Elementary probability notes[*]

Attila Máté
Brooklyn College of the City University of New York

March 12, 2020

Contents

[*] Written for the course Mathematics 2501 (Elementary Probability) at Brooklyn College of CUNY.

1 Sets

A set is a collection. The members of this collection are called its elements; the symbol $x \in A$ indicates that x is an element of the set A. We can describe a set A by listing its elements inside braces { and }; for example,

$$A = \{1, 3, 5, 7, 9\} \tag{1.1}$$

is the set whose elements are the integers 1, 2, 3, 5, 7, and 9. Certain sets commonly occurring in mathematics have standard notation; for example $\mathbb{Z}$ is the set of all integers (positive, negative, or zero), $\mathbb{Q}$ is the set of all rationals, and $\mathbb{R}$ is the set of all real numbers. Sets can also be described by the set-builder operation:

$$S = \{x : \phi(x)\}$$

denotes the set of all things x for which the condition $\phi(x)$ is satisfied. Here the variable x usually has a certain range, i.e., it can assume certain values specified in the context (for example, one might agree that x is a real number, i.e., that x runs over real numbers). Sometimes one can indicate the range in the set-builder operation. For example, the set A in equation (1.1) can also be described as

$$A = \{x \in \mathbb{Z} : 1 \le x < 10 \,\&\, x \text{ is odd}\}; \tag{1.2}$$

here & is the logical "and." That is, for two statements Φ and Ψ, the statement $\Phi \,\&\, \Psi$ is true only in case both Φ and Ψ are true. The fact that the sets described in formulas (1.1) and (1.2) are the same is a special case of the following

Axiom 1.1 (Axiom of Extensionality)**.** Two sets are equal if and only if they have the same elements.

1.1 The empty set

Once one uses the set-builder operation, it is almost inevitable that one encounters a set with no elements; such a set is called the *empty set*, denoted as $\emptyset$. By the Axiom of Extensionality (Axiom 1.1, the empty set is unique, that is, there is only one empty set. With the set-builder operation, one might occasionally write $\emptyset = \{x : x \neq x\}$. With the listing notation, sometimes one writes $\emptyset = \{\}$, noting that nothing is listed between the braces.

1.2 Relations between sets

Definition 1.1. Given two sets A and B, we say that A is a *subset* of B if every element of A is an elemnt also of B.

In this case, we also say that B is a *superset* of A also this term is used less often than the term subset. The symbol $A \subset B$ expresses the statement that A is a subset of B. We also say that B includes A. The symbol $B \supset A$ can also be used. The use of the word “contain” should be used with extreme care, since it is often misused. B contains A should properly mean that A is an element of B (yes, a set can be an elememnt of another set), but it is often misused to mean that A is a subset of B. Such a misuse should absolutely avoided, The best way to say that $x \in A$ is that x is an element of A, or that x belongs to A.

To illustrate the difference between $\in$ and $\subset$, note that $\emptyset \notin \emptyset$, since the empty set has no element, while $\emptyset \subset \emptyset$ is vacuously true. In fact, for every set A, we have $\emptyset \subset A$ is satisfied: given that $\emptyset$ has no element, the no requirements are imposed on A by saying that every element of $\emptyset$ is also an element of A.

The set $\{\emptyset\}$ is the set whose only element is $\emptyset$; it differs from $\emptyset$, since the former has one element, the latter has none. The sets $\{\emptyset\}$ and $\{\{\emptyset\}\}$ both have one elements. but they are not the same sets according to Axiom 1.1, since their elemnts are not the same, as we saw just before.

1.3 Set operations

Given sets A and B, their *union* $A \cup B$ is the set that *contains* (correct use!) the elements of either:[1.1] that is,

$$A \cup B \overset{def}{=} \{x : x \in A \vee x \in B\},$$

where $\vee$ is the symbol for logical “or.” That is, for two statements Φ and Ψ, the statement $\Phi \vee \Psi$ is true if Φ or Ψ is true.[1.2]

The *intersection* $A \cap B$ of the sets A and B is the set that contains only the elements that belong to both A and B. That is,

$$A \cap B \overset{def}{=} \{x : x \in A \,\&\, x \in B\}.$$

[1.1] One might say, “contains the elements of both,” but such use is ambiguous; this is why we clarify what we mean next.

[1.2] In mathematics, logical “or” is always meant in the inclusive sense; that is $\Phi \vee \Psi$ is true if one of Φ and Ψ is true, or if both are true.

The set difference of A minus B, denoted as $A \setminus B$,[1.3] is defined as the set of those elemments of A that are not elements of B:

$$A \setminus B \overset{\mathit{def}}{=} \{x : x \in A \,\&\, x \notin B\}.$$

One might read the right-hand side here as the set of those elements x for which $x \in A$ *but* $x \notin B$. While the word "but" expresses contrast, its meaning in this context is not any different from that of the word "and."

The symmetric difference of the sets A and B is defined as the set

$$A \bigtriangleup B \overset{\mathit{def}}{=} (A \setminus B) \cup (B \setminus A).$$

If A is a set of sets, i.e., a set all whose elements are also sets then $\bigcup A$ is the union of all elements of x. Formally,

$$\bigcup A \overset{\mathit{def}}{=} \{x : \text{ we have } x \in B \text{ for some } B \in A\},$$

or, even more formally,

$$\bigcup A \overset{\mathit{def}}{=} \{x : (\exists y)(y \in A \,\&\, x \in y)\},$$

where $(\exists y)$ is an *existential quantifier*, to be read as "there is a y such that ...." Using *restricted quantifiers*, this can also be written as

$$\bigcup A \overset{\mathit{def}}{=} \{x : (\exists y \in A)(x \in y)\}.$$

Often, mathematicians not trained in logic, and, for an irrational reason, prefer the notation where the elements of A are indexed. That is, let

$$A = \{B_\iota : \iota \in I\};$$

that is, I is a set indexing the elements of A, and the Greek letter ι (iota) is used to indicate that I may not be a set of integers. Assuming that B_ι is a set for all $\iota \in I$, they prefer to use the symbol

$$\bigcup_{\iota \in I} B_\iota,$$

even though the simpler symbol $\bigcup A$ means the same thing.

Similarly, if A is a set of sets, then $\bigcap A$ can be defined as the intersection of all elements of A:

$$\bigcap A \overset{\mathit{def}}{=} \{x : \text{ we have } x \in B \text{ for all } B \in A\},$$

or, even more formally,

$$\bigcap A \overset{\mathit{def}}{=} \{x : (\forall y)(y \in A \to x \in y)\};$$

here $(\forall y)$ is a universal quantifier, to be read as "for all y we have ...," and for two statements Φ and Ψ, $\Phi \to \Psi$ is the conditional, meaning "if Φ is true then Ψ is also true". The only time $\Phi \to \Psi$ is false is in case Φ is true and Ψ is false. Using *restricted quantifiers*, this can also be written as

$$\bigcap A \overset{\mathit{def}}{=} \{x : (\forall y \in A)(x \in y)\}. \tag{1.3}$$

[1.3] Sometime the notation $A - B$ is used, but it should be avoided, since the latter notation is also used with different meanings. In earlier times, typesetting $A \setminus B$ caused extra difficulty for printers, but with computerized typesetting that is no longer an issue.

One needs to be a little careful with using the symbol $\bigcap A$, since it is meaningless in case A is the empty set, since if x in equation (1.3) runs over all sets (as one would naturally expect), then the right-hand side describes the set of all sets, which is meaningless.[1.4] Even if A is the empty set and C is another set, it is reasonable to interpret $C \cap (\bigcap A)$ to be the set C.

1.4 Venn diagrams

Set operations can be illustrated in Venn diagrams. One draws two or three circles in a box, indicating two or three sets, and then shades the results of various set operations. In Figure 1.1, *a)* illustrates the set $A \cap B \cap C$, which is defined as the set $A \cap (B \cap C) = (A \cap B) \cap C$; the equality here can easily be proved, and justifies the omitting of the parentheses in the first expression. Similarly, *a)* illustrates the set $A \cup B \cup C$, which is defined as the set $A \cup (B \cup C) = (A \cup B) \cup C$; the equality here can easily be proved, and justifies the omitting of the parentheses in the first expression. Finally, *a)* illustrates the set $(A \triangle B) \setminus C$. Venn diagrams may be helpful in illustrations, but they should not be used for proofs.

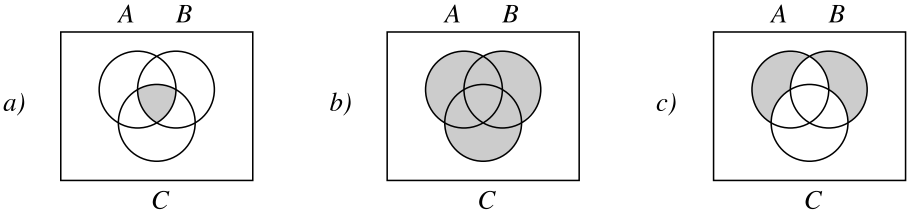

Figure 1.1: Venn diagrams

1.5 The power set of a set

The set of all subsets of a set A is called its power set, and it is usuallt denoed by $\mathcal{P}(A)$. That is,

$$\mathcal{P}(A) \overset{def}{=} \{x : x \subset A\}.$$

2 Probability: the Kolmogorov probability model

We will describe the mathematical model of probability due to Andrey Kolmogorov.

2.1 Outcomes and the sample space

A result of a random experiment is an outcome. For example, when rolling a die, there are six possible outcomes, corresponding to the number rolled. The set of possible outcomes is called the sample space,[2.1] usually denoted by Ω in our discussions.

[1.4] That is, it is meaningless in most versions of axiomatic theory. It make sense in Quine's New Fundation, and axiomatization of set theory of interest to mathematical logicians and to philosophers, but is not commonly used in mathematical practice. See [9].

[2.1] Perhaps this is too simplistic. In certain cases, modeling a random event may not be quite easy, and often it is not clear what the sample space should be.

2.2 Events and probabilities

Events are certain sets of outcomes. That is the set of events $\mathcal{B}$ is a certain subset of the power set $\mathcal{P}(\Omega)$ of Ω. Events will be assigned probabilities: a probability function will be a function $\mathrm{P} : \mathcal{B} \to [0,1]$ mapping $\mathcal{B}$ into the interval $[0,1]$ of real numbers. The set of events and the probability function P will have to satisfy certain properties, to be discussed below.

2.3 Finite uniform probability spaces

One is often faces with a situation that in an experiment there are only finitely many possible outcomes, and each outcome is equally likely. For example, when rolling a fair die, the possible outcomes are 1, 2, 3, 4, 5, 6. The sample space is $\Omega = \{1,2,3,4,5,6\}$. In this case, the set of events $\mathcal{B}$ is the set of all subsets $\mathcal{P}(\Omega)$ of Ω. For example, the event $\{1,2,5\}$ means that the outcome of a roll is 1, 2, or 5. For a finite set A denoting by $|A|$ its number of element, the probability of an event $E \subset \Omega$ will be taken to be

$$\mathrm{P}(E) = \frac{|E|}{|\Omega|}. \tag{2.1}$$

2.3.1 Outcomes and elementary events

As we mentioned, outcomes are elements of the sample space Ω. That is, if $x \in \Omega$, then x is an outcome. An elementary event is the occurrence of an outcome; that is, the elementary event corresponding to x is $\{x\}$; i.e., the one element set contaning only x, called the *singleton* of x, or, more simply, singleton x.[2.2] From a set theoretical point of view, x and $\{x\}$ must be distinguished, so as to be in compliance with the axiom of extensionality (Axiom 1.1). In probability theory, one might be more lax, and sometimes overlook such distinctions (this is a fact of life, not something to be encouraged).

2.4 The algebra of events

We mentioned that the set of events $\mathcal{B}$ is a subset of the power set $\mathcal{P}(\Omega)$ of Ω. We will require that $\mathcal{B}$ satisfy certain reqirements:

Definition 2.1 (Algebra of events)**.** An algebra of events is a set $\mathcal{B} \subset \mathcal{P}(\Omega)$ satisfying the following requirements:

(1) We have $\Omega \in \mathcal{B}$.

(2) If $A, B \in \mathcal{B}$, then $A \setminus B \in \mathcal{B}$.

(3) Assume $A_n \in \mathcal{B}$ for every positive integer n. Then $\bigcup_{n=1}^{\infty} A_n \in \mathcal{B}$.

When stating a definition, one tries to be as economical as possible. In this case, one may need to show that a certain set $\mathcal{B}$ satisfies this definition; the result of the economy is that one has less work to do when doing this. In fact, there is a lot more to this definition then is immediately clear. For example, it follows by Clauses (1) and (2) that $\emptyset = \Omega \setminus \Omega \in \mathcal{B}$. Further, if $A, B \in \mathcal{B}$, then $A \cup B \in \mathcal{B}$. This follows from Clause (3) with $A = A_1$, $B = A_2$, and $A_n = \emptyset$ for $n \geq 2$. We also have

$$A \cap B = \Omega \setminus \big((\Omega \setminus A) \cup (\Omega \setminus B)\big) \in \mathcal{B}$$

[2.2] Note that we are discussing finite uniform probability spaces here. In a more general situation, if $x \in \Omega$, it is not a requirement that singleton x be an event; that is, we may or may not have $\{x\} \in \mathcal{B}$.

according to the above clauses.[2.3] We have a similar result for infinite intersections: if $A_n \in \mathcal{B}$ for every positive integer n, then we also have

$$\bigcap_{n=0}^{\infty} A_n = \Omega \setminus \bigcup_{n=0}^{\infty} (\Omega \setminus A_n) \in \mathcal{B}.$$

A set $\mathcal{B}$ satisfying the properties given in this definition is called a *σ-algebra* of sets. If Clause (3) is only required for a finite number of events, the set is called an algebra of sets. Since we will always require this clause for infinitely many sets, we will use the simpler term algebra, even though we always mean a σ-algebra.

2.5 Probability

Two sets A and B are called *disjoint* if $A \cap B = \emptyset$, i.e., if A and B havo no elements in common. A list of sets are called pairwise disjoint if any two among them are disjoint; i.e., if A_1, A_2, ... are sets (there may be a finite or an infinite number among them) if $A_i \cap A_j = \emptyset$ unless $i = j$; this also means that the same set should not be listed twice, unless it is the empty set (which can be listed any number of times).

Definition 2.2 (Probability function)**.** Given an algebra of events $\mathcal{B}$, a probability function $\mathrm{P} : \mathcal{B} \to [0, 1]$ is a function satisfying the following requirements:

(1) We have $P(\Omega) = 1$.

(2) If $A_n \in \mathcal{B}$ for every positive integer n, and the sets A_n are pairwise disjoint, then

$$\mathrm{P}\Big(\bigcup_{n=1}^{\infty} A_n\Big) = \sum_{n=1}^{\infty} \mathrm{P}(A_n).$$

The property described in Clause (2) is called *σ-additivity*; a similar property stated for a finite number of events is called *finite additivity*, or simply additivity. Intuitively, it is not immediately clear why σ-addivity is needed as opposed to just additivity; this will be clear in the discussion of Russian roulette in Subsection 7.1.[2.4]

There is a lot more in this definition than immediately meets the eye. Clause (2) immediately implies that $\mathrm{P}(\emptyset) = 0$. Indeed, choosing $A_n = \emptyset$ for every n, the equality in this clause would not hold if $\mathrm{P}(\emptyset) > 0$. If $A, B \in \mathcal{B}$ and A and B are disjoint, then $\mathrm{P}(A \cup B) = \mathrm{P}(A) + \mathrm{P}(B)$. This follows from Clause (2) with the choice $A_1 = A$, $A_2 = B$, and $A_n = \emptyset$ for $n \ge 2$. Hence, if $A \in \mathcal{B}$, then $\mathrm{P}(A) + \mathrm{P}(\Omega \setminus A) = \mathrm{P}(\Omega) = 1$; hence, we have

$$\mathrm{P}(\Omega \setminus A) = 1 - \mathrm{P}(A).$$

[2.3] The equality between the sets in the last display is one of the De Morgan identities (click on the link for an explanation and nice Venn diagram illustrations). The sister identity of this is

$$A \cup B = \Omega \setminus \big((\Omega \setminus A) \cap (\Omega \setminus B)\big) \in \mathcal{B}.$$

There are also analogous identities in logc.

[2.4] On account of σ-additivity, one needs to remember that an infinite sum, or sum of a series, is taken to be the limit of its partial sums. While usual addition is commutative, the same is not necessarily true for infinite series – however, it is always true for a series with nonnegative terms, or, more generally, for an absolutely convergent series. However, a conditionally convergent series can always be rearranged to converge to something else or to diverge, according to a theorem of Dirichlet.

The set $\Omega \setminus A$ is called the *complement* [2.5] of A, and one often uses a special notation for it. Following the book [2], we will use the notation

$$A^* \overset{def}{=} \Omega \setminus A.$$

With this notation, we have $\mathrm{P}(A^*) = 1 - \mathrm{P}(A)$.

Definition 2.3 (Probability space)**.** A *probability space* is a triple $(\Omega, \mathcal{B}, \mathrm{P})$, where, as decribed above, Ω is a set of outcomes, $\mathcal{B}$ is an algebra of events on Ω, and $\mathrm{P} : \Omega \to [0, 1]$ is a probability function.

By an abuse of language, one often says, let Ω be a probability space when one in fact means the triple $(\Omega, \mathcal{B}, \mathrm{P})$.

2.5.1 Why the algebra of events

One may ask, why one does not take the set of events all subsets of Ω. In fact, this is what one usually does when Ω is finite or a *countably infinite* set. When Ω is an uncountable set, this is usually impossible. In an important paper [10] Stanislaw Ulam.[2.6] The question whether it is possible to have a probability function on all subsets of an uncountable set leads to deep questions of the foundation of mathematics. Ulam showed that the existence of such an uncoountable set is unprovable with the usual axioms of set theory.

2.6 Nondisjoint events

We have

Lemma 2.1. *Let $(\Omega, \mathcal{B}, \mathrm{P})$ be a probability space, and assume that $A, B \in \mathcal{B}$. Then*

$$\mathrm{P}(A \cup B) = \mathrm{P}(A) + \mathrm{P}(B) - \mathrm{P}(A \cap B). \tag{2.2}$$

Proof. The sets $A \setminus (A \cap B)$, $B \setminus (A \cap B)$, and $A \cap B$ are pairwise disjoint, and we have

$$A \cup B = \big(A \setminus (A \cap B)\big) \cup \big(B \setminus (A \cap B)\big) \cup (A \cap B).$$

Hence

$$\mathrm{P}\big(A \cup B\big) = \mathrm{P}\big(A \setminus (A \cap B)\big) + \mathrm{P}\big(B \setminus (A \cap B)\big) + \mathrm{P}(A \cap B).$$

That is,

$$\begin{aligned} \mathrm{P}\big(A \cup B\big) = &\Big(\mathrm{P}\big(A \setminus (A \cap B)\big) + \mathrm{P}(A \cap B)\Big) \\ &+ \Big(\mathrm{P}\big(B \setminus (A \cap B)\big) + \mathrm{P}(A \cap B)\Big) - \mathrm{P}(A \cap B) \end{aligned} \tag{2.3}$$

Noting that we have

$$\mathrm{P}(A) = \mathrm{P}\big(A \setminus (A \cap B)\big) + \mathrm{P}(A \cap B)$$

[2.5] First, note that the words “complement” and “compliment” have very different meaning (you may look up the word in Wiktionary to see the difference. Second, if A and B are sets and $A \subset B$, then $B \setminus A$ is occasionally called the *relative complement* of A with respect to B. If B is the *universal set* in a discourse, that is, the set of all things under consideration, then $B \setminus A$ is simply called the *complement* of A. In most mathematical discussions, there is no universal set, but in discussing probabilities and events, one might think of Ω as the universal set.

[2.6] Ulam made many other contributions, discussed in the Wikipedia article, to the Manhattan Project (the program to create the first atomic bomb), and he had the basic idea for the Teller–Ulam design, the design on which the hydrogen bomb is based.

and

$$\mathrm{P}(B) = \mathrm{P}\big(B \setminus (A \cap B)\big) + \mathrm{P}(A \cap B),$$

and substuting these into equation (2.3), equation (2.2) follows. □

2.7 Reading

[2, §2.1–2.3, pp. 4–14].

2.8 Homework

[2, Chapter 2, p. 36], 201–207.

3 Combinatorics

In order to discuss finite uniform probability spaces, described in Subsection 2.3, we need to discuss simple combinatorics.

3.1 Permutations: lists of length k of n items

Let n and k be integers with $0 \le k \le n$ A k-permutation of n distinct items a list containing k of given n items, where we distinguish between two lists containing the same items given in different order. We can count these permutations as follows. The first item of the list can be picked in n different ways, depending on which of the items we pick. After this, $n-1$ items remain, so the second item can be picked in $n-1$ different ways. Then, the number of possible lists is $n(n-1)$. When picking the third item, we have to pick from among $n-2$ items, so the number of possible lists containing three items is $n(n-1)(n-2)$. Continuing this, the number of lists containing k items, denoted by P^n_k, is

$$P^n_k = n(n-1)(n-2)\dots(n-k+1) = \prod_{j=0}^{k-1}(n-j), \tag{3.1}$$

For $k = 0$ the product on the right is the empty product, interpreted as 1, saying that there is only one list of length 0, the empty list.

3.1.1 Permutations

For an integer $n \ge 0$, an n-permutation of n items is simply called a permutation of these items. Their number according to equation (3.1) is written as $n!$, read as n-factorial:

$$n! \overset{def}{=} \prod_{j=0}^{n-1}(n-j) = \prod_{j=1}^{n} j = 1 \cdot 2 \cdot \ldots \cdot n.$$

Again, if $n = 0$, the product is the empty product, interpreted as 1. That is, $0! = 1$.

3.1.2 Permutations as mappings

Instead of considering permutations as a rearrangement of a given list, it is often more amenable to a mathematical treatment to consider a permutation of a set A as a one-to-one mapping of the set A onto itself.[3.1] These permutations clearly have the same number as permutations considered as rearrangements. Indeed, given an integer $n > 0$, if $A = \{k : 1 \le k \le n\}$ and $\sigma : A \to A$, then $\sigma(k) = l$ means that in the corresponding rearrangement the number k is placed at the lth place on the list.

3.2 Combinations

Given integers k and n with $0 \le k \le n$, a k-combination of n distinct items is a selection of k of these items, where the order the items are selected does not count. Their number is the same as the number of k-element subsets of an n-element sets (since, in a set, the order in which the elements are listed makes no difference). Given a k-combination, if you permute the selected k items, you get $k!$ lists, which are k-permutations of the n given items. If you take all k-combinations and permute each of them, you get all k-permutations of the n given items. That is, the number of these permutations, P^n_k, is $k!$ times the number of k-combinations of n-items. Denoting by $\binom{n}{k}$ the number of these combinations,[3.2], read as n choose k, according to (3.1), we have

$$\binom{n}{k} = \frac{\prod_{j=0}^{k-1}(n-j)}{k!} = \frac{\prod_{j=0}^{k-1}(n-j)}{\prod_{j=0}^{k-1}(k-j)} = \prod_{j=0}^{k-1}\frac{n-j}{k-j} \tag{3.2}$$

Note that $\binom{n}{0} = 1$, since in this case we pick the empty combination (or the empty subset), so there is only one pick. The formula correctly gives 1 as the empty product. Noting that

$$(n-k)!\prod_{j=0}^{k-1}(n-j) = (n-k)!\prod_{j=n-k+1}^{n} j = \Big(\prod_{j=1}^{n-k} j\Big)\prod_{j=n-k+1}^{n} j = \prod_{j=1}^{n} j = n!,$$

where the parentheses surrounding the first product after the second equation sign is to indicate that the second product is not in the scope of the first product,[3.3] Hence, by (3.1) we have

$$\binom{n}{k} = \frac{\prod_{j=0}^{k-1}(n-j)}{k!} = \frac{(n-k)!\,\prod_{j=0}^{k-1}(n-j)}{(n-k)!\,k!} = \frac{n!}{(n-k)!\,k!}.$$

While the right-hand side is not useful in calculations, it helps us to show that

$$\binom{n}{n-k} = \frac{n!}{k!\,(n-k)!} = \frac{n!}{(n-k)!\,k!} = \binom{n}{k}.$$

3.2.1 The binomial theorem

Let n be a positive integer, and for k with $0 \le k \le n$, let A_k denote the set of all k-element subsets of the set $\{i : 1 \le i \le n\}$. In particular, we have $A_0 = \{\emptyset\}$. The number of elements of A_k is $\binom{n}{k}$.

