mu_n t_n + t_n^2 \sigma_n^2/2) \\ &= \exp\left(\sum_{i=1}^{n} t_i \mu_i + \sum_{i=1}^{n} t_i^2 \sigma_i^2/2 \right). \end{aligned}$$

Note that $E(\mathbf{X})$ is the column vector whose i component is μ_i. Hence,

$$\sum_{i=1}^{n} t_i \mu_i = \mathbf{t}' E(\mathbf{X}).$$

Since the X_i are independent $\mathrm{Cov}(X_i, X_j) = 0$ for $i \neq j$ and so $\mathrm{Var}(\mathbf{X})$ is a diagonal matrix whose i, i term is σ_i^2. Observe also that

$$\sum_{i=1}^{n} t_i^2 \sigma_i^2 / 2 = \frac{1}{2} \mathbf{t}' \mathrm{Var}(\mathbf{X}) \mathbf{t}.$$

Therefore,

$$M_{\mathbf{X}}(\mathbf{t}) = \exp\left(\mathbf{t}' E(\mathbf{X}) + \frac{1}{2} \mathbf{t}' \mathrm{Var}(\mathbf{X}) \mathbf{t} \right).$$

This is the moment generating function of a normal vector. By property (P1) this proves that $\mathbf{X}$ is a normal vector.

An easy consequence of the form of the moment generating function for a normal vector is the following property.

Independence and Covariance

Assume that $\mathbf{X}$ is a normal random vector. Let X_i and X_j be two components of $\mathbf{X}$. Then X_i and X_j are independent if and only if

$$\mathrm{Cov}(X_i, X_j) = 0.$$

We already knew that if X_i and X_j are independent then $\mathrm{Cov}(X_i, X_j) = 0$. What is remarkable here is that the converse holds for normal random vectors. We now prove it. Assume that $\mathbf{X}$ is a normal random vector with n components and such that $\mathrm{Cov}(X_i, X_j) = 0$. Define the vector $\mathbf{Y}$ as

$$\mathbf{Y} = \begin{pmatrix} X_i \\ X_j \end{pmatrix}.$$

Let A be the matrix with 2 rows and n columns such that all components are 0 except for the $(1, i)$ and $(2, j)$ components which are both 1. It is easy to check that

$$\mathbf{Y} = A\mathbf{X}.$$

Therefore, $\mathbf{Y}$ is obtained through a linear transformation of the normal vector $\mathbf{X}$ and is also a normal vector. Hence, the moment generating function of $\mathbf{Y}$ is

$$M_{\mathbf{Y}}(\mathbf{t}) = \exp\left(\mathbf{t}' E(\mathbf{Y}) + \frac{1}{2} \mathbf{t}' \mathrm{Var}(\mathbf{Y}\mathbf{t}) \right)$$

The variance matrix of $\mathbf{Y}$ is easily computed,

$$\mathrm{Var}(\mathbf{Y}) = \begin{pmatrix} \mathrm{Var}(X_i) & \mathrm{Cov}(X_i, X_j) \\ \mathrm{Cov}(X_i, X_j) & \mathrm{Var}(X_j) \end{pmatrix} = \begin{pmatrix} \mathrm{Var}(X_i) & 0 \\ 0 & \mathrm{Var}(X_j) \end{pmatrix}.$$

We use this to get

$$\mathbf{t}'\mathrm{Var}(\mathbf{Y})\mathbf{t} = t_1^2\mathrm{Var}(X_i) + t_2^2\mathrm{Var}(X_j),$$

where t_1 and t_2 are the two components of $\mathbf{t}$. On the other hand

$$\mathbf{t}'E(\mathbf{Y}) = t_1E(X_i) + t_2E(X_j).$$

Hence,

$$\begin{aligned} M_{\mathbf{Y}}(\mathbf{t}) &= \exp(t_1E(X_i) + t_1^2\mathrm{Var}(X_i)/2)\exp(t_2E(X_j) + t_2^2\mathrm{Var}(X_j)/2) \\ &= M_{X_i}(t_1)M_{X_j}(t_2). \end{aligned}$$

By property (P2) this proves that X_i and X_j are independent.

Note that in order to use the property above one must first prove that $\mathbf{X}$ is a normal vector. Showing that the marginal densities are normal is not enough. In the exercises we give an example of two normal random variables whose covariance is 0 and that are not independent. This is so because although the marginal densities are normal the joint density is not.

8.3.3 *The Joint Distribution of the Sample Mean and Variance in a Normal Sample*

Let $X_1, X_2, \ldots, X_n$ be independent, normally distributed with mean μ and standard deviation σ. As we have seen in 5.1

$$\bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

is an unbiased estimator of the mean μ and

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

is an unbiased estimator of σ^2. That is,

$$E(\bar{X}) = \mu \text{ and } E(S^2) = \sigma^2.$$

$\bar{X}$ and S^2 are called the sample mean and the sample variance of the sample $X_1, \ldots, X_n$. We will show that

$$\frac{\bar{X} - \mu}{S/\sqrt{n}}$$

follows a Student distribution with $n - 1$ degrees of freedom. This is an important result in statistics that we have used in 5.3. Let

$$\mathbf{X} = \begin{pmatrix} X_1 \\ X_2 \\ . \\ . \\ X_n \end{pmatrix} \quad \mathbf{D} = \begin{pmatrix} \bar{X} \\ X_1 - \bar{X} \\ X_2 - \bar{X} \\ . \\ X_n - \bar{X} \end{pmatrix}.$$

Note that

$$\mathbf{D} = \begin{pmatrix} 1/n & 1/n & . & . & . & 1/n \\ 1 - 1/n & -1/n & . & . & . & -1/n \\ -1/n & 1 - 1/n & . & . & . & -1/n \\ . & . & . & . & . & . \\ -1/n & -1/n & . & . & . & 1 - 1/n \end{pmatrix} \mathbf{X}.$$

Since its components are normal and independent the vector $\mathbf{X}$ is normal. Thus, $\mathbf{D}$ is the image by a linear transformation of a normal vector. Hence, $\mathbf{D}$ is a normal random vector as well. We now compute for $i = 1, \ldots, n$

$$\mathrm{Cov}(\bar{X}, X_i - \bar{X}) = \mathrm{Cov}(\bar{X}, X_i) - \mathrm{Cov}(\bar{X}, \bar{X}).$$

Recall that

$$\mathrm{Cov}(\bar{X}, \bar{X}) = \mathrm{Var}(\bar{X}) = \frac{\sigma^2}{n}.$$

We now turn to

$$\mathrm{Cov}(\bar{X}, X_i) = \frac{1}{n} \sum_{j=1}^{n} \mathrm{Cov}(X_j, X_i).$$

Note that $\mathrm{Cov}(X_j, X_i) = 0$ for $i \neq j$ since X_i and X_j are independent. Thus,

$$\mathrm{Cov}(\bar{X}, X_i) = \frac{1}{n} \mathrm{Cov}(X_i, X_i) = \frac{\sigma^2}{n}.$$

Therefore,

$$\mathrm{Cov}(\bar{X}, X_i - \bar{X}) = 0$$

and since $\mathbf{D}$ is a normal random vector this is enough to show that, for every $i = 1, \ldots, n$, $\bar{X}$ and $X_i - \bar{X}$ are independent. Since S^2 depends only on the differences $X_i - \bar{X}$ we have proved the following.

The Sample Mean and Sample Variance are Independent

Let $X_1, X_2, \ldots, X_n$ be independent, NORMALLY distributed with mean μ and standard deviation σ. Then, the sample mean

$$\bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

and the sample variance

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$$

are independent.

This independence result relies heavily on the normality of the sample.

We know from 7.2 that a linear combination of independent normal random variables is also a normal variable. Therefore, $\bar{X}$ is normally distributed with mean μ and variance σ^2/n. We now turn to the distribution of S^2. We start with

$$\begin{aligned} \sum_{i=1}^{n} (X_i - \bar{X})^2 &= \sum_{i=1}^{n} (X_i - \mu + \mu - \bar{X})^2 \\ &= \sum_{i=1}^{n} (X_i - \mu)^2 + 2 \sum_{i=1}^{n} (X_i - \mu)(\mu - \bar{X}) + \sum_{i=1}^{n} (\mu - \bar{X})^2. \end{aligned}$$

Note that

$$\sum_{i=1}^{n} (X_i - \mu)(\mu - \bar{X}) = (\mu - \bar{X})(n\bar{X} - n\mu) = -n(\mu - \bar{X})^2,$$

and

$$\sum_{i=1}^{n} (\mu - \bar{X})^2 = n(\mu - \bar{X})^2.$$

Thus,

$$(n-1)S^2 = \sum_{i=1}^{n} (X_i - \bar{X})^2 = \sum_{i=1}^{n} (X_i - \mu)^2 - n(\mu - \bar{X})^2.$$

We divide the preceding equality by σ^2 to get

$$(n-1)S^2/\sigma^2 = \sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma} \right)^2 - n \left(\frac{\mu - \bar{X}}{\sigma} \right)^2.$$

The random variable

$$\sum_{i=1}^{n}\left(\frac{X_i-\mu}{\sigma}\right)^2$$

is the sum of the squares of n independent standard normal random variables. Thus, it is a Chi-square random variable with n degrees of freedom. On the other hand

$$n\left(\frac{\mu-\bar{X}}{\sigma}\right)^2=\left(\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}\right)^2$$

is the square of a standard normal random variable (since the expected value of $\bar{X}$ is μ and its standard deviation is $\sigma/\sqrt{n}$). So it is a Chi-square random variable with one degree of freedom. We are now going to find the distribution of S^2 by using moment generating functions. We rewrite the identity

$$(n-1)S^2/\sigma^2=\sum_{i=1}^{n}\left(\frac{X_i-\mu}{\sigma}\right)^2-n\left(\frac{\mu-\bar{X}}{\sigma}\right)^2$$

as

$$(n-1)S^2/\sigma^2+n\left(\frac{\mu-\bar{X}}{\sigma}\right)^2=\sum_{i=1}^{n}\left(\frac{X_i-\mu}{\sigma}\right)^2.$$

We compute the moment generating functions:

$$E\left[\exp\left(t(n-1)S^2/\sigma^2+tn\left(\frac{\mu-\bar{X}}{\sigma}\right)^2\right)\right]=E\left[\exp\left(t\sum_{i=1}^{n}\left(\frac{X_i-\mu}{\sigma}\right)^2\right)\right].$$

We use also that S^2 and $\bar{X}$ are independent to get

$$E[\exp(t(n-1)S^2/\sigma^2]E\left[\exp\left(tn\left(\frac{\mu-\bar{X}}{\sigma}\right)^2\right)\right]=E\left[\exp\left(t\sum_{i=1}^{n}\left(\frac{X_i-\mu}{\sigma}\right)^2\right)\right].$$

Recall that a Chi-square random variable with k degrees of freedom has a moment generating function $(1-2t)^{-k/2}$. Hence,

$$E\left[\exp\left(tn\left(\frac{\mu-\bar{X}}{\sigma}\right)^2\right)\right]=(1-2t)^{-1/2},$$

and

$$E\left[\exp\left(t\sum_{i=1}^{n}\left(\frac{X_i-\mu}{\sigma}\right)^2\right)\right]=(1-2t)^{-n/2}.$$

Therefore,

$$E[\exp(t(n-1)S^2/\sigma^2](1-2t)^{-1/2}=(1-2t)^{-n/2},$$

and

$$E[\exp(t(n-1)S^2/\sigma^2]=(1-2t)^{-(n-1)/2}.$$

That is, $(n-1)S^2/\sigma^2$ follows a Chi-square distribution with $n-1$ degrees of freedom.

Finally, since

$$\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$$

is a standard normal random variable which is independent of $(n-1)S^2/\sigma^2$ (a Chi-square random variable with $n-1$ degrees of freedom) we have that

$$\frac{\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}}{\sqrt{\frac{(n-1)S^2}{(n-1)\sigma^2}}}=\frac{\bar{X}-\mu}{S/\sqrt{n}}$$

follows a Student distribution with $n-1$ degrees of freedom. We now summarize our results.

Joint Distribution of the Sample Mean and the Sample Variance

Let $X_1, X_2, \ldots, X_n$ be independent, NORMALLY distributed with mean μ and standard deviation σ. Then,

$$(n-1)S^2/\sigma^2$$

follows a Chi-square distribution with $n-1$ degrees of freedom. Moreover,

$$\frac{\bar{X}-\mu}{S/\sqrt{n}}$$

is a Student random variable with $n-1$ degrees of freedom.

Exercises 8.3

1. Show that if $C_1 g(x)$ and $C_2 g(x)$ are both density functions then $C_1 = C_2$.

2. (a) Assume that X and Y are independent Chi-square random variables with degrees of freedom n and m, respectively. Show that $X + Y$ is also a Chi-square random variable with $n + m$ degrees of freedom.

(b) Assume that X and Y are independent and that X and $X + Y$ are Chi-square random variables with degrees of freedom n and m, respectively. Show that Y is also a Chi-square random variable with $m - n$ degrees of freedom.

3. Use a Gamma density to compute the following integral.

$$\int_0^\infty x^5 \mathrm{e}^{-2x} \mathrm{d}x.$$

4. Assume that the vector $\mathbf{X}$ has three independent components X_1, X_2 and X_3 with $\mathrm{Var}(X_i) = \sigma_i^2$.

(a) Write the variance matrix of the vector $\mathbf{X}$.

(b) Let $Y_1 = X_1$, $Y_2 = X_1 + X_2$ and $Y_3 = X_1 + X_2 + X_3$. Find the matrix A such that $\mathbf{Y} = A\mathbf{X}$.

(c) Find the variance matrix of the vector $\mathbf{Y}$.

(d) Are the components of $\mathbf{Y}$ independent?

5. Assume that $\mathrm{Var}(X) = 1$, $\mathrm{Var}(Y) = 2$ and $\mathrm{Cov}(X, Y) = -1$. Compute $\mathrm{Cov}(X - 2Y, X + Y)$.

6. Assume that $\begin{pmatrix} X \\ Y \end{pmatrix}$ is a normal vector with $E(X) = 1$, $\mathrm{Var}(X) = 1$, $E(Y) = -1$, $\mathrm{Var}(Y) = 2$ and $\mathrm{Cov}(X, Y) = -1$. Find the moment generating function of the vector $\begin{pmatrix} X + Y \\ X - Y \end{pmatrix}$.

7. Assume that X_1 and X_2 are independent normally distributed random variables with means μ_1, μ_2 and standard deviations σ_1, σ_2, respectively. Show that $\mathbf{X} = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$ is a normal vector by finding a matrix A and a vector $\mathbf{b}$ such that

$$\mathbf{X} = A\mathbf{Z} + \mathbf{b},$$

where $\mathbf{Z} = \begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix}$ and Z_1 and Z_2 are independent standard normal random variables.

8. Assume that X_1 and X_2 are independent normally distributed random variables with mean μ and standard deviation σ. Let

$$\mathbf{Y} = \begin{pmatrix} X_1 - X_2 \\ X_1 + X_2 \end{pmatrix}.$$

(a) Is $\mathbf{Y}$ a normal vector?
(b) Compute Var($\mathbf{Y}$).
(c) Are $X_1 - X_2$ and $X_1 + X_2$ independent?

9. In this exercise we will construct two normal random variables X and Y such that $\text{Cov}(X, Y) = 0$ and such that X and Y are not independent. Let

$$n(x, y) = \frac{1}{2\pi} \exp(-(x^2 + y^2)/2).$$

That is, n is the joint density of two independent standard normal random variables. Let D_1, D_2, D_3, and D_4 be the interiors of the circles with radius 1 and centered at (2,2), (−2,2), (−2,−2), and (2,−2), respectively. Define

$$f(x, y) = n(x, y) \text{ for } (x, y) \text{ not in } D_1 \cup D_2 \cup D_3 \cup D_4$$
$$f(x, y) = n(x, y) + m \text{ for } (x, y) \text{ in } D_1$$
$$f(x, y) = n(x, y) - m \text{ for } (x, y) \text{ in } D_2$$
$$f(x, y) = n(x, y) + m \text{ for } (x, y) \text{ in } D_3$$
$$f(x, y) = n(x, y) - m \text{ for } (x, y) \text{ in } D_4,$$

where m is a constant small enough so that $f(x, y)$ is always strictly positive.

(a) Show that f is a density.
(b) Show that X and Y are standard normal random variables.
(c) Compute the covariance of X and Y.
(d) Show that X and Y are not independent.

10. Assume that (X, Y) is a normal vector with $\text{Var}(X) = 1$, $\text{Var}(Y) = 2$ and $\text{Cov}(X, Y) = -1$. Find a so that X and $X + aY$ are independent.

11. Show that a Student distribution has an expectation equal to 0.

12. (a) Find the expectation and the variance of a Chi-square random variable with n degrees of freedom.
(b) Find a normal approximation to a Chi-square distribution with n degrees of freedom as n goes to infinity.

13. Find the expected value and the variance of a F distribution with (m, n) degrees of freedom.

14. Let $X_1, X_2, \ldots, X_n$ be random variables and $t_1, t_2, \ldots, t_n$ be constants. Show that

$$\text{Var}\left(\sum_{i=1}^{n} t_i X_i\right) = \sum_{i=1}^{n}\sum_{j=1}^{n} t_i t_j \text{Cov}(X_i, X_j).$$

15. Consider a normal vector $\mathbf{X}$. Assume that

$$\mathbf{X} = A\mathbf{Z},$$

where $\mathbf{Z}$ is a normal vector whose components are independent standard normal random variables and A is an invertible matrix. In this exercise we will compute the density of $\mathbf{X}$.

(a) Show that the density of the vector $\mathbf{Z}$ is

$$f(z_1, \ldots, z_n) = \frac{1}{(2\pi)^{n/2}} \exp\left(-\frac{1}{2}\mathbf{z}'\mathbf{z}\right),$$

where $\mathbf{z}$ is the vector whose components are $z_1, \ldots, z_n$.

(b) Use the change of variables formula to show that the density of vector $\mathbf{X}$ is

$$f(x_1, \ldots, x_n) = \frac{1}{(2\pi)^{n/2}} \det(A^{-1}) \exp\left(-\frac{1}{2}(A^{-1}\mathbf{x})'(A^{-1}\mathbf{x})\right),$$

where $\mathbf{x}$ is the vector whose components are $x_1, \ldots, x_n$ and $\det(A^{-1})$ is the determinant of A^{-1}.

(c) Let Σ be the variance matrix of $\mathbf{X}$. Show that

$$\Sigma = AA',$$

and that Σ is invertible. Recall that for two same size square matrices M and N

$$\det(MN) = \det(M)\det(N)$$

and that matrix M is invertible if and only if $\det(M)$ is not 0. Recall also that $\det(M') = \det(M)$.

(d) Show that the density of vector $\mathbf{X}$ is

$$f(x_1, \ldots, x_n) = \frac{1}{(2\pi)^{n/2}} \det(\Sigma)^{-1/2} \exp\left(-\frac{1}{2}\mathbf{x}'\Sigma^{-1}\mathbf{x}\right).$$

8.4 Conditional Distributions and Expectations

We start with an example for the discrete case.

Example 1. Let X and Y be two random variables with the following joint distribution.

X	0	1	2
Y			
1	1/8	1/8	1/4
2	1/8	0	1/8
3	1/8	1/8	0

By definition of conditional probabilities we have

$$P(X = 0|Y = 1) = \frac{P(X = 0; Y = 1)}{P(Y = 1)} = \frac{1/8}{1/2} = \frac{1}{4},$$

where we read $P(X = 0; Y = 1) = 1/8$ in the table and

$$\begin{aligned} P(Y = 1) &= P(X = 0; Y = 1) + P(X = 1; Y = 1) + P(X = 2; Y = 1) \\ &= \frac{1}{8} + \frac{1}{8} + \frac{1}{4} = \frac{1}{2} \end{aligned}$$

Similarly, we get

$$P(X = 1|Y = 1) = \frac{1/8}{1/2} = \frac{1}{4},$$

and

$$P(X = 2|Y = 1) = \frac{1/4}{1/2} = \frac{1}{2}.$$

The conditional distribution of X given $Y = 1$ is

x	0	1	2
$P(X = x\|Y = 1)$	1/4	1/4	1/2

We now give the definition in the discrete case.

Conditional Distribution, Discrete Case

Let X and Y be two discrete random variables. Assume that y is such that $P(Y = y) > 0$. The conditional distribution of X given $Y = y$ is given by

$$P(X = x|Y = y) = \frac{P(X = x; Y = y)}{P(Y = y)}$$

for all possible values x of the random variable X.

A word on notation we use "," or ";" equivalently between two events. It designates the intersection of the two events.

As the next example illustrates we sometimes know the conditional distribution of a random variable and use it to get the (unconditioned) distribution of the random variable.

Example 2. Assume that the number of customers going into a bank between 2:00 and 3:00 PM has a Poisson distribution with rate λ. Assume that each customer has a probability p of being female. Assume also that arrivals are female or not independently of each other. What is the distribution of the number of female customers between 2:00 and 3:00 p.m.?

Let N be the total number of customers and F be the number of female customers. We now compute the distribution of F. For any fixed positive integer f we have

$$P(F = f) = \sum_{n \geq f} P(N = n; F = f),$$

where $n \geq f$ since there are more arrivals than female arrivals. By definition of the conditional probability

$$P(N = n; F = f) = P(F = f | N = n) P(N = n).$$

Given that $N = n$, F is the number of females among n customers. We are told that each arriving customer has a probability p of being female and that arrivals are female or male independently of each other. Hence, given $N = n$ the number of females F is a binomial distribution with parameters n and p. That is,

$$P(F = f | N = n) = \binom{n}{f} p^f (1 - p)^{n-f}.$$

Since

$$P(N = n) = \exp(-\lambda) \frac{\lambda^n}{n!},$$

we have

$$\begin{aligned} P(N = n; F = f) &= P(F = f | N = n) P(N = n) \\ &= \binom{n}{f} p^f (1 - p)^{n-f} \exp(-\lambda) \frac{\lambda^n}{n!}. \end{aligned}$$

Now,

$$\binom{n}{f} \frac{1}{n!} = \frac{1}{f!(n - f)!}.$$

and by splitting λ^n in $\lambda^f \lambda^{n-f}$ we get

$$P(N = n; F = f) = \frac{(\lambda p)^f}{f!} \exp(-\lambda)(1-p)^{n-f} \frac{\lambda^{n-f}}{(n-f)!}.$$

We use this last formula in the sum

$$P(F = f) = \sum_{n \geq f} P(N = n; F = f) = \frac{(\lambda p)^f}{f!} \exp(-\lambda) \sum_{n \geq f} \frac{((1-p)\lambda)^{n-f}}{(n-f)!}.$$

Finally, note that

$$\sum_{n \geq f} \frac{((1-p)\lambda)^{n-f}}{(n-f)!} = \sum_{k \geq 0} \frac{((1-p)\lambda)^k}{k!} = \exp((1-p)\lambda).$$

Hence,

$$P(F = f) = \frac{(\lambda p)^f}{f!} \exp(-\lambda) \exp((1-p)\lambda) = \frac{(\lambda p)^f}{f!} \exp(-\lambda p).$$

That is, F is also Poisson distributed and its rate is λp.

We now deal with the continuous case.

Conditional Distribution, Continuous Case

Assume that X and Y are continuous random variables with joint density f. Let f_Y be the marginal density of the random variable Y. Let y be a fixed number such that $f_Y(y) > 0$. The conditional density of X given $Y = y$ is defined to be

$$f(x|y) = \frac{f(x,y)}{f_Y(y)}.$$

We will also use the notation $f(x|Y = y)$ for $f(x|y)$ when we will want to emphasize that y is a value of the random variable Y.