[3.1] If A is a finite set and $\sigma : A \to A$ is one-to-one, then σ is also clearly onto A. However, when one considers permutations as mappings, one occasinally thinks of permutations of an infinite set.

[3.2] An older, now deprecated, notation is C^n_k, *not to be used in this course.*

[3.3] The rules are not quite clear where the scope of a product ends; usually, it is ended by a + or − sign unprotected by parentheses. It is not safe to assume that the second product sign would end the scope of the first one (we would tend to think that it does not end the scope).

Assume we are given real numbers x and y_i for $1 \le i \le n$.[3.4] We have

$$\prod_{i=1}^{n}(x+y_i) = \sum_{k=0}^{n}\sum_{S\in A_k} x^{n-k}\prod_{i\in S} y_i. \tag{3.3}$$

The reason this equality holds is that we can evaluate the product on the left by taking all the products that result by picking either x or y_i from each of the factors, and then adding all products resulting this way. When we pick a y_i from k of these factors ($0 \le k \le n$), we have to pick x $n-k$ times, and we obtain the term of the sum corresponding to this k on the right-hand side.[3.5] If we assume that $y_1 = y_2 = \ldots = y_n = y$, then the left-hand side becomes $(x+y)^n$. On the right-hand side, for a fixed k, each of the terms after the second sum becomes $x^{n-k}y^k$. Since there are $\binom{n}{k}$ of these terms for a given k, the sum right-hand side becomes $\sum_{k=0}^{n}\binom{n}{k}x^{n-k}y^k$. That is, we obtain

$$(x+y)^n = \sum_{k=0}^{n}\binom{n}{k}x^{n-k}y^k. \tag{3.4}$$

This is the *binomial theorem.* As we pointed out in the footnote above, the commutativity of the product was used in way in the proof. For example, the binomial theorem is not true for noncommuting square matrices A and B of the same size.

3.3 Lists with repetition, order counts

Let $n, k \ge 0$ be integers, and assume we are given n items. We want to count the number of the lists of length k of these items when items are allowed to repeat. The order in which the items are listed counts. This is a simple task, since there are n ways to pick the first item on the list, then n ways to pick the second, third, etc. items. After k picks, we can create n^k lists. This is the total number of lists that can be created.

3.4 Combinations with repetitions

Let $n, k \ge 0$ be integers, and assume we are given n items. We want to count the number of the lists of length k of these items when items are allowed to repeat, but the order in which the items are listed does not count. Such lists are called k-combinations of n items with repetitions.

The number of these items can be counted as follows. Assume the items to be listed are the numbers $1, 2, 3, \ldots, n$. Define a mapping f from the set of such lists as follows. Let $\lambda = \langle i_1, i_2, i_3, \ldots, i_k\rangle$ be such a list, arranged in nondecreasing order; that is, assume $i_1 \le i_2 \le i_3 \le \ldots \le i_k$. Then put

$$f(\lambda) = \{i_l + l - 1 : 1 \le l \le k\} = \{i_1 + 0, i_2 + 1, i_3 + 2, \ldots, i_k + (k-1)\}.$$

Noting that the numbers listed in the set on the right-hand side are all distinct; in fact, $i_1 + 0 < i_2 + 1 < i_3 + 2 < \ldots < i_k + (k-1)$. It is easy to see that f is a one-to-one mapping from the set of all such lists onto the set of all k-element subsets of the set

$$\{1, 2, 3, \ldots, n+k-1\} = \{m \in \mathbb{Z} : 1 \le m \le n+k-1\}.$$

Since the nuber of these subsets is $\binom{n+k-1}{k}$ according to Subsection 3.2, this is also the number of lists with repetitions we are considering.

[3.4] The quantities x and y_i may be elements of an arbitrary communative ring, instead of being real numbers.

[3.5] Note that in obtaining this equation, the commutativity of multiplication was used in an important way, to ensure that, in each product, the occurrences of x can be moved to the front.

3.5 Reading

[2, § 2.7, pp. 29–32].

4 Discrete sample spaces

A *discrete*[4.1] probability space is a finite or countably infinite probability space in which each subset is an event; that is, it is a triple $(\Omega, \mathcal{B}, \mathrm{P})$ where Ω is finite or countably infinite, and $\mathcal{B} = \mathcal{P}(\Omega)$. According to Clause (2) of Definition 2.2, in order to define the probability function P, it is enough to define $P(\{x\})$ for each $x \in \Omega$. Here we will usually consider finite uniform probability spaces, where probabilities can be calculated by formula (2.1).

4.1 Picking marbles with replacement

In what follows, "marbles" mean colored glass balls, usually red or green, placed in a container, usually called an urn in probability theory. Given the urn with marbles, one randomly picks one of these marbles after thorouly mixing the marbles – of course, the marbles are indistinguishable to the touch. One can pick with replacement, in which the marble picked will be put back in the urn, and without replacement, in which the marble is not put back. First we consider picking with replacement.

Given a red marbles and b green marbles in an urn; here a and b must be positive integers. We pick n marbles with replacement; what is the probability of picking exactly k red marbles; here $0 \le k \le n$. In order to solve this problem, the probability space Ω will be the set of all sequences of picks. There are $a + b$ items, and n picks with repetitions allowed; the number of these picks are $(a + b)^n$ according to Subsection 3.3. That is, $|\Omega| = (a + b)^n$.

As for counting the number of favorable outcomes, we want first to count those outcomes in which we first pick k red marbles, then we pick $n - k$ green marbles. The number of ways we can pick k red marbles is the number of lists of length k that we can form with the a red marbles with repetitions allowed. This is a^k. The number of ways we can then pick $n - k$ green marbles is, similarly, b^{n-k}. The number of ways first picking k red marbles and then $n - k$ green marbles is $a^k b^{n-k}$. The probablility of this happening is

$$\frac{a^k b^{n-k}}{(a+b)^n} = \left(\frac{a}{a+b}\right)^k \left(\frac{b}{a+b}\right)^{n-k}. \tag{4.1}$$

In order to clarify the situation, we first want to consider the simple example when $k = 3$ and $n = 5$. Picking a red marble will be written as R and picking a green marble, G. The sequence $RRRGG$ indicates first picking 3 red marbles, and then 2 green marbles. If we drop the restriction that the red marbles will be picked first, then the possibilities that we pick 3 red marbles and 2 green ones can be described by the sequences $RRRGG$, $RRGRG$, $RRGGR$, $RGRRG$, $RGRGR$, $RGGRR$, $GRRRG$, $GRRGR$, $GRGRR$, $GGRRR$. Each of these possibilities represent mutually exclusive events. For example, $RRGGR$ and $RGRGR$ cannot happen at the same time, since the first sequence indicaes that the second pick is a red marble, while the second sequence indicates that the second pick is green. Finally each of these sequences correspond to the same count a^3b^2 (or, in the general case $a^k b^{n-k}$) sequences of marbles.

[4.1] The words "discrete" and "discreet" have totally different meanings. Look them up in Wiktionary to see the difference.

Next, we want to count the number of sequences formed formed by the letters R and G listed above. This is easy, since each sequence corresponds to a combination of the letters $\{1, 2, 3, 4, 5\}$, since one way to represent these sequences is to list the places where a red letter is located. For example, the sequence $RGRRG$ corresponds to the set $\{1, 3, 4\}$; that is, the number of these sequences ls the number of 3-element subsets of a 5 element set, that is $\binom{5}{3}$. In the general case, this number is the number of k-element subsets of an n-element set; that is, $\binom{n}{k}$.

We obtain the probability of picking k red marbles and $n-k$ green marbles by adding up the probabilities of each particular sequence of picks, since these sequences represent mutually exclusive events. Since each of these probabilities is the same, as given in formula (4.1), and the number of these sequences is $\binom{n}{k}$, we obtain that this probablity is

$$\binom{n}{k}\frac{a^k b^{n-k}}{(a+b)^n} = \binom{n}{k}\left(\frac{a}{a+b}\right)^k\left(\frac{b}{a+b}\right)^{n-k}. \tag{4.2}$$

The second form will be of particular interest to us, since it is a special case of the *binomial distribution*, to be discussed later.

4.2 Picking marbles without replacement

Given a red marbles and b grean marbles in an urn, where a and b are positive integers, we want to pick n marbles without replacement, and the question is, what is the probability that k or these marbles are red. For the question to be meaningful, we must have $0 \le k \le n \le a + b$. The first issue is the selection of the probability space Ω. counting is simplified by not doing this. It is much simpler to consider the set of chosen marbles, and ignore the order in which they were picked; doing this will give the same answer.[4.2]

Since the marbles are not put back, we can simply take as an outcome as the particular marbles picked; noting that a given set of n marbles can be selected in $n!$ ways, when counting all sequence of picks as opposed to just counting the set of marbles picked, both the number $|E|$ of favorable events and the total number of events $|A|$ will get multiplied by $n!$, their ratio in equation (2.1) remains the same, so we choose the simpler way of counting.

That is, in this case, Ω is the set of all picks of n marbles. This number is $\binom{a+b}{n}$. A pick is favorable when this set contains k red marbles and $n-k$ green marbles. The number of sets of k red marbles is $\binom{a}{k}$, and the number of $n-k$ green marbles is $\binom{b}{n-k}$. The total number of sets containing k red marbles and $n-k$ green marbles is the product of these, that is, $\binom{a}{k}\binom{b}{n-k}$. The probability is the number of favorable picks divided by the number of all picks, that is, it is

$$\binom{a}{k}\binom{b}{n-k}\Big/\binom{a+b}{n}. \tag{4.3}$$

4.3 Reading

[2, §2.4, pp. 14–21].

4.4 Homework

[2, Chapter 2, p. 36], 208–212.

[4.2] The reason is that if we consider the sequence of the picks, the number of sets would be multiplied by $n!$, the number of ways the picked marbles can be put in different orders.

5 Conditional probability

Given a probability space Ω and two events A and B with $\mathrm{P}(B) \neq 0$, we define the conditional probability $\mathrm{P}(A \mid B)$, to be read as the probability of A given that B occurs is

$$\mathrm{P}(A \mid B) = \frac{\mathrm{P}(A \cap B)}{\mathrm{P}(B)}. \tag{5.1}$$

As an example, consider the case of rolling a die, and assume we know that the number rolled is ≤ 4, what is the probability that the number rolled is 1, 3, 4, or 5.[5.1] Then, calculating this probability directly, we can take $\Omega_{\text{direct}} = \{1, 2, 3, 4\}$ We take the event to be the set $E_{\text{direct}} = \{1, 3, 4\}$[5.2] so

$$\mathrm{P}_{\text{direct}}(E_{\text{direct}}) = \frac{|E_{\text{direct}}|}{|\Omega_{\text{direct}}|} = \frac{3}{4}.$$

Calculating the same probability as a conditional probability, let the probability space be $\Omega = \{1, 2, 3, 4, 5, 6\}$. Let $A = \{1, 3, 4, 5\}$ and B={1,2,3,4}. Then $A \cap B = \{1, 3, 4\}$ and so $\mathrm{P}(A \cap B) = 3/6$, and $B = \{1, 2, 3, 4\}$, and so $\mathrm{P}(B) = 4/6$, the calculation is simpler if we do not reduce these fractions. Hence

$$\mathrm{P}(A \mid B) = \frac{\mathrm{P}(A \cap B)}{\mathrm{P}(B)} = \frac{3/6}{4/6} = \frac{3}{4};$$

we obtain the same result. We have

Theorem 5.1 (Total probability theorem)**.** *Let Ω be a probability space. Assume $H_i \subset \Omega$ pairwise disjoint events, where $1 \leq i \leq n$ for some integer n or $1 \leq i < \infty$, such that $\bigcup_i H_i = \Omega$, and let $A \subset \Omega$ be another event. Assume, further, that $\mathrm{P}(H_i) \neq 0$ for any i. Then*

$$\mathrm{P}(A) = \sum_i \mathrm{P}(H_i)\,\mathrm{P}(A \mid H_i). \tag{5.2}$$

In case $\mathrm{P}(H_i) = 0$, the conditional probability $\mathrm{P}(A \mid H_i)$ is not defined; this is why the assumption $\mathrm{P}(H_i) \neq 0$ is needed. If we define the product $\mathrm{P}(H_i)\,\mathrm{P}(A \mid H_i)$ to be 0 is case $\mathrm{P}(H_i) = 0$, then this assumption can be dropped. The other assumtions can be somewhat weakened to say that $\mathrm{P}(A_i \cap A_j) = 0$ instead of $A_i \cap A_j = \emptyset$, and $\sum_i \mathrm{P}(H_i) = 1$ instead of $\bigcup_i H_i = \Omega$.

Proof. We have

$$A = A \cap \Omega = A \cap \Big(\bigcup_i H_i\Big) = \bigcup_i (A \cap H_i).$$

Hence,

$$\mathrm{P}(A) = \sum_i \mathrm{P}(A \cap H_i). = \sum_i \mathrm{P}(H_i)\,\mathrm{P}(A \mid H_i);$$

here, the first equation follows by Clause (2) in Definition 2.2, and the second equation holds by the definition of conditional probability given in equation (5.1). □

Theorem 5.2 (Bayes's theorem)**.** *Let Ω be a probability space. Assume $H_i \subset \Omega$ pairwise disjoint, where $1 \leq i \leq n$ for some integer n or $1 \leq i < \infty$, such that $\bigcup_i H_i = \Omega$, and let $A \subset \Omega$ be another*

[5.1] Of course, it is ridiculous to include 5 if we know that the number rolled is ≤ 4, but the question is perfectly meaningful

[5.2] Of course, we do not put 5 into the set E_{direct}, since 5 is not even an element of the probability space Ω_{direct}.

event. Assume, further, that $\mathrm{P}(H_i) \neq 0$ *for any* i*. Then assuming* $\mathrm{P}(A) \neq 0$)*, for any* k *in the same range as* i*, we have*

$$\mathrm{P}(H_k \mid A) = \frac{\mathrm{P}(H_k)\,\mathrm{P}(A \mid H_k)}{\sum_i \mathrm{P}(H_i)\,\mathrm{P}(A \mid H_i)}. \tag{5.3}$$

Proof. By using equation (5.1) twice, we have

$$\mathrm{P}(H_k \mid A) = \frac{\mathrm{P}(H_k \cap A)}{\mathrm{P}(A)} = \frac{\mathrm{P}(H_k)\,\mathrm{P}(A \mid H_k)}{\mathrm{P}(A)}.$$

Using equation (5.2) for $\mathrm{P}(A)$ in the denominator, formula (5.3) follows. □

Problem 5.1. *a).* In a factory, parts are manufactured by three machines, M_1, M_2, and M_3 in proportions $20 : 30 : 50$. The percentages 5%, 2%, and 4% of these parts are defective, respectively. Find the probability that a randomly chosen part is defective.

b). Find the probability that a defective part was manufactured on the second machine.

Solution. a). Write A for the event that a part is defective, and M_i with $i = 1, 2, 3$ for the event that it was manufactured on machine i. We have

$$\mathrm{P}(A) = \sum_{i=1}^{3} \mathrm{P}(A \mid M_i)\,\mathrm{P}(M_i) = .05 \cdot .2 + .02 \cdot .3 + .04 \cdot .5 = .036.$$

b). Using Bayes's Theorem 5.2, we have

$$\mathrm{P}(M_2 \mid A) = \frac{\mathrm{P}(A \mid M_2)\,\mathrm{P}(M_2)}{\sum_{i=1}^{3} \mathrm{P}(A \mid M_i) P(M_i)} = \frac{.02 \cdot .3}{.05 \cdot .2 + .02 \cdot .3 + .04 \cdot .5} = \frac{.006}{.036} = \frac{1}{6} \approx .166667;$$

note that the denominator in the third member is the same as the answer to part *a*).

5.1 Reading

[2, §2.5, pp. 21–25].

5.2 Homework

[2, Chapter 2, p. 36], 213–217.

6 Independence

In common sense, a number of events are independent if they have no influence on one another. The mathematical definition is sommewhat more complicated. We want to give the definiton for any n events. The definition will be recursive in that the the independence of n events will rely on the independence of $n-1$ events

Definition 6.1. Let Ω be a probability space, let $n \geq 1$ be an integer, and let A_i for $1 \leq i \leq n$ be events. If $n = 1$, we will call the event *independent*. If $n > 1$, we will call them *independent* if any $n - 1$ of them are independent, and, in addition, we have

$$\mathrm{P}\Big(\bigcap_{i=1}^{n} A_i\Big) = \prod_{i=1}^{n} \mathrm{P}(A_i). \tag{6.1}$$

For $n = 1$, independence means no restriction on the event A_1, that is, A_1 by itself is always independent; we allowed the case $n = 1$ to allow the recursive definiion and to support some proofs by induction. For $n = 2$, this says that A_1 and A_2 are independent if $\mathrm{P}(A_1 \cap A_2) = \mathrm{P}(A_1)\,\mathrm{P}(A_2)$. If $\mathrm{P}(A_2) \neq 0$, then this is equivalent to the first equation in

$$\mathrm{P}(A_1) = \frac{\mathrm{P}(A_1 \cap A_2)}{\mathrm{P}(A_2)} = \mathrm{P}(A_1 \mid A_2),$$

where the second equation holds in view of (5.1). That is, in this case the the probability of A_1 conditional on A_2 occurring is the same as the unconditional probability of A_1. For $n = 3$, independence means that we have where the second equation holds in view $\mathrm{P}(A_1 \cap A_2 \cap A_3) = \mathrm{P}(A_1)\,\mathrm{P}(A_2)\,\mathrm{P}(A_3)$, $\mathrm{P}(A_1 \cap A_2) = \mathrm{P}(A_1)\,\mathrm{P}(A_2)$, $\mathrm{P}(A_1 \cap A_3) = \mathrm{P}(A_1)\,\mathrm{P}(A_3)$, $\mathrm{P}(A_2 \cap A_3) = \mathrm{P}(A_2)\,\mathrm{P}(A_3)$. It is clear from the definition, that given n events, wether or not they are independent does not depend on the order the events are listed,

6.1 Independence involving complements

Lemma 6.1. *Let Ω be a probability space, let $n \geq 1$ be an integer. Assume the events A_i for $1 \leq i \leq n$ are independent. Then the events A_1^* and A_i for $2 \leq i \leq n$ are independent.*

Proof. For $n = 1$ this is true, since independence imposes no requirements in case $n = 1$. We use induction on n, so assume $n \geq 2$ and the assertion is true for any n' events with $1 \leq n' < n$. Recall that A_1^* denotes the complement of A_1. The only thing we need to show that equation (6.1) remains valid if A_1 is replaced by A_1^*. Writing $B = \bigcap_{i=2}^n A_i$, we have $(A_1 \cap B) \cup (A_1^* \cap B) = B$, and the sets $A_1 \cap B$ and $A_1^* \cap B$ on the left-hand side are disjoint; so $\mathrm{P}(A_1 \cap B) + \mathrm{P}(A_1^* \cap B) = \mathrm{P}(B)$. Hence, using equation (6.1) with $A_1 \cap B = \bigcap_{i=1}^n A_i$ and the analogous equation with $B = \bigcap_{i=2}^n A_i$, we have

$$\begin{aligned}\mathrm{P}(A_1^* \cap B) = \mathrm{P}(B) - P(A_1 \cap B) &= \prod_{i=2}^n \mathrm{P}(A_i) - \prod_{i=1}^n \mathrm{P}(A_i) \\ = \prod_{i=2}^n \mathrm{P}(A_i) - \mathrm{P}(A_1)\prod_{i=2}^n \mathrm{P}(A_i) &= \bigl(1 - \mathrm{P}(A_1)\bigr)\prod_{i=2}^n \mathrm{P}(A_i) = \mathrm{P}(A_1^*)\prod_{i=2}^n \mathrm{P}(A_i).\end{aligned}$$

This shows that equation (6.1) holds with A_1^* replacing A_1, completing the proof. □

Corollary 6.1. *Given a list of independent events, the list of events obtained by replacing any number of them by their complements, the list of events so obtained is independent.*

Proof. When replacing the desired events on the list one by one, the lemma just proved shows that at each step we obtain a list of independent events. □

Problem 6.1. Given $n > 0$ hunters, they shoot at a deer at the same time; each has a probability α of hitting the deer; these events are independent. What is the probability that the deer will be hit at least once.

Solution. For i with $1 \leq i \leq n$ let A_i be the event that the ith hunter hits the deer. Then $\bigcap_{i=1}^n A_i^*$ is the event that none of them hits. Since the events A_i^* are independent according to Corollary 6.1, we have

$$\mathrm{P}\Bigl(\bigcap_{i=1}^n A_i^*\Bigr) = \prod_{i=1}^n \mathrm{P}(A_i^*) = (1 - \alpha)^n.$$

The deer will be hit if this event does not happen; that is, the deer will be hit with probability $1 - (1 - \alpha)^n$.

6.2 Reading

[2, §2.6, pp. 25–29]. The next few pages, [2, § 2.7, pp. 29–32], have been assigned earlier, at the end of Section 3.

6.3 Homework

[2, Chapter 2, p. 36], 218–223. Feel free to attempt the remaining problems, but they are meant to be especially challenging problems, so they are marked with an asterisk. It is OK if you cannot do them.

7 Some classical problems

7.1 Russian roulette

Russian roulette is a game, hopefully only played in a theoretical model and not in practice, in which a single round placed a revolver, a handgun with a rotating cylinder capable of holding six rounds. Two players, A and B, play alternate turns until the losing player gets killed as follows. Before the player puts the gun to his head, the chamber rotated to stop at a random location, then the player puts the gun to his head and pulls the trigger. The player survives only if the gun does not fire, i.e., if the active chamber does not contain a round. If he survives, it is the turn of the other player to do the same. Player A goes first. The mathematical question is, what is the probability that player A gets killed. This question gives a clear illustration why σ-additivity, stated in Clause (2) of Definition 2.2, is important, and finite additivity would not suffice (see the discussion after the definition quoted).

In order to more easily analyze the situation, we reformulate the game with a nonlethal description as follows. The players alternately roll a die, and the player that first rolls a 1 loses, but for easier analysis, assume that the game still goes on indefinitely. Let K_i for $i = 1, 2, 3, \ldots$, be the event that the player on the ith turn rolls a 1. The events K_i are assumed to be independent, and we have $\mathrm{P}(K_i) = 1/6$ for each i.[7.1] Let H_i be the event that a 1 is rolled for the first time in the ith turn. We have

$$\mathrm{P}(H_i) = \mathrm{P}\Big(K_i \cap \bigcap_{j=1}^{i-1} K_j^*\Big) = \mathrm{P}(K_i) \cdot \prod_{j=1}^{i-1} \mathrm{P}(K_j^*) = \frac{1}{6} \cdot \prod_{j=1}^{i-1} \frac{5}{6} = \frac{1}{6}\left(\frac{5}{6}\right)^{i-1};$$

the second equation holds in view of the independence of the events in question (cf. Corollary 6.1). The events H_i are mutually exclusive. Player A loses if H_i occurs for some odd i. That is, the probability that player A loses is

$$\mathrm{P}\Big(\bigcup_{k=0}^{\infty} H_{2k+1}\Big) = \sum_{k=0}^{\infty} \frac{1}{6}\left(\frac{5}{6}\right)^{2k} = \frac{1}{6}\sum_{k=0}^{\infty}\left(\frac{5^2}{6^2}\right)^{k} = \frac{1}{6}\,\frac{1}{1-5^2/6^2} = \frac{6}{36-25} = \frac{6}{11}; \tag{7.1}$$

the third equation uses the sum formula for the geometric series

$$\sum_{n=0}^{\infty} x^n = \frac{1}{1-x}. \tag{7.2}$$

[7.1] This is a reasonable assumption if the players don't cheat.

The reason for reinterpreting the original version with Russian roulette is the difficulty in imagining the event K_i meaning that the player in the ith turn gets killed in case he was killed earlier – it would have been somewht problematic to deal with the independent events K_i. One can still do a similar analysis with the events K_i, involving conditional probabilities, but that is a bit more cumbersome.

7.2 Russian roulette revisited

Here we give a different analysis Russian roulette; for this, we will use the original description. Here, let A be the event the first player gets killed, and B, the second player gets killed. Let T be the event that the first player gets killed in the first round. Noting that $T \subset A$, thus $T \cap A = T$, we have

$$\mathrm{P}(A) = \mathrm{P}(T) + \mathrm{P}(T^* \cap A) = \mathrm{P}(T) + \mathrm{P}(T^*)\,\mathrm{P}(A \mid T^*) = \mathrm{P}(T) + \mathrm{P}(T^*)\,\mathrm{P}(B); \tag{7.3}$$

the third equality holds since $\mathrm{P}(A \mid T^*) = \mathrm{P}(B)$. This is because if the first player does not get killed in the first round then at the beginning of the second round he is in exaclty the same situation as the second player at the beginning of the game: before his turn, the other player will play. Note that $\mathrm{P}(T) = 1/6$, and so $\mathrm{P}(T^*) = 5/6$. Given that $A \cap B = \emptyset$, and $A \cup B \cup (A^* \cap B^*) = \Omega$, we have $\mathrm{P}(A) + \mathrm{P}(B) + \mathrm{P}(A^* \cap B^*) = 1$. Noting that $\mathrm{P}(A^* \cap B^*) = 0$.[7.2] we have $\mathrm{P}(B) = 1 - \mathrm{P}(A)$. With these observation, equation (7.3) becomes

$$\mathrm{P}(A) = \frac{1}{6} + \frac{5}{6}\big(1 - \mathrm{P}(A)\big).$$

Solving this equation, we obtain $\mathrm{P}(A) = 5/11$. This agrees with the result we obtained above. It is interesting to note that this consideration implicitly evaluates the geometric series in (7.1)

7.3 Gambler's ruin

The problem a version of which we will be discusing has a long history; see Gambler's ruin for the history of the problem. In the version of the game we are going to discuss, there are two players, A and B, A having a dollars and B having b dollars, where a and b are positive integers. In each turn, they toss a fair coin, and if it comes out head, B gives A one dollar, and if tail, A gives B one dollar. When one of the players has 0 dollars, he loses the game and the game ends. The player that loses is said to be ruined. The question is, what is the probability of A losing the game.

In answering this question, let P_n denote the probability that A loses the game when he has n dollars; at this point B would have $a + b - n$ dollars. Let A be the event that A loses at this point, having n dollars. Let H and T be the event that the next toss is head or tail. We have

$$P_n = \mathrm{P}(A) = \mathrm{P}(A \cap H) + \mathrm{P}(A \cap T) = \mathrm{P}(H)\,\mathrm{P}(A \mid H) + \mathrm{P}(T)\,\mathrm{P}(A \mid T) = \frac{1}{2}P_{n+1} + \frac{1}{2}P_{n-1},$$

since if the next toss is head, A will have $n + 1$ dollars, and if its a tail, he will have $n - 1$ dollars.

$$P_n = \frac{1}{2}P_{n+1} + \frac{1}{2}P_{n-1}.$$

This is a homogeneous linear recurrence equation, and it is well known how to solve such equations. The method of solving them is discussed in the section on Recurrence equations in [5, Corollary

[7.2] We are not saying that $A^* \cap B^* = \emptyset$, since it is not impossible in principle that both players stay alive for ever, but this event has zero probability. We will leave the formal verification of this to the reader. The problem is that to do this formal analysis, we would have to repeat arguments similar to the preceding subsection.

on p. 129, pdf p. 135]. According to this, such equations are associated with a polynomial equation, called the characteristic equation of the recurrence equation, and the zeros of the former determine the solutions of the latter. The characteristic equation in our case is $\zeta = (1/2)\zeta^2 + (1/2)$, i.e. $\zeta^2 - 2\zeta + 1 = 0$. This can be written as $(\zeta - 1)^2 = 0$, and its only zero is $\zeta = 1$, but this is a double zero, which somewhat complicates the situation. In any case, according to the result just quoted, the general solution is this recurrence equation is

$$P_n = C_1 + C_2 n,$$

where C_1 and C_2 are arbitrary constants. In the present case, these constants can be determined from the initial conditions $P_{a+b} = 0$ and $P_0 = 1$. This is because if A has $a + b$ dollars, then B has lost, since he has 0 dollars, so the probability of A losing is 0. On the other hand, if A has 0 dollars, he has lost the game, so the probability of him losing is 1. Subsituting these initial conditions, we obtain $0 = C_1 + (a + b)C_2$ and $1 = C_1 + C_2 \cdot 0$; that is, $C_1 = 1$ and $C_2 = -1/(a + b)$. Substituting these into the equation describing the solution, we obtain $P_n = 1 - n/(a+b)$. With $n = a$, this gives the probability that A will be ruined at the start of the game:

$$P_a = 1 - \frac{a}{a+b} = \frac{b}{a+b}.$$

7.4 The probability of a union

Lemma 2.1 can be generalized as follows:

Theorem 7.1. *Given an integer $n \geq 1$ and events A_1, A_2, …, A_n, we have*

$$\mathrm{P}\Big(\bigcup_{i=1}^{n} A_i\Big) = \sum_{\substack{S: S\subset\{1,2,\dots,n\}\\ S\neq\emptyset}} (-1)^{|S|+1}\,\mathrm{P}\Big(\bigcap_{i\in S} A_i\Big). \tag{7.4}$$

Here $|S|$ indicates the number of elements of the set S. In words, this formula says that to calculate the probability of a union of events, for all k with $1 \leq k \leq n$, one has to take the probabilities of the intersection of any k of these events, and add this probability if k is odd and subtract it if k is even.

The validity of this result can easily be seen intuitively as follows. Assume $x \in \Omega$ belongs to $k > 0$ among the sets A_1, A_2, …, A_n. When taking the intersection of l of these sets for l with $1 \leq l \leq n$, then such intersections can be taken $\binom{k}{l}$ ways. because we need to select l sets form among those containing x. That is, in the sum of probabilities above, the number of ways x is counted is

$$\sum_{l=1}^{k} \binom{k}{l} (-1)^{l+1} = (-1)\left(\sum_{l=0}^{k} \binom{k}{l} (-1)^{l} - 1\right) = -\Big(\big(1 + (-1)\big)^k - 1\Big) = 1,$$

where the second equation follows by the binomial theorem. This shows that x is counted exactly once, as it should be. While this is not a rigorous proof, in a more advanced framework it can easily be reformulated in terms of integrals and characteristic functions as a rigorous proof.[7.3] Here we will give a proof using induction that goes roughly along the lines of an induction proof of the binomial theorem. In Subsection 3.2.1 we gave a proof of the binomial theorem that does not use

[7.3] Given a a set A, its characteristic function (defined on a set B with $A \subset B$) is a function χ_A such that $\chi_A(x) = 1$ if $x \in A$ and $\chi_A(x) = 0$ if $x \in B \setminus A$.

induction. The ideas of that proof, rather than the ideas outlined above, will be used to give a proof of Theorem 7.1 in Subsection 12.2 based indicator random variables (these are essentially the same as characteristic functions mentioned above, but we will work with the framework of random variables).

Proof. For $n = 1$, the theorem simply says that $\mathrm{P}(A_1) = \mathrm{P}(A_1)$, and for $n = 2$, the result was proved as Lemma 2.1. So assume $n \geq 3$, and assume the result is true if n is replaced by $n - 1$. We have

$$\begin{aligned}\mathrm{P}\Big(\bigcup_{i=1}^{n} A_i\Big) &= \mathrm{P}\Big(A_n \cup \bigcup_{i=1}^{n-1} A_i\Big) = \mathrm{P}(A_n) + \mathrm{P}\Big(\bigcup_{i=1}^{n-1} A_i\Big) - \mathrm{P}\Big(A_n \cap \bigcup_{i=1}^{n-1} A_i\Big) \\ &= \mathrm{P}(A_n) + \mathrm{P}\Big(\bigcup_{i=1}^{n-1} A_i\Big) - \mathrm{P}\Big(\bigcup_{i=1}^{n-1} (A_i \cap A_n)\Big),\end{aligned}$$

where the second equation holds according to Lemma 2.1. Using the induction hypothesis for the second and third terms, we obtain

$$\begin{aligned}\mathrm{P}\Big(\bigcup_{i=1}^{n} A_i\Big) = \mathrm{P}(A_n) &+ \sum_{\substack{S:S\subset\{1,2,\ldots,n-1\}\\ S\neq\emptyset}} (-1)^{|S|+1}\,\mathrm{P}\Big(\bigcap_{i\in S} A_i\Big) \\ &- \sum_{\substack{S:S\subset\{1,2,\ldots,n-1\}\\ S\neq\emptyset}} (-1)^{|S|+1}\,\mathrm{P}\Big(\bigcap_{i\in S} (A_i \cap A_n)\Big)\end{aligned}$$

We have $\bigcap_{i\in S}(A_i \cap A_n) = A_n \cap \bigcap_{i\in S} A_i$, and so

$$\begin{aligned}\mathrm{P}\Big(\bigcup_{i=1}^{n} A_i\Big) = \mathrm{P}(A_n) &+ \sum_{\substack{S:S\subset\{1,2,\ldots,n-1\}\\ S\neq\emptyset}} (-1)^{|S|+1}\,\mathrm{P}\Big(\bigcap_{i\in S} A_i\Big) \\ &- \sum_{\substack{S:S\subset\{1,2,\ldots,n-1\}\\ S\neq\emptyset}} (-1)^{|S|+1}\,\mathrm{P}\Big(A_n \cap \bigcap_{i\in S} A_i\Big).\end{aligned}$$

This can also be written as

$$\mathrm{P}\Big(\bigcup_{i=1}^{n} A_i\Big) = \sum_{\substack{S:S\subset\{1,2,\ldots,n\}\\ S\neq\emptyset \text{ and } n\notin S}} (-1)^{|S|+1}\,\mathrm{P}\Big(\bigcap_{i\in S} A_i\Big) + \sum_{\substack{S:S\subset\{1,2,\ldots,n\}\\ n\in S}} (-1)^{|S|+1}\,\mathrm{P}\Big(\bigcap_{i\in S} A_i\Big),$$

where, in the last term, the $-$ sign was changed to $+$, since previously, n was not counted as an element of S in the corresponging term; furthermore, the term $\mathrm{P}(A_n)$ on the right-hand side was incorporated into the second sum in case $S = \{n\}$. Combining the two sums on the right-hand side into a single sum, we obtain equation (7.4), completing the proof. □

7.5 Rencontre

The word “rencontre” means encounter in French. In this game, $n > 0$ paper slips numbered 1, 2, …, n placed in a hat, and all the paper slips are drawn one after the other. An encounter is the event that for at least one i with $1 \le i \le n$, the slip labeled i is drawn on the ith pick. Let this event be denoted by A_i. In this case, Ω can be taken to be the set of all permutations S_n of the set $\{k : 1 \le k \le n\}$, [7.4] each of these being equally likely.[7.5] Interpreting these permutations as one-to-one mappings of this set onto itself (cf. Subsubsection 3.1.2), $\sigma \in A_i$ means $\sigma(i) = i$. That is,

$$A_i = \{\sigma \in S_n : \sigma(i) = i\}.$$

The probability of a rencontre, $\mathrm{P}\Big(\bigcup_{i=1}^n A_i\Big)$, can be calculated using formula (7.4). To use this formula, note that for $S \subset \{k : 1 \le k \le n\}$, we have

$$\bigcap_{i\in S} A_i = \{\sigma \in S_n : \sigma(i) = i \quad \text{for all} \quad i \in S\}.$$

It is easy to count the number of these permutations: the numbers not belonging to S can be freely permuted, and so this number is $(n-|S|)!$, so the probability of this set is $(n-|S|)!/n!$. Given r with $1 \le r \le n$, in formula (7.4) there are $\binom{n}{r}$ terms for which $|S| = r$; that is, the sum of probabilities associated with these terms is

$$\binom{n}{r} \cdot \frac{(n-r)!}{n!} = \frac{n!}{r!(n-r)!} \cdot \frac{(n-r)!}{n!} = \frac{1}{r!}.$$

These sums must be multiplied by $(-1)^{r+1}$ and added for r with $1 \le r \le n$ to obtain the left-hand side of formula (7.4). That is,

$$\mathrm{P}\Big(\bigcup_{i=1}^n A_i\Big) = \sum_{r=1}^n \frac{(-1)^{r+1}}{r!} = 1 + \sum_{r=n+1}^\infty \frac{(-1)^r}{r!} - \sum_{r=0}^\infty \frac{(-1)^r}{r!}$$

On the right-hand side, the second sum equals $1/e$,[7.6] and the first sum tends to 0 when $n \to \infty$. Thus, for large n, the probability is approximately $1 - 1/e$.

8 Random variables

Intuitively, a random variable is a quantity X that assumes random values. In a more mathematical setting, given a probability space Ω, a random variable is a function $X : \Omega \to \mathbb{R}$, where $\mathbb{R}$ is a set of real numbers.[8.1] Since we will be calculating probabilities involving random variables, we need to impose additional restrictions on X so that certain probabilities are meaningful:

Definition 8.1. Let $(\Omega, \mathcal{B}, \mathrm{P})$ be a probability space. A function $X : \Omega \to \mathbb{R}$ is called a random variable if for all real numbers α we have $\{\omega \in \Omega : X(\omega) \le \alpha\} \in \mathcal{B}$.

[7.4] In fact, the symbol S_n for this set of permutations is standard in elementary group theory.

[7.5] That is, for each $\sigma \in \Omega$, $\mathrm{P}(\{\sigma\}) = 1/n!$. One always needs to rembermber that $\mathrm{P}(\sigma)$ is meaningless: elements of Ω are not events; events are subsets of Ω.

[7.6] Recall that

$$e^x = \sum_{r=0}^\infty \frac{x^r}{r!}$$

for all x.

[8.1] That is, we are discussing real-valued random variables here. In a more general setting, a random variable may assume complex values, matrix values, etc.

This means that is X is a random variable, then for all $\alpha \in \mathbb{R}$ the probability on the right-hand side of the equation

$$F_X(\alpha) \overset{def}{=} \mathrm{P}(\{\omega \in \Omega : X(\omega) \le \alpha\}) \tag{8.1}$$

is meaningful. The function F_X defined by this equation is called the *distribution function*[8.2] of the random variable X.

In many considerations involving random variables, there are different possible choices for the space Ω, and often Ω is not even mentioned in the discussion explicity. For this reason, it is customary to denote the event $\{\omega \in \Omega : X(\omega) \le \alpha\}$ as $[X \le \alpha]$ or more simply as $X \le \alpha$. Given a random variable X and a real number α, the probabilities of the events $X < \alpha$, $X \ge \alpha$, $X > \alpha$, $X = \alpha$, and $X \ne \alpha$ are also defined. To see this, note that

$$\{\omega \in \Omega : X(\omega) < \alpha\} = \bigcup_{n=1}^{\infty} \left\{\omega \in \Omega : X(\omega) \le \alpha - \frac{1}{n}\right\};$$

hence it follows that the probability of the event $X < \alpha$ is defined in view of Clause (3) of Definition 2.1. The event $X \ge \alpha$ is the complement of $X < \alpha$, and the event $X > \alpha$ is the complement of $X \le \alpha$, so these events also have probabilities. Furthen, we have

$$\{\omega \in \Omega : X(\omega) = \alpha\} = \{\omega \in \Omega : X(\omega) \le \alpha\} \cap \{\omega \in \Omega : X(\omega) \ge \alpha\}.$$

Further, for any interval I, the probability of the event $X \in I$ is also defined. For example, if $I = (\alpha, \beta]$, then

$$\{\omega \in \Omega : X(\omega) \in I\} = \{\omega \in \Omega : X(\omega) \le \beta\} \cap \{\omega \in \Omega : X(\omega) > \alpha\}.$$

8.1 Functions of random variables

If X is a random variable on Ω and if $g : \mathbb{R} \to \mathbb{R}$ is a function, then the composition $g \circ X$ is a function defined on Ω with real numbers as values. Whether it is a random variable depends on whether the condition given in Definition 8.1 is satisfied. This certainly happens if g is an nice function.[8.3] The random variable $g \circ X$ is usually denoted as $g(X)$, since, intuitively, one thinks of X as a random quantity rather than a function.

8.2 Reading

[2, §3.1–3.4, pp. 40–46], [2, §5.1–5.2 pp. 80–81] up to Example 2.

8.3 Homework

[2, Chapter 3, p. 61], 301–309, [2, Chapter 5, p. 92], 501–502.

[8.2] Also called the *cumulative distribution function*.

[8.3] The widest class of function g for which this can be true is the class of *Borel measurable* functions; unfortunately, we cannot explain here what this class is. It certainly contains all continuous functions, but, actually, it is a much larger class.

9 The Stieltjes integral

There is a clear analogy between the formulas describing Fourier series and trigonometric interpolation with equidistant nodes. This analogy can be brought out more clearly by rewriting the interpolation formulas with the aid of Stieltjes integrals. The next three definitions describe the Riemann–Stieltjes integral.

Definition 9.1 (Partition)**.** A *partition* of the interval $[a, b]$ is a finite sequence $\langle x_0, x_1, \dots, x_n\rangle$ of points such that

$$P : a = x_0 < x_1 < x_2 < \ldots < x_n = b. \tag{9.1}$$

The *width* or *norm* of a partition is

$$\|P\| \overset{def}{=} \max\{x_i - x_{i-1} : 1 \le i \le n\}.$$

Definition 9.2 (Riemann–Stieltjes sum)**.** Given a partition

$$P : a = x_0 < x_1 < x_2 < \ldots < x_n = b.$$

of the interval $[a, b]$, a *tag* for the interval $[x_{i-1}, x_i]$ with $1 \le i \le n$ is a number $\xi_i \in [x_{i-1}, x_i]$ for each i. A partition with a tag for each interval $[x_{i-1}, x_i]$ is called a *tagged* partition. Given a tagged partition as described, and given the functions f and g on $[a, b]$, the corresponding Riemann–Stieltjes sum is

$$S = \sum_{i=1}^{n} f(\xi_i)\big(g(x_i) - g(x_{i-1})\big).$$

The Riemann–Stieltjes integral

$$\int_a^b f(x)\,dg(x)$$

is defined as the limit of the Riemann–Stieltjes sums S associated with the partition P as $\|P\| \to 0$, independently of the choice of the tags. While not important for our purposes, we will give a rigorous definition (if we take $g(x) = x$ in this definition, this gives the usual definition of the Riemann inegral):

Definition 9.3 (Riemann–Stieltjes integral)**.** If there is a real number A such that for every $\epsilon > 0$ there is a $\delta > 0$ such that for any Riemann–Stieltjes sum S for f and g associated with a partition of width $< \delta$ of $[a, b]$ we have $|A - S| < \epsilon$, then we call A the *Riemann–Stieltjes integral* of f with respect to g on $[a, b]$, and we write $A = \int_a^b f\,dg$. In this case we call f *Riemann–Stieltjes integrable* with respect to g on $[a, b]$.

9.1 More on Stieltjes integrals

The only reason we mentioned Stieltjes integrals is to more closely highlight the analogy between Fourier series and trigonometric interpolation. We will include here some simple results to put Stieltjes integrals in the proper context, even though they are not needed for the discussion below. The first one converts Stieltjes integrals into Riemann integrals in certain cases (but not in the case of interest to us above, when the function playing the role of g is not continuous).

Theorem 9.1. *Assume g is differentiable on $[a,b]$. Assume further that the Riemann integral $\int_a^b f(x)g'(x)\,dx$ and the Riemann–Stieltjes integral $\int_a^b f(x)\,dg(x)$ exist. Then*

$$\int_a^b f(x)\,dg(x) = \int_a^b f(x)g'(x)\,dx.$$

Proof. Let

$$P : a = x_0 < x_1 < x_2 < \ldots < x_n = b.$$

a partition of the interval $[a,b]$. By the mean-value theorem of differentiation, for each i with $1 \le i \le n$ there is a $\xi_i \in [x_{i-1}, x_i]$ such that $g'(\xi_i)(x_i - x_{i-1}) = g(x_i) - g(x_{i-1})$.[9.1] Hence we have

$$\sum_{i=1}^{n} f(\xi_i)\bigl(g(x_i) - g(x_{i-1})\bigr) = \sum_{i=1}^{n} f(\xi_i)g'(\xi_i)(x_i - x_{i-1}).$$

Making $\|P\| \to 0$, the left-hand side tends to $\int_a^b f(x)\,dg(x)$ and the right-hand side tends to $\int_a^b f(x)g'(x)\,dx$, completing the proof. □

9.2 Integration by parts

Theorem 9.2 (Integration by Parts)**.** *Assume the integral $\int_a^b f(x)\,dg(x)$ is defined. Then the integral $\int_a^b g(x)\,df(x)$ is also defined and we have*

$$\int_a^b f(x)\,dg(x) = f(b)g(b) - f(a)g(a) - \int_a^b g(x)\,df(x).$$

Proof. For the proof, we redefine the concept of partition by allowing $P = \langle x_i : 1 \le i \le n\rangle$ to be a nondecreasing sequence. This is a harmless change, since the terms $f(\xi_i)\bigl(g(x_i) - g(x_{i-1})\bigr)$ for which $x_{i-1} = x_i$ do not contribute to the Riemann–Stieltjes sum. Let P be such an arbitrary partition; that is

$$P : a = x_0 \le x_1 \le x_2 \le \ldots \le x_n = b,$$

and let $\xi_i \in [x_{i-1}, x_i]$ be arbitrary tags. We have the identity

$$\sum_{i=1}^{n} f(\xi_i)\bigl(g(x_i) - g(x_{i-1})\bigr) = f(x_n)g(\xi_n) - f(\xi_1)g(x_0) + \sum_{i=1}^{n-1} g(x_i)\bigl(f(\xi_i) - f(\xi_{i+1})\bigr).$$

This is easy to verify; namely, the same terms are added on both sides, in different order. Indeed, for i with $1 \le i \le n$, both sides adds the term $+f(\xi_i)g(x_i)$, except that on the right-hand time for $i = n$ this term is written out separately. Further, both sides adds the terms $-f(\xi_i)g(x_{i-1})$ for i with $1 \le i \le n$, even though on the right-hand side this term is written as $-f(\xi_{i+1})g(x_i)$ for i with $1 \le i \le n-1$, and the term corresponding to $i = 0$, i.e., the term $-f(\xi_1)g(x_0)$, is written out separately. This rearrangement of a sum is called *partial summation* or *Abel rearrangement*, named after the Norwegian mathematician Niels Henrik Abel.

[9.1] In order to apply the mean-value theorem, we need to assume that g is real valued, since the mean-value theorem is not valid for complex-valued functions. The result can nevertheless be proved in case g is complex valued by establishing it separately for the real and the imaginary parts of g.

Making the assumption $a = x_0 = \xi_1 = x_1$ and $x_{n-1} = \xi_n = x_n = b$, the above identity becomes

$$\sum_{i=1}^{n} f(\xi_i)\big(g(x_i) - g(x_{i-1})\big) = f(b)g(b) - f(a)g(a) - \sum_{i=1}^{n-1} g(x_i)\big(f(\xi_{i+1}) - f(\xi_i)\big).$$

Considering

$$P' : a = \xi_1 \le \xi_2 \le \xi_3 \le \ldots \le \xi_n = b$$

with the tags $x_i \in [\xi_i, \xi_{i+1}]$ for i with $1 \le i \le n-1$, the right-hand side contains a Riemann–Stieltjes sum for the integral $\int_a^b g(x)\,df(x)$, and the left-hand side contains a Riemann–Stieltjes sum for the integral $\int_a^b f(x)\,dg(x)$; the fact that we allow $x_{i-1} = x_i$ makes no difference here, since the terms with $x_{i-1} = x_i$ make no contribution to the sum.[9.2] Since $\xi_{i-1} \le x_i \le \xi_i \le x_{i+1} \le \xi_{i+1}$ for all i with $1 \le i \le n-1$, $x_0 = x_1$, and $x_{n-1} = x_n$, we have $\|P\| \le 2\|P'\|$. Hence, making $\|P'\| \to 0$, we also have $\|P\| \to 0$; hence the left-hand side tends to $\int_a^b f(x)\,dg(x)$, since this integral was assumed to exist. So, the right-hand side also has a limit; thus, the integral $\int_a^b g(x)\,df(x)$ also exists, it being the limit of the sum on the right-hand side. This completes the proof of the theorem. □

Using Theorems 9.1 and 9.2, we obtain the integration by parts formula for Riemann integrals:

$$\int_a^b f(x)g'(x)\,dx = f(b)g(b) - f(a)g(a) - \int_a^b f'(x)g(x)\,dx \tag{9.2}$$

under appropriate conditions. We leave the formulation of these condiions to the reader.

9.3 Change of variables

We also have a change of variables (i.e., substitution) formula for Riemann–Stieltjes integrals; it is even simpler than the one for regular Riemann integrals. For this, we need to put

$$\int_b^a f(x)\,dg(x) \overset{def}{=} -\int_a^b f(x)\,dg(x) \qquad (a < b),$$

as is usual in case of Riemann integrals. At this point, we might as well put $\int_a^b f(x)\,dg(x) = 0$ in case $a = b$.