Example 3. Take (X, Y) uniformly distributed on the two dimensional unit disc. That is,

$$f(x,y) = \frac{1}{\pi} \text{ for } x^2 + y^2 \leq 1.$$

Given $y = 1/2$ what is the conditional density of X? In 8.2 we computed the marginal density

$$f_Y(y) = \frac{2}{\pi} \sqrt{1-y^2} \text{ for } y \in [-1, 1].$$

So by definition of the conditional density we have

$$f(x|Y = 1/2) = \frac{f(x, 1/2)}{f_Y(1/2)} = \frac{\frac{1}{\pi}}{\frac{2}{\pi}\sqrt{1-(1/2)^2}} \text{ for } x^2 + (1/2)^2 \leq 1.$$

That is,

$$f(x|Y = 1/2) = \frac{1}{\sqrt{3}} = \frac{\sqrt{3}}{3} \text{ for } -\frac{\sqrt{3}}{2} \leq x \leq \frac{\sqrt{3}}{2}.$$

So X conditioned on $Y = 1/2$ is uniformly distributed on $[-\frac{\sqrt{3}}{2}, \frac{\sqrt{3}}{2}]$. There is nothing special about $Y = 1/2$, of course. One can show that for any y in $[-1, 1]$ the conditional density of X given $Y = y$ is uniform on $[-\sqrt{1-y^2}, \sqrt{1+y^2}]$.

Example 4. Assume that X is picked uniformly in $(0, 1)$ and then Y is picked uniformly in $(0, X)$. What is the distribution of (X, Y)?

From the information above we get that the conditional distribution of Y given $X = x$ is uniform in $(0, x)$. That is,

$$f(y|x) = \frac{1}{x} \text{ for } y \in (0, x).$$

We use the formula for conditional density to get

$$f(x, y) = f(y|x) f_X(x) = \frac{1}{x} \text{ for } 0 < y < x < 1,$$

where we are using that $f_X(x) = 1$ for x in $[0, 1]$ and 0 otherwise. We see that (X, Y) is not uniformly distributed. Is Y uniformly distributed? We compute

$$f_Y(y) = \int_y^1 f(x, y)\mathrm{d}x = \int_y^1 \frac{1}{x}\mathrm{d}x = -\ln y \text{ for } y \in (0, 1).$$

Hence, Y is not uniformly distributed either.

8.4.1 Conditional Expectations

Example 5. In Example 2 the conditional distribution of F given $N = n$ is a binomial distribution with parameters n and p. Hence, its expected value denoted by $E(F|N = n)$ is

$$E(F|N = n) = np.$$

Now, n is a value of the random variable N. This allows us to introduce a new random variable denoted by $E(F|N)$ and defined as follows. The random variable

$E(F|N)$ is $E(F|N = n) = np$ when $N = n$. $E(F|N)$ is called the conditional expectation of F given N. In this example we have

$$E(F|N) = Np.$$

Note that $E(F|N) = g(N)$ where g is deterministic (i.e., nonrandom), indeed $g(x) = px$. This is a general fact: the conditional expectation of a random variable X given Y is always a function of Y.

Example 6. Consider Example 4. The random variable X is uniformly distributed on $(0, 1)$. Given $X = x$ then Y is uniformly distributed on $(0, x)$. Since the expected value of an uniform distribution on $(0, x)$ is $x/2$ we have

$$E(Y|X = x) = \frac{x}{2}.$$

Hence,

$$E(Y|X) = \frac{X}{2}.$$

Conditional Expectation

Let X and Y be two random variables. Let $E(X|Y = y)$ be the expected value of the distribution of X conditioned on $Y = y$. The random variable $E(X|Y)$ is defined to be $E(X|Y = y)$ on the event $Y = y$. It is called the conditional expectation of X given Y. Moreover, $E(X|Y) = g(Y)$ where

$$g(y) = \frac{1}{f_Y(y)} \int x f(x, y) \mathrm{d}x \text{ if } X \text{ and } Y \text{ are continuous}$$

and

$$g(y) = \frac{1}{P(Y = y)} \sum_x x P(X = x, Y = y) \text{ if } X \text{ and } Y \text{ are discrete.}$$

This is not the most general definition of conditional expectation but it will be enough for our purposes. Provided the conditional distribution is well defined (i.e., $f_Y(y) > 0$ in the continuous case and $P(Y = y) > 0$ in the discrete case) and the corresponding expected values exist our definition is fine. We now check the existence of the function g. Assume that X and Y are continuous random variables with joint density f, the computation for discrete random variables is similar and we omit it. By the definitions of expectation and conditional distribution we get

$$E(X|Y = y) = \int x f(x|y) \mathrm{d}x = \int x \frac{f(x, y)}{f_Y(y)} \mathrm{d}x = \frac{1}{f_Y(y)} \int x f(x, y) \mathrm{d}x.$$

As expected the expression above depends on y only. Define g by

$$g(y) = \frac{1}{f_Y(y)} \int x f(x, y) \mathrm{d}x.$$

For every y such that $f_Y(y) > 0$ we have

$$E(X|Y = y) = g(y).$$

By the definition of the conditional expectation we get $E(X|Y) = g(Y)$. This proves the existence of the function g and also gives an explicit expression for g that will be useful below.

We now turn to some of the properties of conditional expectations.

P1. We have that

$$E[E(X|Y)] = E(X).$$

First note that since $E(X|Y)$ is a random variable it makes sense to compute its expectation. We now prove this formula in the continuous case. We have $E(X|Y) = g(Y)$ for the function g defined above. Recall that, if the expected value exists we have

$$E(g(Y)) = \int g(y) f_Y(y) \mathrm{d}y.$$

Hence,

$$E[E(X|Y)] = E(g(Y)) = \int g(y) f_Y(y) \mathrm{d}y.$$

We now use the expression of g above to get

$$\begin{aligned} E[E(X|Y)] &= \int g(y) f_Y(y) \mathrm{d}y = \int \frac{1}{f_Y(y)} \int x f(x, y) \mathrm{d}x f_Y(y) \mathrm{d}y \\ &= \int \int x f(x, y) \mathrm{d}x \mathrm{d}y. \end{aligned}$$

Now use that

$$\int f(x, y) \mathrm{d}y = f_X(x)$$

to have

$$E[E(X|Y)] = \int x f_X(x) \mathrm{d}x = E(X)$$

which completes the computation. Note that we have freely interchanged the integration order. This is fine provided the function we integrate is regular enough. That will always be the case in our examples.

Example 7. We go back to Example 4. We have noted that $E(Y|X) = X/2$. Therefore,

$$E[E(Y|X)] = E(X/2) = \frac{1}{4},$$

since X is uniformly distributed in $(0, 1)$ and $E(X) = 1/2$. We now check property $P1$ by computing $E(Y)$ directly. We have computed

$$f_Y(y) = -\ln y \text{ for } y \in (0, 1).$$

Doing an integration by parts we get

$$E(Y) = \int_0^1 -y \ln y \mathrm{d}y = -\frac{1}{2} y^2 \ln y \Big]_0^1 + \frac{1}{2} \int_0^1 y \mathrm{d}y = \frac{1}{4} y^2 \Big]_0^1 = \frac{1}{4}.$$

We do have $E[E(Y|X)] = E(X)$.

Example 8. Let N and F be two discrete random variables. Assume that the conditional distribution of F given $N = n$ is binomial with parameters n and p. Hence, $E(F|N = n) = np$ and $E(F|N) = pN$. By P1 we get

$$E(F) = E[E(F|N)] = E(pN) = pE(N).$$

The point of this example is that sometimes conditional distributions can be used to compute (unconditional) expectations.

P2. Assume that X and Y are independent. Then,

$$E(X|Y) = E(X).$$

Property P2 is intuitively reasonable. If X and Y are independent then conditioning on Y should not change the distribution of X and hence its expected value. To prove P2 we use the function g again. Assuming that X and Y are continuous random variables we know that

$$E(X|Y = y) = g(y),$$

where

$$g(y) = \frac{1}{f_Y(y)} \int x f(x, y) \mathrm{d}x.$$

Since X and Y are independent

$$f(x, y) = f_X(x) f_Y(y),$$

and

$$g(y) = \frac{1}{f_Y(y)} \int x f(x, y) \mathrm{d}x = \frac{1}{f_Y(y)} \int x f_X(x) f_Y(y) \mathrm{d}x$$
$$= \frac{1}{f_Y(y)} f_Y(y) \int x f_X(x) \mathrm{d}x.$$

That is,

$$g(y) = \int x f_X(x) \mathrm{d}x = E(X).$$

The function g is the constant $E(X)$. Therefore, $E(X|Y)$ is also this constant and P2 is proved.

P3. The conditional expectation is linear. That is, let U, V, and W be random variables and a and b be constants then

$$E(aU + bV|W) = aE(U|W) + bE(V|W).$$

P3 is a consequence of the linearity of the expectation. We omit this proof.

P4. Let h be a function, X and Y random variables then

$$E(h(Y)X|Y) = h(Y)E(X|Y).$$

This last property can be quite useful in the computation of conditional expectations, we will use it below. We now give the steps to prove it. We have

$$E(h(Y)X|Y = y) = E(h(y)X|Y = y) = h(y)E(X|Y = y),$$

where the first equality is intuitively clear but requires a proof. One way to do this is to compute the joint distribution of $h(Y)X$ and Y and then the conditional distribution of $h(Y)X$ given $Y = y$. The second equality just uses the linearity of conditional expectations since $h(y)$ is a constant. This yields P4.

P5. Let h be a function and Y a random variable then

$$E(h(Y)|Y) = h(Y).$$

This is not really a surprise. Conditioning on Y does not change the distribution of $h(Y)$. We now prove P5. Let $X = 1$ in P4, then

$$E(h(Y)|Y) = h(Y)E(1|Y).$$

A constant is independent of any other random variable, hence by P2

$$E(1|Y) = E(1) = 1.$$

This concludes the proof of P5.

8.4.1.1 Prediction and Conditional Expectations

One of the main questions in statistics is, given a sample, how to find an optimal estimator. In Bayes' estimation in particular we are concerned with how to find a function of Y (representing the sample) which is closest to X (representing the parameter we are estimating). This turns out to be $E(X|Y)$.

Prediction and Conditional Expectation

Let X and Y be two random variables such that X has a finite second moment. We look for the minimum of

$$E[(X - h(Y))^2]$$

over all functions h such that $h(Y)$ has a finite second moment. The minimum is attained for

$$h(Y) = E(X|Y).$$

In words, the best predictor X based on Y is the conditional expectation $E(X|Y)$. Note that our definition of "best" is with respect to the mean quadratic distance $E[(X - h(Y))^2]$ (this is why we need X and $h(Y)$ to have a second moment). If we pick a different distance we may end up with a different optimal predictor. We now prove this result. Define g as

$$g(Y) = E(X|Y).$$

We have

$$\begin{aligned} E[(X - h(y))^2] &= E[(X - g(Y) + g(Y) - h(y))^2] \\ &= E[(X - g(y))^2] + E[(g(Y) - h(y))^2] \\ &\quad +2E[(X - g(y))(g(Y) - h(y))] \end{aligned}$$

We will show that the double product is 0. Once this is done, since $E[(g(Y) - h(y))^2] \geq 0$ (the expected value of a positive random variable is positive) we have

$$E[(X - h(y))^2] \geq E[(X - g(y))^2],$$

for all h. This shows that for any h, $E[(X - h(y))^2]$ is larger than $E[(X - g(y))^2]$ and we will be done.

We now show that the double product is 0. By P4

$$E[(h(Y) - g(Y))X|Y] = (h(Y) - g(Y))E(X|Y) = (h(Y) - g(Y))g(Y).$$

By taking expectations on both sides and using (P1) we have

$$E[(h(Y) - g(Y))X] = E[(h(Y) - g(Y))g(Y)].$$

We move the right-hand side to the left and use the linearity of expectations to get

$$E[(h(Y) - g(Y))(X - g(Y))] = 0.$$

This proves that the double product is 0 and we are done.

Example 9. Consider Example 4 again. The random variable X is uniformly distributed on $(0, 1)$. Given $X = x$ then Y is uniformly distributed on $(0, x)$. What is the best predictor of Y of the form $h(X)$?

We know that the best predictor of Y based on X is $E(Y|X)$. Since

$$E(Y|X) = \frac{X}{2}$$

the best predictor of Y is $X/2$.

Example 10. Let $\begin{pmatrix} X \\ Y \end{pmatrix}$ be a normal vector. What is $E(X|Y)$?

In a first step we find a so that $X - aY$ and Y are independent. First note that the vector $\begin{pmatrix} X - aY \\ Y \end{pmatrix}$ can be obtained from $\begin{pmatrix} X \\ Y \end{pmatrix}$ by a linear transformation

$$A = \begin{pmatrix} 1 & -a \\ 0 & 1 \end{pmatrix}.$$

Hence, $\begin{pmatrix} X - aY \\ Y \end{pmatrix}$ is also a normal vector. We know that coordinates of a random vector are independent if and only if they are uncorrelated. Hence, we need to find a such that

$$\mathrm{Cov}(X - aY, Y) = 0.$$

Recall that Cov is a bilinear operator (i.e., linear in each coordinate). Therefore,

$$\mathrm{Cov}(X - aY, Y) = \mathrm{Cov}(X, Y) - a\mathrm{Cov}(Y, Y) = 0.$$

Since $\mathrm{Cov}(Y, Y) = \mathrm{Var}(Y)$ we get

$$a = \frac{\mathrm{Cov}(X, Y)}{\mathrm{Var}(Y)}.$$

We now turn to the computation of the conditional expectation. Using that $X - aY$ and Y are independent, by P2 we have

$$E(X - aY|Y) = E(X - aY).$$

We use the linearity of expectations and conditional expectations on both sides to get

$$E(X|Y) - aE(Y|Y) = E(X) - aE(Y).$$

Since $E(Y|Y) = Y$,

$$E(X|Y) = a(Y - E(Y)) + E(X).$$

Hence, the best predictor of X based on Y is a linear function of Y. It is a remarkable fact that we get a linear function given the vast set of functions at our disposal.

Exercises 8.4

1. For the distribution in Example 1,

(a) Compute the conditional distribution of X given $Y = 2$.
(b) Compute the conditional distribution of Y given $X = 0$.

2. Assume that X and Y are discrete independent random variables. Show that for all k and j

$$P(X = j, X + Y = k) = P(X = j)P(Y = k - j).$$

3. Assume that X and Y are independent binomial random variables with parameters (n, p) and (ℓ, p), respectively.

(a) Show that for $0 \leq j \leq k$

$$P(X = j|X + Y = k) = \frac{\binom{n}{j}\binom{\ell}{k-j}}{\binom{n+\ell}{k}}.$$

(Recall the sum of two independent binomials with the same p is also a binomial).

(b) Use (a) to prove that

$$\sum_{j=0}^{k} \binom{n}{j}\binom{\ell}{k-j} = \binom{n+\ell}{k}.$$

4. Let X and Y be two discrete random variables. Assume that y is such that $P(Y = y) > 0$. Show that

$$\sum_x P(X = x|Y = y) = 1,$$

where the sum is over all possible values x of X.

5. Consider Example 3. Fix y in $[-1, 1]$. Show that the conditional density of X given $Y = y$ is uniform on $[-\sqrt{1-y^2}, \sqrt{1+y^2}]$.

6. Consider Example 4. Compute the conditional density of X given $Y = y$ for a fixed y in $(0, 1)$.

7. The joint density of (X, Y) is given by

$$f(x, y) = \exp(-y) \text{ for } 0 < x < y.$$

(a) Check that f is indeed a density.
(b) Fix $y > 0$. Show that the conditional density of X given $Y = y$ is uniform in $(0, y)$.
(c) Compute the conditional density of Y given $X = x$ for a fixed $x > 0$.

8. Assume that X and Y are continuous random variables. Show that for each y such that $f_Y(y) > 0$ we have

$$\int_{-\infty}^{+\infty} f(x|y)\mathrm{d}x = 1.$$

9. Assume that X and Y are independent and Poisson distributed with rates λ and μ, respectively. Fix a positive integer n.

(a) Show that for any positive integer $k \leq n$

$$P(X = k|X + Y = n) = \binom{n}{k} \left(\frac{\lambda}{\lambda+\mu}\right)^k \left(1 - \frac{\lambda}{\lambda+\mu}\right)^{n-k}.$$

(b) What is the conditional distribution of X given $X + Y = n$?
(c) What is $E(X|X + Y)$?

10. Consider X and Y with joint density

$$f(x, y) = x + y \text{ for } 0 < x < 1 \text{ and } 0 < y < 1.$$

(a) What is $E(X|Y)$?
(b) What is $E(X)$?

11. 1. Consider X and Y with joint density

$$f(x, y) = 8xy \text{ for } 0 < y < x < 1.$$

(a) What is $E(X|Y)$?
(b) What is $E(XY^2|Y)$?

12. Consider Example 4. Show that the best predictor of X based on Y is

$$-\frac{1}{\ln Y}(1 - Y).$$

13. Assume that X and Y are discrete random variables.

(a) Show that

$$E(X|Y) = g(Y),$$

where

$$g(y) = \frac{1}{P(Y = y)} \sum_x xP(X = x, Y = y).$$

(b) Prove P1 in the discrete case.
(c) Prove P2 in the discrete case.

14. Consider a sequence $X_1, \ldots, X_n, \ldots$ of i.i.d. discrete random variables with expected value μ. Let N be a random variable taking values on natural numbers and independent of $X_1, \ldots, X_n, \ldots$. Define Y as

$$Y = \sum_{i=1}^{N} X_i,$$

so Y is a sum of a random number of random variables.

(a) Show that

$$P(Y = k|N = n) = P(X_1 + \cdots + X_n = k).$$

(b) Use (a) to show that

$$E(Y|N = n) = n\mu.$$

(c) Compute $E(Y)$.

15. Let X and Y be two random variables such that $E(X|Y)$ is a constant c.

(a) Show that $c = E(X)$.
(b) Show that

$$E(XY) = E[E(XY|Y)] = E[YE(X|Y)].$$

(c) Use that $E(X|Y)$ is a constant in (b) to get

$$E(XY) = E(X|Y)E(Y)$$

and conclude that

$$E(XY) = E(X)E(Y).$$

That is, X and Y are uncorrelated.

16. Assume that X is a Bernoulli random variable and that Y is a discrete random variable. Show that

$$E(X|Y = y) = P(X = 1|Y = y).$$

17. Let (X, Y) be a normal vector with $E(X) = 1$, $E(Y) = 2$, $\text{Cov}(X, Y) = -1$, $\text{Var}(X) = 1$ and $\text{Var}(Y) = 3$. What is $E(X|Y)$?

18. Let X and Y be independent and having the same distribution.

(a) Show that

$$E(X|X + Y) = E(Y|X + Y).$$

(b) Explain why

$$E(X + Y|X + Y) = X + Y.$$

(c) Use (a) and (b) to show that

$$E(X|X + Y) = \frac{1}{2}(X + Y).$$

Chapter 9
Finding and Comparing Estimators

9.1 Finding Estimators

In Chap. 5 we were concerned with the problem of estimating the mean of a certain distribution. Assume that $X_1, X_2, \ldots, X_n$ is an i.i.d. sample of the distribution of interest. Let μ be the mean of this distribution. Then, a natural way to estimate μ is to use

$$\hat{\mu} = \frac{1}{n}(X_1 + X_2 + \cdots + X_n) = \bar{X}.$$

This is what we did in Chap. 5. However, is $\bar{X}$ the best way to estimate μ? What does "best" mean? What if I am interested in estimating not the mean but something else? This chapter will be concerned with such questions. We start by defining estimators.

Estimators

Let $X_1, X_2, \ldots, X_n$ be an i.i.d. sample of a given distribution with parameter θ. An estimator of θ is any function of the sample $X_1, X_2, \ldots, X_n$. The function cannot depend on θ.

We now introduce our first method to find an estimator.

9.1.1 The Method of Moments

We consider an example first.

R.B. Schinazi, *Probability with Statistical Applications*,
DOI 10.1007/978-0-8176-8250-7_9, © Springer Science+Business Media, LLC 2012

Example 1. Consider the exponential distribution with parameter θ. It has the probability density $f(x) = \theta \exp(-\theta x)$. Its first moment is $1/\theta$ (see Example 7 in Sect. 2.3). Hence,

$$\theta = \frac{1}{\mu_1}$$

and the method of moment estimator (or m.m.e.) for θ is

$$\hat{\theta} = \frac{1}{\hat{\mu}_1} = \frac{1}{\bar{X}}.$$

For instance, assume that we have the following 20 observations 2.9983, 0.2628, 0.9335, 0.6655, 2.2192, 1.4359, 0.6097, 0.0187, 1.7226, 0.5883, 0.9556, 1.5699, 2.5487, 1.3402, 0.1939, 0.5204, 2.7406, 2.4878, 0.5281, 2.2410. The sample average is 1.329 and $\hat{\theta} = 0.7524$.

We summarize the method below.

The Method of Moments Estimator

For any natural number $k \geq 1$, $\mu_k = E(X^k)$ is called the kth moment of the random variable X (if it exists!). Hence, μ_1 is the expectation. Assume that we want to estimate the parameter θ of the distribution of X. The method of moments consists in computing as many moments of X as necessary in order to have an equation in θ that can be solved as a function of the moments of X. Then each moment $E(X^k)$ that appears in the solution is estimated using

$$\hat{\mu}_k = \frac{1}{n}\left(X_1^k + X_2^k + \cdots + X_n^k\right).$$

Example 2. Let X be a discrete random variable uniformly distributed over $\{1, 2, \ldots, \theta\}$. That is,

$$P(X = m) = 1/\theta \text{ for } m \text{ in } \{1, 2, \ldots, \theta\}.$$

As computed in Sect. 2.3 we have

$$\mu_1 = \frac{1+\theta}{2}.$$

Therefore,

$$\theta = 2\mu_1 - 1$$

and the m.m.e. in this case is

$$\hat{\theta} = 2\hat{\mu}_1 - 1 = 2\bar{X} - 1.$$

Assume that we have the following observations: 5, 2, 1, 3, 2, 2, 2, 6, 2, 3. The average is 2.8 and $\hat{\theta} = 4.6$.

Example 3. Let X be a normal random variable with mean μ and variance σ^2. We have $E(X) = \mu$. Hence, the method of moments estimator of μ is $\bar{X}$. In order to have the estimator for $\theta = \sigma^2$ we need another equation. Namely,

$$\theta = \sigma^2 = E(X^2) - E(X)^2.$$

Hence, the method of moments estimator of θ is

$$\hat{\theta} = \hat{\mu}_2 - \hat{\mu}_1^2.$$

We now show that this is a different estimator from the estimator S^2 we have used so far. Recall that

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2.$$

Expanding the squares gives

$$S^2 = \frac{1}{n-1} \left(\sum_{i=1}^{n} X_i^2 - 2\bar{X} \sum_{i=1}^{n} X_i + n(\bar{X})^2 \right).$$

Since

$$\sum_{i=1}^{n} X_i = n\bar{X}$$

we get

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} X_i^2 - \frac{n}{n-1} (\bar{X})^2 = \frac{n}{n-1} (\hat{\mu}_2 - \hat{\mu}_1^2),$$

which is indeed different from $\hat{\theta}$.