Theorem 9.3. *Assume the integral $\int_a^b f(x)\,dg(x)$ exists, and let $h : [A, B] \to [a, b]$ be a nondecreasing or nonincreasing function onto $[a, b]$. Then the integral $\int_A^B f\big(h(t)\big)\,dg\big(h(t)\big)$, exists and we have*

$$\int_A^B f\big(h(t)\big)\,dg\big(h(t)\big) = \int_{h(A)}^{h(B)} f(x)\,dg(x).$$

Note that $h(A) = a$ and $h(B) = b$ in case h is nondecreasing, and $h(A) = b$ and $h(B) = a$ in case h is nonincreasing. As for the proof, it is fairly direct and straightforward except that it involves simple results about uniform continuity, and so we omit the proof.[9.3] Readers familiar with uniform continuity can easily construct a proof.

[9.2] The equality $\xi_i = \xi_{i+1}$ is possible, whether or not we allow the possibility that $x_{i-1} = x_i$. This causes no trouble, just as allowing $x_{i-1} = x_i$ causes no trouble.

[9.3] A function h satisfying the requirements of Theorem 9.3 is necessarily continuous, and so also uniformly continuous.

Problem 9.1. Let f be a function on $[-1,1]$ that is continuous at 0, and let g be the function that is

$$g(x) = \begin{cases} 0 & \text{if } -1 \le x < 0, \\ 1 & \text{if } 0 \le x \le 1. \end{cases}$$

Show that

$$\int_{-1}^{1} f(x)\,dg(x) = f(0).$$

Solution. Let

$$P : -1 = x_0 < x_1 < x_2 < \ldots < x_n = 1$$

be a partition and let $\xi_i \in [x_{i-1}, x_i]$ be a tag for each i with $1 \le i \le n$. Let $k = k(P)$ with $1 \le k \le n$ be such that $x_i < 0$ for $i < k$ and $x_i \ge 0$ for $i \ge k$; clearly, k depends on the partition P. Then

$$g(x_i) - g(x_{i-1}) = \begin{cases} 1 & \text{if } i = k, \\ 0 & \text{if } i \ne k. \end{cases} \qquad (1 \le i \le n).$$

Hence

$$S(P) = \sum_{i=1}^{n} f(\xi_i)\bigl(g(x_i) - g(x_{i-1})\bigr) = f(\xi_k) = f(\xi_{k(P)});$$

$S(P)$ depends on also on the tags, not just on P, but this dependence is not indicated. Making $\|P\| \to 0$, we have $\xi_{k(P)} \to 0$. Since f is continuous at 0, we have

$$\int_{-1}^{1} f(x)\,dg(x) = \lim_{\|P\|\to 0} S(P) = \lim_{\|P\|\to 0} f(\xi_{k(P)}) = f(0).$$

9.4 Improper Stieltjes integrals

In probability theory, Stieltjes integrals are often used on the interval $(-\infty, \infty)$. The above definition only allows the use of finite intervals. For infinite intervals, the integral is taken as the limit of integrals on finie intervals. Such integrals are called improper Stieltjes integrals. We have

$$\int_{a}^{\infty} f(x)\,dg(x) \overset{def}{=} \lim_{b\to+\infty} \int_{a}^{b} f(x)\,dg(x),$$

$$\int_{-\infty}^{b} f(x)\,dg(x) \overset{def}{=} \lim_{a\to-\infty} \int_{a}^{b} f(x)\,dg(x),$$

and

$$\int_{-\infty}^{\infty} f(x)\,dg(x) \overset{def}{=} \lim_{\substack{a\to-\infty \\ b\to+\infty}} \int_{a}^{b} f(x)\,dg(x).$$

In the last limit, $a \to -\infty$ and $b \to +\infty$ independently.[9.4]

[9.4] That is, the limit is approached when a is a very large negative number and b is a very large positive number, but there is no connection between a and b; for example, it would be wrong to assume that $a = -b$. We will not give a precise definition. Suffice it to say that instead of definiting the last integral as a double limit, we can define it as

$$\int_{-\infty}^{0} f(x)\,dg(x) + \int_{0}^{\infty} f(x)\,dg(x).$$

10 Expectation

Intuitively, the expectation of a random variable will be the average value of a random variable, weighted by the probability of assuming this or nearby values. A defintition that works for a large class of distribuions. is as follows.

Definition 10.1. Let $H_i \subset \Omega$ pairwise disjoint nonempty events, where $1 \le i \le n$ for some integer n or $1 \le i < \infty$, such that $\bigcup_i H_i = \Omega$. Let X be a random variable such that X is constant on each H_i. Let $\omega_i \in H_i$ for each i. Then the expectation $\mathrm{E}(X)$ of X is defined as

$$\mathrm{E}(X) = \sum_i X(\omega_i)\,\mathrm{P}(H_i); \tag{10.1}$$

in case the sum on the right-hand side is an infinite sum, we need to assume the sum is absolutely convergent; if it is not, $\mathrm{E}(X)$ will not be defined.

The assumption of absolute convergence is necessary in view of footnote 2.4 on p. 10. There it is explained the the sum on the right-hand side of equation (10.1) is not defined without the assumption of absolute convergence, since there is no clear way to specify the order of the summands. Another issue that arises with the above definition is its soundness; that is, one needs to show that it actually defines something. This is an issue at present, since given a random variable X for which such events H_i can be found, the choice of these events is not unique. What needs to be proven is that if one chooses a different collection K_j of events with the same property, the corresponding sum in equation (10.1) gives the same value. This is not very difficult to do, but we will omit a proof here. Given a general random variable X, the expectation of X can be defined by approximating X with random variables for which the expectation is determined by the above definition – we omit the specific details.

This only gives the expectation of certain large class of random variables, to be called discrete (see a discussion below, in Section 11) it is possible to define expectation for a much larger class by approximating those random variables by members of the class for which it was defined above. The general definition suggested by this line of thought is a kind of integral:

$$\mathrm{E}(X) = \int_\Omega X(\omega)\,d\mathrm{P}(\omega). \tag{10.2}$$

Unfortunately, the precise definition of the integral on the right is beyond the level of the course. Therefore, we have to take a less elegant approach, and define expectation in terms of an improper Stieltjes integral (see Subsection 9.4).

10.1 Expectation via Stieltjes integrals

Let X be a random variable with distribution function F_X (see formula (8.1)) and let g be a nice function; we will consider the expectation of the random variable $g(X)$; see Subsection 8.1. For the sake of simplicity, we will asssume that g is continuos, though this assumption can be relaxed. We first assume that X is bounded; that is, there are a and b such that $a < X < b$.[10.1]

First note that if X were to satisfy the conditions of Definition 10.1, the expectation of $g(X)$ could be calculated by the equation

$$\mathrm{E}\big(g(X)\big) = \sum_i g\big(X(\omega_i)\big)\,\mathrm{P}(H_i);$$

[10.1] Here a may be a huge negative number, and b may be a huge positive number, but they must be finite.

correspoinding to equation (10.1), since $g(X)$ is constant on each H_i. Assuming this is not the case, take a partition P of the interval $[a, b]$ an in (9.1), where we assumed that $a < X < b$. Taking tags ξ_i as in Definition 9.2 the Riemann Stieltjes sum

$$S = \sum_{i=1}^{n} g(\xi_i)\big(F_X(x_i) - F_X(x_{i-1})\big),$$

where F_X is the distribution function (cf. (8.1)) of the random variable X, approximate the Riemann-Stieltjes integral

$$\int_a^b g(x)\, dF_X(x) \tag{10.3}$$

Now, if we define the sets H_i as

$$H_i = \{\omega \in \Omega : x_{i-1} < X \le x_i\},$$

then $\mathrm{P}(H_i) = F_X(x_i) - F_X(x_{i-1})$ according to (8.1), and we define the random variable X_S with $X_S = \xi_i$ for $1 \le X \le n$. then the Riemann sum S can also be written as

$$S = \sum_{i=1}^{n} g(X_S)\, \mathrm{P}(H_i).$$

According to Definition 10.1, we have $\mathrm{E}\big(g(X_S)\big) = S$ for the Riemann-Stieltjes sums S. If the integral in (10.3) exists, this integral will be considered the expectation $\mathrm{E}\big(g(X)\big)$ of $g(X)$. If X is not bounded,[10.2] we need to take an improper Riemann–Stieltjes integral

$$\mathrm{E}\big(g(X)\big) = \int_{-\infty}^{\infty} g(x)\, dF_X(x), \tag{10.4}$$

assuming that this integral is absolutely convergent, i.e., that the integral

$$\int_{-\infty}^{\infty} |g(x)|\, dF_X(x),$$

also exists. The assumption of absolute convergence is required since absolute convergence is important in Definition 10.1.

Note that in the above considerations the random variables X_S approximate the random variable X, and in fact, the equation

$$\mathrm{E}(X) = \int_{-\infty}^{\infty} x\, dF_X(x) \tag{10.5}$$

(this is equation (10.4) with $g(x) = x$) will be taken as the definition of $\mathrm{E}(X)$. As for equation (10.4), this should be taken as a consequence of the definition (quite a useful consequence in applications below), though it is not easy to derive the latter equation from the definition given in (10.3) In

[10.2] We may also allow g not to be bounded, but we would still want to require that g is bounded on finite integrals. More precisely, it is necessary to require that the Riemann–Stieltjes integral exists. The function F_X is clearly nondecreasing, and it is known that if g is continuous and F_X is nondecreasing, then the integral $\int_a^b g\, dF_X$ exists, but it may exist also under less stringent assumptions.

fact, this equation is not especially suited for theoretical discussions. For example, to derive the equation $\mathrm{E}(X+Y)=\mathrm{E}(X)+\mathrm{E}(Y)$ from this definition required methods of multivariate calculus. The definition (10.2) is much more suited for such discussion - unfortunately, that definition requires more mathematical backgound than is supposed in these notes.

In equation (10.5) it is no longer necessary to explicitly assume absolute convergence, since this is implicitly ensured by the definition of the improper Stieltjes integral in Subsection 9.4 that the in the limit $a \to -\infty$ and $b \to \infty$ a and b change independently (most importantly, it is not assumed that $a=-b$. If the integral in this equation is not convergent, then $\mathrm{E}(X)$ will not be defined.

10.2 Reading

[2, §6.1–6.2, pp. 95-101], up to (c) The Expectation of a Function of Several rv's.

11 Discrete and continuous random variables

To overcome the difficulty presented by not being able to use the definition of expectation given in equation (10.2), we will need to a traditional distinction of discrete and continous random random variables. While not all random variables fall in either of these two classes, this will be sufficient for us to discuss a wide range of material. The distinction between discreta and continuous random variables is unnecessary in a more advanced presentation.

11.1 Discrete random variables

A *discrete*[11.1] random variable X is one that only assumes countably many values. This is exactly the kind of random variable covered by Definition 10.1. Instead of using the distribution function, it is customary to use the *probability function* for such a variable. Given a discrete random variable, this is the function

$$p_X(x) = \mathrm{P}(X = x). \tag{11.1}$$

This function can be defined for all real x, but it will be nonzero only for the countably many values $x \in \mathrm{ra}(X)$, [11.2] i.e., only if X actually assumes the value x (and even then, it may be zero). The expectation of a discrete X can be calculated as

$$\mathrm{E}(X) = \sum_{x \in \mathrm{ra}(X)} x p_X(x). \tag{11.2}$$

This corresponds to formula (10.1) when writing x_i for the elememnts of $\mathrm{ra}(X)$ and $H_i = \{\omega \in \Omega : X(\omega) = x_i\}$, assuming this sum is absolutely convergent.[11.3] If $g : \mathbb{R} \to \mathbb{R}$ is a function, then we can calculate the expectation of $g(X)$ as

$$\mathrm{E}(g(X)) = \sum_{x \in \mathrm{ra}(X)} g(x) p_X(x),$$

again assuming absolute convergence. This equation is also covered by Definition 10.1.

[11.1] The word "discrete" has been discussed before. See footnote 4.1 on p. 15.

[11.2] $\mathrm{ra}(X)$ is the range of the function X, i.e., the set of values asssumed by the random variable X.

[11.3] The choice of H_i in Definition 10.1 is of course not unique, but the one we gave here is a suitable choice.

11.2 Continuous random variables

A random variable X is called continuous if its distribution function F_X is continuous. We need to impose more restrictions of the distribution function for us to be able to calculate the expectation of X without Riemann–Stieltjes integrals. In addition to assuming that F_X is continous, we will assume that it is differentiable, perhaps with the exceptions of a few points (to be discussed later). The derivative of F'_X is called the *density function* of the random variable X:

$$f_X(x) \overset{def}{=} F'_X(x) = \frac{d}{dx}\,\mathrm{P}(\{\omega \in \Omega : X(\omega) \le x\}. \tag{11.3}$$

The issue with the definition of the density function is that for the density function to be useful, we need to have

$$F_X(x) = \int_{-\infty}^{x} f_X(t)\,dt \tag{11.4}$$

for all $x \in \mathbb{R}$. If $f_X(x) = F'_X(x)$ for all $x \in \mathbb{R}$ and f_X is continuous everywhere then this equation is satisfied, but it is known to hold also in other situation. For example, if f_X is Riemann integrable on all finite intervals, the condition is satisfied even if f_X is not continuous everywhere. But there are other situations when a density function f_X satisfying equation (11.4) exists even when F_X is not differentiable everywhere.

According to Theorem 9.1 the expectation of X given by formula (10.5) can be calculated as

$$\mathrm{E}(X) = \int_{-\infty}^{\infty} x f_X(x)\,dx \tag{11.5}$$

assuming that f_X is continuous. Even if f_X is continuous, this integral may not exist; but then the Rieamann-Stieltjes integral in (10.5) does not exist, either, and $\mathrm{E}(X)$ is not defined. Similarly, given a nice function $g : \mathbb{R} \to \mathbb{R}$, according to formula (10.4), the expectation of $g(X)$ can be calculated as

$$\mathrm{E}\big(g(X)\big) = \int_{-\infty}^{\infty} g(x) f_X(x)\,dx. \tag{11.6}$$

Here, in addition to the continuity of f_X, we need to assume that this integral is absolutely convergent – the reason is similar to the situation with formulas (10.4) and (10.5), where we needed to assume absolute convergence for the former, while for the latter this was ensured by the definition of the improper integral (note that $f_X \ge 0$ since F_X is nondecreasing). The assumption on the continuity can be somewhat relaxed. For example, we may even allow $f_X = F'_X$ to be not defined at finitely many points.[11.4]

11.3 Homework

[2, Chapter 3, p. 61], 309–320. I did not mention quartiles in the notes, though they occur in the problems. To deal with these problems, you need to look up the definition of quartiles and the related examples in the textbook. While quartiles are of little importance in the course for now, this is something you can easily do on your own. See [2, §3.6, pp. 53–53]. The rest of this chapter will be assigned as reading later in the notes.

[11.4] In a Riemann integral, the integrand must be defined at every point of the interval of integration, so at a point where f_X is not defined, we need to assign an arbitrary value to f_X – this will not influence the value of the integral. If f_X has infinite discontinuity at finitely many points, it may still be possible to use these integrals as improper integrals (in a Riemann integral, the integrand must be bounded).

12 Properties of expectation

12.1 Linearity of expectation

If X and Y are random variables and c is a real number such that the expectations of X and Y exist, then we have

$$\mathrm{E}(X + cY) = \mathrm{E}(X) + c\,\mathrm{E}(Y). \tag{12.1}$$

This follows immediately from equation (10.2); unfortunately, the integral in this equation cannot be explained in these notes. While equation (10.5) could be taken as a definition of expectation, proving the above equation by using this equation is too cumbersome, and requires multivariate calculus. The approach we are going to take is to note that it is sufficient to approach the result for discrete random variables, since all random variables can be approximated by discrete random variables; in fact, in the proof of (10.4) and (10.5) we used the discrete random variable X_S to approximate X.

For discrete variables, formula (10.2) is not the one to use, Instead, we can use equation (10.1) in Definition 10.1 with sets H_i that work for both of the discrete random variables X and Y. Such sets can be defined as the nonempty ones among all the sets

$$\{\omega : X(\omega) = x \quad \text{and} \quad Y(\omega) = y\}$$

for $x, y \in \mathbb{R}$. Since X and Y assume only countable many values, the number of these sets is countable. Labeling them as H_i and taking $\omega_i \in H_i$ for some or all positive integers i (depending on whether there are finitely many or infinitely many among these sets), according to equation (10.1) we have

$$\begin{aligned}\mathrm{E}(X + cY) &= \sum_i \bigl(X(\omega_i) + cY(\omega_i)\bigr)\,\mathrm{P}(H_i)\\ &= \sum_i X(\omega_i)\,\mathrm{P}(H_i) + c\sum_i Y(\omega_i)\,\mathrm{P}(H_i) = \mathrm{E}(X) + c\,\mathrm{E}(Y),\end{aligned}$$

establishing equation (12.1) for discrete rancom variables.

12.1.1 Expectiation of a constant

A constant a can can be considered as a random variable Y such that $Y(\omega) = a$ all the time. We naturally have $\mathrm{E}(Y) = a$ in this case. While this remark may seem frivolous, using equation (12.1) it shows that

$$\mathrm{E}(X + a) = \mathrm{E}(X) + a. \tag{12.2}$$

12.2 Probability of union, revisited

Given an event A, its indicator random variable I_A is the random variable defined as

$$I_A(\omega) = \begin{cases} 1 & \text{if } \omega \in A, \\ 0 & \text{if } \omega \notin A. \end{cases} \tag{12.3}$$

It is immediate that $\mathrm{E}(I_A) = \mathrm{P}(A)$.[12.1] We are going to give a new proof of Theorem 7.1. Indeed, writing I_i for the indicator variable of the event A_i for i with $1 \le i \le n$ in that theorem, then we have

$$\prod_{i=1}^{n} \bigl(1 + (-1)I_i\bigr) = \sum_{k=0}^{n} \sum_{\substack{S : S \subset \{1,2,\dots,n\} \\ |S|=k}} \prod_{i \in S} (-1)I_i;$$

indeed, this is just equation (3.3) with $x = 1$ and $y_i = (-1)I_i$; here there was no reason here to break up the sum according to the cardinality of S, but we wanted to show the parallel with equation (3.3). This can be written more appropriately for our purposes as

$$1 - \prod_{i=1}^{n} (1 - I_i) = 1 + \sum_{S : S \subset \{1,2,\dots,n\}} (-1)^{|S|+1} \prod_{i \in S} I_i = \sum_{\substack{S : S \subset \{1,2,\dots,n\} \\ S \ne \emptyset}} (-1)^{|S|+1} \prod_{i \in S} I_i;$$

here, the minus sign before the product is accounted for by the 1 in the exponent of $(-1)^{|S|+1}$ after the first equation. Further, the term corresponding to $S = \emptyset$ cancels against the 1 added to the sum – so this 1 does not appear on the right-hand side. Taking expectations and using equation (12.1), this equation becomes

$$\mathrm{E}\Bigl(1 - \prod_{i=1}^{n} (1 - I_i)\Bigr) = \sum_{\substack{S : S \subset \{1,2,\dots,n\} \\ S \ne \emptyset}} (-1)^{|S|+1}\, \mathrm{E}\Bigl(\prod_{i \in S} I_i\Bigr). \tag{12.4}$$

Note that

$$\bigcup_{i=1}^{n} A_i = \Omega - \bigcap_{i=1}^{n} A_i^*,$$

we have

$$1 - \prod_{i=1}^{n} \bigl(1 - I_i(\omega)\bigr) = 1 \qquad \text{if and only if} \qquad \omega \in \bigcup_{i=1}^{n} A_i;$$

furhtermore,

$$\prod_{i \in S} I_i(\omega) = 1 \qquad \text{if and only if} \qquad \omega \in \bigcap_{i \in S} A_i.$$

Using these, equation (12.4) becomes identical to equation (7.4), giving an alternate proof of Theorem 7.4.

12.3 Independent random variables

Given $n \ge 1$, the random variables X_1, X_2, …, X_n are called independent, if for all intervals $I_i \subset \mathbb{R}$ $(1 \le i \le n)$, the events $\{\omega \in \Omega : X_i(\omega) \in I_i\}$, $1 \le i \le n$, are independent. Here the intervals can be closed, open, and the one-point interval $[a, a] = \{a\}$ for $a \in \mathbb{R}$ is allowed. We have

Theorem 12.1. *Let $n \ge 1$ be an integer, and let X_1, X_2, …, X_n be independent random variables. Then*

$$\mathrm{E}\Bigl(\prod_{i=1}^{n} X_i\Bigr) = \prod_{i=1}^{n} \mathrm{E}(X_i). \tag{12.5}$$

[12.1] The easy calculation is shown below in equation (14.1), where indicator variables are discussed in more detail.

We will prove this only for $n = 2$, and only in case the variables X_i are discrete. The extension to larger values of n is fairly routine, and the extension to arbitrary random variables can be done similarly as in the proof of equation (12.1)

Proof. Assume X and Y are independent discrete variables. We have

$$\mathrm{E}(XY) = \sum_{x,y\in\mathbb{R}} xy\,\mathrm{P}(\{\omega \in \Omega : X(\omega) = x \text{ and } Y(\omega) = y\});$$

in view of equation (10.1). This is because the sets

$$\{\omega \in \Omega : X(\omega) = x \text{ and } Y(\omega) = y\}$$

are pairwise disjoint (for different among these sets X or Y will assume different values, so those sets cannot have any element in common), there are only countably many nonempty ones among them since X and Y only assume countably many values, and those can be taken as H_i in equation (10.1). As X and Y are independent, we have

$$\mathrm{P}(\{\omega \in \Omega : X(\omega) = x \text{ and } Y(\omega) = y\}) = \mathrm{P}(\{\omega \in \Omega : X(\omega) = x\}) \cdot \mathrm{P}(\{\omega \in \Omega : Y(\omega) = y\}).$$

Hence we have

$$\begin{aligned} \mathrm{E}(XY) &= \sum_{x,y\in\mathbb{R}} xy\,\mathrm{P}(\{\omega \in \Omega : X(\omega) = x\}) \cdot \mathrm{P}(\{\omega \in \Omega : Y(\omega) = y\}) \\ &= \sum_{x\in\mathbb{R}} x\,\mathrm{P}(\{\omega \in \Omega : X(\omega) = x\}) \cdot \sum_{x\in\mathbb{R}} y\,\mathrm{P}(\{\omega \in \Omega : Y(\omega) = y\} = \mathrm{E}(X)\,\mathrm{E}(Y), \end{aligned}$$

where the last equation also holds by equation (10.1). □

The following lemma is almost immediate.

Lemma 12.1. *Let $n \geq 1$, and assume the random variables X_i for i with $1 \leq i \leq n$ are independent. Let a_i, $1 \leq i \leq n$ be real numbers. Then the random variables $X_i + a_i$, $1 \leq i \leq n$ are also independent.*

Proof. Let I_i be an interval. We have $X_i + a_i \in I_i$ if and only if $X_i \in \{x - a_i : x \in I_i\}$. Noting that the set $\{x - a_i : x \in I_i\}$ is also an interval, the result follows from the definition of independence. □

12.4 Homework

[2, Chapter 6, p. 112], 601–610, [2, Chapter 6, p. 128], 701.

13 Variance

Given a random variable X, its variance is defined as

$$\mathrm{V}(X) = \mathrm{E}\Big(\big(X - \mathrm{E}(X)\big)^2\Big). \tag{13.1}$$

As we noted before, the expection of a random variable is not always defined (see the discussion following equation (10.5).) *A fortiori*,[13.1] its variance is not always defined either. The variance of a random variable measures the deviation of a random variable from its expectation.

[13.1] Latin, meaning "even more so," or "with a stronger reason."

Some simple observations are in order. Let X be a random variable that has a variance, and let $c \in \mathbb{R}$. Then

$$\mathrm{V}(X+c) = V(X). \tag{13.2}$$

Indeed, writing $Y = X + c$, we have $\mathrm{E}(Y) = \mathrm{E}(X) + c$; hence $Y - \mathrm{E}(Y) = X - \mathrm{E}(X)$. Hence the assertion is an immediate consequence of equation (13.1). Given X and c as before, we also have

$$\mathrm{V}(cX) = c^2 V(X). \tag{13.3}$$

In showing this, in view of (13.2) we may assume that $\mathrm{E}(X) = 0$ (since we may replace X with $Y = X - \mathrm{E}(X)$), and $\mathrm{E}(Y) = 0$. Then we also have $\mathrm{E}(cX) = c\,\mathrm{E}(X)$, and so

$$\mathrm{V}(cX) = \mathrm{E}\big((cX)^2\big) = \mathrm{E}(c^2X^2) = c^2\,\mathrm{E}(X^2) = c^2\,\mathrm{V}(X),$$

as we wanted to show.

13.1 Moments; the calculation of variance

Given a positive integer n, the nth moment of a random variable X is defined as $\mathrm{E}(X^n)$. The second moment has an important use in calculating the variance. Indeed, writing $\mu = \mathrm{E}(X)$, we have

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

For the last equation, note that μ^2 is just a number, which can be taken as a random variable assuming a constant value, so its expectation is just μ^2. Given that $\mu = \mathrm{E}(X)$, this shows with

$$\mathrm{V}(X) = \mathrm{E}(X^2) - \big(\mathrm{E}(X)\big)^2, \tag{13.4}$$

13.2 Pairwise independent variables

Given an integer $n > 0$, the random variables X_1, X_2,…, X_n are called *pairwise independent* if any two of them are independent.

It is important to note that pairwise independent random variables are not necessarily independent, as the following example due to Sergei Bernstein shows. Let X_1, X_2, and X_3 be three random variables be defined as follows. Tossing a fair coin twice, let $X_1 = 1$ if the first toss is a head, 0 otherwise, let $X_2 = 1$ if the second toss is a head, 0 otherwise, let $X_3 = 1$ if exaclty one of the two tosses is a head, 0 otherwise. We leave it to the reader to show that these variables are pairwise independent, but not independent.

The following theoem has important applications:

Theorem 13.1. *Let $n > 0$ be an integer, and assume the random variables variables X_i for i with $1 \le i \le n$ pairwise independent each of which has variance. Then*

$$\mathrm{V}\Big(\sum_{i=1}^{n} X_i\Big) = \sum_{i=1}^{n} \mathrm{V}(X_i). \tag{13.5}$$

Proof. We may assume that $\mathrm{E}(X_i) = 0$ for each i, otherwise we may replace X_i with $X_i - \mathrm{E}(X_i)$ in view of equation (13.2); further, this replacement does not affect the assumption of pairwise independence in view of Lemma 12.1. With this assumption we have

$$\mathrm{V}\Big(\sum_{i=1}^{n} X_i\Big) = \mathrm{E}\left(\Big(\sum_{i=1}^{n} X_i\Big)^2\right) = \mathrm{E}\Big(\sum_{i=1}^{n}\sum_{j=1}^{n} X_iX_j\Big) = \sum_{i=1}^{n}\sum_{j=1}^{n} \mathrm{E}(X_iX_j).$$

For $i = j$ the terms on the right-hand side become $\mathrm{E}(X_i^2)$. When $i \neq j$, noting that X_i and X_j are independent, by Theorem 12.1 for these terms we have $\mathrm{E}(X_i X_j) = \mathrm{E}(X_i)\,\mathrm{E}(X_j) = 0$, since we assumed that $\mathrm{E}(X_i) = 0$ for each i. Thus we have

$$\mathrm{V}\Big(\sum_{i=1}^{n} X_i\Big) = \sum_{i=1}^{n} \mathrm{E}(X_i^2) = \sum_{i=1}^{n} \mathrm{V}(X_i),$$

as we wanted to show. □

13.3 Reading

[2, §7.1–7.2, pp. 117–119],

13.4 Homework

[2, Chapter 6, p. 128], 702.

14 Some discrete distributions

14.1 One-point distribution

This is the distribution of the random variable X for which $X(\omega) = a$ with some a for all $\omega \in \Omega$. For his variable, we have

$$p_X(a) = 1.$$

for its probability function.

14.2 Two-point distribution

If the random variable X assumes two values, a and b. Its probability function is

$$p_X(a) = p, \qquad p_X(b) = q,$$

for some p and q with $0 \le p \le 1$ and $q = 1 - p$.

14.3 Indicator variables

Indicator variables have been described before, in equation (12.3). Given an event A, its indicator variable is defined as the variable $I_A = 1$ is A occurs, and $I_A = 0$ otherwise. Let $p = \mathrm{P}(A)$; then

$$p_X(1) = p, \qquad p_X(0) = q,$$

where $q = 1 - p$. We have

$$\mathrm{E}(I_A) = 1 \cdot p + 0 \cdot q = p. \tag{14.1}$$

Noting that $I_A^2 = I_A$ we also have $\mathrm{E}(I_A^2) = p$, and using equation (13.4), we obtain

$$\mathrm{V}(I_A) = p - p^2 = p(1 - p) = pq, \tag{14.2}$$

14.4 Uniform distribution

In this distribution, the variable X that assumes the values 1, 2, ..., n ($n > 0$) with equal probability is said to have uniform distibution. We have

$$p_X(i) = \frac{1}{n} \qquad (1 \le i \le n).$$

14.5 Geometric distribution

Given two numbers real p and q with $0 < p < 1$ and $q = 1 - p$, a random variable with geometric distribution assumes the values k for all integers k with $k \ge 0$ such that

$$p_X(k) = q^k p \qquad (k \ge 0).$$

Note that the total probability is 1, as it should be for all probability functions:

$$\sum_{k=0}^{\infty} q_X(k) = \sum_{k=0}^{\infty} q^k p = p \sum_{k=0}^{\infty} q^k = p\,\frac{1}{1-q} = 1,$$

where the third equation holds in view of the sum formula of the geometric series, and the last equation holds because $p = 1 - q$.

14.6 For the first time (fft) distribution

This is almost the same as the geometric distribution, but here X assumes the values k, where k is an integer with $k \ge 1$, and given numbers real p and q with $0 < p < 1$ and $q = 1 - p$, we have

$$p_X(k) = q^{k-1} p \qquad (k \ge 1).$$

This random variable can be realized by the following experiment. Assume in repeated independent trials of an experiment, each time the probability of success is p, and numbering these trials by 1, 2, 3, ..., let $X = k$ if the experiment suceeds first on the kth trial. The expectation of X can be calculated as follows.

$$\begin{aligned}
\mathrm{E}(X) &= \sum_{k=1}^{\infty} k p_X(k) = \sum_{k=1}^{\infty} k p q^{k-1} = p \sum_{k=1}^{\infty} \frac{d}{dq} q^k = p \frac{d}{dq} \sum_{k=1}^{\infty} q^k \\
&= p \frac{d}{dq} \Big(-1 + \sum_{k=0}^{\infty} q^k \Big) = p \frac{d}{dq} \left(-1 + \frac{1}{1-q} \right) \\
&= p\,\frac{1}{(1-q)^2} = (1-q)\,\frac{1}{(1-q)^2} = \frac{1}{1-q};
\end{aligned}$$

for the third equation, we made use of the theorem that power series can be termwise differentiated inside the interval of convergence.

For calculating the variance, we will make use of equation (13.4); instead of directly calculating the second moment $\mathrm{E}(X^2)$, it will, however, be more natural to first calculate $\mathrm{E}\big(X(X-1)\big)$, and note that $\mathrm{E}(X^2) = \mathrm{E}(X) + \mathrm{E}\big(X(X-1)\big)$. We have

$$\begin{aligned}\mathrm{E}\big(X(X-1)\big) &= \sum_{k=1}^{\infty} k(k-1)p_X(k) = \sum_{k=2}^{\infty} k(k-1)p_X(k) = \sum_{k=2}^{\infty} k(k-1)pq^{k-1}\\ &= qp\sum_{k=2}^{\infty} k(k-1)q^{k-2} = qp\sum_{k=2}^{\infty}\frac{d^2}{dq^2}q^k = qp\frac{d^2}{dq^2}\sum_{k=2}^{\infty}q^k\\ &= qp\frac{d^2}{dq^2}\Big(-1-q+\sum_{k=0}^{\infty}q^k\Big) = qp\frac{d^2}{dq^2}\left(-1-q+\frac{1}{1-q}\right)\\ &= qp\,\frac{2}{(1-q)^3} = q(1-q)\,\frac{2}{(1-q)^3} = \frac{2q}{(1-q)^2};\end{aligned}$$

here the second equation holds since $k-1=0$ for $k=1$. Thus,

$$\mathrm{E}(X^2) = \mathrm{E}(X) + \mathrm{E}\big(X(X-1)\big) = \frac{1}{1-q} + \frac{2q}{(1-q)^2} = \frac{1-q+2q}{(1-q)^2} = \frac{1+q}{(1-q)^2}$$

Hence, using equation (13.4), we have

$$\mathrm{V}(X^2) = \mathrm{E}(X^2) - \big(\mathrm{E}(X)\big)^2 = \frac{1+q}{(1-q)^2} - \frac{1}{(1-q)^2} = \frac{q}{(1-q)^2}.$$

14.7 Reading

[2, §3.5, pp. 47–48], Material from [2, §3.6–3.8, pp. 49-60] will be covered later in the notes.

15 Binomial distribution

Let A be an event of probability p indicating the success of a certain trial. We want to repeat the trial n times ($n > 0$) independently, and let A_i indicate that the ith trial is successful.[15.1] Write $q = 1-p$. Given k with $0 \le k \le n$, the event[15.2]

$$\Big(\bigcap_{i=1}^{k} A_i\Big) \cap \Big(\bigcap_{i=k+1}^{n} A_i^*\Big)$$

indicate that the first k trials are successful, and the rest of them are not. Since these events are independent (cf. Corollary 6.1), and the probability of A_i is p, so that of A_i^* is $1-p=q$, the probability of this event is $p^k q^{n-k}$.

Let X be the random variable assuming values k with $0 \le k \le n$ such that $X = k$ indicates that out of n independent repetition of the trials just described, exactly k is successful, and let

[15.1] Some caution is needed here. The probability space corresponding to a single trial is not the same as the probability space corresponding to n repetitions of the trial. For example, rolling a die can be described by the probability space $\Omega = \{i : 1 \le i \le 6\}$, but rolling a die twice needs to be described by the probability space $\Omega' = \{(i,j) : 1 \le i, j \le 6\}$. In fact, Ω' is the Cartesian product $\Omega' = \Omega \times \Omega$. Similarly, the probability space corresponding to the n independent repetitions of the above trial can be described by a Cartesian product. We will omit the technical details.

[15.2] For $k = 0$, the intersection for $i = 1$ to k indicates the empty intersection, which naturally would be meaningless, since it indicates the set of all sets, which is not permitted. However, note that the notation is allowed in a context when the intersection is taken with another set (see the comment at the end of Subsection 1.3), and then the resulting set is just the other set in the intersection, i.e. the intersection from $k+1$ to n. A similar discussion applied to the case $k = n$.

$1 \le i_1 < i_2 < \cdots < i_k \le n$ describe the places where the trial is successul.[15.3] For each fixed choice $\{i_l : 1 \le l \le k\}$ of the k-element set the probability of such a result of the trials is the same as above, where successes occurred in the beginning. For different k-element sets these events are mutually exclusive, since if i is an element of one of the sets but not the others, this indicates a success at the ith place in one sequence of trials, and failure in the other sequence. Since there are $\binom{n}{k}$ such k-element sets, the total probability of k successful outcomes, we have

$$p_X(k) = \mathrm{P}(X = k) = \binom{n}{k} p^k q^{n-k} \qquad (0 \le k \le n). \tag{15.1}$$

A variable X having this distribution is said to have a *binomial distribution* with parameters n and p; in symbols, $X \sim \mathrm{Bin}(n,p)$.

15.1 Binomial variable as a sum of indicator variables

For i with $1 \le i \le n$, let I_i be the random variable defined as

$$I_i(\omega) = \begin{cases} 1 & \text{if } \omega \in A_i, \\ 0 & \text{if } \omega \notin A_i; \end{cases}$$

that is, $I_i = 1$ if the ith trial is successful and $I_i = 0$ otherwise. Then we have for the binomial variable X:

$$X = \sum_{i=1}^{n} I_i.$$

Hence, according to equation (14.1), we have

$$\mathrm{E}(X) = \sum_{i=1}^{n} \mathrm{E}(I_i) = np. \tag{15.2}$$

Similarly, given that the random variables I_i, independent, by equation (14.2) and by Theorem 13.1 we have

$$\mathrm{V}(X) = \sum_{i=1}^{n} \mathrm{V}(I_i) = npq. \tag{15.3}$$

15.2 Direct determination of the expectation and the variance of the binomial distribution

While the direct determination of the expectation and the variance of a binomial variable is more complicated than using indicator variables, the method is worth studying since it shows important techniques of calculation. First note that for k and n with $1 \le k \le n$ we have

$$k\binom{n}{k} = k\frac{n!}{k!(n-k)!} = k\frac{n(n-1)!}{k(k-1)!(n-k)!} = n\frac{(n-1)!}{(k-1)!(n-k)!} = n\binom{n-1}{k-1}. \tag{15.4}$$

Using this for the binomial variable X, we have

$$\mathrm{E}(X) = \sum_{k=0}^{n} k p_X(k) = \sum_{k=1}^{n} k p_X(k) = \sum_{k=1}^{n} k\binom{n}{k} p^k q^{n-k} = \sum_{k=1}^{n} n\binom{n-1}{k-1} p^k q^{n-k};$$

[15.3] The argument that follows is similar to the one used in Subsection 4.1 to describe picking marbles with replacement.

in the second equality, the zero term corresponding to $k = 0$ was omitted from the summation, and in the fourth equality, we used equation (15.4). Hence, we have

$$\mathrm{E}(X) = np\sum_{k=1}^{n}\binom{n-1}{k-1}p^{k-1}q^{(n-1)-(k-1)} = np\sum_{l=0}^{n-1}\binom{n-1}{l}p^{l}q^{(n-1)-l} = np(p+q)^{n-1} = np;$$

here, in the second equation, we replaced the summation variable k with $l = k - 1$. The third equation was obtained by the binomial theorem, and, for the fourth, note that $p + q = 1$.

The calculation of the variance is using similar ideas. As in Subsection 14.6, to this end, we calculate $\mathrm{E}\big(X(X-1)\big)$. we have

$$\begin{aligned}\mathrm{E}\big(X(X-1)\big) &= \sum_{k=0}^{n}k(k-1)p_X(k) = \sum_{k=2}^{n}k(k-1)p_X(k)\\ &= \sum_{k=2}^{n}k(k-1)\binom{n}{k}p^kq^{n-k} = \sum_{k=2}^{n}n(n-1)\binom{n-2}{k-2}p^kq^{n-k};\end{aligned}$$

here, in the second equality, we omitted the first two terms that are 0, and in fourth equality, we used equation (15.4) twice. Further, we have

$$\begin{aligned}\mathrm{E}\big(X(X-1)\big) &= n(n-1)p^2\sum_{k=2}^{n}\binom{n-2}{k-2}p^{k-2}q^{(n-2)-(k-2)}\\ &= n(n-1)p^2\sum_{l=0}^{n-2}\binom{n-2}{l}p^lq^{(n-2)-l} = n(n-1)p^2(p+q)^{n-2} = n(n-1)p^2;\end{aligned}$$

here, similarly as before in the second equation, we replaced the summation variable k with $l = k-2$. The third equation was obtained by the binomial theorem, and, for the fourth, note that $p+q = 1$.

Hence, we have

$$\begin{aligned}\mathrm{E}(X^2) &= \mathrm{E}(X) + \mathrm{E}\big(X(X-1)\big) = np + n(n-1)p^2\\ &= np - np^2 + n^2p^2 = np(1-p) + n^2p^2 = npq + n^2p^2.\end{aligned}$$

and

$$\mathrm{V}(X) = \mathrm{E}(X^2) - \big(\mathrm{E}(X)\big)^2 = npq + n^2p^2 - n^2p^2 = npq.$$

15.3 Reading

[2, §9.1–9.2, pp. 92–151]. Material in the rest of §9.2 on or after p. 152 will be discussed later in the notes.

16 Some continuous distributions

16.1 Uniform distribution

Given an interval $[a, b]$, a random variable X is said to be uniformly distributed on this interval if X assumes values only in this interval, and the probability of X assuming a value in a subinterval of this interval is proportional to the length of the interval. The density function of X can be described as follows:

$$f_X(x) = \begin{cases} 1/(b-a) & \text{if } a \le x \le b, \\ 0 & \text{otherwise.} \end{cases}$$

As for the distribution function, for $x \in [a, b]$ we have

$$F_X(x) = \int_{-\infty}^{x} f_X(t)\,dt = \int_{a}^{x} \frac{1}{b-a}\,dt = \frac{1}{b-a} t\Big|_{t=a}^{x} = \frac{x-a}{b-a}.$$

That is,

$$F_X(x) = \begin{cases} 0 & \text{if } -\infty < x < a, \\ (x-a)/(b-a) & \text{if } a \le x \le b, \\ 1 & \text{if } b < x. \end{cases}$$

Note that $f_X(x) = F'_X(x)$ for all x different from a and b. For $x = a$ and $x = b$ $F'_X(x)$ is not defined; at these points, $f_X(x)$ could be assigned any value, and it will not make any difference; the choice we made above was convenient, but arbitrary.

As for the expectation of X, we have

$$\mathrm{E}(X) = \int_{-\infty}^{\infty} x f_X(x)\,dx = \int_{a}^{b} x\,\frac{1}{b-a}\,dx = \frac{1}{2(b-a)} x^2\Big|_{x=a}^{b} = \frac{b^2-a^2}{2(b-a)} = \frac{a+b}{2}.$$

As for the second moment, we have

$$\begin{aligned} \mathrm{E}(X^2) &= \int_{-\infty}^{\infty} x^2 f_X(x)\,dx = \int_{a}^{b} x^2\,\frac{1}{b-a}\,dx = \frac{1}{3(b-a)} x^3\Big|_{x=a}^{b} = \frac{b^3-a^3}{3(b-a)} \\ &= \frac{(b-a)(a^2+ab+b^2)}{3(b-a)} = \frac{1}{3}\,(a^2+ab+b^2). \end{aligned}$$

Hence, for the variance, we obtain

$$\begin{aligned} \mathrm{V}(X) &= \mathrm{E}(X^2) - \big(\mathrm{E}(X)\big)^2 = \frac{1}{3}\,(a^2+ab+b^2) - \frac{1}{4}\,(a^2+2ab+b^2) \\ &= \frac{1}{12}\,(4a^2+4ab+4b^2-3a^2-6ab-3b^2) = \frac{1}{12}\,(a^2-2ab+b^2) = \frac{1}{12}\,(b-a)^2. \end{aligned}$$

16.2 Cauchy distribution

Given a real number $a > 0$, the random variable X with density function

$$f_X(x) = \frac{a}{\pi}\,\frac{1}{a^2+x^2} \qquad (x \in \mathbb{R}).$$

is said to have Cauchy distribution. As for the factor a/π, it is there to make sure that

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

Indeed, we have

$$\int_{-\infty}^{\infty} \frac{1}{a^2+x^2}\,dx = \frac{1}{a} \lim_{\substack{A\to-\infty\\ B\to\ \ \infty}} \int_A^B \frac{1}{1+(x/a)^2}\,\frac{1}{a}\,dx = \frac{1}{a} \lim_{\substack{A\to-\infty\\ B\to\ \ \infty}} \int_{A/a}^{B/a} \frac{1}{1+t^2}\,dt$$
$$= \frac{1}{a} \lim_{\substack{A\to-\infty\\ B\to\ \ \infty}} \left(\arctan\frac{B}{a} - \arctan\frac{A}{a}\right) = \frac{\pi}{a}.$$

The Cauchy distribution has important applications in physics, but our main interest in it is that it is an example for a distribution that has no expectation. Indeed, the integral expressing the expectation is

$$\mathrm{E}(X) = \int_{-\infty}^{\infty} x\,\frac{1}{a^2+x^2}\,dx;$$

this integral, however, is divergent. This is easily shown by using the comparison test for the integral on the interval $[0,\infty)$. Furthermore, the integral can easily be calculated on finite intervals by making he substitution $t = a^2 + x^2$, when $dt = 2x\,dx$, and a direct calculation shows that the integral on the interval $[a,b]$ diverges when we make $a \to -\infty$ and $b \to \infty$.[16.1]

16.3 Exponential distribution

Assume a certain equipment is placed in service at time $x = 0$, and for any small Δx the probability of failure in the interval $[x, x+\Delta x]$ is approximately $\lambda\Delta x$, assuming it has not failed before.[16.2] The random variable X will have value x if the equipment fails at time x. Clearly, $x \ge 0$, since the equipment was not in service before time 0. The random variable X so described is said to have an *exponential distribution.*

16.3.1 The distribution function of the exponential distribution

Given $x > 0$, to find the probability of $X > x$, let $\Delta x > 0$ be small. We have

$$\frac{\mathrm{P}(X > x+\Delta x)}{\mathrm{P}(X > x)} = \frac{\mathrm{P}(X > x+\Delta x \,\&\, X > x)}{\mathrm{P}(X > x)}$$
$$= \mathrm{P}(X > x+\Delta x \mid X > x) = 1 - \mathrm{P}(X \le x+\Delta x \mid X > x)$$
$$= 1 - \mathrm{P}(x < X \le x+\Delta x \mid X > x) \approx 1 - \lambda\Delta x;$$

here the second equation holds in view of the definition of conditional probability in equation (5.1). [16.3] The fourth equation uses the fact that for any two events we have $\mathrm{P}(A \mid B) = \mathrm{P}(A \cap B \mid B)$, also an immediate consequence of equation (5.1). The approximate equation at the end holds in view of the description of the random variable X. Writing $G(x) = \mathrm{P}(X > x)$, this equation can be written as

$$\frac{G(x+\Delta x) - G(x)}{G(x)} = \frac{G(x+\Delta x)}{G(x)} - 1 \approx -\lambda\Delta x,$$

[16.1] When $a = -b$, the integral on the interval $[a,b]$ is zero, but when taking the limit, a and b must vary independently.

[16.2] We used a closed interval here, but since the probability of failure at any single point of time is 0, it makes no difference whether the interval is open or closed. We need to say Δx is small, since the dependence on the length of the interval is not linear, so this relation can only hold in the limit when $\Delta x \to 0$.

[16.3] We used the logic symbol & for "and" rather than the intersection symbol $\cup$, since the events are described as relations rather than as sets.

or else, as

$$\frac{G(x+\Delta x)-G(x)}{\Delta x} \approx -\lambda G(x).$$

Making $\Delta x \to 0$, this gives the equation

$$G'(x) = -\lambda G(x).$$

This is a differential equation; we can solve it as follows. Writing $y = G(x)$, this equation says $dy/dx = -\lambda y$, i.e.,

$$\frac{dy}{y} = -\lambda dx.$$

Integrating this, we obtain

$$\int \frac{dy}{y} = \int (-\lambda) dx,$$

that is,

$$\ln |y| = -\lambda x + C.$$

for some constant C. The absolute value is not needed since y denotes a probability, so $0 \le y \le 1$. This means that $y = e^{-\lambda x + C}$. That is,

$$\mathrm{P}(X > x) = G(x) = e^{-\lambda x + C}.$$

Noting that we have $G(0) = \mathrm{P}(X > 0) = 1$, this means that $e^C = 1$, and so $C = 0$. Thus, we have $G(x) = e^{-\lambda x}$. Therefore,

$$F_X(x) = \mathrm{P}(X \le x) = 1 - G(x) = 1 - e^{-\lambda x}.$$

$$F_X(x) = \begin{cases} 0 & \text{if } -\infty < x \le 0, \\ 1 - e^{-\lambda x} & \text{if } x > 0. \end{cases}$$

We have $f_X(x) = F'_X(x)$, that is

$$f_X(x) = \begin{cases} 0 & \text{if } -\infty < x < 0, \\ \lambda e^{-\lambda x} & \text{if } x \ge 0. \end{cases}$$

At $x = 0$ the derivative $F'_X(x)$ is not defined, and the value of $f_X(0)$ is chosen arbitrarily; this will have no effect on any of the calculations.