Very little is used about the normal distribution in this example. In fact, the method of moments estimator of μ is always $\bar{X}$. However, the estimator for the variance may be different from $\hat{\mu}_2 - \hat{\mu}_1^2$. This may happen for instance if the distribution has only one parameter such as in Examples 1 and 2. See the following example and the exercises.

Example 4. Consider a Poisson distribution with parameter λ. We have $E(X) = \lambda$. Hence, the method of moments estimator for λ is $\bar{X}$. But the variance of a Poisson is also λ (see Sect. 3.3). Hence, the m.m.e. for the variance is this case is also $\bar{X}$.

Example 5. Let X be distributed according to a Gamma distribution with parameters r and λ. We have (see Sect. 7.1)

$$E(X) = \mu_1 = \frac{r}{\lambda} \text{ and } E(X^2) = \mu_2 = \frac{r(r+1)}{\lambda^2}.$$

We need to solve these two equations in r and λ. We have

$$\mu_2 = \left(\frac{r}{\lambda}\right)^2 + \frac{r}{\lambda^2}.$$

We substitute $\frac{r}{\lambda}$ by μ_1 to get

$$\mu_2 = \mu_1^2 + \frac{\mu_1}{\lambda}.$$

Hence,

$$\lambda = \frac{\mu_1}{\mu_2 - \mu_1^2},$$

and

$$r = \lambda\mu_1 = \frac{\mu_1^2}{\mu_2 - \mu_1^2}.$$

This yields the following method of moments estimates

$$\hat{\lambda} = \frac{\hat{\mu}_1}{\hat{\mu}_2 - \hat{\mu}_1^2} \text{ and } \hat{r} = \frac{\hat{\mu}_1^2}{\hat{\mu}_2 - \hat{\mu}_1^2}.$$

9.1.2 The Method of Maximum Likelihood

I have two seemingly identical coins. One is biased and it shows tails with probability 3/4 and the other one is honest and shows tails with probability 1/2. I pick one of the coins at random and toss it three times. I get two tails and one heads. Is this the biased coin? I cannot answer this question with certainty but I can decide what this coin is more *likely* to be given the results of my experiment. If it is the biased coin the probability of getting two tails and one heads is (the number of tails in three tosses is a binomial random variable)

$$3\left(\frac{3}{4}\right)^2\left(\frac{1}{4}\right) = \frac{27}{64} \sim 0.42.$$

On the other hand if this is the honest coin then the probability of getting two tails and one heads is

$$3\left(\frac{1}{2}\right)^2\left(\frac{1}{2}\right) = \frac{3}{8} \sim 0.38.$$

Hence, based on our sample it is more likely that we picked the biased coin than the honest coin.

We now rephrase this question a little more formally. Let p be the probability of tails of the coin we picked. We know that p can be either 3/4 or 1/2. We toss the coin 3 times and this yields an i.i.d. sample. We get two tails and one heads in our sample. The probability (or likelihood) of such a result is $3p^2(1-p)$. The maximum likelihood occurs for $p = 3/4$. In this particular example the parameter can take only two possible values. In general, the parameter can take any value in a given interval. In this example the maximum likelihood estimator is $\hat{p} = 3/4$.

We next state an important definition. First, a word on notation. If X is a random variable we denote by x an observation of this random variable. For instance, if I roll a fair die then X can be defined as the face shown and $x = 5$ is a particular observation. It is important to use upper case letters for the random variables and lower case for observations.

The Likelihood Function

Let $X_1, X_2, \ldots, X_n$ be an i.i.d. sample of a given distribution with parameter θ. If the distribution is discrete then the likelihood function L of the sample is defined as

$$L(\theta; x_1, \ldots, x_n) = P(X_1 = x_1|\theta)P(X_2 = x_2|\theta)\ldots P(X_n = x_n|\theta).$$

If the distribution is continuous with density function f then the likelihood function L of the sample is defined as

$$L(\theta; x_1, \ldots, x_n) = f(x_1|\theta)\ldots f(x_n|\theta).$$

A word on notation. We use $f(x|\theta)$ and $P(X = x|\theta)$ to emphasize that these functions depend on θ. More precisely, given θ we know these functions. This is why we use the conditional notation "$|\theta$." We now summarize the maximum likelihood method.

The Maximum Likelihood Estimator

Let $X_1, X_2, \ldots, X_n$ be an i.i.d. sample of a given distribution with parameter θ. Let L be the likelihood function of this sample. Given the observations $x_1, x_2, \ldots, x_n$, the maximum likelihood estimator (or m.l.e.) of θ (if it exists!) is the value $\hat{\theta}$ that maximizes L as a function of θ. In other words, for all possible θ we must have

$$L(\theta; x_1, \ldots, x_n) \leq L(\hat{\theta}; x_1, \ldots, x_n).$$

Example 6. Consider the exponential distribution with parameter θ. In Example 1 we showed that the m.m.e. of θ is

$$\hat{\theta}_m = \frac{1}{\hat{\mu}_1} = \frac{1}{\bar{X}}.$$

We now find the m.l.e. of θ. Assume that we have an i.i.d. sample $X_1, X_2, \ldots, X_n$. The likelihood function is

$$\begin{aligned} L(\theta; x_1, \ldots, x_n) &= f(x_1|\theta) \ldots f(x_n|\theta) \\ &= \theta \exp(-\theta x_1) \ldots \theta \exp(-\theta x_n) = \theta^n \exp(-\theta(x_1 + \cdots + x_n)) \end{aligned}$$

provided $x_1, x_2, \ldots, x_n$ are all strictly positive. The function L is 0 otherwise. Given $x_1, \ldots, x_n$ we want to find the maximum of L as a function of θ only. L is clearly differentiable in θ for any θ in $(0, \infty)$. Hence, if it has a maximum then the derivative of L should be 0 at that point. Instead of looking for a maximum for L it is more convenient and equivalent to look for a maximum for $\ln L$. We have

$$\ln L(\theta) = n \ln \theta - \theta(x_1 + \cdots + x_n).$$

We take the derivative with respect to θ to get

$$\frac{\mathrm{d}}{\mathrm{d}\theta} \ln L(\theta) = \frac{n}{\theta} - (x_1 + \cdots + x_n).$$

Hence, this derivative is 0 if and only if $\theta = 1/\bar{x}$ (the lower case x is not a typo, we are averaging on the observations). The fact that the derivative is 0 at $\theta = 1/\bar{x}$ does not necessarily imply that there is a maximum there. To determine whether we have a maximum we know from Calculus that there are at least two tests that we may use: the first derivative test or the second derivative test. We compute the second derivative

$$\frac{\mathrm{d}^2}{\mathrm{d}\theta^2} \ln L(\theta) = \frac{-n}{\theta^2},$$

which is strictly negative for all θ and in particular for $\theta = 1/\bar{x}$. By the second derivative test $\ln L$ and therefore L has a maximum at $\theta = 1/\bar{x}$. Hence, we have found the m.l.e. of θ it is

$$\hat{\theta}_l(x_1, \ldots, x_n) = \frac{1}{\bar{x}}.$$

A more compact way to write this is

$$\hat{\theta}_l = \frac{1}{\bar{X}}.$$

The m.m.e. and the m.m.l. coincide in this case.

Example 7. Recall that the Poisson distribution with parameter λ has the following distribution

$$P(X = x|\lambda) = \exp(-\lambda)\frac{\lambda^x}{x!}$$

for all positive integers x. Hence, the likelihood function is

$$\begin{aligned} L(\lambda; x_1, \ldots, x_n) &= P(X_1 = x_1|\lambda)P(X_2 = x_2|\lambda)\ldots P(X_n = x_n|\lambda) \\ &= \exp(-n\lambda)\frac{\lambda^{x_1+\cdots+x_n}}{x_1!\ldots x_n!}. \end{aligned}$$

Therefore,

$$\ln L(\lambda) = -n\lambda + (x_1 + \cdots + x_n)\ln\lambda - \ln(x_1!\ldots x_n!).$$

This function is differentiable for all $\lambda > 0$. We have

$$\frac{\mathrm{d}}{\mathrm{d}\lambda}\ln L(\lambda) = -n + \frac{x_1 + \cdots + x_n}{\lambda}.$$

Hence, this derivative is 0 only for $\lambda = \bar{x}$. It is easy to check that the second derivative is always strictly negative and therefore the m.l.e. for λ is $\hat{\lambda} = \bar{X}$. Once again the m.m.e. and m.l.e. coincide.

The next example shows that the m.l.e. and m.m.e. need not coincide.

Example 8. Let X be a discrete random variable uniformly distributed over $\{1, 2, \ldots, \theta\}$. In Example 2 we have shown that the m.m.e. is $\hat{\theta}_m = 2\hat{\mu}_1 - 1 = 2\bar{X} - 1$. We now find the m.l.e. for θ that we denote by $\hat{\theta}_l$. Assume that we have the i.i.d. sample $X_1, X_2, \ldots, X_n$ with the corresponding observation $x_1, x_2, \ldots, x_n$. The x_i are natural numbers in $\{1, 2, \ldots, \theta\}$. By the definition the likelihood function L is

$$L(\theta; x_1, \ldots, x_n) = P(X_1 = x_1|\theta)P(X_2 = x_2|\theta)\ldots P(X_n = x_n|\theta) = \frac{1}{\theta^n}.$$

Given $x_1, x_2, \ldots, x_n$ we need to decide whether the function L (as a function of θ only) has a maximum and if so where? In order to do so we rewrite L as a function of θ only. We have

$$L(\theta) = \frac{1}{\theta^n} \text{ if } \theta \geq x_1, \theta \geq x_2, \ldots, \theta \geq x_n.$$

If the condition $\theta \geq x_1, \theta \geq x_2, \ldots, \theta \geq x_n$ fails it means that one of the observation is strictly larger than θ. That cannot happen and therefore if this

condition fails we have $L(\theta) = 0$. It is easy to see that the condition $\theta \geq x_1, \theta \geq x_2, \ldots, \theta \geq x_n$ is actually equivalent to $\theta \geq \max(x_1, \ldots, x_n)$. Let

$$x_{(n)} = \max(x_1, \ldots, x_n).$$

We have

$$L(\theta) = \frac{1}{\theta^n} \text{ if } \theta \geq x_{(n)}$$

and $L(\theta) = 0$ if $\theta < x_{(n)}$. Note that

$$\ln L(\theta) = -n \ln \theta$$

and the derivative is $-n/\theta$ which is never 0. This expression is negative for all positive θ. Hence, $\ln L$ is decreasing for all θ in the domain of $\ln L$ which is $(x_{(n)}, \infty)$. Therefore the maximum value for $\ln L$ is attained when θ is minimum. This minimum value of θ is $x_{(n)}$. Any θ less than that results in $L = 0$. Hence, we have found the m.l.e. It is

$$\hat{\theta}_l = X_{(n)} = \max(X_1, \ldots, X_n).$$

Observe that it is quite different from the m.m.e. Taking the numerical values from Example 2 we get $\hat{\theta}_l = 6$ while the m.m.e. was shown to be Sect. 4.6.

Example 9. Consider an i.i.d. sample of a normal distribution with mean μ and variance σ^2. In this case we have two unknown parameters. Recall that the density is

$$f(x|\mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right).$$

Hence, the likelihood function is

$$L(\mu, \sigma; x_1, \ldots, x_n) = \frac{1}{\sqrt{2\pi}^n \sigma^n} \exp\left(-\frac{(x_1-\mu)^2 + \cdots + (x_n-\mu)^2}{2\sigma^2}\right).$$

Given the observations $x_1, \ldots, x_n$ we have

$$\ln L(\mu, \sigma) = -n \ln \sqrt{2\pi} - n \ln \sigma - \frac{(x_1-\mu)^2 + \cdots + (x_n-\mu)^2}{2\sigma^2}.$$

The partial derivatives of $\ln L$ are

$$\frac{\mathrm{d}}{\mathrm{d}\mu} \ln L = \frac{1}{\sigma^2} \sum_{i=1}^{n} (x_i - \mu)$$

and

$$\frac{\mathrm{d}}{\mathrm{d}\sigma}\ln L = -\frac{n}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^{n}(x_i - \mu)^2.$$

We now solve the system of equations

$$\frac{\mathrm{d}}{\mathrm{d}\mu}\ln L = 0$$
$$\frac{\mathrm{d}}{\mathrm{d}\sigma}\ln L = 0$$

The first equation gives the solution

$$\hat{\mu} = \bar{x}.$$

Multiplying the second equation by σ^3 yields

$$-n\sigma^2 + \sum_{i=1}^{n}(x_i - \mu)^2 = 0.$$

Substituting $\bar{x}$ for μ in the second equation gives

$$\hat{\sigma} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

We need now to check that $\ln L$ is maximum for $(\hat{\mu}, \hat{\sigma})$. We use a second derivative test for a two variables function. From Calculus we know that if at the point $(\hat{\mu}, \hat{\sigma})$ we have

$$\frac{\mathrm{d}^2}{\mathrm{d}\mu^2}\ln L < 0 \text{ and } \frac{\mathrm{d}^2}{\mathrm{d}\mu^2}\ln L \frac{\mathrm{d}^2}{\mathrm{d}\sigma^2}\ln L - \left(\frac{\mathrm{d}^2}{\mathrm{d}\mu\mathrm{d}\sigma}\ln L\right)^2 > 0$$

then $\ln L$ has a maximum at $(\hat{\mu}, \hat{\sigma})$. We now compute the second partial derivatives. We have

$$\frac{\mathrm{d}^2}{\mathrm{d}\mu^2}\ln L(\mu, \sigma) = -\frac{n}{\sigma^2}$$

which is always negative. We have

$$\frac{\mathrm{d}^2}{\mathrm{d}\sigma^2}\ln L = \frac{n}{\sigma^2} - \frac{3}{\sigma^4}\sum_{i=1}^{n}(x_i - \mu)^2.$$

We compute the value of the last function at the point $(\hat{\mu}, \hat{\sigma})$ and we get

$$\frac{\mathrm{d}^2}{\mathrm{d}\sigma^2} \ln L(\hat{\mu}, \hat{\sigma}) = \frac{n}{\hat{\sigma}^2} - \frac{3}{\hat{\sigma}^4} n\hat{\sigma}^2 = -\frac{2n}{\hat{\sigma}^2}.$$

We also need

$$\frac{\mathrm{d}^2}{\mathrm{d}\mu \mathrm{d}\sigma} \ln L(\mu, \sigma) = -\frac{2}{\sigma^3} \sum_{i=1}^{n} (x_i - \mu).$$

To compute the value of the preceding function at the point $(\hat{\mu}, \hat{\sigma})$ we substitute μ by $\bar{x}$. Since $\sum_{i=1}^{n} (x_i - \bar{x}) = 0$ (why?) we have

$$\frac{\mathrm{d}^2}{\mathrm{d}\mu \mathrm{d}\sigma} \ln L(\hat{\mu}, \hat{\sigma}) = 0.$$

Hence, at $(\hat{\mu}, \hat{\sigma})$ we get

$$\frac{\mathrm{d}^2}{\mathrm{d}\mu^2} \ln L \frac{\mathrm{d}^2}{\mathrm{d}\sigma^2} \ln L - \left(\frac{\mathrm{d}^2}{\mathrm{d}\mu \mathrm{d}\sigma} \ln L \right)^2 = \left(-\frac{n}{\hat{\sigma}^2} \right) \left(-\frac{2n}{\hat{\sigma}^2} \right) > 0.$$

This concludes the proof that $\ln L$ has a maximum at $(\hat{\mu}, \hat{\sigma})$. This is the m.l.e. of (μ, σ). Note that it is the same as the m.m.e.

Exercises 9.1

1. (a) Show that the estimator for the variance $\hat{\theta} = \hat{\mu}_2 - \hat{\mu}_1^2$ found in Example 3 can also be written as

$$\hat{\theta} = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})^2.$$

(b) Show that S^2 is different from $\hat{\theta}$.

2. Find the method of moments estimator of the variance for exponential distribution with parameter θ.

3. Let X be a discrete random variable uniformly distributed over $\{1, 2, \ldots, \theta\}$. Find the method of moments estimator of the variance of X.

4. Assume that an i.i.d. sample of a Gamma distribution with parameters r and λ is 5.95, 7.84, 3.24, 4.71, 5.12, 10.09, 4.18, 3.92, 6.36, 2.51. Use Example 4 to get numerical estimates for r and λ.

5. Let U be a continuous random variable uniformly distributed on $[-\theta, \theta]$. Find the method of moments estimator for θ.

6. Let U be a continuous random variable uniformly distributed on $[0, \theta]$.

(a) Find the method of moments estimator for θ.
(b) Compute the m.m.e. given the following observations 0.1158, 0.7057, 1.6263, 0.0197, 0.2778, 0.4055, 0.3974, 1.2076, 0.5444, 0.3976.

7. Show that the condition $\theta \geq x_1, \theta \geq x_2, \ldots, \theta \geq x_n$ is equivalent to $\theta \geq \max(x_1, \ldots, x_n)$.

8. Let f be a strictly positive function. Show that f is maximum at a if and only if $\ln f$ is maximum at a.

9. Show that $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$.

10. Recall that a Bernoulli random variable X has the following distribution. $P(X = 1|p) = p$ and $P(X = 0|p) = 1 - p$.

(a) Find the m.m.e. of p.
(b) Show that the distribution of X can be written as

$$P(X = x|p) = p^x(1 - p)^{1-x} \text{ for } x = 0 \text{ or } 1.$$

(c) Use (b) to find the m.l.e. of p.

11. The Pareto distribution has density

$$f(x|a, \theta) = \theta a^{\theta} x^{-\theta-1} \text{ for } x \geq a.$$

Take $a = 1$ and assume that θ is an unknown parameter somewhere in $(1, \infty)$.

(a) Check that f is indeed a probability density.
(b) Find the m.m.e. of θ.
(c) Find the m.l.e. of θ.

12. Consider the continuous uniform distribution in $[0, \theta]$.

(a) Find the m.l.e. of θ.
(b) Use the observations of Exercise 6 to compute the m.l.e.

13. Consider the continuous uniform density on $[\theta - 1/2, \theta + 1/2]$.

(a) Given observations $x_1 \ldots x_n$ show that the likelihood function is

$$L(\theta) = 1 \text{ if } \max(x_1, \ldots, x_n) - 1/2 \leq \theta \leq \min(x_1, \ldots, x_n) + 1/2$$

and 0 otherwise.
(b) Show that $\theta_l = \frac{1}{2}(\min(x_1, \ldots, x_n) + \max(x_1, \ldots, x_n))$ is an m.l.e.

(c) Show that there are infinitely many m.l.e.'s of θ in this case.
(d) Find the m.m.e. of θ.

14. Consider a geometric distribution with parameter p. That is,

$$P(X = x|p) = (1-p)^{x-1}p \text{ for } x = 1, 2, \ldots$$

(a) Find the m.m.e. of p.
(b) Find the m.l.e. of p.

15. Let $x_1, x_2, \ldots, x_n$ be n real numbers. Let f be the function

$$f(\theta) = \sum_{i=1}^{n}(x_i - \theta)^2.$$

Show that f attains its minimum at

$$\theta = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

16. Let $x_1, x_2, \ldots, x_{2n+1}$ be $2n+1$ real numbers. Define the function f by

$$f(\theta) = \sum_{i=1}^{2n+1} |x_i - \theta|.$$

Let m be the median of the sample $x_1, x_2, \ldots, x_{2n+1}$. That is, there are as many observations below m as there are above m. The median is unique because our sample size is odd. We will show in this exercise that f is minimum for $\theta = m$. Let $a > 0$.

(a) Show that if $x_i \leq m$ then $|x_i - (m+a)| = |x_i - m| + a$; if $m < x_i < m + a$ then $|x_i - (m+a)| = -|x_i - m| + a$; if $x_i \geq m + a$ then $|x_i - (m+a)| = |x_i - m| - a$.
(b) Let A be the set of indices i such that $x_i \leq m$, B be the set of indices i such that $m < x_i < m + a$ and C be the set of indices i such that $x_i \geq m + a$. Show that

$$f(m+a) = \sum_{i \in A}(|x_i - m| + a) + \sum_{i \in B}(-|x_i - m| + a) + \sum_{i \in C}(|x_i - m| - a).$$

(c) Use (b) to get

$$f(m+a) = f(m) + a(|A| + |B| - |C|) - 2\sum_{i \in B}|x_i - m|,$$

where $|A|$ is the number of elements in A.

(d) Using that if i is in B then $|x_i - m| < a$ we get

$$f(m + a) \geq f(m) + a(|A| - |B| - |C|).$$

(e) Show that $|A| \geq |B| + |C|$ and conclude that

$$f(m + a) \geq f(m).$$

(f) Redo the steps above to get $f(m-a) \geq f(m)$ and conclude that f is minimum at m.

17. Consider the Laplace distribution with density

$$f(x|\theta) = \frac{1}{2}\exp(-|x - \theta|).$$

(a) For an odd sample size find the m.l.e. of θ (use Exercise 16).
(b) Find the m.m.e. of θ.
(c) Use the following observations to compute the m.m.e. and the m.l.e. of θ: 4.1277, −0.2371, 3.4503, 2.7242, 4.0894, 3.0056, 3.5429, 2.8466, 2.5044, 2.0306, −0.1741.

9.2 Comparing Estimators

In order to compare two estimators of the parameter θ we will measure how close each estimator is to the parameter we want to estimate. If estimator 1 is closer to the parameter than estimator 2 then estimator 1 is said to be better than estimator 2. Hence, "better" depends heavily on what the measure of closeness is. There are many possible choices to measure closeness but by far the most used is the mean square error (or quadratic mean distance) that we now define. Assume that $\hat{\theta}$ is an estimator for θ. The mean square error between θ and $\hat{\theta}$ is

$$\mathrm{d}(\hat{\theta}, \theta) = E((\hat{\theta} - \theta)^2).$$

Note that $\hat{\theta}$ is a random variable (it will change from sample to sample) so $(\hat{\theta} - \theta)^2$ is also a random variable. To measure how close $\hat{\theta}$ is to θ we would like something nonrandom. This is why we take the expectation of $(\hat{\theta} - \theta)^2$ to define $\mathrm{d}(\hat{\theta}, \theta)$. Another natural choice for $\mathrm{d}(\hat{\theta}, \theta)$ is $E(|\hat{\theta} - \theta|)$. The problem with this choice is that it is a lot less convenient mathematically. This is why we will stick with the mean square error.

Example 1. Let μ and σ^2 be the mean and variance of a given distribution. Assume we use $\bar{X}$ to estimate μ. Then (see Sect. 4.1)

$$\mathrm{d}(\bar{X}, \mu) = E((\bar{X} - \mu)^2) = \mathrm{Var}(\bar{X}) = \frac{\sigma^2}{n}.$$

If we use X_1 (the first observation) to estimate μ then

$$\mathrm{d}(X_1, \mu) = E((X_1 - \mu)^2) = \mathrm{Var}(X_1) = \sigma^2.$$

Hence, for all $n \geq 2$, $\bar{X}$ is a better estimator of μ than X_1 is. This is not surprising given that $\bar{X}$ uses a lot more information about the sample than X_1 does.

The following property of mean square errors is often useful.