16.3.2 Alternative derivation of the distribution function

Given $x > 0$, to find the probability of $X > x$, let n be a large positive integer, and divide the interval $(0, x]$ into n parts

$$I_i = \left(\frac{i-1}{n}x, \frac{i}{n}x\right] \qquad (1 \le i \le n).$$

For i with $1 \le i \le n$ we have

$$\begin{aligned} \mathrm{P}\left(X > \frac{ix}{n} \,\middle|\, X \ge \frac{(i-1)x}{n}\right) &= \mathrm{P}\left(X \notin I_i \,\middle|\, X \ge \frac{(i-1)x}{n}\right) \\ &= 1 - \mathrm{P}\left(X \in I_i \,\middle|\, X \ge \frac{(i-1)x}{n}\right) \approx 1 - \lambda \frac{x}{n} \end{aligned}$$

Thus,

$$\begin{aligned}\mathrm{P}\left(X > \frac{ix}{n}\right) &= \mathrm{P}\left(X > \frac{ix}{n} \,\&\, X \ge \frac{(i-1)x}{n}\right)\\ &= \mathrm{P}\left(X > \frac{ix}{n}\,\middle|\, X \ge \frac{(i-1)x}{n}\right)\mathrm{P}\left(X > \frac{(i-1)x}{n}\right) \approx \left(1-\lambda\frac{x}{n}\right)\mathrm{P}\left(X > \frac{i-1}{n}x\right);\end{aligned}$$

The first equation holds because if $X > ix/n$ then also $X > (i-1)x/n$, so saying that both events occur just means that the first event occurs.[16.4] As for the second equation, that holds in view of the definition of conditional probability in equation (5.1). Using these equations for i with $1 \le i \le n$, and noting that for $i = 1$ we have $\mathrm{P}(X > (i-1)x/n) = \mathrm{P}(X > 0) = 1$, we obtain

$$P(X > x) \approx \left(1 - \lambda\frac{x}{n}\right)^n.$$

Making $n \to \infty$, we obtain

$$P(X > x) = \lim_{n\to\infty}\left(1 - \lambda\frac{x}{n}\right)^n = e^{-\lambda x};$$

see [7] for a discussion of the exponential function, where this limit is taken to be the definition of the exponential function. Thus, for $x > 0$, we have $F_X(x) = \mathrm{P}(X \le x) = 1 - \mathrm{P}(X > x) = 1 - e^{-\lambda x}$.

16.3.3 The expectation and the variance of the exponential distribution

To calculate the expectation of X is not difficult. We have

$$\mathrm{E}(X) = \int_{-\infty}^{\infty} x f_X(x)\,dx = \int_0^\infty \lambda x e^{-\lambda x}\,dx = \lim_{A\to\infty}\int_0^A \lambda x e^{-\lambda x}\,dx.$$

Using integration by parts with $f(x) = x$ and $g'(x) = \lambda e^{-\lambda x}$ (cf. equation (9.2)), when we have $f'(x) = 1$, and $g(x) = -e^{-\lambda x}$, we obtain

$$\begin{aligned}\mathrm{E}(X) &= \lim_{A\to\infty}\left(-xe^{-\lambda x}\Big|_{x=0}^{A} + \int_0^A e^{-\lambda x}\,dx\right) = \lim_{A\to\infty}\left(-Ae^{-\lambda A} - \frac{1}{\lambda}e^{-\lambda x}\Big|_{x=0}^{A}\right)\\ &= \lim_{A\to\infty}\left(-Ae^{-\lambda A} - \frac{1}{\lambda}e^{-\lambda A} + \frac{1}{\lambda}\right).\end{aligned}$$

Noting that $\lim_{A\to\infty} Ae^{-\lambda A} = 0$ and $\lim_{A\to\infty} e^{-\lambda A} = 0$,[16.5] we obtain that

$$\mathrm{E}(X) = \int_0^\infty \lambda x e^{-\lambda x}\,dx = \frac{1}{\lambda}. \tag{16.1}$$

As for the second moment, we have

$$\mathrm{E}(X^2) = \int_{-\infty}^{\infty} x^2 f_X(x)\,dx = \int_0^\infty \lambda x^2 e^{-\lambda x}\,dx = \lim_{A\to\infty}\int_0^A \lambda x^2 e^{-\lambda x}\,dx.$$

[16.4] We used the logic symbol & for "and" rather than the intersection symbol ∪, since the events are described as relations rather than as sets.

[16.5] To establish the first limit, we have

$$\lim_{A\to\infty} Ae^{-\lambda A} = \lim_{A\to\infty}\frac{A}{e^{\lambda A}}. = \lim_{A\to\infty}\frac{1}{\lambda e^{\lambda A}} = 0,$$

where we used l'Hospital's rule to obtain the second equation.

Integrating by parts with $f(x) = x^2$ and $g'(x) = \lambda e^{-\lambda x}$ (cf. equation (9.2)), when we have $f'(x) = 2x$, and $g(x) = -e^{-\lambda x}$, we arrive at

$$\mathrm{E}(X^2) = \lim_{A\to\infty} \Big(-x^2 e^{-\lambda x} \Big|_{x=0}^{A} + \int_0^A 2xe^{-\lambda x}\, dx \Big) = -\lim_{A\to\infty} A^2 e^{-\lambda A} + \int_0^\infty 2xe^{-\lambda x}\, dx$$

The limit on the right-hand side is 0,[16.6] and the integral is $2/\lambda^2$ according to (16.1). Thus, $E(X^2) = 2/\lambda^2$. Hence,

$$\mathrm{V}(X) = \mathrm{E}(X^2) - \big(\mathrm{E}(X)\big)^2 = \frac{2}{\lambda^2} - \frac{1}{\lambda^2} = \frac{1}{\lambda^2}.$$

16.4 Reading

[2, §3.7, pp. 53–56], up to (c) Normal Distribution, [2, §3.8, pp. 59–60].

17 The normal distribution

The normal distribution is the distribution most widely used in statistics. This is because of the Central Limit Theorem, to be discussed below, in Subsection 17.3.

17.1 The standard normal distribution

The random variable X has standard normal distribution if its density function is

$$f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}. \tag{17.1}$$

the scaling factor $1/\sqrt{2\pi}$ is there to ensure that $\int_{-\infty}^{\infty} f_X(x)\, dx = 1$. That this is indeed so, one can verify by using the formula

$$\int_{-\infty}^{\infty} e^{-x^2}\, dx = \sqrt{\pi}; \tag{17.2}$$

see [8, Formula (2.3), Subsection 2.1, p. 6].[17.1]

17.1.1 The expectation of the standard normal distribution

We have

$$\mathrm{E}(X) = \int_{-\infty}^{\infty} \frac{x}{\sqrt{2\pi}} e^{-x^2/2}\, dx = \int_{-\infty}^{0} \frac{x}{\sqrt{2\pi}} e^{-x^2/2}\, dx + \int_{0}^{\infty} \frac{x}{\sqrt{2\pi}} e^{-x^2/2}\, dx = 0;$$

the last equation holds since the integrals after the second equation cancel each other, as can be seen by substituting $t = -x$ in one of them.[17.2]

[16.6] Use l'Hospital's rule twice for the limit $\lim_{A\to\infty} A^2/e^{\lambda A}$.

[17.1] The evaluation of this integral requires multivariate calculus.

[17.2] We would cation against directly using the change of variable rule on improper integrals. While this would be possible to do, one needs to know the precise rules when this is allowed. It is usually preferable to rewrite the improper integral on an infinite integral as a limit of an integral on a finite interval. Note that when using the equation $\int_{-\infty}^{\infty} = \lim_{\substack{a\to-\infty\\ b\to\infty}} \int_a^b$, a and b must vary independently; we definitely cannot assume that $a = -b$. The integral can also be fully evaluated using the substitution $t = -x^2/2$ without splitting it up, but it is easier to point out the cancelation.

17.1.2 The variance of the standard normal distribution

For the variance, we have

$$\begin{aligned}\mathrm{V}(X)=\mathrm{E}(X^2)&=\int_{-\infty}^{\infty}\frac{x^2}{\sqrt{2\pi}}e^{-x^2/2}\,dx=2\int_{0}^{\infty}\frac{x^2}{\sqrt{2\pi}}e^{-x^2/2}\,dx\\&=-\frac{2}{\sqrt{2\pi}}\lim_{A\to\infty}\int_0^A x\,(-xe^{-x^2/2})\,dx.\end{aligned}$$

Using integration by parts with $f(x)=x$ and $g'(x)=-xe^{-x^2/2}$ (cf. equation (9.2)), when we have $f'(x)=1$, and $g(x)=e^{-x^2/2}$, we obtain

$$\begin{aligned}\mathrm{V}(X)=\mathrm{E}(X^2)&=-\frac{2}{\sqrt{2\pi}}\lim_{A\to\infty}\left(xe^{-x^2/2}\Big|_{x=0}^{A}-\int_0^A e^{-x^2/2}\,dx\right)\\&=-\frac{2}{\sqrt{2\pi}}\lim_{A\to\infty}\left(Ae^{-A^2/2}-\int_0^A e^{-x^2/2}\,dx\right)=\frac{2}{\sqrt{2\pi}}\int_0^\infty e^{-x^2/2}\,dx\\&=\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}}\,e^{-x^2/2}\,dx=\int_{-\infty}^{\infty}f_X(x)\,dx=1;\end{aligned}$$

the fifth equation holds since the integral from 0 to ∞ is the same as the integral from $-\infty$ to 0. The integrand after this equation is the same as the density function, and so the integral is 1.

17.1.3 The distribution function of the standard normal distribution

The distribution function of the standard normal distribution is often denoted by Φ:

$$\Phi(x)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x}e^{-t^2/2}\,dt. \tag{17.3}$$

In view of the importance of the standard normal distribution, most statistics books have tables of the function Φ; for example [2, p. 324].

17.2 The general normal distribution

Let Y be a random variable with standard normal distribution, and let $\sigma>0$ and μ be real numbers. Then the random variable $X=\sigma Y+\mu$ is said to have general normal distribution denoted as $\mathcal{N}(\mu,\sigma^2)$. We have $\mathrm{E}(X)=\mu$ and $\mathrm{V}(X)=\sigma^2$ according to equations (12.1), (13.2), and (13.3).

17.2.1 The distribution function of the general normal distribution

We have

$$\begin{aligned}F_X(x)&=\mathrm{P}(X\le x)=\mathrm{P}(\sigma Y+\mu\le x)=\mathrm{P}\left(Y\le\frac{x-\mu}{\sigma}\right)=\Phi\left(\frac{x-\mu}{\sigma}\right)\\&=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{(x-\mu)/\sigma}e^{-t^2/2}\,dt=\lim_{A\to-\infty}\int_A^{(x-\mu)/\sigma}\frac{1}{\sqrt{2\pi}}e^{-t^2/2}\,dt.\end{aligned}$$

Making the change of variables $u=\sigma t+\mu$, we have $du=\sigma\,dt$, i.e., $dt=(1/\sigma)\,du$; further, we have $t=(u-\mu)/\sigma$. As for the limits, for $t=A$ we have $u=\sigma A+\mu$, and for $t=(x-\mu)/\sigma$, we have $u=x$. Thus,

$$F_X(x)=\lim_{A\to-\infty}\int_{A\sigma+\mu}^{x}\frac{1}{\sqrt{2\pi}}\,e^{-(u-\mu)^2/(2\sigma^2)}\,\frac{1}{\sigma}\,du=\frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{x}e^{-(u-\mu)^2/(2\sigma^2)}\,du.$$

17.2.2 The density function of the general normal distribution

For the density function of x we have

$$f_X(x) = \frac{d}{dx} F_X(x) = \frac{1}{\sqrt{2\pi}\sigma}\, e^{-(x-\mu)^2/(2\sigma^2)}.$$

17.3 The Central Limit Theorem

The Central Limit Theorem states that if X_1, X_2, X_3, …, is an infinite sequence of *independent, identically distributed* (abbreviated as i. i. d.[17.3]) random variables with $\mathrm{E}(X_i) = \mu$ and $\mathrm{V}(X_i) = \sigma^2$ (i.e., we assume that the expectation and the variance of these variables exists)[17.4] Then the Central Limit Theorem asserts that the distribution of the average

$$\bar{X}_n \overset{def}{=} \frac{1}{n}\sum_{i=1}^{n} X_i$$

approximates a normal distribution. To state this precisely, note that

$$\mathrm{E}(\bar{X}_n) = \frac{1}{n}\sum_{i=1}^{n} \mathrm{E}(X_i) = \frac{1}{n}\sum_{i=1}^{n} \mu = \mu$$

and

$$\mathrm{V}(\bar{X}_n) = \frac{1}{n^2}\sum_{i=1}^{n} \mathrm{V}(X_i) = \frac{1}{n^2}\sum_{i=1}^{n} \sigma^2 = \frac{1}{n}\sigma^2.$$

according to equations (12.1), (13.3), and Theorem 13.1. Using also equation (13.2), we obtain that for $Y_n = \sqrt{n}\,(\bar{X}_n - \mu)/\sigma$ we have $\mathrm{E}(Y_n) = 0$ and $\mathrm{V}(Y_n) = 1$. The way Y_n results from $\bar{X}_n$ is called *standardization*, and Y_n is called the *standard* variable corresponding to $\bar{X}_n$. The Lindeberg–Lévy Central Limit Theorem can be formulated as follows.

Theorem 17.1 (Lindeberg–Lévy Central Limit Theorem)**.** *Let $\sigma > 0$ and μ be real numbers, let X_1, X_2, X_3, …, be an infinite sequence of independent, identically distributed random variables with $\mathrm{E}(X_i) = \mu$ and $\mathrm{V}(X_i) = \sigma^2$ for each $i > 0$, and for each $n > 0$ write*

$$\bar{X}_n \overset{def}{=} \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Then, for all real x we have

$$\lim_{n\to\infty} \mathrm{P}\left(\frac{\sqrt{n}\,(\bar{X}_n - \mu)}{\sigma} < x\right) = \Phi(x). \tag{17.4}$$

There are many kinds of convergence in probability theory. The convergence described in this theorem is called convergence *in distribution.* That is, the standardized variable on the left-hand side of side of equation (17.4) is said to converge in distribution to the standard normal variable. The proof of this theorem is above the level of this course.

[17.3] An important abbreviation to remember, since it is used frequently.

[17.4] The existence of the second moment of a random variable ensures that both its expectation and its variance exist, though the proof may not be appropriate for the level of this course.

17.4 Normal approximation to the binomial distribution

Given that a binomial variable $X \sim \mathrm{Bin}(n, p)$ is the sum of n i.i.d. indicator random variables, as mentioned in Subsection 15.1, it satisfies the conditions of the Central Limit Theorem 17.1. Given that $\mathrm{E}(X) = np$ and $\mathrm{V}(X) = npq$, the corresponding standardized variable is $(X - np)/\sqrt{np(1-p)}$. Hence, according to the Central Limit Theorem, we have

$$\mathrm{P}\left(\frac{X - np}{\sqrt{np(1-p)}} \le t\right) \approx \Phi(t) \qquad (t \in \mathbb{R}).$$

With $t = (x - np)/\sqrt{np(1-p}$, this gives

$$\mathrm{P}\left(\frac{X - np}{\sqrt{np(1-p)}} \le \frac{x - np}{\sqrt{np(1-p)}}\right) = \Phi\left(\frac{x - np}{\sqrt{np(1-p)}}\right) \qquad (x \in \mathbb{R}).$$

That is,

$$F_X(x) = \mathrm{P}(X \le x) \approx \Phi\left(\frac{x - np}{\sqrt{np(1-p)}}\right) \qquad (x \in \mathbb{R}); \tag{17.5}$$

this holds because the inequality on the left-hand side of the previous formula is equivalent to $X \le x$. This approximation is of practical use when $n \ge 30$. Given that $X \in [0, n]$, one would not use this approximation for x outside this interval. For x close to 0, better approximations are available, based on the Poisson distribution.

17.4.1 Continuity correction

The binomial variable $X \sim \mathrm{Bin}(n, p)$ can assume only integer values, whereas the normal variable approximating it is a continous variable. Given an integer k with $0 \le k < n$, the events $X \le k$ and $X < k + 1$ are the same. So, when one wants to approximate the probability $\mathrm{P}(X \le k)$ by using formula equation (17.5), one gets a more accurate result if in this equation, instead of taking $x = k$, one takes $x = k + 1/2$. Doing so is called the *continuity correction*.

17.5 Reading

[2, §8.1–8.4, pp. 130–139],

18 The Poisson distribution

The Poisson distribution describes the distribution of the following random variable X. Assume a certain event A may happen several times in the time interval $[0, 1]$; the occurrences are independent, and in any time interval of length δ the probability of A occurring is approximately $\lambda\delta$, where $\lambda > 0$ is a given constant, the approximation is better for small δ, it becoming exact as $\delta \to 0$. That, for an intervals $I \subset [0, 1]$ we have

$$\lim_{|I| \to 0} \frac{\mathrm{P}(A \text{ occurs in } I)}{|I|} = \lambda,$$

where $|I|$ denotes the length of the interval I. Then X denotes the number of occurrences of A in the interval $[0, 1]$. X can assume the values 0, 1, 2, ...

To find the probability $\mathrm{P}(X = k)$ for each $k \ge 0$, for a large integer n, divide the interval $[0, 1)$ into n equal parts:[18.1]

$$I_i = \left[\frac{i-1}{n}, \frac{i}{n}\right) \qquad (1 \le i \le n).$$

The probability of A occurring in I_i is approximately λ/n; these events being independent. So the probability of A occurring in k of these intervals follows the Binomial distribution $\mathrm{Bin}(n, \lambda/n)$:

$$\begin{aligned}\mathrm{P}(X = k) &\approx \binom{n}{k}\left(\frac{\lambda}{n}\right)^k \left(1 - \frac{\lambda}{n}\right)^{n-k} = \frac{\prod_{i=0}^{k-1}(n-i)}{k!}\frac{\lambda^k}{n^k}\left(1 - \frac{\lambda}{n}\right)^n \left(1 - \frac{\lambda}{n}\right)^{-k} \\ &= \frac{\lambda^k}{k!}\left(1 - \frac{\lambda}{n}\right)^n \left(1 - \frac{\lambda}{n}\right)^{-k} \frac{\prod_{i=0}^{k-1}(n-i)}{n^k} = \frac{\lambda^k}{k!}\left(1 - \frac{\lambda}{n}\right)^n \left(1 - \frac{\lambda}{n}\right)^{-k} \prod_{i=0}^{k-1}\frac{n-i}{n}\end{aligned}$$

Making $n \to \infty$ we obtain the exact formula for the probability function of X. Noting that we have

$$\lim_{n\to\infty}\left(1 + \frac{x}{n}\right)^n = e^x$$

for any real x (see [7]), we find that

$$\mathrm{P}(X = k) = \lim_{n\to\infty}\frac{\lambda^k}{k!}\left(1 - \frac{\lambda}{n}\right)^n \left(1 - \frac{\lambda}{n}\right)^{-k} \prod_{i=0}^{k-1}\frac{n-i}{n} = \frac{\lambda^k}{k!}e^{-\lambda},$$

given that the limit of the last two factors is 1.

In the calculation we neglected the possibility that the event A may occur more than once in an interval I_i. However, the probability for A to occur l times in an interval I_i is about λ^i/n^l, so[18.2] using the formula

$$\sum_{n=0}^{\infty} x^n = \frac{1}{1-x} \qquad (|x| < 1)$$

for the geometric series, we can calculate the probability for A to occur more than in the interval I_i for a single value of i is about

$$\sum_{l=2}^{\infty}\frac{\lambda^l}{n^l}\frac{\lambda^l}{n^l} = \frac{\lambda^2}{n^2}\sum_{l=2}^{\infty}\frac{\lambda^{l-2}}{n^{l-2}} = \frac{\lambda^2}{n^2}\sum_{l=2}^{\infty}\frac{\lambda^{l-2}}{n^{l-2}} = \frac{\lambda^2}{n^2}\sum_{m=0}^{\infty}\frac{\lambda^m}{n^m} = \frac{\lambda^2}{n^2}\frac{1}{1-\lambda/n} = \frac{\lambda^2}{n(n-\lambda)},$$

where the third equation was obtained by replacing the summation variable l by $m + 2$, in which case $m = l - 2$ and the summation runs from $m = 0$ to $m = \infty$. So the probability that A occurs at least once in one of the intervals I_i for some $1 \le i \le n$ is at most $\lambda^2/(n - \lambda)$; actually, it is somewhat less, since the events A occurring more than once in I_i for different values of i are not mutually exclusive.

Note that the probabilities $\mathrm{P}(X = k)$ for $k = 0, 1, 2, \ldots$ add up to 1, as they should:

$$\sum_{k=0}^{\infty}\mathrm{P}(X = k) = \sum_{k=0}^{\infty}\frac{\lambda^k}{k!}e^{-\lambda} = e^{-\lambda}\sum_{k=0}^{\infty}\frac{\lambda^k}{k!} = e^{-\lambda}e^{\lambda} = 1,$$

[18.1] The occurrence of A at any particular time (i.e., in an interval of length zero) is zero, so we omit the point 1 from the interval $[0, 1]$ so as to simplify the calculation.

[18.2] We are intentionally using n as the summation variable to stress the point that the summation variable has no meaning outside the sum, so this use of n does not conflict of the use of n denoting the number of interval. We would have a problem if we needed to use that n inside the summation. Further note that 0^0 is meaningless, but in situations where the sum involves x^n, for $n = 0$ and $x = 0$ we take x^n to be 1 by a wide-spread mathematical convention.

where the penultimate[18.3] equality holds in view of the Taylor series of e^λ. The expectation and the variance of the Poisson distribution can be calculated as the limit of the approximating binomial distribution. Here we give direct calculation.

$$\mathrm{E}(X) = \sum_{k=0}^{\infty} k\,\frac{\lambda^k}{k!}e^{-\lambda} = \sum_{k=1}^{\infty} k\,\frac{\lambda^k}{k!}e^{-\lambda} = \sum_{k=1}^{\infty} \frac{\lambda^k}{k!/k}e^{-\lambda} = \sum_{k=1}^{\infty} \lambda\frac{\lambda^{k-1}}{(k-1)!}e^{-\lambda};$$

here in the second equation we changed the lower limit of summation from $k = 0$ to $k = 1$, since the term for $k = 0$ is 0 anyway (since it is multiplied by k. For the third equation, we moved k to the denominator as $1/k$, and for the fourth equation we used the fact that $k! = (k-1)!\,k$ for $k \ge 1$ ($0! = 1$ by definition) in the form $k!/k = (k-1)!$. We also wrote λ^k as $\lambda\,\lambda^{k-1}$.

Next, we will move λ and $e^{-\lambda}$ outside from the scope of the summation sign, and replace the summation variable k with $l + 1$, in which case $l = k - 1$; so we will sum from $l = 0$ to $l = \infty$. We obtain

$$\mathrm{E}(X) = \lambda e^{-\lambda}\sum_{l=0}^{\infty}\frac{\lambda^l}{l!} = \lambda e^{-\lambda}e^{\lambda} = \lambda,$$

where for the second equation we used the Taylor series of e^λ.

The variance $\mathrm{V}(X)$ of X can be calculated from the formula $\mathrm{V}(X) = \mathrm{E}(X^2) - \big(\mathrm{E}(X)\big)^2$, where, to calculate the first term on the right-hand side, we will use the formula $\mathrm{E}(X^2) = \mathrm{E}\big(X(X-1)\big) + \mathrm{E}(X)$. We have

$$\begin{aligned}\mathrm{E}\big((X(X-1)\big) &= \sum_{k=0}^{\infty} k(k-1)\,\frac{\lambda^k}{k!}e^{-\lambda} = \sum_{k=2}^{\infty} k(k-1)\,\frac{\lambda^k}{k!}e^{-\lambda}\\ &= \sum_{k=2}^{\infty}\frac{\lambda^k}{k!/\big(k(k-1)\big)}e^{-\lambda} = \sum_{k=2}^{\infty}\lambda^2\frac{\lambda^{k-2}}{(k-2)!}e^{-\lambda};\end{aligned}$$

Similarly to the calculation of $\mathrm{E}(X)$, we omitted the terms for $k = 0$ and $k = 1$, which are zero in view of their being multiplied by $k(k-1)$. After the third equation, we moved $k(k-1)$ to the to the denominator, and then noted that $k!/\big(k(k-1)\big) = (k-2)!$ for $k \ge 2$. We also replaced λ^k with $\lambda^2\,\lambda^{k-2}$. We will next replace the summation variable k with $l + 2$, in which case $l = k - 2$; so we will sum from $l = 0$ to $l = \infty$. We obtain

$$\mathrm{E}\big((X(X-1)\big) = \lambda^2 e^{-\lambda}\sum_{l=0}^{\infty}\frac{\lambda^l}{l!} = \lambda^2 e^{-\lambda}e^{\lambda} = \lambda^2,$$

where for the second equation we used the Taylor series of e^λ. Thus,

$$\mathrm{E}(X^2) = \mathrm{E}\big(X(X-1)\big) + \mathrm{E}(X) = \lambda^2 + \lambda.$$

Hence

$$\mathrm{V}(X) = \mathrm{E}(X^2) - \big(\mathrm{E}(X)\big)^2 = \lambda^2 + \lambda - \lambda^2 = \lambda.$$

The description of the Poisson distribution in terms of events in the time interval $[0, 1]$ can naturally be restated in terms of distances, areas, volumes, or weight. For example, assuming infinitely small chocolate chips, given the average density of chips in a unit amount of dough, one can describe the probability that there are k chocolate chips in a cookie using a unit amount of dough.