The Mean Square Error Formula

Assume that $\hat{\theta}$ is an estimator of the parameter θ. The corresponding mean square error is defined as

$$\mathrm{d}(\hat{\theta}, \theta) = E((\hat{\theta} - \theta)^2).$$

The following formula is useful in computing the mean square error:

$$E((\hat{\theta} - \theta)^2) = \mathrm{Var}(\hat{\theta}) + (E(\hat{\theta}) - \theta)^2.$$

In order to prove this formula it is important to remember that the expectation is linear. That is, $E(aX) = aE(X)$ and $E(X + Y) = E(X) + E(Y)$ where X and Y are random variables and a is a real number. It is also important to realize that $\hat{\theta}$ is a random variable while $E(\hat{\theta})$ and θ are numbers. Moreover, the expected value of a number is the number itself. We have

$$\begin{aligned}(\hat{\theta} - \theta)^2 &= (\hat{\theta} - E(\hat{\theta}) + E(\hat{\theta}) - \theta)^2 \\ &= (\hat{\theta} - E(\hat{\theta}))^2 + 2(\hat{\theta} - E(\hat{\theta}))(E(\hat{\theta}) - \theta) + (E(\hat{\theta}) - \theta)^2.\end{aligned}$$

Note that

$$E((\hat{\theta} - E(\hat{\theta}))^2) = \mathrm{Var}(\hat{\theta})$$

and since $(E(\hat{\theta}) - \theta)^2$ is a number it is equal to its expectation. Hence, the formula is proved provided the expectation of the cross product above is 0. We now show that.

$$(\hat{\theta} - E(\hat{\theta}))(E(\hat{\theta}) - \theta) = \hat{\theta}E(\hat{\theta}) - \hat{\theta}\theta - E(\hat{\theta})^2 + E(\hat{\theta})\theta.$$

We now take expectations across the equality above to get

$$E((\hat{\theta} - E(\hat{\theta}))(E(\hat{\theta}) - \theta)) = E(\hat{\theta}E(\hat{\theta})) - E(\hat{\theta}\theta) - E(\hat{\theta})^2 + E(\hat{\theta})\theta.$$

We have

$$E(\hat{\theta}E(\hat{\theta})) = E(\hat{\theta})E(\hat{\theta}) = E(\hat{\theta})^2$$

and $E(\hat{\theta}\theta) = \theta E(\hat{\theta})$. Therefore,

$$E((\hat{\theta} - E(\hat{\theta}))(E(\hat{\theta}) - \theta)) = 0$$

and the formula is proved.

In applying the mean square error formula we will sometimes need to compute the variance of the sample average. We now recall this useful formula.

The Expectation and Variance of the Sample Average

Assume that $X_1, \ldots, X_n$ is an i.i.d. sample of a distribution with mean μ and variance σ^2. Then,

$$E(\bar{X}) = \mu \text{ and } \mathrm{Var}(\bar{X}) = \frac{\sigma^2}{n}.$$

Example 2. Consider a random variable X having the uniform distribution on $[0, \theta]$. The expected value of X is $\theta/2$ (see Sect. 2.3). Hence, given an i.i.d. sample the m.m.e. of θ is $\hat{\theta}_m = 2\bar{X}$. Therefore, $E(\hat{\theta}_m) = \theta$. In this case, the mean square error formula for $\hat{\theta}_m$ is reduced to

$$\mathrm{d}(\hat{\theta}_m, \theta) = \mathrm{Var}(\hat{\theta}_m) = 4\mathrm{Var}(\bar{X}),$$

where we are using that the Variance is a quadratic operator (see Sect. 2.4). Since $\mathrm{Var}(X_1) = \theta^2/12$ (see Sect. 2.4) we have

$$\mathrm{d}(\hat{\theta}_m, \theta) = \frac{\theta^2}{3n}.$$

There is another rather natural estimator of θ. Since the range of possible values is $[0, \theta]$ another candidate to estimate θ is the largest observation in the sample. Moreover, it turns out that the largest observation in the sample is the maximum likelihood estimator of θ (for a very similar computation see Example 8 in Sect. 9.1). For a sample $X_1, \ldots, X_n$ we denote by $X_{(n)}$ the largest observation. For instance, if we have the three observations $X_1 = 1.1$, $X_2 = 0.7$, $X_3 = 0.9$ then $X_{(3)} = 1.1$. We will now compute the mean square error for $X_{(n)}$ in order to decide whether it is a better estimator than $2\bar{X}$. We first compute the distribution function $F_{(n)}$ of $X_{(n)}$. For x in $[0, \theta]$ we have

$$F_{(n)}(x) = P(X_{(n)} \leq x) = P(X_1 \leq x, \ldots, X_n \leq x),$$

where we are using the fact that the largest observation is less than x if and only if every observation is less than x. We now use that the observations are independent to get

$$F_{(n)}(x) = P(X_1 \leq x) \dots P(X_n \leq x) = F(x)^n,$$

where we used also that the observations are identically distributed. We denote their common distribution function by F. We have for x in $[0, \theta]$

$$F(x) = P(X_1 \leq x) = \int_0^x \frac{1}{\theta} \mathrm{d}x = \frac{x}{\theta}.$$

Hence,

$$F_{(n)}(x) = \left(\frac{x}{\theta}\right)^n.$$

By taking the derivative (with respect to x) of $F_{(n)}$ we get the density function of $X_{(n)}$ denoted by $f_{(n)}$. For x in $[0, \theta]$

$$f_{(n)}(x) = \frac{n}{\theta^n} x^{n-1}.$$

We now use this density function to compute the mean square error of $X_{(n)}$. Starting with the expected value we have

$$E(X_{(n)}) = \int_0^\theta x \frac{n}{\theta^n} x^{n-1} \mathrm{d}x = \frac{n}{n+1}\theta.$$

For the second moment,

$$E(X_{(n)}^2) = \int_0^\theta x^2 \frac{n}{\theta^n} x^{n-1} \mathrm{d}x = \frac{n}{n+2}\theta^2.$$

Hence, the variance of $X_{(n)}$ is

$$\mathrm{Var}(X_{(n)}) = \frac{n}{n+2}\theta^2 - \left(\frac{n}{n+1}\theta\right)^2 = \frac{n}{(n+2)(n+1)^2}\theta^2.$$

According to the mean square error formula we also need to compute

$$(E(X_{(n)}) - \theta)^2 = \left(\frac{n}{n+1}\theta - \theta\right)^2 = \frac{1}{(n+1)^2}\theta^2.$$

Finally,

$$\begin{aligned} \mathrm{d}(X_{(n)}, \theta) &= \mathrm{Var}(X_{(n)}) + (E(X_{(n)}) - \theta)^2 \\ &= \frac{n}{(n+2)(n+1)^2}\theta^2 + \frac{1}{(n+1)^2}\theta^2 = \frac{2}{(n+2)(n+1)}\theta^2 \end{aligned}$$

It is easy to check that for all $n \geq 3$ we have

$$\frac{2}{(n+2)(n+1)} < \frac{1}{3n}.$$

That is, for $n \geq 3$ and every θ

$$\mathrm{d}(X_{(n)}, \theta) < \mathrm{d}(2\bar{X}, \theta).$$

Therefore, $X_{(n)}$ is a better estimator than $2\bar{X}$ for all $n \geq 3$.

Now that we know how to compare two estimators a natural question is to search for the best estimator. It turns out that there is no such thing as a "best" estimator unless we restrict our search. To see why assume that $\hat{\theta}_1$ is the "best" estimator of θ. This means that for any other estimator $\hat{\theta}_2$ we must have

$$\mathrm{d}(\hat{\theta}_1, \theta) \leq \mathrm{d}(\hat{\theta}_2, \theta)$$

for all possible θ. Let I be the set of all possible θ. Let θ_0 be a fixed value in I and let $\hat{\theta}_2 = \theta_0$. This is a terrible estimator: it does not use the sample information at all. It is always the constant θ_0. However,

$$\mathrm{d}(\hat{\theta}_2, \theta_0) = 0.$$

That is, it is a perfect estimator when $\theta = \theta_0$. So if we want the inequality $\mathrm{d}(\hat{\theta}_1, \theta) \leq \mathrm{d}(\hat{\theta}_2, \theta)$ to hold for all θ we need to have $\mathrm{d}(\hat{\theta}_1, \theta_0) = 0$ as well. But this implies that $\hat{\theta}_1 = \theta_0$ when $\theta = \theta_0$. This must be true for all θ_0 in I. That is $\hat{\theta}_1 = \theta$ all θ in I. In other words, the best estimator would have to be perfect. It would give us the true value of the parameter every time! Since an estimator is a function of the sample there is no such thing as a perfect estimator. Hence, there is no "best" estimator $\hat{\theta}_1$.

What we can do, however, is look for the best estimator under some constraint. A natural class of estimators that turns out to be very fruitful mathematically (under the mean square error criterion we are using) is the class of unbiased estimators that we now define.

Unbiased Estimators

The estimator $\hat{\theta}$ is said to be an unbiased estimator of θ if $E(\hat{\theta}) = \theta$.

Example 3. We go back to the continuous uniform distribution on $[0, \theta]$ of example 2. Recall that the m.m.e. is $\hat{\theta}_m = 2\bar{X}$ and that $E(\hat{\theta}_m) = \theta$. Hence, $\hat{\theta}_m$ is an unbiased estimator of θ. On the other hand the m.l.e. estimator is $X_{(n)}$ the largest observation in the sample. We computed $E(X_{(n)}) = \frac{n}{n+1}\theta$. Therefore, $X_{(n)}$ is a biased estimator of θ. However, we proved in Example 2 that $X_{(n)}$ is a better estimator than $\hat{\theta}_m$. So by looking only for unbiased estimators we may very well miss better estimators.

Observe that for unbiased estimators the mean square error is really only the variance of the estimator.

The Mean Square Error for an Unbiased Estimator

Assume that $\hat{\theta}$ is an *unbiased* estimator of the parameter θ. Then,

$$\mathrm{d}(\hat{\theta}, \theta) = E((\hat{\theta} - \theta)^2) = \mathrm{Var}(\hat{\theta}).$$

It turns out that unbiased estimation is particularly well understood for a certain class of probability distributions. We will define this class. We first need a definition.

Support of a Probability Distribution

The support of a probability distribution is the set of x's such that $f(x|\theta)$ (continuous case) or of the probability distribution $P(X = x|\theta)$ (discrete case) is strictly positive.

For instance, a Bernoulli distribution with parameter p is

$$P(X = x|p) = p^x(1-p)^{1-x}$$

for $x = 0$ and $x = 1$. For any other x, $P(X = x|p) = 0$. The support of a Bernoulli distribution is therefore $\{0, 1\}$.

Exponential Family of Probability Distributions

A family of probability distributions with parameter θ is said to be an exponential family of probability distributions if the logarithm of the density

function $f(x|\theta)$ (continuous case) or of the probability distribution $P(X = x|\theta)$ (discrete case) can be written as

$$a(\theta)t(x) + b(\theta) + r(x).$$

Moreover, the support of the distribution must be the same for each θ. In addition, the set I of all possible values for θ must be an open interval of real numbers and the functions a and b must have continuous second derivatives.

The term "exponential family" is widely used but is unfortunate. If $y > 0$ then y is the exponential of $\log y$: a positive number is always the exponential of its logarithm. The point here is that $f(x|\theta)$ is the exponential of $a(\theta)t(x)+b(\theta)+r(x)$ for some functions a, b, r and t. This is not as general one may think at first glance. For instance, the function x^θ cannot be written as $a(\theta)t(x) + b(\theta) + r(x)$. Hence, a family of densities of the type $C(\theta)\exp(-x^\theta)$ is not an exponential family! However, most of the models we have encountered so far can be written as $\exp(a(\theta)t(x) + b(\theta) + r(x))$. What will prevent some of them from being an exponential family is the additional condition that the support of the distribution does not depend on θ. See for example the uniform density below.

Example 4. Consider a family of Bernoulli random variable with parameter p. That is,

$$P(X = x|p) = p^x(1-p)^{1-x},$$

where $x = 0$ or 1. We have

$$\ln(P(X = x|p)) = x\log p + (1-x)\log(1-p) = x\log\left(\frac{p}{1-p}\right) + \log(1-p),$$

for $x = 0$ or 1. We can set $a(p) = \log(\frac{p}{1-p})$, $b(p) = \log(1-p))$, $t(x) = x$ and $r(x) = 0$. The support of the distribution is $\{0, 1\}$ and does not depend on p. The possible values for p are in $I = (0, 1)$, an open interval, and the functions a and b are infinitely differentiable on I. This shows the collection of Bernoulli distributions with parameter p is an exponential family.

Example 5. Consider the family of normal distributions with mean μ and standard deviation 1. We have

$$f(x|\mu) = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{(x-\mu)^2}{2}\right) \text{ for } x \text{ in } (-\infty, +\infty).$$

Thus,

$$\ln(f(x|\mu)) = \mu x - \mu^2/2 - \log(\sqrt{2\pi}) - x^2/2.$$

We set $a(\mu) = \mu$, $b(\mu) = \mu^2/2$, $t(x) = x$, $r(x) = -\log(\sqrt{2\pi}) - x^2/2$. The support of the distribution is the whole real line and therefore does not depend on μ. The functions a and b are clearly infinitely differentiable on the whole real line I (μ can be any real number). Therefore, the collection of normal distributions with mean μ and standard deviation 1 is an exponential family.

Example 6. Consider X the family of uniform distribution on $[0, \theta]$. We have

$$f(x|\theta) = \frac{1}{\theta} \text{ for } x \text{ in } [0, \theta]$$

and $f(x|\theta) = 0$ for all other x. In this case the support of the distribution depends on θ (it is $[0, \theta]$). Hence, the family of uniform distributions is NOT an exponential family.

Exercises 9.2

1. Find two random variables X and Y such that

$$E|X - 1| < E|Y - 1|$$

and

$$E((X - 1)^2) > E((Y - 1)^2).$$

Hence, X is closer to 1 under the first distance and Y is closer to 1 under the second distance.

2. Consider the continuous uniform distribution on $[0, \theta]$ of Example 2. We computed $E(X_{(n)}) = \frac{n}{n+1}\theta$. Let $T_n = \frac{n+1}{n}X_{(n)}$

(a) Show that T_n is an unbiased estimator of θ.
(b) Is T_n a better estimator than $X_{(n)}$?

3. Consider a family of geometric random variables with parameter p. That is,

$$P(X = x|p) = p(1 - p)^{x-1},$$

where $x \geq 1$ is a positive integer. Show that this is an exponential family.

4. Consider the family of normal distributions with mean μ and standard deviation σ. We have

$$f(x|\mu) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x - \mu)^2}{2\sigma^2}\right).$$

(a) Assume that σ is known. Show that this is an exponential family with parameter μ.
(b) Assume that μ is known. Show that this is an exponential family with parameter σ.

5. We know that (see Example 2) that $\hat{\theta}_m = 2\bar{X}$ is an unbiased estimator of θ with variance $\frac{\theta^2}{3n}$. The purpose of this exercise is to show that even if we limit ourselves to estimators of the form $a\bar{X}$ then $\hat{\theta}_m$ is not the best choice. That is, there are better choices than $a = 2$.

(a) Show that

$$\mathrm{d}(a\bar{X}, \theta) = a^2 \frac{\theta^2}{12n} + \theta^2 \left(\frac{a}{2} - 1\right)^2.$$

(b) Show that $\mathrm{d}(a\bar{X}, \theta)$ is minimum for $a_0 = \frac{6n}{3n+1}$. The estimator $a_0\bar{X}$ is the best in the class of estimators $a\bar{X}$.
(c) Show that the estimator found in (b) is biased.

6. Consider the family of normal distributions with mean 0 and standard deviation σ. Our parameter is $\theta = \sigma^2$.

(a) Show that

$$\hat{\theta} = \frac{1}{n} \sum_{i=1}^{n} X_i^2$$

is an unbiased estimator of θ.
(b) Show that

$$\mathrm{Var}(\hat{\theta}) = 2\frac{\sigma^4}{n}.$$

(You may want to use that the fourth moment of a standard normal is 3).

7. Consider a family of exponential distributions with expectation α. That is,

$$f(x|\alpha) = \frac{1}{\alpha} \mathrm{e}^{\frac{-1}{\alpha}x} \text{ for } x > 0.$$

(a) Show that $\bar{X}$ is an unbiased estimator of α.
(b) Show that $\frac{n}{n+1}\bar{X}$ is the best estimator of α in the class of all estimators $a\bar{X}$.

8. Consider the family of distributions with density

$$f(x|\theta) = \frac{1}{2\theta} \exp\left(-\frac{|x|}{\theta}\right) \text{ for } x \in (-\infty, +\infty).$$

(a) Is this an exponential family of distributions?
(b) Compute $E(|X|)$.
(c) Find an unbiased estimator for θ.

9. Consider the family of Pareto distributions

$$f(x|\theta) = \frac{\theta}{(1+x)^{\theta+1}} \text{ for } x > 0,$$

where θ is in $(1, \infty)$.

(a) Is this an exponential family of distributions?
(b) Compute $E(X)$.
(c) $\bar{X}$ is an unbiased estimator of what parameter?

10. In this exercise we give an example where all the estimators are biased. Assume we have a sample of size 1 of a binomial distribution with parameters $(10, p)$. We would like an unbiased estimator for $\log p$. By contradiction assume that there is such an estimator $\hat{p}$. Given the observation x (which is an integer between 0 and 10), $\hat{p}$ must be a function of x. Let g be this function. That is, $\hat{p} = g(x)$.

(a) Show that

$$E(\hat{p}) = \sum_{x=0}^{10} g(x) \binom{10}{x} p^x (1-p)^{10-x}.$$

(b) Explain why $E(\hat{p})$ cannot be $\log p$ whatever the choice of g is. Hence, any estimator of $\log p$ is necessarily biased.

9.3 Sufficient Statistics

So far we have seen methods for finding and comparing estimators. In this section we will see how it is sometimes possible to improve an estimator or even to find a minimum variance unbiased estimator (m.v.u.e.). The central idea is sufficiency.

Sufficient Statistics

Let $X_1, X_2, \ldots, X_n$ be an i.i.d. sample of a distribution with parameter θ. A statistic S (i.e., a function of $(X_1, X_2, \ldots, X_n)$) is said to be sufficient for θ if the conditional distribution of $X_1, X_2, \ldots, X_n$ given $S = s$ does not depend on θ.

The idea behind this definition is that all the information about θ is contained in the sufficient statistic S. There is no need to look at the whole sample $(X_1, X_2, \ldots, X_n)$, it is *sufficient* to only look at S.

Example 1. Consider $X_1, X_2, \ldots, X_n$ an i.i.d. sample of the Bernoulli distribution with $P(X_i = 1|\theta) = \theta$. Is X_1 a sufficient statistic?

For $x_1, \ldots, x_n$ a sequence in $\{0, 1\}$ we compute

$$P(X_1 = x_1 \ldots X_n = x_n | X_1 = x_1) = \frac{P(X_1 = x_1 \ldots X_n = x_n)}{P(X_1 = x_1)}.$$

In fact the notation should be $P(X_1 = x_1 \ldots X_n = x_n|\theta; X_1 = x_1)$ instead of $P(X_1 = x_1 \ldots X_n = x_n|X_1 = x_1)$. To simplify the notation we omit θ. Recall that $P(X_i = x_i) = \theta^{x_i}(1-\theta)^{1-x_i}$. Therefore,

$$P(X_1 = x_1 \ldots X_n = x_n | X_1 = x_1) = \frac{\theta^{x_1}(1-\theta)^{1-x_1} \ldots \theta^{x_n}(1-\theta)^{1-x_n}}{\theta^{x_1}(1-\theta)^{1-x_1}}.$$

Hence, if we let $t = \sum_{i=2}^{n} x_i$

$$P(X_1 = x_1 \ldots X_n = x_n | X_1 = x_1) = \theta^t (1-\theta)^{n-t}.$$

Clearly the conditional distribution of $X_1, X_2, \ldots, X_n$ given $X_1 = x_1$ depends on θ. Therefore, X_1 is not a sufficient statistic.

Let $S = \sum_{i=1}^{n} X_i$, we are going to show that S is a sufficient statistic. Since S is a sum of i.i.d. Bernoulli random variables it is a binomial random variable and we have for $s = 0, 1, \ldots, n$

$$P(S = s) = \binom{n}{s} \theta^s (1-\theta)^{n-s}.$$

Note that if $s = \sum_{i=1}^{n} x_i$ then $\{S = s\}$ is a subset of $\{X_1 = x_1 \ldots X_n = x_n\}$. Therefore,

$$P(X_1 = x_1 \ldots X_n = x_n | S = s) = \frac{P(X_1 = x_1 \ldots X_n = x_n)}{P(S = s)}$$

and

$$P(X_1 = x_1 \ldots X_n = x_n | S = s) = \frac{\theta^s (1-\theta)^{n-s}}{\binom{n}{s} \theta^s (1-\theta)^{n-s}} = \frac{1}{\binom{n}{s}}.$$

This time all the θ's cancel and the conditional distribution of $X_1, X_2, \ldots, X_n$ given $S = s$ does not depend on θ. The statistic S is indeed sufficient. What is remarkable here is that a single number S summarizes all the information regarding θ which is contained in the (usually) very long vector $(X_1, X_2, \ldots, X_n)$.

The definition of sufficiency does not give a method to find a sufficient statistic. The next result does exactly that.

A Factorization Criterion for Sufficiency

Let $X_1, X_2, \ldots, X_n$ be an i.i.d. sample of a distribution with parameter θ. A statistic S is sufficient if and only if the likelihood function L can be factored as follows

$$L(\theta; x_1, \ldots, x_n) = g(\theta; S(x_1, \ldots, x_n))h(x_1, \ldots, x_n),$$

where g and h are positive functions.

In words, the statistic S is sufficient if and only if L can be factored in a function of S and θ and a function that does not depend on θ.

Example 2. Consider $X_1, X_2, \ldots, X_n$ an i.i.d. sample of the exponential distribution with parameter θ. The likelihood function is

$$L(\theta; x_1, \ldots, x_n) = \theta^n \exp(-\theta(x_1 + x_2 + \cdots + x_n)),$$

for positive reals $x_1, \ldots, x_n$. Let $S(x_1, \ldots, x_n) = x_1 + \cdots + x_n$ and

$$g(\theta; s) = \theta^n \exp(-\theta s).$$

Set $h(x_1, \ldots, x_n) = 1$ if all x_i are positive and 0 otherwise. Note that h does not depend on θ. We have

$$L(\theta; x_1, \ldots, x_n) = g(\theta; S(x_1, \ldots, x_n))h(x_1, \ldots, x_n).$$

This proves that $S = \sum_{i=1}^{n} X_i$ is a sufficient statistic.

Example 3. Consider $X_1, X_2, \ldots, X_n$ an i.i.d. sample of the uniform distribution in $[0, \theta]$. The likelihood function is

$$L(\theta; x_1, \ldots, x_n) = \frac{1}{\theta^n} \text{ if all } x_i \in [0, \theta]$$

and $L(\theta; x_1, \ldots, x_n) = 0$ otherwise. Let $x_{(n)}$ be the largest of all x_i's. We can rewrite the function L as

$$L(\theta; x_1, \ldots, x_n) = \frac{1}{\theta^n} \text{ if } 0 \leq x_{(n)} \leq \theta,$$

and $L = 0$ otherwise. Let $S(x_1, \ldots, x_n) = x_{(n)}$ and

$$g(\theta; s) = \frac{1}{\theta^n} \text{ if } s \leq \theta.$$

Set $h(x_1, \ldots, x_n) = 1$ if all x_i are positive and 0 otherwise. We have

$$L(\theta; x_1, \ldots, x_n) = g(\theta; S(x_1, \ldots, x_n))h(x_1, \ldots, x_n).$$

The factorization criterion shows that $S = X_{(n)}$ is a sufficient statistic for θ.

Next we give an example of sufficient statistics for a two dimensional parameter.

Example 4. Consider the family of normal distributions with mean μ and standard deviation σ. Both μ and σ are unknown and therefore we have a two dimensional parameter $\theta = (\mu, \sigma)$. The likelihood function in this case is

$$L(\theta; x_1, \ldots, x_n) = \frac{1}{\sigma^n (2\pi)^{n/2}} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2\right).$$

After expanding the squares inside the exponential we get

$$L(\theta; x_1, \ldots, x_n) = \frac{1}{\sigma^n (2\pi)^{n/2}} \exp\left(-\frac{1}{2\sigma^2} \left(\sum_{i=1}^{n} x_i^2 - 2\mu \sum_{i=1}^{n} x_i + n\mu^2\right)\right).$$

Let the function g be

$$g(s, t, \mu, \sigma) = \frac{1}{\sigma^n (2\pi)^{n/2}} \exp\left(-\frac{1}{2\sigma^2}(s - 2\mu t + n\mu^2)\right).$$

It is easy to check that

$$L(\theta; x_1, \ldots, x_n) = g\left(\sum_{i=1}^{n} x_i^2, \sum_{i=1}^{n} x_i, \mu, \sigma\right).$$

It turns out that the factorization criterion stated above holds for multi dimensional statistics. Hence, the two dimensional statistic

$$\left(\sum_{i=1}^{n} X_i^2, \sum_{i=1}^{n} X_i\right)$$

is sufficient for (μ, σ).

We now state that exponential families always have a sufficient statistic.

Exponential Families and Sufficiency

Recall that a family of probability distributions with parameter θ is an exponential family if and only if its logarithm can be written as

$$a(\theta)t(x) + b(\theta) + r(x) \text{ for } x \in A,$$

where the support of the distribution A does not depend on θ. Consider an i.i.d. sample $X_1, X_2, \ldots, X_n$ of such a family then

$$S = \sum_{i=1}^{n} t(X_i)$$

is sufficient.

This result is easy to prove. For, the likelihood function for an i.i.d. sample is

$$L(\theta; x_1, \ldots, x_n) = \exp\left(a(\theta)\sum_{i=1}^{n} t(x_i) + nb(\theta) + \sum_{i=1}^{n} r(x_i)\right) \text{ for } x \in A.$$

Let $S(x_1, \ldots, x_n) = \sum_{i=1}^{n} t(X_i)$ and

$$g(\theta; s) = \exp(a(\theta)s + nb(\theta))$$

Set $h(x_1, \ldots, x_n) = \exp\left(\sum_{i=1}^{n} r(x_i)\right)$ if all x_i are in A and 0 otherwise. We have

$$L(\theta; x_1, \ldots, x_n) = g(\theta; S(x_1, \ldots, x_n))h(x_1, \ldots, x_n).$$

By the factorization criterion S is a sufficient statistic.

Example 5. Consider the family of normal distributions with mean μ and standard deviation 1. We have shown in Sect. 9.2 that this is an exponential family with $t(x) = x$. Hence,

$$S = \sum_{i=1}^{n} t(X_i) = \sum_{i=1}^{n} X_i$$

is sufficient for μ.

Sufficiency turns out to be a very useful criterion to get the best unbiased estimator. We first define "best."

Minimum Variance Unbiased Estimator

Consider an exponential family with parameter θ. Let $\hat{\theta}$ be an *unbiased* estimator of θ such that $\mathrm{Var}(\hat{\theta})$ is smaller than the variance of all other unbiased of θ. Then, $\hat{\theta}$ is said to be a minimum variance unbiased estimator (m.v.u.e) of θ. In particular, no unbiased estimator has a lower variance than $\hat{\theta}$.

Next we give a criterion to find an m.v.u.e based on a sufficient statistic. This is a particular case of the Lehmann–Scheffé Theorem.

M.V.U.E. and Sufficiency

Consider an exponential family of probability distributions with parameter θ. Let $R = g(S)$ be an *unbiased* estimator of $\tau(\theta)$ where S is a sufficient statistic and g is a function. Then R is a minimum variance unbiased estimator (m.v.u.e.) of $\tau(\theta)$.

Hence, it is enough to have an unbiased estimator which is a function of S to have an optimal unbiased estimator.

Example 6. Consider the Poisson distribution with mean θ. That is,

$$P(X = x|\theta) = \exp(-\theta)\frac{\theta^x}{x!} = \exp(-\theta + x\log\theta - \log x!),$$

where θ is in $I = (0, \infty)$. Set $a(\theta) = \log(\theta)$, $b(\theta) = -\theta$, $t(x) = x$ and $r(x) = -\log x!$. Note that $P(X = x|\theta) > 0$ if and only if x is a positive integer or 0. Hence the support of the distribution does not depend on θ. Moreover, a and b are infinitely differentiable on I. Therefore, the family of Poisson distributions is exponential. Hence,

$$S = \sum_{i=1}^{n} X_i$$

is sufficient for θ. Since θ is the mean of this distribution we know that $\bar{X}$ is an unbiased estimator of θ. On the other hand $\bar{X} = g(S)$ where $g(x) = x/n$. Therefore, $\bar{X}$ is an m.v.u.e. This is the best unbiased estimator where "best" refers to the usual quadratic distance.

Assume now that we would like an m.v.u.e. for $\exp(-\theta)$ (instead of θ) for the Poisson distribution. It is easy to think of an unbiased estimator but not so easy to find one which is a function of S. The next result gives as a method to do just that.

Conditional Expectation, M.V.U.E. and Sufficiency

Consider an exponential family of probability distributions with parameter θ. Let R be an *unbiased* estimator of $\tau(\theta)$, let S be a sufficient statistic. Then the conditional expectation $E(R|S)$ is a minimum variance unbiased estimator (m.v.u.e.) of $\tau(\theta)$.

This is a rather remarkable result. Take *any* unbiased estimator of $\tau(\theta)$, by taking the conditional expectation with respect to the sufficient statistic we get an m.v.u.e.!

Example 7. Consider again the family of Poisson distributions with parameter θ. Let $\tau(\theta) = \exp(-\theta)$. We want to find an m.v.u.e. for $\tau(\theta)$. The first step is to find an unbiased estimator. For an i.i.d. sample $X_1 \dots X_n$ let $R = 1$ if $X_1 = 0$ and $R = 0$ if $X_1 > 0$. This is not a great estimator: we will estimate $\exp(-\theta)$ by 0 or 1 depending on what X_1 is and ignore the rest of the sample! However, it is an unbiased estimator and to get an m.v.u.e. it is enough to compute the conditional expectation of R with respect to a sufficient statistic. Note that

$$P(R = 1) = P(X_1 = 0) = \tau(\theta) = \exp(-\theta).$$

Since R is a Bernoulli random variable $E(R) = P(R = 1)$. Hence, R is an unbiased estimator of $\tau(\theta)$. On the other hand we know that

$$S = \sum_{i=1}^{n} X_i$$

is a sufficient statistic. We now compute $E(R|S)$. Since R is a Bernoulli random variable we have for any positive integer s

$$E(R|S = s) = P(R = 1|S = s) = \frac{P(X_1 = 0; S = s)}{P(S = s)}.$$

Note that

$$\{X_1 = 0; S = s\} = \{X_1 = 0\} \cap \left\{ \sum_{i=2}^{n} X_i = s \right\},$$

two independent events. Moreover, recall that a sum of independent Poisson random variables is a Poisson random variable whose rate is the sum of the rates. Therefore,

$$P\left(\sum_{i=2}^{n} X_i = s \right) = \exp(-(n-1)\theta) \frac{((n-1)\theta)^s}{s!}$$

and

$$P(S = s) = \exp(-n\theta) \frac{(n\theta)^s}{s!}.$$

Hence,

$$E(R|S = s) = \frac{\exp(-\theta) \exp(-(n-1)\theta)((n-1)\theta)^s/s!}{\exp(-n\theta)(n\theta)^s/s!} = \left(\frac{n-1}{n} \right)^s.$$

Therefore, an m.v.u.e. for $\exp(-\theta)$ is $(\frac{n-1}{n})^S$.

We have chosen to concentrate on exponential families in this Section because this is where most of the applications are and the theory is simpler. Next, we state a result that illustrates why sufficiency is a powerful tool in general and not only for exponential families.

Rao–Blackwell Theorem

Consider a distribution family depending on θ and assume that it has a sufficient statistic S. Let R be an estimator of θ and let $R' = E(R|S)$. Then R' is a better estimator than R in that

$$\mathrm{d}(R', \theta) \leq \mathrm{d}(R, \theta),$$

for all θ.

In words, one always improves an estimator by computing its conditional expectation with respect to a *sufficient* statistic. The proof is based on the following property of conditional expectation.

Lemma 1. *For any random variables X and Y we have*

$$\mathrm{Var}[E(X|Y)] \leq \mathrm{Var}(X).$$

That is, the variance of the conditional expectation is less than the variance of the (unconditioned) variable. We now prove the Lemma.

By definition

$$\mathrm{Var}(X) = E(X^2) - E(X)^2.$$

By Property P1 (Sect. 8.4) we have

$$\mathrm{Var}(X) = E(E(X^2|Y)) - E(E(X|Y))^2.$$

We now subtract and add $E(E(X|Y)^2)$ to get

$$\mathrm{Var}(X) = E[E(X^2|Y) - E(X|Y)^2] + E[E(X|Y)^2] - E[E(X|Y)]^2.$$

Observe now that

$$E[E(X|Y)^2] - E[E(X|Y)]^2 = \mathrm{Var}[E(X|Y)],$$

so

$$\mathrm{Var}(X) = E[E(X^2|Y) - E(X|Y)^2] + \mathrm{Var}[E(X|Y)].$$

Conditioned on $Y = y$

$$E(X^2|Y = y) - E(X|Y = y)^2 = \mathrm{Var}(X|Y = y).$$

A variance is of course always positive so for all y

$$E(X^2|Y = y) - E(X|Y = y)^2 \geq 0$$

and the corresponding random variable $E(X^2|Y) - E(X|Y)^2$ is positive as well. Therefore, $\mathrm{Var}(X)$ is the sum of the positive term $E[E(X^2|Y) - E(X|Y)^2]$ and $\mathrm{Var}[E(X|Y)]$. Hence,

$$\mathrm{Var}(X) \geq \mathrm{Var}[E(X|Y)].$$

The proof of the Lemma is complete.

It is now easy to complete the proof of Rao–Blackwell Theorem. Recall that

$$\mathrm{d}(R, \theta) = E((R - \theta)^2).$$

and the formula

$$\mathrm{d}(R, \theta) = \mathrm{Var}(R) + (E(R) - \theta)^2.$$

Since $E(R') = E(R)$ (why?) in order to compare $\mathrm{d}(R, \theta)$ to $\mathrm{d}(R', \theta)$ we only need to compare $\mathrm{Var}(R)$ to $\mathrm{Var}(R')$. According to the Lemma

$$\mathrm{Var}(R) \geq \mathrm{Var}(R')$$

and so this completes the proof of the Theorem. However, the attentive reader will have noticed that we have not used the main hypothesis of the Theorem: sufficiency of S! In fact the hypothesis has been implicitly used. We have defined $R' = E(R|S)$. The problem is that R' could very well depend on θ. If R' depends on θ it is not an estimator of θ! It is because S is sufficient that R' does not depend on θ and is therefore an estimator of θ. The sufficiency hypothesis is crucial in providing a true estimator of θ.

Exercises 9.3

1. Consider the family of normal distributions with mean 0 and standard deviation σ.

(a) Show that $\sum_{i=1}^{n} X_i^2$ is a sufficient statistic for σ.
(b) Find an m.v.u.e. for σ^2.

2. Consider the family of normal distributions with mean μ and standard deviation $\sigma = 1$. Find an m.v.u.e. for μ.

3. Consider $X_1, X_2, \ldots, X_n$ an i.i.d. sample of the exponential distribution with parameter θ. That is, the density is

$$f(x|\theta) = \theta \mathrm{e}^{-\theta x} \text{ for } x > 0$$

and 0 otherwise.

(a) Show that $\bar{X}$ is an m.v.u.e. of $1/\theta$.
(b) Recall that a sum of n i.i.d. random variables with parameter θ has a Γ distribution with parameters n and θ and density

$$g(s|\theta) = \frac{1}{(n-1)!}\theta^n s^{n-1}\mathrm{e}^{-\theta s}.$$

Use g to compute the expected value of

$$\frac{1}{\sum_{i=1}^n X_i}.$$

(c) Show that

$$\hat{\theta} = \frac{n-1}{\sum_{i=1}^n X_i}$$

is an m.v.u.e. of θ.
(d) Show that the variance of $\hat{\theta}$ is $\theta^2/(n-2)$.
(e) Explain why $1/\bar{X}$ is also a natural choice to estimate θ. Compare its mean square error to the one of $\hat{\theta}$.

4. Consider a family of Poisson distributions with parameter θ. We want to estimate $\tau(\theta) = \exp(-\theta) = P(X = 0)$. For an i.i.d. sample $X_1 \ldots X_n$ let $R = 1$ if $X_1 = 0$ and $R = 0$ if $X_1 > 0$. In Example 6 we have computed

$$R' = E(R|S) = \left(\frac{n-1}{n}\right)^S,$$

where S is the sum of the X_i. We have shown that R' is an m.v.u.e. of $\tau(\theta)$. Here are the values of a sample of size 50:

0 2 4 2 5 2 3 1 0 2 2 1 0 1 0 2 2 2 1 2 0 4 0 3 1 1 4 1 3 4 1 3 1 4 0 1 0 2 0 5 1 3 2 2 3 2 2 3 3 0

(a) Compute R and R' for this sample.
(b) Let B be the number of 0's in the sample. Show that B is a binomial random variable with parameters n and $\tau(\theta)$.
(c) Show that $B' = B/n$ is an unbiased estimator of $\tau(\theta)$. Compute B' for the sample above.
(d) Compute the variance of B'.
(e) Compute the variance of R' and compare it to the variance of B'.

5. Consider again the family of Poisson distributions with parameter θ. Let $\tau(\theta) = \exp(-\theta)$.

(a) Explain why $\exp(-\bar{X})$ is a reasonable choice to estimate $\tau(\theta)$.
(b) Show that n goes to infinity $(\frac{n-1}{n})^S$ approaches $\exp(-\bar{X})$.
(c) Compute the variance of $\exp(-\bar{X})$.

6. Consider the family of normal distributions with mean μ and known standard deviation σ. Is $\bar{X}$ an m.v.u.e. of μ?

7. Consider a family of exponential distributions with expectation α. That is,

$$f(x|\alpha) = \frac{1}{\alpha} \exp\left(-\frac{1}{\alpha}x\right).$$

Show that $\bar{X}$ is an m.v.u.e. of α.

8. Consider the family of distributions with density

$$f(x|\theta) = \mathrm{e}^{-(x-\theta)} \text{ if } x \geq \theta$$

and $f(x|\theta) = 0$ for $x < \theta$.

(a) Is this an exponential family of distributions?
(b) Find a sufficient statistic for θ.

9. Consider an i.i.d. sample of random variables with density

$$\frac{\theta}{2} \exp(-\theta|x|).$$

(a) Show that this is an exponential family of distributions.
(b) Compute $E|X|$.
(c) Find a function τ such that you have an m.v.u.e. for $\tau(\theta)$.

10. Consider an i.i.d. sample of random variables with density

$$\theta x^{-\theta-1} \text{ for } x \geq 1.$$

Find a sufficient statistic.

11. Consider an i.i.d. sample of Γ random variables. The density is

$$g(s|\theta) = \frac{1}{(k-1)!}\theta^k s^{k-1}\mathrm{e}^{-\theta s}$$

where k is a known positive integer and θ is the unknown parameter. Show that

$$\left(\Pi_{i=1}^n X_i, \sum_{i=1}^n X_i\right)$$

is a sufficient statistic.

12. Consider X, Y i.i.d. with density $\theta\mathrm{e}^{-\theta x}$ for $x \geq 0$. Let $T = X + Y$.

(a) Show that the conditional distribution of $X|T = t$ is

$$f(x|t) = \frac{1}{t} \text{ for } 0 < x < t.$$

(b) How come there is no θ in $f(x|t)$?
(c) Compute $P(X > 2|T = t)$.

13. Consider a sequence $X_1, \ldots, X_n, \ldots$ of i.i.d. random variables with expected value μ and variance equal to 1. Let $R = E(\bar{X}|X_1)$.

(a) Show that

$$R = \frac{1}{n}X_1 + \frac{n-1}{n}\mu.$$

(b) Compute $\text{Var}(R)$ and compare it to $\text{Var}(\bar{X})$.
(c) Explain why R is not an estimator of μ.

9.4 Bayes' Estimators

So far θ has always been an unknown parameter. In this section we will use instead the Bayesian approach for which θ is the value of a random variable. First a word on notation: we have consistently used upper case letters (such as X) for a random variable and the corresponding lower case letter (x) for a possible value. The upper case letter corresponding to θ is Θ. We will give ourselves a so-called *prior* distribution for the random variable Θ and θ will be a possible value of Θ. We use the word "prior" because it is a distribution we pick before (i.e., prior) having a sample of observations. Once we have a sample we compute the conditional distribution of Θ given the sample. This is the so-called *posterior* distribution of Θ. We start with an example.

Example 1. Consider a family of Bernoulli distributions with parameter θ. Assume we know nothing about θ except that it is in $[0, 1]$. We give ourselves an uniform distribution for Θ: $f(\theta) = 1$ for all θ in $[0, 1]$. The choice of an uniform (a flat distribution) shows that θ could be anywhere in $[0, 1]$. Given $\Theta = \theta$ the Bernoulli probability distribution is

$$f(x|\theta) = \theta^x(1-\theta)^{1-x}$$

for $x = 0$ or 1. Given an i.i.d. sample $X_1, \ldots, X_n$ the posterior distribution of Θ is by definition the distribution of Θ given $X_1, \ldots, X_n$. We now compute it. First we compute the probability distribution of the vector $(X_1, X_2, \ldots, X_n, \Theta)$. Let

$$s = \sum_{i=1}^{n} x_i.$$

Then,

$$f(x_1, \ldots, x_n, \theta) = f(x_1, \ldots, x_n|\theta)f(\theta) = \theta^{x_1}(1-\theta)^{1-x_1}\ldots\theta^{x_n}(1-\theta)^{1-x_n}$$

$$= \theta^s(1-\theta)^{n-s} \text{ for } \theta \in [0, 1]$$

We are now ready to compute the posterior distribution of Θ.

$$f(\theta|x_1,\ldots,x_n) = \frac{f(x_1,\ldots,x_n,\theta)}{f(x_1,\ldots,x_n)} = \frac{\theta^s(1-\theta)^{n-s}}{f(x_1,\ldots,x_n)} \text{ for } \theta \in [0,1],$$

where $f(x_1,\ldots,x_n)$ is the density of the vector $(X_1,\ldots,X_n)$. We will now try to identify the posterior distribution with as few computations as possible. Recall that the beta density with parameters a and b is

$$\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1} \text{ for } \theta \in [0,1].$$

Letting $s = a-1$ and $n-s = b-1$ we see that the posterior looks like a beta distribution with parameters $a = s+1$ and $b = n-s+1$. Except that instead of having $\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}$ we have $\frac{1}{f(x_1,\ldots,x_n)}$. Are these two quantities equal? Since $\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1}$ is a probability density its integral is 1 and therefore

$$\int_0^1 \theta^{a-1}(1-\theta)^{b-1}\mathrm{d}\theta = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}.$$

But the integral of the posterior distribution is also 1. Hence,

$$\int_0^1 \frac{\theta^{a-1}(1-\theta)^{b-1}}{f(x_1,\ldots,x_n)}\mathrm{d}\theta = \frac{1}{f(x_1,\ldots,x_n)}\int_0^1 \theta^{a-1}(1-\theta)^{b-1}\mathrm{d}\theta = 1,$$

and

$$\int_0^1 \theta^{a-1}(1-\theta)^{b-1}\mathrm{d}\theta = f(x_1,\ldots,x_n).$$

Therefore,

$$f(x_1,\ldots,x_n) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}.$$

This proves that the posterior distribution is indeed a beta distribution with parameters $a = s+1$ and $b = n-s+1$ where s is the sum of the x_i and n is the size of the sample. This is in fact a general method: if you can identify the posterior distribution up to a term constant in θ (such as $\frac{1}{f(x_1,x_2,\ldots,x_n)}$ in this example) then that is the distribution you are looking for.

Here is an i.i.d. sample of 50 Bernoulli random variables. 0 0 1 0 0 1 0 0 0 0 1 0 0 0 0 1 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 1 0 1 0 1 1 0 0 0 0 1 0 0 0 0 1 0 0 0 0 0. The size of this sample is $n = 50$ and the sum of the x_i is $s = 11$. Hence, with an uniform prior distribution for Θ we get a posterior which is a beta distribution with parameters $a = s+1 = 12$ and $b = n-s+1 = 40$. See Fig. 9.1 below.

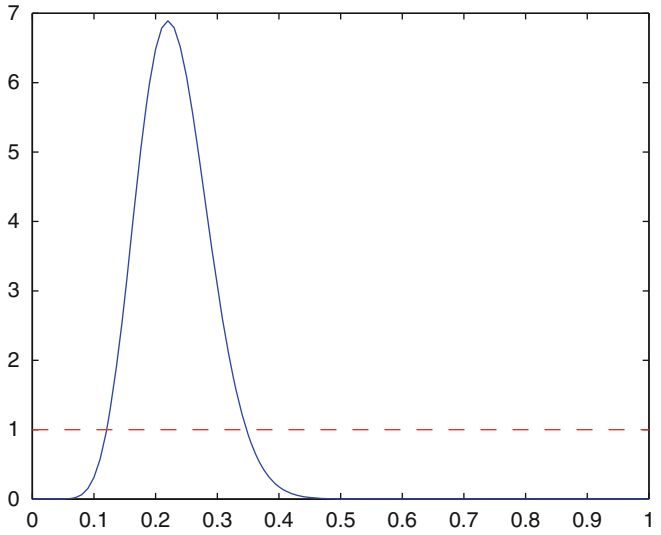


Fig. 9.1 Observe how we go from a flat prior distribution to a rather pointed posterior distribution

We summarize the method below.

Prior and Posterior Distributions

Consider θ as the value of a random variable Θ rather than a fixed number. Let $f(\theta)$ be the prior distribution of Θ. Let $X_1, X_2, \ldots, X_n$ be an i.i.d. sample of a distribution that depends on θ. The posterior distribution of Θ is the conditional distribution of Θ given the observations $x_1, \ldots, x_n$ of the sample.

We should use the posterior distribution to estimate θ. For instance, in Example 1 we could use the beta distribution to compute the probability that Θ is between 0.1 and 0.3, say. But in many situations we would like to have a number to work with. A natural choice for a Bayes' estimator of θ is the expected value of the posterior distribution. This turns out to also be the optimal choice in the sense defined below.

Bayes' Estimator

The Bayes' estimator T^* of θ is

$$T^* = E(\Theta | X_1 \ldots, X_n).$$

That is, T^* is the conditional expectation of Θ given $(X_1, \ldots, X_n)$.