[18.3] The one before the last one.

18.1 Reading

The Poisson distribution is discussed in the textbook as an approximation to the binomial distribution (see [2, p. 153 middle]) and also in its own right [2, p. 49, p. 153 middle, pp. 158–160, upto but not including Theorem 7].

18.2 Homework

The homework that can be found in the book [2] has been posted at the webside as the file website for the course as the file homework.pdf. The homework most closely related to the material presented here can be found at the end of Chapter 9. Homework from earlier chapters have been dealt with in the past.

19 The gamma distribution

19.1 The gamma fuction

The gamma function is defined by the integral

$$\Gamma(t) \overset{def}{=} \int_0^\infty x^{t-1}e^{-x}\,dx \qquad (t>0); \tag{19.1}$$

This is an improper integral for two reasons. First, the upper limit is $+\infty$, second, for the values of t with $0<t<1$ the integrand becomes infinite at $x=0$. The easiest way to interpret this integral by splitting it up at, say $x=1$ in case $0<t<1$:

$$\Gamma(t)=\int_0^1 x^{t-1}e^{-x}\,dx+\int_1^\infty x^{t-1}e^{-x}\,dx=\lim_{\epsilon\searrow 0}\int_\epsilon^1 x^{t-1}e^{-x}\,dx+\lim_{A\to\infty}\int_1^\infty x^{t-1}e^{-x}\,dx$$

here $\epsilon\searrow 0$ means that ϵ tends to zero from the right,[19.1] The second integral on the right-hand side is convergent for all values of t,[19.2] The first integral on the right-hand side is convergent for only for $t>0$. [19.3] An important property of the gamma function is that

$$\Gamma(t+1)=t\Gamma(t) \qquad (t\ne 0,-1,-2,-3\ldots). \tag{19.2}$$

It is not hard to prove this formula for $t>0$.[19.4] Indeed, using the integration by parts formula

$$\int_a^b f(x)g'(x)\,dx = f(x)g(x)\Big|_{x=a}^b-\int_a^b f'(x)g(x)\,dx \tag{19.3}$$

[19.1] The notation suggests that ϵ needs to decrease in order to reach 0.

[19.2] This is not hard to show by using the comparison test for integrals, which is similar to the comparison test for series. For large values of x, the integrand is less than $e^{-x/2}$.

[19.3] The comparison test used near 0 shows that the first integral is convergent if and only if

$$\int_0^1 x^{t-1}\,dx$$

is convergent, since $e^{-1}\le e^{-x}<1$ for $0<x\le 1$; this integral is convergent for $t>0$, and it is divergent for $t\le 0$. The gamma function can be extended to all values of t, real or complex, with the exception of the values 0, -1, -2, -4, ... If one considers complex values of t, then the integral is convergent whenever $\Re x>0$. We will, however, not be concerned with complex values of t in these notes,

[19.4] In fact, defining $\Gamma(t)$ for all complex values of t with $\Re(t)>0$ by formula (19.1), one way to extend $\Gamma(t)$ for any other values in its domain by using formula (19.2).

with $f(x) = e^{-x}$ and $g'(x) = tx^{t-1}$, when $f'(x) = -e^{-x}$ and $g(x) = x^t$. For small $\epsilon > 0$ and large positive A we have

$$\int_{\epsilon}^{A} e^{-x}\, tx^{t-1}\, dx = e^{-x}x^t \Big|_{x=\epsilon}^{A} + \int_{\epsilon}^{A} e^{-x}\, x^t\, dt.$$

When $\epsilon \searrow 0$, i.e., ϵ tends to 0 from the right,[19.5] and $A \to \infty$, the left-hand side tends to $t\Gamma(t)$, the first term on the right tends to 0, and the second term tends to $\Gamma(t+1)$, verifying equation (19.2).

For $t = 1$, the integral in formula (19.1) is easily evaluated. For positive A, we have

$$\Gamma(1) = \int_0^\infty e^{-x}\, dx = \lim_{A\to\infty} \int_0^A e^{-x}\, dx = \lim_{A\to\infty} \Big(-e^{-x} \Big|_{x=0}^{A} \Big) = \lim_{A\to\infty} \big(-e^{-A}) - (-e^0) \big) = 1.$$

This shows that $\Gamma(1) = 1 = 0!$, and so, using equation (19.2) repeatedly we can see that for any $n \geq 1$ the equation $\Gamma(n+1) = n!$ holds. Thus, the gamma function can be considered an extension of the factorial for noninteger arguments.

19.2 The gamma distribution

The gamma distribution has two parameters. The continuous random variable X with values $X \geq 0$ has $\Gamma(\alpha, \theta)$ distribution for positive real numbers α and θ if we have

$$f_X(x) = \begin{cases} \frac{1}{\Gamma(\alpha)\theta^\alpha}\, x^{\alpha-1} e^{-x/\theta} & \text{if} \quad x > 0, \\ 0 & \text{if} \quad x \leq 0. \end{cases} \tag{19.4}$$

Using the defintion of the gamma function given in equation (19.1), we can see that $\int_1^\infty f_X(x)\, dx = 1$. The special case of $\alpha = 1$ gives the exponential distribution. See the Wikipedia page of the gamma distribution at for graphs of the density functionof the gamma distribution. Here α is called the shape parameter, and θ is called the scale parameter.[19.6]

19.3 Sums of independent gamma distributions of the same scale

We have the following

Theorem 19.1. *Let n be a positive integer, and let $\theta > 0$ and $\alpha_i > 0$ be real numbers for $1 \leq i \leq n$, and let X_i be independent random variables for $1 \leq i \leq n$. and assume that X_i has distribution $\Gamma(\alpha_i, \theta)$. Then the random variable $X = \sum_{i=1}^n X_i$ has $\Gamma\Big(\sum_{i=1}^n \alpha_i, \theta\Big)$ distribution.*

We are not in position to prove this result, since one needs multivariate calculus to handle the distribution of sums of random variables.

19.4 Reading

For the gamma distribution, see [2, p. 58 and p. 228].

[19.5] We only need to take limit $\epsilon \searrow 0$ if $0 < t < 1$. For $t \geq 0$, we can simply take $\epsilon = 0$.

[19.6] The meaning of the parameters is that by changing the shape parameter, he shape of the density function changes, while the change of the scale parameter essentially spreads out the same shape by stretching it in the x direction, and contractiong it in the y direction as θ increases.

20 The chi squared distribution

Chi refers to the Greek letter χ.[20.1] As we will show below, in Subsection 20.1, if X has $\mathcal{N}(0,1)$, i.e., standard normal, distribution, then X^2 has $\Gamma(1/2,2)$ distribution; this distribution is also called $\chi^2(1)$ distribution, that is, chi squared distribution with *degree of freedom* 1. If X_i are independent standard normal variables, then the distribution of $\sum_{i=1}^n X_i^2$ is called $\chi^2(n)$ distribution, that is, the chi squared distribution of degree of freedom n. According to Theorem 19.1, this distribution, being the sum of n independent random variables each having $\Gamma(1/2,2)$ distribution, is identical to the $\Gamma(n/2,2)$ distribution.

In using Theorem 19.1, we made use of the result that if the random variables X_1, X_2, …, X_n are independent, and f_1, f_2, …, f_n are nice real-valued functions on the set of real numbers, then the random variables $f_1(X_1)$, $f_2(X_2)$, $f_3(X_3)$, …, $f_n(X_n)$ are also independent.[20.2] We only used this in the case $f_i(x)=x^2$ for all i, but the more general statement is also true.

20.1 The square of the standard normal distribution

Let X be a standard normal variable, i.e., a variable having distribution $\mathcal{N}(0,1)$. We will calculate the density function of $Y=X^2$. Recall that the density function of X is

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

Next we will calculate the distribution function of $Y=X^2$. For an arbitrary $t\ge 0$ We have

$$\begin{aligned}F_Y(t)&=\mathrm{P}(Y\le t)=\mathrm{P}(X^2\le t)=\mathrm{P}(-\sqrt{t}\le X\le\sqrt{t})\\&=\mathrm{P}(-\sqrt{t}\le X\le 0)+\mathrm{P}(0\le X\le\sqrt{t})=2\,\mathrm{P}(0\le X\le\sqrt{t});\end{aligned}$$

the last equation holds since the density function of X is even, i.e., $f_X(t)=f_X(-t)$. Thus,

$$F_Y(t)=2\int_0^{\sqrt{t}}f_X(x)\,dx=\frac{2}{\sqrt{2\pi}}\int_0^{\sqrt{t}}e^{-x^2/2}\,dx=\frac{\sqrt{2}}{\sqrt{\pi}}\int_0^u e^{-x^2/2}\,dx,$$

where $u=\sqrt{t}$.

For the density function of Y we have

$$f_Y(t)=\frac{d}{dt}F_Y(t)=\frac{du}{dt}\frac{d}{du}F_Y(u^2)=\frac{d\sqrt{t}}{dt}\frac{d}{du}\frac{\sqrt{2}}{\sqrt{\pi}}\int_0^u e^{-x^2/2}\,dx,$$

where in the second equations we used that $t=u^2$, and we also used the chain rule of differentiation. Noting that, according to the Fundamental Theorem of Calculus, the derivative of an integral with respect to its upper limit is the integrand at the upper limit, we obtain

$$f_Y(t)=\frac{1}{2\sqrt{t}}\frac{\sqrt{2}}{\sqrt{\pi}}e^{-u^2/2}=\frac{1}{\sqrt{\pi}\,2^{1/2}}t^{-1/2}e^{-t/2}=\frac{1}{\sqrt{\pi}\,2^{1/2}}t^{1/2-1}e^{-t/2};$$

[20.1] This is the lower case chi. The upper case letter looks the same as the capital X.

[20.2] The function f_i needs to be *Borel measurable* to ensure that $f(X_i)$ is a random variable; however, a definition of Borel measurable is beyond the scope of this course. Suffice it to say that assuming the functions f_i are continuous is certainly sufficient.

we wrote the right-hand side in a form where one can recognize that this density is identical to the density funcion of the $\Gamma(1/2, 2)$ distribution given in formula (19.4) with $\alpha = \theta = 1/2$, except for the value of $\Gamma(1/2)$.

Since the constant in a density function must be so chosen that the integral of the density function on $(-\infty, \infty)$ is 1, the constants in the two density functions must be equal; so we can conclude that $\Gamma(1/2) = \sqrt{\pi}$ and that the square Y of a standard normal distribution has $\Gamma(1/2, 2)$ distribution. This distribution is also called the $\chi^2(1)$ distribution.

20.2 Reading

The χ^2 distribution is discussed in [2, pp. 227–229].

21 Chebyshev inequality and the law of large numbers

Let Xbe a random variable for which the expectation $\mathrm{E}(X^2)$ exists.[21.1] Then, writing $m = \mathrm{E}(X)$, we have

$$\mathrm{V}(X) = \mathrm{E}\big((X-m)^2\big) \ge \epsilon^2\,\mathrm{P}(|X-m| \ge \epsilon).$$

The last inequality holds because $(X-m)^2$ is always nonnegative.

This is explained in the book [2, p. 122] for a continuous variable. Here we give a different explanation. First, if for random variables U and V we always have $U \le V$ then $\mathrm{E}(U) \le \mathrm{E}(V)$, assuming that the expectations exist. This is intuitively obvious, since the expectation is thought of the average value of a random variable in repeated experiments, so if U is always less than V, then the average of of U is less than that of V.[21.2] So let Y be the random variable

$$Y = \begin{cases} \epsilon^2 & \text{if} \quad |X-m| \ge \epsilon, \\ 0 & \text{otherwise.} \end{cases}$$

Then Y is a discrete variable with values ϵ^2 and 0, so

$$\mathrm{E}(Y) = \epsilon^2\,\mathrm{P}(|X-m| \ge \epsilon).$$

Since $Y \le (X-m)^2$, this justifies the above inequality.

Rearranging the above inequality and writing $\sigma^2 = V(X)$, we have

$$\mathrm{P}(|X-m| \ge \epsilon) \le \sigma^2/\epsilon^2. \tag{21.1}$$

This is called Chebyshev's inequality.

21.1 The law of large numbers

We have the following

[21.1] $\mathrm{E}(X^2)$ is called the second moment of X. If $\mathrm{E}(X^2)$ exists, then its expectation $\mathrm{E}(X)$ and variance $\mathrm{V}(X)$ also exists. To prove this rigorously is beyond the level of this course.

[21.2] This description of the expectation is somewhat circular, but in the Kolmogorov probability model, discussed at the beginning of the course, the expectation of a random variable has a rigorous mathematical definition. Using this, one can prove that $\mathrm{E}(U) \le \mathrm{E}(V)$. The closest we can come to the definition of expectation is to give formulas how to calculate the expectation of a continuous and of a discrete variable. In a more advanced treatment, there is no distinction between discrete and continuous variables, and the expectation is defined as an integral.

Theorem 21.1 (The law of large numbers)**.** *Let $n > 0$ be an integer, and let $\sigma \ge 0$ and m be real numbers, and let X_1, X_2, X_3, …, be pairwise*[21.3] *independent random variables with* $\mathrm{E}(X_i) = m$ *and* $\mathrm{V}(X_i) = \sigma^2$ *for i with $1 \le i < \infty$, and let $\epsilon > 0$. Writing*

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i,$$

we have

$$\lim_{n\to\infty} \mathrm{P}(|\bar{X}_n - m| \ge \epsilon) = 0.$$

Proof. We have

$$\mathrm{V}(n\bar{X}_n) = \mathrm{V}\Big(\sum_{i=1}^{n} X_i\Big) = \sum_{i=1}^{n} V(X_i) = \sum_{i=1}^{n} \sigma^2 = n\sigma^2, \tag{21.2}$$

where the second equation holds because the variables X_i are pairwise independent (one does not need full independence to ensure this equation). Hence, recalling that for a random variable X and for a real number α we have $\mathrm{V}(\alpha X) = \alpha^2\,\mathrm{V}(X)$, we have

$$V(\bar{X}_n) = \mathrm{V}\left(\frac{1}{n}\,n\bar{X}_n\right) = \frac{1}{n^2}\,\mathrm{V}(n\bar{X}_n) = \frac{1}{n}\sigma^2. \tag{21.3}$$

Hence, according to Chebyshev's inequality (21.1) we have

$$\mathrm{P}(|\bar{X}_n - m| \ge \epsilon) \le \frac{\sigma^2/n}{\epsilon^2} \to 0$$

as $n \to \infty$, completing the proof. □

21.2 Reading

Chebyshev's inequality and the law of large numbers is discussed in [2, pp. 120–122].

22 Statistical theory: point estimation

A good introduction to statistical theory is given in [2, Chapter 10–11, pp. 169–178]. The description is definitely worth reading, but it is not mathematically deep. Here we concentrate on the mathematical aspects.

22.1 Point estimation: general theory

In a statistical investigation one wants to find a the value of a nonrandom quantity θ that influences the observation; for example, the quantity θ may be a physical constant, but more examples are given in [2, loc. cit.].[22.1] The result of each observation is a list $\mathbf{x}$ of numbers of a given length (that is, each measurement gives k numbers (such as, say, the weight and height of an object gives

[21.3] That is, any two of them are independent. The book [2, Theorem 5, p. 121] makes the stronger assumption that these random variables are independent, but that assumption is never used. The assumtion of parwise independence is enough to ensure that equation (21.2) holds.

[22.1] Latin, loco citato, translated as in the place cited.

a list of length 2). If n independent observations are made, we have a sequence $\mathbf{x}_1$, $\mathbf{x}_2$, $\mathbf{x}_3$, …, $\mathbf{x}_n$, observaions. This sequence is called a sample. In the mathematical model of the observation, $\mathbf{x}_i$ is the result of evaluating a vector-valued[22.2] random variable $\mathbf{X}_i$. [22.3] The reason for considering $\mathbf{X}_i$ a random variable is so that we can use methods of probability theory to analyse the experiment. The nature of randomness in evaluating $\mathbf{X}_i$ may lie in the behavior of the measuring instrument, or in the selection of the item to be measured from a large number of similar items, etc. We regard the observations as independent, since it is hardly reasonable to make a large number of observations, if an observation depends on another observation. So, in the mathematical model, the parameter θ is estimated as a function $\theta^* = \theta^*(\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3, \dots \mathbf{X}_n)$. Instead of a single parameter θ, the experiment may involve a list of parameters $\boldsymbol{\theta}$, and one may consider a list $\boldsymbol{\theta}^*$ of functions approximating the list of parameters.

22.2 Estimation of the mean

All this is a bit too abstract, so we consider a concrete situation. In a large region, let m denote the average weight of all adult males. This is a fixed quantity, at least at a given time. In order to estimate the m it is usually not feasible to weigh all adult males and take the average (i.e., arithmetic mean). Instead, one randomly selects n individuals, and find that their weights are x_1, x_2, x_3, …, x_n, and one takes

$$m^*(x_1, x_2, \dots, x_n) = \bar{x}_n \overset{def}{=} \frac{1}{n}\sum_{i=1}^{n} x_i \tag{22.1}$$

as an approximation of m.[22.4] In order to know how good this way of approximating θ is, one uses a probabilistic model of the experiment in which x_i is replaced with a random variable X_i. It is assumed that $\mathrm{E}(X_i) = m$; in fact, one assumes that these variables are independent and identically distributed. One can study the behavior of the random variable

$$m^*(X_1, X_2, \dots, X_n) = \bar{X}_n \overset{def}{=} \frac{1}{n}\sum_{i=1}^{n} X_i. \tag{22.2}$$

Such use of x_i referring to the measurement and X_i to the corresponding random variable is common.

One finds that

$$\begin{aligned} \mathrm{E}\big(m^*(X_1, X_2, \dots, X_n)\big) &= \mathrm{E}(\bar{X}_n) = \mathrm{E}\Big(\frac{1}{n}\sum_{i=1}^{n} X_i\Big) \\ &= \frac{1}{n}\sum_{i=1}^{n} \mathrm{E}(X_i) = \frac{1}{n}\sum_{i=1}^{n} m = \frac{1}{n}\,nm = m. \end{aligned} \tag{22.3}$$

Such an estimate is called unbiased. In general, an estimate $\theta^*(X_1, X_2, \dots, X_n)$ is called *unbiased* if

$$\mathrm{E}\big(\theta^*(X_1, X_2, \dots, X_n)\big) = \theta.$$

Assuming that $\mathrm{V}(X_i) = \sigma^2$ exists, we have for every $\epsilon > 0$,

$$\lim_{n\to\infty} \mathrm{P}\left(|m^*(X_1, X_2, \dots, X_n) - m| \ge \epsilon\right) = \lim_{n\to\infty} \mathrm{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_n - m\right| \ge \epsilon\right) = 0$$

[22.2] That is, the value if a list of numbers, rather just a single number.

[22.3] In what follows, we will consider only a simplified model, in which each $\mathbf{X}_i = X_i$ is a real-valued random variable.

[22.4] The notation $\bar{x}_n$ is customary in statistics for the expression on the right-hand side.

according to Theorem 21.1. Such an etimate is called consistent. In general, $\theta^*(X_1, X_2, \dots, X_n)$ is called *consistent* if for every $\epsilon > 0$ the relation

$$\lim_{n\to\infty} \mathrm{P}\left(|\theta^*(X_1, X_2, \dots, X_n) - \theta| \ge \epsilon\right) = 0.$$

holds.

22.3 Estimation of the variance when the mean is known

We recall that the variance can be expressed as

$$\mathrm{V}(X) = \mathrm{E}\left(\left(X - \mathrm{E}(X)\right)^2\right) \tag{22.4}$$

Assuming that the value of $m = \mathrm{E}(X)$, estimating the variance comes down to estimating the value of

$$\mathrm{V}(X) = \mathrm{E}\left((X - m)^2\right).$$

That is, we need to estimate the mean of the random variable $(X - m)^2$. This is the same problem we discussed in Subsection 22.2. So, we can just use the solution described there. That is,

$$\sigma^{*2} = \frac{1}{n}\sum_{i=1}^{n}(X_i - m)^2 \tag{22.5}$$

is an unbiased and consistent estimate of σ^2.

22.4 Estimation of the variance when the mean is not known

When the variance is not known, one is tempted to use equation (22.5) with m replaced by the estimate given for m in formula (22.2), but it turns out that this would lead to an estimate that is biased. The following is an unbiased estimate:

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2, \tag{22.6}$$

where $\bar{X}_n$ is given in equation (22.2) to estimate m. To see this, we will evaluate the expectation of $(n-1)S^2$. Writing $m = \mathrm{E}(X)$,[22.5] we have

$$\begin{aligned}(n-1)S^2 &= \sum_{i=1}^{n}(X_i - \bar{X}_n)^2 = \sum_{i=1}^{n}\left((X_i - m) - (\bar{X}_n - m))\right)^2 \\ &= \sum_{i=1}^{n}(X_i - m)^2 - \sum_{i=1}^{n}2(\bar{X}_n - m)(X_i - m) + \sum_{i=1}^{n}(\bar{X}_n - m)^2 \\ &= \sum_{i=1}^{n}(X_i - m)^2 - 2(\bar{X}_n - m)\sum_{i=1}^{n}(X_i - m) + n(\bar{X}_n - m)^2,\end{aligned}$$

where the las equation holds since in the expression preceding it, the factor $2(\bar{X}_n - m)$ in the second sum does not depend on i, and in the third sum the *summand*,[22.6] the terms do not depend on i.

[22.5] The fact that m is not known will not affect this calculation.

[22.6] The quantity being summed.

Noting that the second sum on the right-hand side equals $n(\bar{X}_n - m)$ according to equation (22.2), we have

$$\begin{aligned}(n-1)S^2 &= \sum_{i=1}^{n}(X_i - m)^2 - 2n(\bar{X}_n - m)^2 + n(\bar{X}_n - m)^2,\\ &= \sum_{i=1}^{n}(X_i - m)^2 - n(\bar{X}_n - m)^2.\end{aligned}$$

Hence

$$\begin{aligned}\mathrm{E}\left((n-1)S^2\right) &= \mathrm{E}\Big(\sum_{i=1}^{n}(X_i - m)^2 - n(\bar{X}_n - m)^2\Big)\\ &= \sum_{i=1}^{n}\mathrm{E}\left((X_i - m)^2\right) - n\,\mathrm{E}\left((\bar{X}_n - m)^2\right) = \sum_{i=1}^{n}\mathrm{V}(X_i) - n\,\mathrm{V}(\bar{X}_n);\end{aligned} \tag{22.7}$$

the last equation holas since $\mathrm{E}(X_i) = \mathrm{E}(\bar{X}_n) = m$ (cf. (22.4)), Given that the variables X_i are independent[22.7] we have

$$\mathrm{V}(n\bar{X}_n) = \mathrm{V}\Big(\sum_{i=1}^{n}X_i\Big) = \sum_{i=1}^{n}\mathrm{V}(X_i) = \sum_{i=1}^{n}\sigma^2 = n\sigma^2,$$

where the third equation holds since $\mathrm{V}(X_i) = \sigma^2$. Therefore,

$$\mathrm{V}(\bar{X}_n) = \mathrm{V}\left(\frac{1}{n}\,n\bar{X}_n\right) = \frac{1}{n^2}\,\mathrm{V}(n\bar{X}_n) = \frac{1}{n^2}\,n\sigma^2 = \frac{1}{n}\sigma^2. \tag{22.8}$$

Substituting this into the right-hand side of (22.7), we obtain that

$$\mathrm{E}\left((n-1)S^2\right) = n\sigma^2 - \sigma^2 = (n-1)\sigma^2.$$

Hence, $\mathrm{E}(S^2) = \sigma^2$, showing that S^2 is indeed an unbiased estimate for σ^2.

22.5 Estimation of the standard deviation when the mean is not known

When $\mathrm{E}(X_i)$ is not known, one uses the square root of formula (22.6) to estimate the standard deviation:

$$S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2}, \tag{22.9}$$

This is not an unbiased estimate, but it is useful in practice.