Recall from Sect. 8.4 that the minimum of

$$E[(\Theta - h(X_1, \ldots, X_n)^2]$$

over all the functions h is attained by

$$h(X_1 \ldots, X_n) = E(\Theta | X_1 \ldots, X_n).$$

Example 2. We go back to Example 1. We showed that with an uniform prior we get a beta posterior with parameters $a = s + 1$ and $b = n - s + 1$ where n is the size of the sample and s is the sum of the observations. To get the Bayes' estimator we need the expected value of a beta distribution. We do the computation now. Assume that B has a beta distribution with parameters a and b.

$$E(B) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \int_0^1 \theta \theta^{a-1}(1-\theta)^{b-1} \mathrm{d}\theta = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \int_0^1 \theta^a (1-\theta)^{b-1} \mathrm{d}\theta.$$

The integrand (up to a constant) is the density of a beta with parameters $a + 1$ and b. Hence,

$$\int_0^1 \theta^a (1-\theta)^{b-1} \mathrm{d}\theta = \frac{\Gamma(a+1)\Gamma(b)}{\Gamma(a+b+1)}.$$

Therefore,

$$E(B) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \frac{\Gamma(a+1)\Gamma(b)}{\Gamma(a+b+1)} = \frac{\Gamma(a+b)}{\Gamma(a+b+1)} \frac{\Gamma(a+1)}{\Gamma(a)}.$$

Recall that

$$\Gamma(x) = (x-1)\Gamma(x-1)$$

for all $x > 1$. Hence, we have

$$E(B) = \frac{a}{a+b}.$$

Therefore, the Bayes' estimator in this example is

$$\frac{\sum_{i=1}^n X_i + 1}{n+2}.$$

This is close but not identical to the m.v.u.e. which was found to be $\bar{X}$. We now compare these two estimators for the sample of Example 1. Recall that $n = 50$ and $s = 11$. Hence, the m.v.u.e. is $11/50 = 0.22$ and the Bayes' estimator is

$12/52 = 0.23$. The sample was actually from a Bernoulli with $\theta = 0.25$ so the Bayes' estimate is slightly closer to the true value in this case.

Example 3. Let $X_1, \ldots, X_n$ be an i.i.d. sample with a normal distribution with mean θ and variance 1. Let Θ have a standard normal (i.e., mean 0 and variance 1) prior distribution. Compute the Bayes' estimate.

We have the prior

$$f(\theta) = \frac{1}{\sqrt{2\pi}} \exp(-\theta^2/2)$$

and given θ the distribution of an observation is

$$f(x|\theta) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(x-\theta)^2\right).$$

Therefore,

$$f(x_1, \ldots, x_n|\theta) = \frac{1}{(\sqrt{2\pi})^n} \exp\left(-\frac{1}{2}\sum_{i=1}^{n}(x_i-\theta)^2\right).$$

Hence, the posterior distribution is:

$$\begin{aligned} f(\theta|x_1, \ldots, x_n) &= \frac{f(x_1, \ldots, x_n|\theta) f(\theta)}{f(x_1, \ldots, x_n)} \\ &= \frac{1}{f(x_1, \ldots, x_n)} \frac{1}{(\sqrt{2\pi})^n} \exp\left(-\frac{1}{2}\sum_{i=1}^{n}(x_i-\theta)^2\right) \\ &\quad \times \frac{1}{\sqrt{2\pi}} \exp(-\theta^2/2) \\ &= \frac{1}{f(x_1, \ldots, x_n)} \frac{1}{(\sqrt{2\pi})^{n+1}} \exp\left(-\frac{1}{2}\sum_{i=1}^{n}(x_i-\theta)^2 - \frac{1}{2}\theta^2\right) \end{aligned}$$

The important part of this expression is the part that contains θ. As remarked earlier if we can identify that part as a known distribution we will be done and we will not have to compute $f(x_1, \ldots, x_n)$. To that purpose we concentrate on the terms containing θ. By expanding the square we have

$$\frac{1}{2}\sum_{i=1}^{n}(x_i-\theta)^2 + \frac{1}{2}\theta^2 = \frac{1}{2}\left(\sum_{i=1}^{n} x_i^2 - 2\theta\sum_{i=1}^{n} x_i + n\theta^2 + \theta^2\right).$$

Hence,

$$\frac{1}{2}\sum_{i=1}^{n}(x_i-\theta)^2+\frac{1}{2}\theta^2=\frac{n+1}{2}\left(\theta^2-\frac{2}{n+1}\theta\sum_{i=1}^{n}x_i+\frac{1}{n+1}\sum_{i=1}^{n}x_i^2\right).$$

We now "complete the square" to get

$$\frac{1}{2}\sum_{i=1}^{n}(x_i-\theta)^2+\frac{1}{2}\theta^2=\frac{n+1}{2}\left(\theta-\frac{1}{n+1}\sum_{i=1}^{n}x_i\right)^2 - \frac{n+1}{2}\left(\frac{1}{n+1}\sum_{i=1}^{n}x_i\right)^2+\frac{1}{2}\sum_{i=1}^{n}x_i^2.$$

We plug this side computation into the posterior distribution to get

$$f(\theta|x_1,\ldots,x_n)=g(x_1,\ldots,x_n)\exp\left(-\frac{n+1}{2}\left(\theta-\frac{1}{n+1}\sum_{i=1}^{n}x_i\right)^2\right),$$

where

$$g(x_1,\ldots,x_n)=\frac{1}{f(x_1,\ldots,x_n)}\frac{1}{(\sqrt{2\pi})^{n+1}} \times\exp\left(\frac{n+1}{2}\left(\frac{1}{n+1}\sum_{i=1}^{n}x_i\right)^2-\frac{1}{2}\sum_{i=1}^{n}x_i^2\right).$$

We state the exact expression of g for the sake of completeness but we only need to know that it does not depend on θ. We are now ready to identify the posterior distribution as a normal distribution. Recall that a normal distribution with mean μ and variance σ^2 has density

$$\frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{1}{2\sigma^2}(\theta-\mu)^2\right).$$

Therefore, the posterior distribution is a normal distribution with mean $\frac{1}{n+1}\sum_{i=1}^{n}x_i$ and variance $1/(n+1)$. Hence, the Bayes' estimate for θ is

$$\frac{1}{n+1}\sum_{i=1}^{n}X_i.$$

Example 4. Here is an i.i.d. sample for a normal distribution with unknown mean and variance 1.

1.5377 2.8339 −1.2588 1.8622 1.3188 −0.3077 0.5664 1.3426 4.5784 3.7694 −0.3499 4.0349 1.7254 0.9369 1.7147 0.7950 0.8759 2.4897 2.4090 2.4172 1.6715 −0.2075 1.7172 2.6302 1.4889 2.0347 1.7269 0.6966 1.2939 0.2127

According to Example 3, with a standard normal prior we get a posterior with mean $\frac{1}{n+1}\sum_{i=1}^{n} x_i$ and variance $1/(n+1)$. Here $n = 30$ and $\sum_{i=1}^{n} x_i = 46.5569$. Hence, the Bayes' estimate is 1.5018. On the other hand the m.v.u.e. in this case is $\bar{X}$ which for this sample is 1.5519.

What is the probability that Θ is between 1.3 and 1.5? This can be computed using the posterior distribution which is normally distributed with mean 1.5018 and variance 1/31. Hence,

$$P(1.3 < \Theta < 1.5 | x_1 \dots x_{30}) = P(-1.12 < Z < -0.01) = 0.36,$$

where Z is a standard normal variable.

Exercises 9.4

1. Use the method of Example 2 to compute the variance of a beta distribution with parameters a and b.

2. In Example 1, estimate the probability that Θ is between 0.2 and 0.3 using the posterior distribution. (You will need to compute the integral numerically.)

3. Consider a family of Bernoulli distributions as in Example 1. Instead of taking an uniform prior as we did there take a prior which is a beta with parameters $a = b = 2$.

(a) Sketch on the same graph the density of an uniform and the density of a beta with parameters 2 and 2.
(b) Compute the posterior distribution.
(c) Compute the Bayes' estimator on the sample of Example 1.

4. Sketch the graphs of the prior and posterior densities of Example 3. Use the sample of Example 4.

5. Let $X_1, \dots, X_n$ be an i.i.d. sample of a Poisson distribution with mean θ. Let Θ have an exponential with parameter 1 prior distribution. That is, its density is

$$f(\theta) = \exp(-\theta) \text{ for } \theta > 0.$$

(a) Show that the posterior distribution is a Gamma with parameters $1 + \sum_{i=1}^{n} x_i$ and $n + 1$.
(b) Find the Bayes' estimator.

6. (a) Show that for $a > 1$ and $b > 0$ we have

$$\int_0^\infty x^{a-1} \exp(-bx) = \frac{\Gamma(a)}{b^a}.$$

(b) Use (a) to compute the expected value of $\exp(-X)$ where X has a Gamma distribution with parameters a and b.
(c) In Exercise 5 we have shown that for a Poisson distribution if the prior is exponentially distributed then the posterior has a Gamma distribution. Use this result and (b) to find a Bayes' estimator for $\exp(-\theta)$ where θ is the mean of a Poisson distribution with mean θ.

7. Recall that the first two moments of a Gamma distribution with parameters a and b are a/b and $a(a+1)/b^2$, respectively.

(a) Find a and b given that the first moment is 1 and the second is 3.
(b) Use a prior Gamma distribution with the parameters above to find a Bayes' estimate for the mean of a Poisson distribution.

8. Let $X_1, \ldots, X_n$ be an i.i.d. sample with a normal distribution with mean θ and variance 1. Let Θ have a normal mean a and variance 1 prior distribution. Compute the Bayes' estimate. (You may want to imitate the computation of Example 3.)

9. Let $X_1, \ldots, X_n$ be an i.i.d. sample of a normal distribution with mean 0 and variance $1/\theta$. Note that for computational convenience our parameter is the inverse of the variance. Let Θ have an exponential mean 1 prior distribution.

(a) Show that the posterior distribution has density

$$\frac{1}{f(x_1, \ldots, x_n)} \frac{1}{(\sqrt{2\pi})^n} \theta^{n/2} \exp\left(-\frac{1}{2}\theta \sum_{i=1}^{n} x_i^2\right) \exp(-\theta)$$

for $\theta > 0$, where $f(x_1, \ldots, x_n)$ is the density of the vector $(X_1, \ldots, X_n)$.
(b) Show that the distribution in (a) is a Gamma distribution and identify the parameters.
(c) Find the Bayes' estimate.

10. Consider an i.i.d. sample from an uniform distribution on $[0, \theta]$. Assume that the prior distribution of Θ is uniform on $[0, 1]$.

(a) Show that the density of $(X_1, \ldots, X_n, \Theta)$ is

$$f(x_1, \ldots, x_n, \theta) = \frac{1}{\theta^n} \text{ for } \theta \in (x_{(n)}, 1),$$

where $x_{(n)}$ is the largest value of the sample.

(b) Show that the density of $(X_1, \ldots, X_n)$ is

$$f(x_1, \ldots, x_n) = \int_{x_{(n)}}^{1} \frac{1}{\theta^n} \mathrm{d}\theta.$$

(c) Using (a) and (b) compute the posterior distribution.
(d) Compute the Bayes' estimator.

Chapter 10
Multiple Linear Regression

10.1 The Least Squares Estimate

The main purpose of this chapter is to predict the value of some variable Y based on a given set of variables X_i where $i = 1, 2, \ldots, p-1$ and p is an integer larger than or equal to 2. The following example will be examined in detail throughout this chapter.

Example 1. For a given country let Y be the under 5 infant mortality (number of children dead by age 5 per 1,000 births), X_1 the percentage of children vaccinated against measles, X_2 the percentage of children vaccinated against diphtheria, tetanus, and pertussis infections (a single vaccine called DPT3 takes care of the three infections), X_3 the percentage of the population that has access to clean water and let X_4 be the percentage of population that have access to sanitation (sewage system, septic tanks, and so on). In this case we would have $p - 1 = 4$, that is, $p = 5$.

The simplest model to predict Y using $X_1, X_2, \ldots, X_{p-1}$ is a *linear* model:

$$Y = b_0 + b_1 X_1 + b_2 X_2 + \cdots + b_{p-1} X_{p-1},$$

where b_0, b_1, b_2, $\ldots$, b_{p-1} are constants to be estimated. Note that this is a generalization of what we did in Chap. 6. There we had only one explanatory variable X and therefore we had $p = 2$. When $p = 2$ the model is called a simple regression model. For $p \geq 3$ we call it a *multiple linear regression model*.

The first step is to estimate the constants b_i, $i = 0, 1, 2, \ldots, p-1$. The estimates will be based on observations. Assume that we have a sample of n observations:

$$(y_i, x_{i,1}, x_{i,2}, \ldots, x_{i,p-1}) \text{for} i = 1, 2, \ldots n,$$

where the y_i are observations of Y, the $x_{i,1}$ are observations of the variable X_1, the $x_{i,2}$ are observations of the variable X_2 and so on. To avoid double indices and other

R.B. Schinazi, *Probability with Statistical Applications*,
DOI 10.1007/978-0-8176-8250-7_10, © Springer Science+Business Media, LLC 2012

cumbersome notations it is better to formulate the model in terms of matrices. Let **Y**, **b**, and **X** be

$$\mathbf{Y} = \begin{pmatrix} y_1 \\ y_2 \\ . \\ . \\ y_n \end{pmatrix} \quad \mathbf{b} = \begin{pmatrix} b_0 \\ b_1 \\ . \\ . \\ b_{p-1} \end{pmatrix} \quad \mathbf{X} = \begin{pmatrix} 1 & x_{1,1} & x_{1,2} & \dots & x_{1,p-1} \\ 1 & x_{2,1} & x_{2,2} & \dots & x_{2,p-1} \\ 1 & x_{3,1} & x_{3,2} & \dots & x_{3,p-1} \\ . & . & . & \dots & . \\ 1 & x_{n,1} & x_{n,2} & \dots & x_{n,p-1} \end{pmatrix}.$$

Note that the first column of **X** corresponds with the constant in the linear model, the second column corresponds with the variable X_1, the third column with the variable X_2, and so on. Observe also that **Y** is a column vector with n components, **b** is a column vector with p components and that **X** is a $n \times p$ matrix. We can rewrite the multiple linear regression in matrix form as

$$\mathbf{Y} = \mathbf{Xb}.$$

Example 2. Going back to Example 1 here is some of the data extracted from the databases of the World Health Organization

http://www.who.int/whosis/whostat/2010/en/index.html

Afghanistan	257	75	85	48	37
Albania	14	98	99	97	98
Algeria	41	88	93	83	95
Andorra	4	98	99	100	100
Angola	220	79	81	50	57
.	.	.	.	.	.
.	.	.	.	.	.
Zimbabwe	96	66	62	82	44

We will actually be working with a list of 163 countries. Hence, the table above has 163 rows. The complete table can be downloaded from my webpage

http://www.uccs.edu/~rschinaz/.

The first column corresponds to infant mortality, the second to measles vaccination, the third to DPT3 vaccination, the fourth to clean water access and the fifth to sanitation access. Hence,

$$\mathbf{X} = \begin{pmatrix} 1 & 75 & 85 & 48 & 37 \\ 1 & 98 & 99 & 97 & 98 \\ 1 & 88 & 93 & 83 & 95 \\ 1 & 98 & 99 & 100 & 100 \\ 1 & 79 & 81 & 50 & 57 \\ . & . & . & . & \end{pmatrix} \quad \mathbf{Y} = \begin{pmatrix} 257 \\ 14 \\ 41 \\ 4 \\ 220 \\ . \end{pmatrix}$$

Going back to the model

$$\mathbf{Y} = \mathbf{Xb},$$

what is the best estimate of $\mathbf{b}$? As we have seen already "best" is relative to the criterion we choose. We will use the least squares criterion. That is, we want to find an estimate of $\mathbf{b}$ that we denote by $\hat{\mathbf{b}}$ that minimizes the quantity

$$S(\mathbf{b}) = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_{i,1} - b_2 x_{i,2} - \cdots - b_{p-1} x_{i,p-1})^2.$$

We first compute the partial derivative of S with respect to b_0:

$$\frac{\mathrm{d}}{\mathrm{d}b_0} S = \sum_{i=1}^{n} 2(-1)(y_i - b_0 - b_1 x_{i,1} - b_2 x_{i,2} - \cdots - b_{p-1} x_{i,p-1}).$$

We set the partial derivative equal to 0 and we get the equation

$$nb_0 + b_1 \sum_{i=1}^{n} x_{i,1} + b_2 \sum_{i=1}^{n} x_{i,2} + \cdots + b_{p-1} \sum_{i=1}^{n} x_{i,p-1} = \sum_{i=1}^{n} y_i.$$

Fix j in $\{1, 2 \ldots, p-1\}$ and take the partial derivative with respect to b_j:

$$\frac{d}{db_j} S = \sum_{i=1}^{n} 2(-x_{i,j})(y_i - b_0 - b_1 x_{i,1} - b_2 x_{i,2} - \cdots - b_{p-1} x_{i,p-1}).$$

Setting this partial derivative equal to 0 yields for $j = 1, 2, \ldots, p-1$:

$$b_0 \sum_{i=1}^{n} x_{i,j} + b_1 \sum_{i=1}^{n} x_{i,j} x_{i,1} + \cdots + b_{p-1} \sum_{i=1}^{n} x_{i,j} x_{i,p-1} = \sum_{i=1}^{n} y_i x_{i,j}.$$

Hence, we have the following system of p equations and p unknowns b_0, $b_1, \ldots, b_{p-1}$. The first equation is

$$nb_0 + b_1 \sum_{i=1}^{n} x_{i,1} + b_2 \sum_{i=1}^{n} x_{i,2} + \cdots + b_{p-1} \sum_{i=1}^{n} x_{i,p-1} = \sum_{i=1}^{n} y_i,$$

and the other $p-1$ equations are

$$b_0 \sum_{i=1}^{n} x_{i,j} + b_1 \sum_{i=1}^{n} x_{i,j} x_{i,1} + \cdots + b_{p-1} \sum_{i=1}^{n} x_{i,j} x_{i,p-1} = \sum_{i=1}^{n} y_i x_{i,j},$$

for $j = 1, 2, \ldots, p-1$.

We will use matrices to reduce this system of equations to a single matrix equation. Recall that $\mathbf{X}'$ is the transpose of matrix $\mathbf{X}$: we exchange rows and columns. Hence,

$$\mathbf{X}' = \begin{pmatrix} 1 & 1 & 1 & \dots & 1 \\ x_{1,1} & x_{2,1} & x_{3,1} & \dots & x_{n,1} \\ x_{1,2} & x_{2,2} & x_{3,2} & \dots & x_{n,2} \\ . & . & . & \dots & . \\ x_{1,p-1} & x_{2,p-1} & x_{3,p-1} & \dots & x_{n,p-1} \end{pmatrix}$$

and

$$\mathbf{X}'\mathbf{X} = \begin{pmatrix} 1 & 1 & 1 & \dots & 1 \\ x_{1,1} & x_{2,1} & x_{3,1} & \dots & x_{n,1} \\ x_{1,2} & x_{2,2} & x_{3,2} & \dots & x_{n,2} \\ . & . & . & \dots & . \\ x_{1,p-1} & x_{2,p-1} & x_{3,p-1} & \dots & x_{n,p-1} \end{pmatrix} \begin{pmatrix} 1 & x_{1,1} & x_{1,2} & \dots & x_{1,p-1} \\ 1 & x_{2,1} & x_{2,2} & \dots & x_{2,p-1} \\ 1 & x_{3,1} & x_{3,2} & \dots & x_{3,p-1} \\ . & . & . & \dots & . \\ 1 & x_{n,1} & x_{n,2} & \dots & x_{n,p-1} \end{pmatrix}$$

After multiplication we get

$$\mathbf{X}'\mathbf{X} = \begin{pmatrix} n & \sum_{i=1}^n x_{i,1} & \sum_{i=1}^n x_{i,2} & \dots & \sum_{i=1}^n x_{i,p-1} \\ \sum_{i=1}^n x_{i,1} & \sum_{i=1}^n x_{i,1}^2 & \sum_{i=1}^n x_{i,1}x_{i,2} & \dots & \sum_{i=1}^n x_{i,1}x_{i,p-1} \\ . & . & . & \dots & . \\ . & . & . & \dots & . \\ \sum_{i=1}^n x_{i,p-1} & \sum_{i=1}^n x_{i,p-1}x_{i,1} & \sum_{i=1}^n x_{i,p-1}x_{i,2} & \dots & \sum_{i=1}^n x_{i,p-1}^2 \end{pmatrix}$$

Note that $\mathbf{X}'$ is a $p \times n$ matrix (p rows, n columns) and that $\mathbf{X}$ is a $n \times p$ matrix. Therefore, $\mathbf{X}'\mathbf{X}$ is a square $p \times p$ matrix. It is now easy to check that the system of equations above can be written as

$$\mathbf{X}'\mathbf{X}\mathbf{b} = \mathbf{X}'\mathbf{Y}.$$

These are the so-called *normal* equations. If $\mathbf{X}'\mathbf{X}$ is an invertible matrix the equation has a unique solution

$$\hat{\mathbf{b}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}.$$

The Least Squares Estimate

Assume that we have a sample of n observations of the variables $(Y, X_1, \dots, X_{p-1})$:

$$(y_i, x_{i,1}, x_{i,2}, \dots, x_{i,p-1}) \text{ for } i = 1, 2, \dots, n.$$

The best linear approximation of Y by $(X_1, X_2, \dots, X_{p-1})$ is

$$Y = \hat{b}_0 + \hat{b}_1 X_1 + \cdots + \hat{b}_{p-1} X_{p-1},$$

where

$$\hat{\mathbf{b}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}.$$

The solution $\hat{\mathbf{b}}$ is the "best" with respect to the least squares criterion. That is, the function

$$S(\mathbf{b}) = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_{i,1} - b_2 x_{i,2} - \cdots - b_{p-1} x_{i,p-1})^2$$

is minimum for $\mathbf{b} = \hat{\mathbf{b}}$.

At this point we only know that S has a critical point at $\hat{\mathbf{b}}$. That is, all the partial derivatives of S at $\hat{\mathbf{b}}$ are 0. Since S is differentiable everywhere we know that if it has a minimum at $\hat{\mathbf{b}}$ then $\hat{\mathbf{b}}$ must be a critical point. However, we still need to check that a minimum actually occurs at $\hat{\mathbf{b}}$ and we will do this after the next example.

Example 3. We go back to the World Health Organization (WHO) data. From Example 2 the matrix $\mathbf{X}$ is a 163×5 matrix.

$$\mathbf{X} = \begin{pmatrix} 1 & 75 & 85 & 48 & 37 \\ 1 & 98 & 99 & 97 & 98 \\ 1 & 88 & 93 & 83 & 95 \\ 1 & 98 & 99 & 100 & 100 \\ 1 & 79 & 81 & 50 & 57 \\ . & . & . & . & \end{pmatrix}.$$

Hence, $\mathbf{X}'$ is a 5×163 matrix.

$$\mathbf{X}' = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & . \\ 75 & 98 & 88 & 98 & 79 & . \\ 85 & 99 & 93 & 99 & 81 & . \\ 48 & 97 & 83 & 100 & 50 & . \\ 37 & 98 & 95 & 100 & 57 & . \end{pmatrix},$$

and $\mathbf{X}'\mathbf{X}$ is a 5×5 matrix

$$\mathbf{X}'\mathbf{X} = \begin{pmatrix} 163 & 14034 & 14181 & 13947 & 11429 \\ 14034 & 1239626 & 1250968 & 1226664 & 1028337 \\ 14181 & 1250968 & 1266817 & 1238697 & 1038277 \\ 13947 & 1226664 & 1238697 & 1240959 & 1046217 \\ 11429 & 1028337 & 1038277 & 1046217 & 950577 \end{pmatrix}.$$

The determinant of $\mathbf{X}'\mathbf{X}$ is approximately 2.7729×10^{19}, (clearly not 0!) and the matrix can be inverted. The inverse matrix of $\mathbf{X}'\mathbf{X}$ is approximately (rounding to the fifth decimal)

$$(\mathbf{X}'\mathbf{X})^{-1} = 10^{-5} \begin{pmatrix} 35713 & -171 & -126 & -229 & 146 \\ -171 & 26 & -22 & -2 & 0 \\ -126 & -22 & 23 & 0 & 0 \\ -229 & -2 & 0 & 7 & -3 \\ 146 & 0 & 0 & -3 & 2 \end{pmatrix}.$$

Finally, we can use the normal equations to get

$$\hat{\mathbf{b}} = \begin{pmatrix} \hat{b}_0 \\ \hat{b}_1 \\ \hat{b}_2 \\ \hat{b}_3 \\ \hat{b}_4 \end{pmatrix} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = \begin{pmatrix} 284.51 \\ -0.19113 \\ -0.31886 \\ -1.7055 \\ -0.59177 \end{pmatrix}.$$

In other words, the least square method tells us that the best linear approximation of Y using the variables X_1, X_2, X_3 and X_4 is given by the equation

$$Y = 284.5 - 0.19X_1 - 0.32X_2 - 1.7X_3 - 0.59X_4.$$

We now prove that the function S actually has a minimum at $\hat{\mathbf{b}}$. We will use the following properties. For any matrices $\mathbf{A}$ and $\mathbf{B}$ with appropriate dimensions we have

$$(\mathbf{AB})' = \mathbf{B}'\mathbf{A}',$$
$$(\mathbf{A}')' = \mathbf{A},$$

and

$$(\mathbf{A} + \mathbf{B})' = \mathbf{A}' + \mathbf{B}'.$$

We will express S as a matrix product. First observe that if $\mathbf{a}$ is a column vector with n components $a_1, a_2, \ldots, a_n$ then

$$\mathbf{a}'\mathbf{a} = \sum_{i=1}^{n} a_i^2.$$

Moreover, note that $\mathbf{Y} - \mathbf{Xb}$ is a column vector with n components. For $i = 1, \ldots, n$ the i component is $y_i - b_0 - b_1 x_{i,1} - b_2 x_{i,2} - \cdots - b_{p-1} x_{i,p-1}$. Hence,

$$S(\mathbf{b}) = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_{i,1} - b_2 x_{i,2} - \cdots - b_{p-1} x_{i,p-1})^2 = (\mathbf{Y} - \mathbf{Xb})'(\mathbf{Y} - \mathbf{Xb}).$$

We subtract and add $\mathbf{X}\hat{\mathbf{b}}$ to get

$$S(\mathbf{b}) = (\mathbf{Y} - \mathbf{X}\mathbf{b})'(\mathbf{Y} - \mathbf{X}\mathbf{b}) = (\mathbf{Y} - \mathbf{X}\hat{\mathbf{b}} + \mathbf{X}(\hat{\mathbf{b}} - \mathbf{b}))'(\mathbf{Y} - \mathbf{X}\hat{\mathbf{b}} + \mathbf{X}(\hat{\mathbf{b}} - \mathbf{b})).$$

Expanding the right-hand side yields

$$\begin{aligned} S(\mathbf{b}) = {} & (\mathbf{Y} - \mathbf{X}\hat{\mathbf{b}})'(\mathbf{Y} - \mathbf{X}\hat{\mathbf{b}}) + (\mathbf{X}(\hat{\mathbf{b}} - \mathbf{b}))'(\mathbf{X}(\hat{\mathbf{b}} - \mathbf{b})) \\ & + (\mathbf{Y} - \mathbf{X}\hat{\mathbf{b}})'(\mathbf{X}(\hat{\mathbf{b}} - \mathbf{b})) + (\mathbf{X}(\hat{\mathbf{b}} - \mathbf{b}))'(\mathbf{Y} - \mathbf{X}\hat{\mathbf{b}}). \end{aligned}$$

We will show now that the last two terms are 0. Note that

$$[(\mathbf{Y} - \mathbf{X}\hat{\mathbf{b}})'(\mathbf{X}(\hat{\mathbf{b}} - \mathbf{b}))]' = (\mathbf{X}(\hat{\mathbf{b}} - \mathbf{b}))'(\mathbf{Y} - \mathbf{X}\hat{\mathbf{b}}).$$

Hence it is enough to show that

$$(\mathbf{Y} - \mathbf{X}\hat{\mathbf{b}})'(\mathbf{X}(\hat{\mathbf{b}} - \mathbf{b})) = \mathbf{0}.$$

We have

$$(\mathbf{Y} - \mathbf{X}\hat{\mathbf{b}})'(\mathbf{X}(\hat{\mathbf{b}} - \mathbf{b})) = (\mathbf{Y}' - \hat{\mathbf{b}}'\mathbf{X}')(\mathbf{X}(\hat{\mathbf{b}} - \mathbf{b})) = (\mathbf{Y}'\mathbf{X} - \hat{\mathbf{b}}'\mathbf{X}'\mathbf{X})(\hat{\mathbf{b}} - \mathbf{b}).$$

We use now that $\hat{\mathbf{b}}$ is a solution of the normal equations. That is,

$$\mathbf{X}'\mathbf{X}\hat{\mathbf{b}} = \mathbf{X}'\mathbf{Y}.$$

Therefore,

$$\hat{\mathbf{b}}'\mathbf{X}'\mathbf{X} = \mathbf{Y}'\mathbf{X}$$

and

$$(\mathbf{Y} - \mathbf{X}\hat{\mathbf{b}})'(\mathbf{X}(\hat{\mathbf{b}} - \mathbf{b})) = (\mathbf{Y}'\mathbf{X} - \hat{\mathbf{b}}'\mathbf{X}'\mathbf{X})(\hat{\mathbf{b}} - \mathbf{b}) = (\mathbf{Y}'\mathbf{X} - \mathbf{Y}'\mathbf{X})(\hat{\mathbf{b}} - \mathbf{b}) = \mathbf{0}.$$

Hence,

$$S(\mathbf{b}) = (\mathbf{Y} - \mathbf{X}\hat{\mathbf{b}})'(\mathbf{Y} - \mathbf{X}\hat{\mathbf{b}}) + (\mathbf{X}(\hat{\mathbf{b}} - \mathbf{b}))'(\mathbf{X}(\hat{\mathbf{b}} - \mathbf{b})).$$

Observe now that $\mathbf{a}'\mathbf{a} \geq 0$ for any column vector $\mathbf{a}$. Therefore,

$$(\mathbf{X}(\hat{\mathbf{b}} - \mathbf{b}))'(\mathbf{X}(\hat{\mathbf{b}} - \mathbf{b})) \geq \mathbf{0}.$$

This implies that for any column vector $\mathbf{b}$ with n components we have

$$S(\mathbf{b}) \geq (\mathbf{Y} - \mathbf{X}\hat{\mathbf{b}})'(\mathbf{Y} - \mathbf{X}\hat{\mathbf{b}}) = \mathbf{S}(\hat{\mathbf{b}}).$$

That is, S has a minimum at $\hat{\mathbf{b}}$. Moreover, this minimum is unique since S is differentiable everywhere and has a unique critical point at $\hat{\mathbf{b}}$.

In Chap. 6 we have dealt with simple linear regression. That is, the case $p = 2$. In the next example we check that our matrix computations yield the same formula we had in Chap. 6.

Example 4. Consider the case $p = 2$. That is, the model is

$$Y = b_0 + b_1 X.$$

We have n observations of (Y, X) denoted by (y_i, x_i) for $i = 1, 2, \ldots, n$. The matrix notation for this particular case is

$$\mathbf{Y} = \begin{pmatrix} y_1 \\ y_2 \\ . \\ . \\ y_n \end{pmatrix} \quad \mathbf{b} = \begin{pmatrix} b_0 \\ b_1 \end{pmatrix} \quad \mathbf{X} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ . & . \\ . & . \\ 1 & x_n \end{pmatrix}.$$

We have

$$\mathbf{X}'\mathbf{X} = \begin{pmatrix} n & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{pmatrix}.$$

Recall from linear algebra that a matrix

$$\mathbf{A} = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

is invertible if and only if $det(\mathbf{A}) = ad - bc \neq 0$. If so the inverse matrix is

$$\mathbf{A}^{-1} = \frac{1}{\det(\mathbf{A})} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}.$$

We go back to $\mathbf{X}'\mathbf{X}$. Its determinant is

$$n \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i \right)^2.$$

It turns out that this determinant is 0 if and only if all the x_i are equal (see the exercises). Assuming that this is not the case we have

$$(\mathbf{X}'\mathbf{X})^{-1} = \frac{1}{\det(\mathbf{X}'\mathbf{X})} \begin{pmatrix} \sum_{i=1}^n x_i^2 & -\sum_{i=1}^n x_i \\ -\sum_{i=1}^n x_i & n \end{pmatrix}.$$

We also have

$$\mathbf{X}'\mathbf{Y} = \begin{pmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n x_i y_i \end{pmatrix}.$$

By the normal equations

$$\hat{\mathbf{b}} = \begin{pmatrix} \hat{b}_o \\ \hat{b}_1 \end{pmatrix} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

$$= \frac{1}{\det(\mathbf{X}'\mathbf{X})} \begin{pmatrix} \sum_{i=1}^n x_i^2 \sum_{i=1}^n y_i - (\sum_{i=1}^n x_i)(\sum_{i=1}^n x_i y_i) \\ n \sum_{i=1}^n x_i y_i - (\sum_{i=1}^n x_i)(\sum_{i=1}^n y_i) \end{pmatrix}.$$

It is now easy to check that the formula is the same as the one used in Chap. 6.

Exercises 10.1

1. Give an example of a function with a critical point and no minimum or maximum.

2. Let **a** and **b** be two column vectors with n components each.

(a) Is $\mathbf{ab}'$ a number, a matrix? Specify the dimension if it is a matrix.
(b) Same question for $\mathbf{a}'\mathbf{b}$.
(c) What is the (i, j) term of $\mathbf{ab}'$?

3. Show that if **a** is a column vector then $\mathbf{a}'\mathbf{a} \geq 0$.

4. In Example 3 the least square method tells us that the best linear approximation of Y using the variables X_1, X_2, X_3, and X_4 is given by the equation

$$Y = 284.5 - 0.19X_1 - 0.32X_2 - 1.7X_3 - 0.59X_4.$$

(a) Explain why we should expect the coefficients of X_1, X_2, X_3, X_4 to be negative.
(b) In your opinion what are the most important variables in explaining Y? Explain your reasoning.

5. Compute the best linear approximation of Y using only the variables X_3 and X_4 for Example 3.

6. Compute the best linear approximation of Y using only the variable X_4 for Example 3.

7. A square matrix P is said to be idempotent if $P^2 = P \times P = P$.

(a) Give an example of a 2×2 matrix which is idempotent.
(b) Let I be the unit matrix (it has a 1 on every diagonal entry and a 0 everywhere else) with the same dimension as P. Show that if P is idempotent so is $I - P$.
(c) If P is idempotent compute P^n for every natural n.

8. (a) Find a matrix P such that

$$\hat{\mathbf{Y}} = P\mathbf{Y}.$$

(b) Show that P is idempotent.

9. In this exercise we prove that the determinant in Example 4 is not 0 unless all the x_i are equal.

Let $x_1, x_2, \ldots, x_n$ and $y_1, x_2, \ldots, y_n$ be two fixed finite sequences of real numbers. Define the function R by

$$R(u) = \sum_{i=1}^{n} (ux_i + y_i)^2.$$

(a) Show that

$$R(u) = u^2 \sum_{i=1}^{n} x_i^2 + 2u \sum_{i=1}^{n} x_i y_i + \sum_{i=1}^{n} y_i^2.$$

(b) Show that $R(v) = 0$ for some v if and only if for every $i = 1, \ldots, n$ we have $y_i = -vx_i$. If that is the case the sequences x_i and y_i are said to be proportional.
(c) Show that if the sequences x_i and y_i are not proportional then $R(u) > 0$ for every u.
(d) Show that if the sequences x_i and y_i are not proportional then

$$\left(\sum_{i=1}^{n} x_i y_i \right)^2 < \sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i^2.$$

This is Cauchy's inequality. To prove the inequality note that R is a second degree polynomial (as a function of u). For it to be always strictly positive its discriminant needs to be negative.
(e) Use Cauchy's inequality with $y_i = 1$ for all i to show that the determinant in Example 4 is not 0 unless all the x_i are equal.

10. Let A be a matrix.

(a) Show that $(A')' = A$.
(b) A matrix B is said to symmetric if $B' = B$. Give an example of a 3×3 matrix which is symmetric.
(c) Let B be a matrix. Show that $B'B$ is a symmetric matrix.

11. Let A be an invertible matrix.

(a) Show that A' is also invertible and $(A')^{-1} = (A^{-1})'$. (Start with $AA^{-1} = I$ where I is the identity matrix and then take transposes on both sides of the equality).
(b) Show that if A is symmetric (i.e., $A' = A$) then so is A^{-1}.

10.2 Statistical Inference

The least squares method used in the preceding section gives a linear equation to explain how Y varies as a function of one or several variables X. In this section we will use statistical inference to decide how good this equation is. In order to do so we need an underlying probability model that we now formulate. We will assume that the variable Y we wish to explain is random and that the explanatory variables $X_1, \ldots, X_p$ are deterministic (i.e., nonrandom). Assume that we have a sample of n observations:

$$(y_i, x_{i,1}, x_{i,2}, \ldots, x_{i,p-1}) \text{ for } i = 1, 2, \ldots, n,$$

where the y_i are observations of Y, the $x_{i,1}$ are observations of the variable X_1, the $x_{i,2}$ are observations of the variable X_2 and so on. Recall the notation from Sect. 10.1:

$$\mathbf{Y} = \begin{pmatrix} y_1 \\ y_2 \\ . \\ . \\ y_n \end{pmatrix} \quad \mathbf{b} = \begin{pmatrix} b_0 \\ b_1 \\ . \\ . \\ b_{p-1} \end{pmatrix} \quad \mathbf{X} = \begin{pmatrix} 1 & x_{1,1} & x_{1,2} & \ldots & x_{1,p-1} \\ 1 & x_{2,1} & x_{2,2} & \ldots & x_{2,p-1} \\ 1 & x_{3,1} & x_{3,2} & \ldots & x_{3,p-1} \\ . & . & . & \ldots & . \\ 1 & x_{n,1} & x_{n,2} & \ldots & x_{n,p-1} \end{pmatrix}.$$

We now state our assumptions. They will be in force for the rest of this chapter.

The Model

We assume the following model

$$\mathbf{Y} = \mathbf{Xb} + \mathbf{e},$$

where $\mathbf{e}$ is a column vector with n components e_i. The random vector $\mathbf{e}$ is assumed to be *normal* and

$$E(\mathbf{e}) = \mathbf{0},$$

where $\mathbf{0}$ is the 0 vector with n components. Moreover, $\mathbf{e}$ has a variance matrix

$$\text{Var}(\mathbf{e}) = \sigma^2 \mathbf{I}_n,$$

where $\mathbf{I}_n$ is the $n \times n$ identity matrix (it has a 1 at each diagonal entry and 0's everywhere else) and $\sigma > 0$ is a parameter that will need to be estimated.

There are several important consequences of these assumptions that we now review.

1. Using that $\mathbf{X}$ and $\mathbf{b}$ are not random we have

$$E(\mathbf{Y}) = \mathbf{E}(\mathbf{Xb}) + \mathbf{E}(\mathbf{e}) = \mathbf{Xb}.$$

   Hence, the model $\mathbf{Y} = \mathbf{Xb} + \mathbf{e}$ implies that $E(\mathbf{Y})$ is a linear function of the variables X and $\mathbf{e}$ represents the random fluctuations of $\mathbf{Y}$ around its expected value.
2. Since $\text{Var}(\mathbf{e})$ is assumed to be diagonal we have that all covariances $\text{Cov}(e_i, e_j)$ for $i \neq j$ are 0. Using that $\mathbf{e}$ is a normal vector this implies that e_i and e_j are independent for all $i \neq j$.
3. For every i we have $\text{Var}(e_i) = \sigma^2$. That is, all the e_i are assumed to have the same variance.
4. Since $\mathbf{e}$ is a normal vector the vector $\mathbf{Y}$

$$\mathbf{Y} = \mathbf{Xb} + \mathbf{e},$$

   is also normal and its variance is

$$\text{Var}(\mathbf{Y}) = \text{Var}(\mathbf{e}) = \sigma^2 \mathbf{I}_n.$$

In particular, Y_i and Y_j are independent for all $i \neq j$.

When doing a linear regression the first task is to test whether $b_1 = b_2 = \cdots = b_{p-1} = 0$. If we cannot reject $b_1 = b_2 = \cdots = b_{p-1} = 0$ we can conclude that the model is not adequate. That is, our assumption that $E(Y)$ is a linear function of the X's is not adequate. We now construct such a test.

Let

$$\mathbf{1} = \begin{pmatrix} 1 \\ 1 \\ . \\ . \\ 1 \end{pmatrix}$$

be a column vector with n components all equal to 1 and let

$$\bar{\mathbf{Y}} = \begin{pmatrix} \bar{y} \\ \bar{y} \\ . \\ . \\ \bar{y} \end{pmatrix}$$

be a column vector with n components all equal to $\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$. We now define the sums of squares that will determine how fit the model is.

$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2 = (\mathbf{Y} - \bar{\mathbf{Y}})'(\mathbf{Y} - \bar{\mathbf{Y}}),$$

where *SST* is called the total sum of squares. Let

$$\hat{\mathbf{Y}} = \mathbf{X}\hat{\mathbf{b}}.$$

That is, $\hat{\mathbf{Y}}$ is the vector of y's predicted by the least squares method. If $\hat{\mathbf{Y}}$ is close enough to $\mathbf{Y}$ (the observed $y's$) then the model is probably adequate. To measure closeness we define another sum of squares

$$SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = (\mathbf{Y} - \hat{\mathbf{Y}})'(\mathbf{Y} - \hat{\mathbf{Y}}).$$

The sum of squares *SSE* is called the residual error sum of squares. This is so because each $y_i - \hat{y}_i$ represents the "error" made by the linear approximation. The smaller *SSE* is compared to *SST* the better the model is. The third sum of squares is defined by

$$SSR = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 = (\hat{\mathbf{Y}} - \bar{\mathbf{Y}})'(\hat{\mathbf{Y}} - \bar{\mathbf{Y}}),$$

the sum of squares due to regression. Next we state the formula relating the different sums of squares.

Partitioning the Total Sum of Squares

We have the following partitioning of the total sum of squares:

$$SST = SSE + SSR.$$

We now prove this formula. We subtract and add $\hat{\mathbf{Y}}$ to get

$$SST = (\mathbf{Y} - \bar{\mathbf{Y}})'(\mathbf{Y} - \bar{\mathbf{Y}}) = (\mathbf{Y} - \hat{\mathbf{Y}} + \hat{\mathbf{Y}} - \bar{\mathbf{Y}})'(\mathbf{Y} - \hat{\mathbf{Y}} + \hat{\mathbf{Y}} - \bar{\mathbf{Y}}).$$

We expand the product to get

$$SST = (\mathbf{Y} - \hat{\mathbf{Y}})'(\mathbf{Y} - \hat{\mathbf{Y}}) + (\mathbf{Y} - \hat{\mathbf{Y}})'(\hat{\mathbf{Y}} - \bar{\mathbf{Y}}) + (\hat{\mathbf{Y}} - \bar{\mathbf{Y}})'(\mathbf{Y} - \hat{\mathbf{Y}}) + (\hat{\mathbf{Y}} - \bar{\mathbf{Y}})'(\hat{\mathbf{Y}} - \bar{\mathbf{Y}}).$$

The first term in the r.h.s. is *SSE* and the last term is *SSR*. So we only need to show that the two middle terms are 0. Note that the second term is the transpose of the third. Hence it is enough to show that the third term is 0. We now do that.

$$(\hat{\mathbf{Y}} - \bar{\mathbf{Y}})'(\mathbf{Y} - \hat{\mathbf{Y}}) = \hat{\mathbf{Y}}'(\mathbf{Y} - \hat{\mathbf{Y}}) - \bar{\mathbf{Y}}'(\mathbf{Y} - \hat{\mathbf{Y}}). \tag{10.1}$$

Recall the normal equations

$$\mathbf{X}'\mathbf{X}\hat{\mathbf{b}} = \mathbf{X}'\mathbf{Y}.$$

Since

$$\mathbf{X}'\mathbf{Y} - \mathbf{X}'\mathbf{X}\hat{\mathbf{b}} = \mathbf{X}'(\mathbf{Y} - \mathbf{X}\hat{\mathbf{b}}) = \mathbf{X}'(\mathbf{Y} - \hat{\mathbf{Y}})$$

we get

$$\mathbf{X}'(\mathbf{Y} - \hat{\mathbf{Y}}) = \mathbf{0}, \tag{10.2}$$

where $\mathbf{0}$ is a column vector with all its p components equal to 0. Going back to (10.1) and using (10.2) we have

$$\hat{\mathbf{Y}}'(\mathbf{Y} - \hat{\mathbf{Y}}) = (\mathbf{X}\hat{\mathbf{b}})'(\mathbf{Y} - \hat{\mathbf{Y}}) = \hat{\mathbf{b}}'\mathbf{X}'(\mathbf{Y} - \hat{\mathbf{Y}}) = \hat{\mathbf{b}}'\mathbf{0} = 0.$$

Finally, we need to show that the second term in the r.h.s of (10.1) is also 0. Recall that the first row of $\mathbf{X}'$ has only 1's. Hence the first component of the column vector $\mathbf{X}'(\mathbf{Y} - \hat{\mathbf{Y}})$ is the sum of residuals $\sum_{i=1}^{n}(y_i - \hat{y}_i)$. Using (10.2) we get

$$\sum_{i=1}^{n}(y_i - \hat{y}_i) = 0.$$

Recalling that $\bar{\mathbf{Y}}$ is a column vector whose components are all $\bar{y}$ we get

$$\bar{\mathbf{Y}}'(\mathbf{Y} - \hat{\mathbf{Y}}) = \sum_{i=1}^{n}\bar{y}(y_i - \hat{y}_i) = \bar{y}\sum_{i=1}^{n}(y_i - \hat{y}_i) = \bar{y} \times 0 = 0.$$

This completes the proof of the formula $SST = SSR + SSE$.

Example 1. Here are the three sums of squares for Example 3 in 10.1. We have

$$SST = 537,638,$$

$$SSE = 129,914$$

and

$$SSR = 407,724.$$

Note that $SSE + SSR = 537,638$ which is equal to SST as it should.

A first measure of the goodness of fit of the linear model is the following statistic.

The R^2 Statistic

The R^2 statistic is defined by

$$R^2 = \frac{SSR}{SST}.$$

The coefficient R is always in $[0, 1]$. The closer it is to 1 the better the linear model fits the data.

Example 2. We compute R^2 for our example. For example 1 we have

$$R^2 = \frac{SSR}{SST} = \frac{407,724}{537,638} = 0.76$$

and taking the square root we get $R = 0.87$ which is pretty high and shows a good fitness of the linear model.

Remark 2. It is easy to artificially increase R^2 making the model appear better than it is. Every time we add an explanatory variable R^2 increases. So if we add enough explanatory variables (even if they have nothing to do with the model!) we can get as close to 1 as we want. This, of course, is not recommended. The point of a mathematical model is to explain something in the simplest possible way. Hence, one should strive to keep p as low as possible.

We are now ready to test whether the model is adequate.

Testing the Model

To test

$$H_0 : b_1 = b_2 = \cdots = b_p = 0$$

against

$$H_a : \text{at least one } b_i \text{ for } 1 \leq i \leq p \text{ is not } 0$$

we use the statistic

$$F = \frac{SSR/(p-1)}{SSE/(n-p)},$$

where $p - 1$ is the number of explanatory variables $X_1, X_2, \ldots, X_{p-1}$ and n is the number of observations in the sample. Under the null hypothesis H_0, F follows an F distribution with degrees $(p - 1, n - p)$. The P-value of the test is given by

$$P = P(F(p-1, n-p) > F).$$

Recall that $\mathbf{e}$ is normal with mean $\mathbf{0}$ and variance $\sigma^2\mathbf{I_n}$. Under the null hypothesis *SSR* and *SSE* are independent and are distributed according to a Chi-Square distribution with degrees $p - 1$ and $n - p$, respectively. This is why F follows an F distribution with degrees $p - 1$ and $n - p$. The proof of this fact as well as the other proofs we omit in this chapter are very well done in Chaps. 2 and 3 in *Linear Models* by S.R.Searle, 1971, John Wiley.

Example 3. We now perform the test on our example. We have $p = 5$ and $n = 163$. The statistic F is

$$F = \frac{SSR/(p-1)}{SSE/(n-p)} = \frac{SSR/4}{SSE/158} = 124.$$

The P value

$$P(F(4, 158) > 124)$$

is very small: the F table yields

$$P(F(4, 100) > 5.02) = 0.001 \text{ and } P(F(4, 200) > 4.81) = 0.001.$$

Therefore, $P(F(5, 158) > 210)$ is much smaller than 0.001. Hence we reject the null hypothesis. This tells us only that at least one of the b_i for $1 \leq i \leq 5$ is significantly different from 0. We are now going to construct a test to test individual b_i's. For a given i, if we cannot reject $b_i = 0$ it means that the corresponding X_i does not contribute significantly in explaining the variable Y.

Recall from last section that the least squares estimate of $\mathbf{b}$ is

$$\hat{\mathbf{b}} = \mathbf{AY},$$

where

$$\mathbf{A} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$$

is a nonrandom matrix. Hence

$$E(\hat{\mathbf{b}}) = \mathbf{A}E(\mathbf{Y}) = \mathbf{AXb} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Xb} = \mathbf{b}.$$

This proves the following result.

Unbiased Estimator

Under the assumptions of the model the least squares estimate

$$\hat{\mathbf{b}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

is an unbiased estimator of $\mathbf{b}$.

We now turn to $\text{Var}(\hat{\mathbf{b}})$. Recall from Sect. 8.3 that if $\mathbf{A}$ is nonrandom matrix and $\mathbf{T}$ is a random matrix then

$$\text{Var}(\mathbf{AT}) = \mathbf{A}\text{Var}(\mathbf{T})\mathbf{A}'.$$

Writing again

$$\hat{\mathbf{b}} = \mathbf{AY},$$

where

$$\mathbf{A} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$$

we get

$$\text{Var}(\hat{\mathbf{b}}) = \mathbf{A}\text{Var}(\mathbf{Y})\mathbf{A}' = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\sigma^2\mathbf{I}_n[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']'.$$

Observe now that $\mathbf{X}'\mathbf{X}$ is a symmetric matrix (i.e., it is equal to its transpose). The inverse (if it exists) of a symmetric matrix is also symmetric hence

$$[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']' = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1},$$

and

$$\text{Var}(\hat{\mathbf{b}}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}.$$

We state the result we just proved.

Variance of the Estimator

Under the assumptions of the model we have

$$\text{Var}(\hat{\mathbf{b}}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}.$$

From the result above we see that the variance of $\hat{b}_i$ (the i component of $\hat{\mathbf{b}}$) is

$$\text{Var}(\hat{b}_i) = \sigma^2 c_{ii}$$

where c_{ii} is the i term in the diagonal of $(\mathbf{X}'\mathbf{X})^{-1}$. In order to test b_i we still need an estimate for σ^2.

Estimating the Variance

An unbiased estimator of σ^2 is

$$\hat{\sigma^2} = \frac{SSE}{n-p}.$$

We will prove this in the exercises.

Example 4. For our example

$$\hat{\sigma^2} = \frac{SSE}{n-p} = \frac{129,914}{158} = 822,$$

and an estimate for σ is then $\sqrt{822} = 28$.

We are now ready to test individual b_i's.

The Distribution of $\hat{b}_i$

Under the normal assumptions of the model for $i = 0, 1, \ldots, p-1$

$$\frac{\hat{b}_i - b_i}{\sqrt{\hat{\sigma}^2 c_{ii}}}$$

follows a Student distribution with $n - p$ degrees of freedom, where c_{ii} is the i term in the diagonal of $(\mathbf{X}'\mathbf{X})^{-1}$.

The property above is a consequence of the following facts: $\hat{\mathbf{b}}$ is a normal vector (why?), $\hat{\mathbf{b}}$ and *SSE* are independent and *SSE* has a Chi-square distribution with $n-p$ degrees of freedom. The proofs can be found in Searle (1971).

We now apply this property to test individual b_i's in our example.

Example 5. Recall that X_1 is the percentage of children vaccinated against measles. We perform the test

$$H_0 : b_1 = 0$$
$$H_a : b_1 \neq 0.$$

From Example 3 in 10.1 we have

$$(\mathbf{X}'\mathbf{X})^{-1} = 10^{-5} \begin{pmatrix} 35713 & -171 & -126 & -229 & 146 \\ -171 & 26 & -22 & -2 & 0 \\ -126 & -22 & 23 & 0 & 0 \\ -229 & -2 & 0 & 7 & -3 \\ 146 & 0 & 0 & -3 & 2 \end{pmatrix}.$$

Since we numbered our b_i's starting at $i = 0$ the rows and columns of the matrix also start at 0. Hence, we read $c_{11} = 26 \times 10^{-5}$.

$$\sqrt{\hat{\sigma}^2 c_{11}} = \sqrt{822 \times 26 \times 10^{-5}} = 0.46.$$

Hence,

$$\frac{\hat{b}_1}{\sqrt{\hat{\sigma}^2 c_{11}}} = \frac{-0.19}{0.46} = -0.41,$$

where $\hat{b}_1$ was computed in Example 3 in 10.1. Therefore, the P value for this two-sided test is

$$P = 2P(t(158) < -0.41) = 0.68.$$

We cannot reject the null hypothesis. Hence, the measles vaccination rate does not appear to have a significant role in the rate of infant mortality.

10.2.1 Geometric Interpretation

As seen above a consequence of the normal equations is (2)

$$\mathbf{X}'(\mathbf{Y} - \hat{\mathbf{Y}}) = \mathbf{0}.$$

This shows that the dot product of each column of $\mathbf{X}$ with the residual vector $\mathbf{Y} - \hat{\mathbf{Y}}$ is 0. That is the residual vector is perpendicular to each column vector of $\mathbf{X}$ and therefore to the vector space spanned by the column vectors of $\mathbf{X}$. Since $\hat{\mathbf{Y}} = \mathbf{X}\hat{\mathbf{b}}$ we also know that $\hat{\mathbf{Y}}$ belongs to the vector space spanned by the column vectors of $\mathbf{X}$. This implies that the predicted vector $\hat{\mathbf{Y}}$ is the orthogonal projection of $\mathbf{Y}$ on this vector space.

Remark 3. The different tests we described in this section all depend on the assumptions about the model. In practice it is important to check these assumptions as we did in 6.2 for the simple linear regression. In particular one should perform an analysis of the residuals $y_i - \hat{y}_i$. More precisely, one should check whether the residuals appear to have equal variance, are uncorrelated and are normally distributed. See for instance Chap. 7 in *A second course in statistics* (Fifth edition) by W. Mendenhall and T. Sincich.

Exercises 10.2

1. (a) In Example 5 above we tested the relevance of X_1 in the model. Test the relevance of the other variables.
(b) Consider the model explaining Y with only variables X_3 and X_4. Compute the new $(\mathbf{X}'\mathbf{X})^{-1}$, $\hat{\mathbf{b}}$, $\hat{\sigma^2}$.
(c) Compute R^2 for the model in (b) and compare it to the R^2 for the full model that was computed in Example 2.
(d) For the model in (b) test whether $b_3 = b_4 = 0$ and test also whether the individual b_i's are 0.
(e) Consider a model explaining Y with only the variable X_4. Compare this model to the model in (b).
(f) Discuss the practical implications of your findings.

2. (a) Use the World Health Organization data at http://www.who.int/whosis/whostat/2010/en/index.html to get a linear equation explaining life expectancy at birth using the following explanatory variables: under 1 mortality rate, under 5 mortality rate and adult mortality rate.

(b) Perform all the relevant tests.
(c) Based on the tests performed in (b) decide whether you should eliminate one or more explanatory variables. If so compare the full model to the reduced model.
(d) Discuss the practical implications of your findings.

3. Show that $\hat{\mathbf{b}}$ is a normal vector.

4. (a) Show that

$$SST - SSR = \mathbf{Y}'\mathbf{Y} - \hat{\mathbf{Y}}'\hat{\mathbf{Y}} - \mathbf{2}(\mathbf{Y} - \hat{\mathbf{Y}})'\bar{\mathbf{Y}}.$$

(b) Use (a) to conclude that

$$SST - SSR = \mathbf{Y}'\mathbf{Y} - \hat{\mathbf{Y}}'\hat{\mathbf{Y}}.$$

(Recall that the sum of residuals is 0).
(c) Use (b) to show that

$$\mathbf{Y}'\mathbf{Y} = (\mathbf{Y} - \hat{\mathbf{Y}})'(\mathbf{Y} - \hat{\mathbf{Y}}) + \hat{\mathbf{Y}}'\hat{\mathbf{Y}}.$$

5. (a) Use that

$$\mathbf{X}'(\mathbf{Y} - \hat{\mathbf{Y}}) = \mathbf{0}$$

to get

$$\hat{\mathbf{Y}}'(\mathbf{Y} - \hat{\mathbf{Y}}) = \mathbf{0}.$$

(b) Give a geometrical interpretation of (b).

6. Recall that

$$R^2 = \frac{SSR}{SST} \text{ and } F = \frac{SSR/(p-1)}{SSE/(n-p)}.$$

Show that

$$F = \frac{R^2/(p-1)}{(1-R^2)(n-p)}.$$

7. In this exercise we prove that $SSE/(n-p)$ is an unbiased estimator of σ^2.

(a) Show that

$$\hat{\mathbf{Y}} = \mathbf{P}\mathbf{Y},$$

where

$$\mathbf{P} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}.$$

(b) Show that

$$SSE = \mathbf{Y}'(\mathbf{I_n} - \mathbf{P})\mathbf{Y},$$

where $\mathbf{I_n}$ is the $n \times n$ identity matrix. (Use 4 (a)).

(c) Show that

$$E(\mathbf{Y}'(\mathbf{I_n} - \mathbf{P})\mathbf{Y}) = E(\mathbf{Y}')(\mathbf{I_n} - \mathbf{P})E(\mathbf{Y}) + \sigma^2 \text{trace}(\mathbf{I_n} - \mathbf{P}).$$

(Use that if $\mathbf{T}$ is a random vector and A is a nonrandom matrix then $E(\mathbf{T}'A\mathbf{T}) = E(\mathbf{T}')AE(\mathbf{T}) + \text{trace}(A\text{Var}(\mathbf{T}))$, where the trace of a matrix is the sum of its diagonal terms.)

(d) Show that

$$(\mathbf{I_n} - \mathbf{P})E(\mathbf{Y}) = 0$$

and hence

$$E(\mathbf{Y}'(\mathbf{I_n} - \mathbf{P})\mathbf{Y}) = \sigma^2 \text{trace}(\mathbf{I_n} - \mathbf{P}).$$

(e) Show that

$$\text{trace}(\mathbf{P}) = \text{trace}(\mathbf{I_p}) = p.$$

(You may use that $\text{trace}(AB) = \text{trace}(BA)$ for any same size square matrices A and B).

(f) Use (e) to show that

$$E(\mathbf{Y}'(\mathbf{I_n} - \mathbf{P})\mathbf{Y}) = \sigma^2(n - p).$$

(g) Show that $SSE/(n - p)$ is an unbiased estimator of σ^2.

Further Reading

Probability

The following two references are at a slightly higher level than this book. They cover additional topics and examples in probability. They are:

The essentials of probability by R. Durrett (The Duxbury Press)

Probability by J. Pitman (Springer Verlag).

An introduction to probability theory and its applications by W. Feller (Volume I, third edition, Wiley) is at a substantial higher level than this book. It has influenced several generations of probabilists and covers hundreds of interesting topics. It is a GREAT book.

Statistics

A very good elementary introduction to statistics is *Introduction of the practice of statistics* by D. Moore and G. McCabe (second edition, Freeman). A more mathematical approach to statistics is contained in *Probability and Statistics* by K. Hastings (Addison-Wesley).

At an intermediate level the reader may read *Introduction to the Theory of statistics* by A.M. Mood, F.A. Graybill and D.C. Boes (third edition, McGraw Hill) and *Mathematical statistics and data analysis* by J.A. Rice (third edition, Thomson).

For simple and multiple linear regression *A second course in statistics* (Fifth edition) by W. Mendenhall and T. Sincich is a good text focussing on the applied side of things. For the theory *Linear Models* by S.R. Searle, 1971, John Wiley is a good text. The reader will find all the proofs that we omitted in Chap. 10.

To find the mathematical proofs that were omitted in Chap. 9 as well as many other results the reader may consult *The theory of statistical inference* by S. Zacks (Wiley). This is an advanced text.

R.B. Schinazi, *Probability with Statistical Applications*,
DOI 10.1007/978-0-8176-8250-7, © Springer Science+Business Media, LLC 2012

Common Distributions

Discrete Distributions

Bernoulli with parameter p:

$$P(X = 0) = 1 - p \text{ and } P(X = 1) = p$$
$$E(X) = p$$
$$\text{Var}(X) = p(1 - p)$$
$$M_X(t) = 1 - p + pe^t.$$

Binomial with parameters n and p:

$$P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k} \text{ for } k = 0, 1, \ldots, n.$$
$$E(X) = np$$
$$\text{Var}(X) = np(1 - p)$$
$$M_X(t) = (1 - p + pe^t)^n.$$

Geometric with parameter p:

$$P(X = k) = (1 - p)^{k-1} p \text{ for } k = 0, 1, \ldots.$$
$$E(X) = \frac{1}{p}$$
$$\text{Var}(X) = \frac{1 - p}{p^2}$$
$$M_X(t) = \frac{pe^t}{1 - (1 - p)e^t}.$$

R.B. Schinazi, *Probability with Statistical Applications*,
DOI 10.1007/978-0-8176-8250-7, © Springer Science+Business Media, LLC 2012

Poisson with parameter λ:

$$P(X = k) = \mathrm{e}^{-\lambda}\frac{\lambda^k}{k!} \text{ for } k = 0, 1, \ldots.$$

$$E(X) = \lambda$$

$$\mathrm{Var}(X) = \lambda$$

$$M_X(t) = \exp(\lambda(\mathrm{e}^t - 1)).$$

Continuous Distributions

Beta with parameters (a, b):

$$f(x) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}x^{a-1}(1-x)^{b-1} \text{ for } 0 < x < 1$$

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

$$\mathrm{Var}(X) = \frac{ab}{(a+b)^2(a+b+1)}$$

Exponential with parameter a:

$$f(x) = a\mathrm{e}^{-ax} \text{ for } x > 0$$

$$E(X) = \frac{1}{a}$$

$$\mathrm{Var}(X) = \frac{1}{a^2}$$

$$M_X(t) = \frac{a}{a-t} \text{ for } t < a.$$

Gamma with parameters (r, λ):

$$f(x) = \frac{\lambda^r}{\Gamma(r)}x^{r-1}\mathrm{e}^{-\lambda x} \text{ for } x > 0$$

$$E(X) = \frac{r}{\lambda}$$

$$\mathrm{Var}(X) = \frac{r}{\lambda^2}$$

$$M_X(t) = \left(\frac{\lambda}{\lambda - t}\right)^r \text{ for } t < \lambda.$$

Normal with mean μ and standard deviation σ:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \mathrm{e}^{-\frac{1}{2\sigma^2}(x-\mu)^2} \text{ for } -\infty < x < +\infty$$

$$E(X) = \mu$$

$$\mathrm{Var}(X) = \sigma^2$$

$$M_X(t) = \mathrm{e}^{\mu t + \frac{1}{2}\sigma^2 t^2}.$$

Standard normal:

$$f(x) = \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{1}{2}x^2} \text{ for } -\infty < x < +\infty$$

$$E(X) = 0$$

$$\mathrm{Var}(X) = 1$$

$$M_X(t) = \mathrm{e}^{\frac{1}{2}t^2}.$$

Uniform on $[a, b]$:

$$f(x) = \frac{1}{b-a} \text{ for } a < x < b$$

$$E(X) = \frac{a+b}{2}$$

$$\mathrm{Var}(X) = \frac{(b-a)^2}{12}.$$

Normal Table

The table below gives $P(0 < Z < z)$ for a standard normal random variable Z. For instance $P(0 < Z < 0.43) = 0.1664$.

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0199	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.1480	0.1517
0.4	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1772	0.1808	0.1844	0.1879
0.5	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.2190	0.2224
0.6	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2517	0.2549
0.7	0.2580	0.2611	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2823	0.2852
0.8	0.2881	0.2910	0.2939	0.2967	0.2995	0.3023	0.3051	0.3078	0.3106	0.3133
0.9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.3340	0.3365	0.3389
1.0	0.3413	0.3438	0.3461	0.3485	0.3508	0.3531	0.3554	0.3577	0.3599	0.3621
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.3770	0.3790	0.3810	0.3830
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.3980	0.3997	0.4015
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319
1.5	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4535	0.4545
1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633
1.8	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4699	0.4706
1.9	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.4750	0.4756	0.4761	0.4767
2.0	0.4772	0.4778	0.4783	0.4788	0.4793	0.4798	0.4803	0.4808	0.4812	0.4817
2.1	0.4821	0.4826	0.4830	0.4834	0.4838	0.4842	0.4846	0.4850	0.4854	0.4857
2.2	0.4861	0.4864	0.4868	0.4871	0.4875	0.4878	0.4881	0.4884	0.4887	0.4890
2.3	0.4893	0.4896	0.4898	0.4901	0.4904	0.4906	0.4909	0.4911	0.4913	0.4916
2.4	0.4918	0.4920	0.4922	0.4925	0.4927	0.4929	0.4931	0.4932	0.4934	0.4936
2.5	0.4938	0.4940	0.4941	0.4943	0.4945	0.4946	0.4948	0.4949	0.4951	0.4952

(continued)

R.B. Schinazi, *Probability with Statistical Applications*,
DOI 10.1007/978-0-8176-8250-7, © Springer Science+Business Media, LLC 2012

(continued)

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
2.6	0.4953	0.4955	0.4956	0.4957	0.4959	0.4960	0.4961	0.4962	0.4963	0.4964
2.7	0.4965	0.4966	0.4967	0.4968	0.4969	0.4970	0.4971	0.4972	0.4973	0.4974
2.8	0.4974	0.4975	0.4976	0.4977	0.4977	0.4978	0.4979	0.4979	0.4980	0.4981
2.9	0.4981	0.4982	0.4982	0.4983	0.4984	0.4984	0.4985	0.4985	0.4986	0.4986
3.0	0.4987	0.4987	0.4987	0.4988	0.4988	0.4989	0.4989	0.4989	0.4990	0.4990

Student Table

The table below gives t_a such that $P(|t(n)| < t_a) = a$ where $t(n)$ is a Student distribution with n degrees of freedom. For instance, we read that $P(|t(5)| < 1.48) = 0.8$.

n	$a = 0.6$	$a = 0.7$	$a = 0.8$	$a = 0.9$	$a = 0.95$
1	1.38	1.96	3.08	6.31	12.71
2	1.06	1.39	1.89	2.92	4.30
3	0.98	1.25	1.64	2.35	3.18
4	0.94	1.19	1.53	2.13	2.78
5	0.92	1.16	1.48	2.02	2.57
6	0.91	1.13	1.44	1.94	2.45
7	0.90	1.12	1.41	1.89	2.36
8	0.89	1.11	1.40	1.86	2.31
9	0.88	1.10	1.38	1.83	2.26
10	0.88	1.09	1.37	1.81	2.23
11	0.88	1.09	1.36	1.80	2.20
12	0.87	1.08	1.36	1.78	2.18
13	0.87	1.08	1.35	1.77	2.16
14	0.87	1.08	1.35	1.76	2.14
15	0.87	1.07	1.34	1.75	2.13
16	0.86	1.07	1.34	1.75	2.12
17	0.86	1.07	1.33	1.74	2.11
18	0.86	1.07	1.33	1.73	2.10
19	0.86	1.07	1.33	1.73	2.09
20	0.86	1.06	1.33	1.72	2.09
21	0.86	1.06	1.32	1.72	2.08
22	0.86	1.06	1.32	1.72	2.07
23	0.86	1.06	1.32	1.71	2.07
24	0.86	1.06	1.32	1.71	2.06
25	0.86	1.06	1.32	1.71	2.06
∞	0.84	1.04	1.28	1.64	2.01

R.B. Schinazi, *Probability with Statistical Applications*,
DOI 10.1007/978-0-8176-8250-7, © Springer Science+Business Media, LLC 2012

Chi-Square Table

The table below gives χ_a such that $P(\chi(n) < \chi_a) = a$ where $\chi(n)$ is a Chi-Square distribution with n degrees of freedom. For instance, we read that $P(\chi(6) < 1.64) = 0.05$.

n	$a = 0.01$	$a = 0.05$	$a = 0.90$	$a = 0.95$	$a = 0.99$
1	0.00	0.00	2.71	3.84	6.63
2	0.02	0.10	4.61	5.99	9.21
3	0.11	0.35	6.25	7.81	11.34
4	0.30	0.71	7.78	9.49	13.28
5	0.55	1.15	9.24	11.07	15.09
6	0.87	1.64	10.64	12.59	16.81
7	1.24	2.17	12.02	14.07	18.48
8	1.65	2.73	13.36	15.51	20.09
9	2.09	3.33	14.68	16.92	21.67
10	2.56	3.94	15.99	18.31	23.21
11	3.05	4.57	17.28	19.68	24.72
12	3.57	5.23	18.55	21.03	26.22
13	4.11	5.89	19.81	22.36	27.69
14	4.66	6.57	21.06	23.68	29.14
15	5.23	7.26	22.31	25.00	30.58
16	5.81	7.96	23.54	26.30	32.00
17	6.41	8.67	24.77	27.59	33.41
18	7.01	9.39	25.99	28.87	34.81
19	7.63	10.12	27.20	30.14	36.19
20	8.26	10.85	28.41	31.41	37.57
21	8.90	11.59	29.62	32.67	38.93
22	9.54	12.34	30.81	33.92	40.29
23	10.20	13.09	32.01	35.17	41.64
24	10.86	13.85	33.20	36.42	42.98
25	11.52	14.61	34.38	37.65	44.31

R.B. Schinazi, *Probability with Statistical Applications*,
DOI 10.1007/978-0-8176-8250-7, © Springer Science+Business Media, LLC 2012

Index

R.B. Schinazi, *Probability with Statistical Applications*,
DOI 10.1007/978-0-8176-8250-7, © Springer Science+Business Media, LLC 2012