If U and V are random variables, and $U = \sqrt{V}$, then $\mathrm{E}(U) \ne \sqrt{\mathrm{E}(V)}$ unless U is *almost surely*[22.8] constant. To see this, one needs to express the same statement in a different way:

If U and V are random variables and $V = U^2$, then $\mathrm{E}(V) \ne \big(\mathrm{E}(U)\big)^2$ unless U is almost surely constant. Indeed, if $\mathrm{E}(V) = \big(\mathrm{E}(U)\big)^2$ then

$$V(U) = \mathrm{E}(U^2) - \big(\mathrm{E}(U)\big)^2 = \mathrm{E}(V) - \mathrm{E}(V) = 0,$$

[22.7] Pairwise independence is enough for this.

[22.8] “Almost surely” is a common expression in probability theory, used to mean “with probability 1.”

which can happen only when U is constant. Indeed, $V(U) = \mathrm{E}\left(\left(U - \mathrm{E}(U)\right)^2\right)$, and so, we need only to observe that, with $Z = U - \mathrm{E}(U)$, if $\mathrm{E}(Z^2) = 0$ then $Z = 0$ almost surely.

The proof is easy with the aid of Chebyshev's inequality; see (21.1); it is given in the solution of one of the problems in [8]. The problem is currently given as Problem 5.1 on p. 22, at the end of the section on Differencing and other transformations of time series (the Contents with clickable links is given at the beginning of the manuscript) and the solution is given toward the end of the manuscript on p. 126.[22.9]

22.6 Reading

An introduction to statistical theory is given in [2, Chapter 20, pp. 167–178]; this maybe skimmed quickly to understand what statistics is about. There is no time to discuss this this in the course.

Point estimation is the subject of [2, Chapter 12, starting on p. 190]. We focus only on the estimation of the mean and the variance, [2, Section 12.4, pp. 196–198]. The rest of Chapter 12 beginning with the method of maximum likelihood [2, p. 198] is beyond the scope of this course.

22.7 Homework

See the website for the course as the file homework.pdf for homework related to Chapter 12.

23 Interval estimation: an introduction

For simplicity, we will first consider a single random variable X with a given family of distributions depending on a single parameter θ. For example, consider X having exponential distribution $\mathrm{Exp}(\theta)$. That is, the distribution function of X is

$$F_X(x) = \mathrm{P}(X \le x) = \begin{cases} 1 - e^{-x/\theta} & \text{if} \quad x > 0, \\ 0 & \text{if} \quad x \le 0. \end{cases} \tag{23.1}$$

We make a single measurement x of X; then the question is, what does this say about θ?

As an illustration, we will discuss [2, Problem 1302, p. 249], where the lifetime of a certain kind of lightbulb is considered. The lifetime of a lightbulb will be modeled by the exponential distribution with θ as an unknown parameter. The measurement x will be the actual lifetime of a single randomly selected lightbulb.

23.1 Abstract discussion

From this single measurement we certainly cannot tell what θ is, but we might make a statement that we think it likely, or we are fairly sure, that θ is in the interval (a_1, a_2) for certain numbers a_1 and a_2; we might even characterize how sure we are by saying that we are 95% sure that $a_1 < \theta < a_2$. What sense does this make, given that θ is a fixed but unknown quantity? Clearly, our choice of a_1 and a_2 must depend on x, the value of our measurement, since it would make no sense the make

[22.9] If I ever teach a course on Time Series, these locations in the manuscript are likely to change as a result of adding more material. It is possible in to to set up in LaTeX (which is used to typeset the present manuscript) floating links between manuscript that keep up with changes in the target manuscript of the link, but only if the source (the present manuscript in our case) is typeset after changes typeset after the change. I only do this for manuscripts that have close connections (the time series manuscript is closely tied to another manuscript I wrote on R programming for time series).

a measurement and not base our answer on the outcome of this measurement. That is, a_1 and a_2 must be functions of the measurement, so our estimation of θ need to be of form $a_1(x) < \theta < a_2(x)$, or in the probability model $a_1(X) < \theta < a_2(X)$. We would say that

$$a_1(X) < \theta < a_2(X)$$

with 95% *confidence* if

$$\mathrm{P}\big(a_1(X) < \theta < a_2(X)\big) = .95. \tag{23.2}$$

The interval (a_1, a_2) is called a 95% *confidence interval* for θ. We introduced the word "confidence" so we don't have to use the word "probability" instead, since the phrase $a_1 < \theta < a_2$ with 95% probability would create the false impression that θ assumes various random values. This is not the case. θ is an unknown but fixed quantity. What assume random values are a_1 and a_2; this should have been clear when we said that a_1 and a_2 depend on the outcome x of the measurement.

Using the event complementary to the one in equation (23.2), that equation can also be written as the equation

$$.05 = \mathrm{P}\big(\theta \le a_1(X) \text{ or } \theta \ge a_2(X)\big) = \mathrm{P}\big(\theta \le a_1(X)\big) + \mathrm{P}\big(\theta \ge a_2(X)\big);$$

here the equation holds since the two events on the right-hand side are mutually exclusive (because $a_1(X) < a_2(X)$). To satisfy the first equation, one usually chooses equitably between the two events on the right-hand side:

$$\mathrm{P}\big(\theta \le a_1(X)\big) = .025 \quad \text{and} \quad \mathrm{P}\big(\theta \ge a_2(X)\big) = .025. \tag{23.3}$$

One usually wants to write the two events in $\theta \le a_1(X)$ and $\theta \ge a_2(X)$ occurring in (23.3) as $X \le b_1(\theta)$ and $X \ge b_2(\theta)$, where the first of these events is the same as one of the two events above, and the second one is the same as the other.[23.1] That is, the equations in (23.3) can be rewritten as

$$\mathrm{P}\big(X \le b_1(\theta)\big) = .025 \quad \text{and} \quad \mathrm{P}\big(X \ge b_2(\theta)\big) = .025 \tag{23.4}$$

This is not always possible.

In fact, for this to be possible, a_1 and a_2 must both be increasing or both be decreasing. This condition is also sufficient if these functions are also continuous. In case both are increasing, b_2 needs to be the inverse of a_1 and b_1, the inverse of a_2, and in case both are decreasing, b_1 needs to be the inverse of a_1, and b_2, the inverse of a_2.

23.2 Example: estimating the lifetime of lightbulbs

Rather than pursuing the abstract discussion of the general situation, we will focus on the example we started with involving the exponential distribution given in equation (23.1). If in this equation one substitutes $y = x/\theta$, one obtains the equation[23.2]

$$F_X(\theta y) = \mathrm{P}(X \le \theta y) = \begin{cases} 1 - e^{-y} & \text{if} \quad y > 0, \\ 0 & \text{if} \quad y \le 0. \end{cases}$$

[23.1] That is, either the first of the latter two events corresponds to the first of the former two events, or the first of the latter events corresponds two the second of the former two events.

[23.2] Note that $\theta > 0$ is assumed in the exponential distribution, so that $x > 0$ holds if and only if $y > 0$ holds.

Note that the right-hand side here does not explicitly depend on θ. This will make it easy to satisfy the equations in (23.3) and (23.4). Writing $G(y) = F_X(\theta y)$, we need to values for y_1 and y_2 such that

$$G(y_1) = .025 \qquad \text{and} \qquad G(y_2) = 1 - .025 = .975$$

Then we have

$$\text{P}(X/y_1 \le \theta) = \text{P}(X \le \theta y_1) = G(y_1) = .025, \tag{23.5}$$

and[23.3]

$$\text{P}(X/y_2 > \theta) = \text{P}(X > \theta y_2) = 1 - \text{P}(X \le \theta y_2) = 1 - G(y_2) = .025 = G(y_1) = .025, \tag{23.6}$$

The event that neither of the events described on the left-hand sides of formulas (23.5) and (23.6) occurs is the event

$$X/y_2 \le \theta < X/y_1;$$

the first inequality says that the event in (23.6) does not occur, and the second inequality says that the event in (23.5) does not occur. That is

$$\text{P}(X/y_2 \le \theta < X/y_1) = 1 - \text{P}(X/y_1 \le \theta) - \text{P}(X/y_2 > \theta) = 1 - .025 - .025 = .95. \tag{23.7}$$

The interval $(X/y_2, X/y_1)$ is called a level 95% confidence interval for θ (in the probability model), or, if x is an observation of X (i.e., the lifetime of a single randomly selected light bulb of the time being considered), then the confidence interval is $(x/y_2, x/y_1)$ is called a level 95% confidence interval for θ (as calculated).[23.4]

23.3 A numerical illustration

Assuming a randomly chosen lightbulb lasts 1000 hours, we will give a numerical estimate for θ involving the given type of lightbulbs. To solve the equation $G(y) = A$ for y given A with $0 < A < 1$ is a simple matter. The equation can be written as $1 - e^{-y} = A$, i.e., $e^{-y} = 1 - A$, that is $-y = \ln(1 - A)$.

So, we obtain $y = -\ln(1 - A)$. With $A = .025$ we obtain $y_1 \approx 0.02531780$ and with $A = .975$ we obtain $y_2 \approx 3.68887945$. With $x = 1000$ the confidence interval $(x/y_2, x/y_1)$ gives the interval

$$(271.085031, 39497.8902).$$

This is a level 95% confidence interval for θ. It is important to remember that θ is a fixed but unknown quantity. But if one takes different lightbulbs, they will have different lifetimes, and they give different confidence intervals. But on average, 95% of the time, the interval obtained will contain θ, and 5% of the time we will obtain an interval that does not contain θ. Note that the expectation of a random variable with $\text{Exp}(\theta)$ distribution is θ, so an estimation of θ amounts to estimating the average lifetime of a lightbulb of the given type.

This example is of not much practical value, since one would hardly want to use the lifetime of a single lightbulb to predict the value of θ. But even the use of two lightbulbs would cause severe

[23.3] Noting that X is a continuous variable, we have $\text{P}(X = x) = 0$ for any value of x. So, next, we could write $X \ge \theta y_2$ instead of $X > \theta y_2$, but logically, the latter is correct, because we want to write the complement of the event $X \le \theta y_2$.

[23.4] Whether one takes a closed or open, or semiclosed interval is immaterial, since $\text{P}(X = t) = 0$ for any value of t. If one tries to correctly reflect the inequality on the left-hand side of (23.7), then one would write $[X/y_2, X/y_1)$ for the confidence interval.

technical complications, For example, if X_1 and X_2 are independent, and have the same $\mathrm{Exp}(\theta)$ distribution, then the determination of the distribution of $(X_1 + X_2)/2$ would require methods of multivariable calculus. On the other hand the example is a very good illustration of how confidence intervals are constructed.

If one wants to make a conclusion by examining a large number of lightbulbs, one would take the sum

$$\sum_{i=1}^{n} X_i,$$

where the X_i are independent identically distributed random variables, each representing a lightbulb. In view of the Central Limit Theorem, the distribution of this sum can be approximated by a normal distribution. At that point, the issue is to find a a confidence interval for the mean of a normal variable.

23.4 Reading

Interval estimation is discussed in [2, Chapter 13, starting on p. 222]. The first few pages of this chapter, up to p. 226 is recommended as light reading, but most of the material in this chapter related to what is discussed in the forthcoming sections in these notes. The main item discussed in this section of the notes is related to the solution of [2, Problem 1302, p. 249].

24 Confidence interval for the mean of a normal variable with σ^2 known

24.1 The case of a single normal variable

The method outlined for the exponential distribution is applicable in general, *mutatis mutandis*.[24.1] Let X be a random variable with normal distribution $\mathcal{N}(\theta, \sigma^2)$, where σ^2 is known. Given a small $\alpha > 0$[24.2] we show how to construct a confidence interval confidence level $1 - \alpha$. As we know, the variable $Y = (X - \theta)/\sigma$ is a standard normal variable. Given a real number ξ with $0 < \xi < 1$, write let the quantity λ_ξ be so determined that

$$\xi = \mathrm{P}(Y > \lambda_\xi) = 1 - \Phi(\lambda_\xi). \tag{24.1}$$

Looking at it geometrically, λ_ξ cuts off and area of size ξ on the right tail under the standard normal curve (i.e., the density function of the $\mathcal{N}(0, 1)$ distribution). Then, noting that the density function

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

is even (i.e., $f_Y(x) = f_Y(-x)$), we have

$$\mathrm{P}(Y < -\lambda_\xi) = \mathrm{P}(Y > \lambda_\xi) = \xi.$$

Thus, cutting off an area of size $\alpha/2$ at the two tails of the standard normal distribution, we have

$$\mathrm{P}(-\lambda_{\alpha/2} \le Y \le \lambda_{\alpha/2}) = 1 - \mathrm{P}(Y < -\lambda_{\alpha/2}) - \mathrm{P}(Y > \lambda_{\alpha/2}) = 1 - \frac{\alpha}{2} - \frac{\alpha}{2} = 1 - \alpha,$$

[24.1] After making the necessary changes.

[24.2] The usual values of α in statistics are .05, .01, or smaller assuming $\alpha < 1$ (this ensures that the two "tails" we are cutting off do not overlap, i.e., the two probabilities after the first equation concern disjoint events). Since, we have $Y = (X - \theta)/\sigma$, we have

$$\mathrm{P}\left(-\lambda_{\alpha/2} \le \frac{X-\theta}{\sigma} \le \lambda_{\alpha/2}\right) = 1 - \alpha. \tag{24.2}$$

The inequality within the scope of the probability symbol $\mathrm{P}(\cdot)$ can be written in a form that localizes θ to an interval. Indeed, multiplying all three members of the inequality by σ, we obtain

$$-\lambda_{\alpha/2}\,\sigma \le X - \theta \le \lambda_{\alpha/2}\,\sigma;$$

Multiplying this inequality by -1, all the inequality signs turn around:

$$\lambda_{\alpha/2}\,\sigma \ge -X + \theta \ge -\lambda_{\alpha/2}\,\sigma.$$

Adding X to all three members, this becomes

$$X + \lambda_{\alpha/2}\,\sigma \ge \theta \ge X - \lambda_{\alpha/2}\,\sigma.$$

This inequality can also be written as

$$X - \lambda_{\alpha/2}\,\sigma \le \theta \le X + \lambda_{\alpha/2}\,\sigma.$$

Thus, formula (24.2) can also be written as

$$\mathrm{P}(X - \lambda_{\alpha/2}\,\sigma \le \theta \le X + \lambda_{\alpha/2}\,\sigma) = 1 - \alpha. \tag{24.3}$$

Thus a level $1 - \alpha$ confidence interval for θ of an $\mathcal{N}(\theta, \sigma^2)$ variable when σ^2 is known is $[X - \lambda_{\alpha/2}\,\sigma, X + \lambda_{\alpha/2}\,\sigma]$ in the probability model. If the observed value of X is x, this interval becomes $[x - \lambda_{\alpha/2}\,\sigma, x + \lambda_{\alpha/2}\,\sigma]$. As before, it is unimportant whether we take a closed or open interval here.

24.2 A random sample from a normal distribution

Give a random sample from a normal distribution $\mathcal{N}(\theta, \sigma^2)$, that is given identically distributed independent random variables X_1, X_2, X_3, …, X_n with distribution $\mathcal{N}(\theta, \sigma^2)$, assuming that σ^2 is known, we would like to find a confidence interval for θ. For this, we will consider the variable

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

It is known that $\bar{X}_n$ is also normally distributed.[24.3] We have $E(\bar{X}_n) = \theta$ in view of equation (22.3) and $\mathrm{V}(\bar{X}_n) = \sigma^2/n$ according to equation (22.8); that is, the standard deviation of $\bar{X}_n$ is $\sigma/\sqrt{n}$. So we can obtain a confidence interval interval for θ by replacing X with $\bar{X}_n$ and σ with $\sigma/\sqrt{n}$ in (24.3):

$$\mathrm{P}\left(\bar{X}_n - \lambda_{\alpha/2}\,\frac{\sigma}{\sqrt{n}} \le \theta \le \bar{X}_n + \lambda_{\alpha/2}\,\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha. \tag{24.4}$$

24.3 Reading

Confidence interval for the mean of the normal distribution is discussed in [2, p. 230–231].

[24.3] To prove this one needs multivariate calculus, so it it beyond the level of the present course. The sum of independent normal variables is normal even if the variables being summed are not identically distributed.

25 Confidence interval for the mean of a normal variable with σ^2 unknown

When we have a relatively large sample (perhaps $n \geq 120$) X_1, X_2, X_3, $\dots$, X_n, from a normal distribution $\mathcal{N}(\theta, \sigma^2)$ with θ and σ^2 unknown, we can first estimate σ^2 as given in Subsection 22.4, and then find a confidence interval with the method described in Subsection 24.2. This does not work in case of small n, since the test variable (to be described) does not have normal distribution.

25.1 The Student t-distribution

The Student t-distribution, developed by William Sealy Gosset to handle the statistics of small samples of normal distribution. He used the pseudonym Student, as required by his employer, the Irish Guinness Brewery, for publishing his results. Given independent random variables X and Y, where X is a standard normal variable, and Y has a χ^2 distribution of degree of freedom f, then

$$Z = \frac{X}{\sqrt{Y/f}} \tag{25.1}$$

has a Student t-distribution of degree of freedom f. The density function of Z us

$$f_Z(x) = \frac{\Gamma\left(\frac{f+1}{2}\right)}{\sqrt{f\pi}\,\Gamma\left(\frac{f}{2}\right)} \left(1 + \frac{x^2}{f}\right)^{-(f+1)/2}.$$

The is discussed in [2, pp. 329–330]; tables for the t-distribution essential for contructing confidence intervals and for statistical testing are given in [2, p. 326]. When appropriate computer programs, such as Maxima, are available, these tables are not necessary.

25.2 Constructing a confidence interval for the mean

Given a sample X_1, X_2, $\dots$, X_n of the distribution $\mathcal{N}(\theta, \sigma^2)$ with θ and σ^2 unknown, we take the variable

$$T = \frac{\bar{X}_n - \theta}{S/\sqrt{n}} \tag{25.2}$$

as the variable to be used for the test, where $\bar{X}_n$ is given in (22.2) and S, in (22.9). We will explain why this variable has Student t-distibution with degree of freedom $n-1$, though a full proof is beyond the scope of these notes. We have

$$T = \frac{(\bar{X}_n - \theta)/(\sigma/\sqrt{n})}{\sqrt{S^2/\sigma^2}} = \frac{(\bar{X}_n - \theta)/(\sigma/\sqrt{n})}{\sqrt{\sigma^{-2}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2/(n-1)}}.$$

Here the numerator is a standard normal variable. From the discussion above, it is clear that the mean of the numerator is 0 and its variance is 1. That it is a normal variable follows from the fact that it is the sum of identically distributed normal variables is a normal variable at the beginning of Subsection 24.2. The variable

$$Y = \sigma^{-2} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2 \tag{25.3}$$

has χ^2 distribution of degree of freedom $n-1$. We recall that the χ^2 distribution of degree of freedom $n-1$ is the distribution of the sum of the squares of $n-1$ independent standard normal variables, but here we add n variables and they are not independent.[25.1] Finally, the variable Y defined in the last displayed formula and the variable $\bar{X}_n$ are independent; this we are not in a position to prove.[25.2]

That is, the variable T fits the description given in (25.1) with $f = n-1$; thus, T has student distribution of degree of freedom $n-1$. The densitiy function of the student distribution looks somewhat like that of the normal distribution, but the tails are somewhat thicker. The construction of the confidence interval follows the method used in case of the normal distribution. Given ξ with $0 < \xi < 1$ and given a random variable X having Student t-distribution of degree of freedom f, let $t_\xi(f)$ be so determined that

$$P\big(X > t_\xi(f)\big) = \xi. \tag{25.4}$$

This is analogous to the definition λ_ξ in (24.1). Given that the Student t-distribution is symmetric about the origin (that is, its density function is even), we also have $P\big(X < -t_\xi(f)\big) = \xi$. Hence, analogously to equation (confint: restricting normalized var) we have

$$\mathrm{P}\left(-t_{\alpha/2}(n-1) \le \frac{\bar{X}_n - \theta}{S/\sqrt{n}} \le t_{\alpha/2}(n-1)\right) = 1-\alpha. \tag{25.5}$$

Rearranging this in a way we used to obtain formula (24.4), we obtain the level $1-\alpha$ confidence interval for θ:

$$\mathrm{P}\left(\bar{X}_n - t_{\alpha/2}(n-1)\,\frac{S}{\sqrt{n}} \le \theta \le \bar{X}_n + t_{\alpha/2}(n-1)\,\frac{S}{\sqrt{n}}\right) = 1-\alpha. \tag{25.6}$$

This is of course written with the probablity model in mind. Using observed values, the level $1-\alpha$ confidence interval can be written as

$$\left(\bar{x}_n - t_{\alpha/2}(n-1)\,\frac{s}{\sqrt{n}}, \bar{x}_n + t_{\alpha/2}(n-1)\,\frac{s}{\sqrt{n}}\right),$$

where the lower case letters are observed values, or calculated from obsereved values, of the corresponding upper case random variables.

25.3 Reading

The Student t-distribution is discussed in [2, pp. 229–230]. The confidence interval for the mean of a normal variable when σ^2 is unknown is discussed in [2, pp. 231–232].

26 Confidence interval for the variance from a normal sample

Let X_1, X_2, X_3, …, X_n, independent random variables with distribution $\mathcal{N}(\theta, \sigma^2)$, where θ and σ^2 are unknown. Then, defining $\bar{X}_n$ as in (22.2), the random variable

$$Y = \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$$

[25.1] One can argue intuitively that by deducting $\bar{X}_n$ from each of the independent variables X_i we are taking away one degree of freedom, so the degrees of freedom is $n-1$, but such an explanation is really not satisfactory; one needs a proof.

[25.2] For the curious: one can prove fairly easily that the joint distribution of $\bar{X}_n$ and Y is a multivariate normal distribution; further, their correlation is 0. According to known results, this implies that $\bar{X}_n$ and Y are independent. See the section on the multivariate normal distribution in [8]; currently Subsection 2.7.

has χ^2-distribution with degree of freedom $n-1$, as mentioned right after formula (25.3). Proceeding analogously to the definition of λ_ξ in (24.1 (cf. also formula (25.4) Let U be a random variable having χ^2 distribution with degree of freedom f, given ξ with $0<\xi<1$ let χ^2_ξ be such that

$$P\big(U > \chi^2_\xi(f)\big) = \xi. \tag{26.1}$$

Now, the χ^2-distribution in not symmetric about the origin,[26.1] so cuting off the values of Y at both tail ends will look different it was in case of the normal or the Student t-distributions. Noting that

$$\mathrm{P}\big(U < \chi^2_{1-\xi}(f)\big) = \mathrm{P}\big(U \le \chi^2_{1-\xi}(f)\big) = 1 - \mathrm{P}\big(U > \chi^2_{1-\xi}(f)\big) = 1-(1-\xi) = \xi.$$

Analogously to equations (24.2) and (25.5), we have

$$\mathrm{P}\left(\chi^2_{1-\alpha/2}(n-1) \le \frac{1}{\sigma^2}\sum_{i=1}^n (X_i - \bar{X}_n)^2 \le \chi^2_{\alpha/2}(n-1)\right) = 1-\alpha. \tag{26.2}$$

We will rewrite the inequaly

$$\chi^2_{1-\alpha/2}(n-1) \le \frac{1}{\sigma^2}\sum_{i=1}^n (X_i - \bar{X}_n)^2 \le \chi^2_{\alpha/2}(n-1)$$

in the scope of the probability $\mathrm{P}(\cdot)$. Noting that each member of this inequality is positive, we can take reciprocals. The inequalities will turn around:

$$\frac{1}{\chi^2_{1-\alpha/2}(n-1)} \ge \frac{\sigma^2}{\sum_{i=1}^n (X_i - \bar{X}_n)^2} \ge \frac{1}{\chi^2_{\alpha/2}(n-1)}$$

Reversing the sides, this can also be written as

$$\frac{1}{\chi^2_{\alpha/2}(n-1)} \le \frac{\sigma^2}{\sum_{i=1}^n (X_i - \bar{X}_n)^2} \le \frac{1}{\chi^2_{1-\alpha/2}(n-1)}.$$

Multiplying all three members by $\sum_{i=1}^n (X_i - \bar{X}_n)^2$, we obtain

$$\frac{1}{\chi^2_{\alpha/2}(n-1)}\sum_{i=1}^n (X_i - \bar{X}_n)^2 \le \sigma^2 \le \frac{1}{\chi^2_{1-\alpha/2}(n-1)}\sum_{i=1}^n (X_i - \bar{X}_n)^2.$$

This inequality is equivalent to the inequality in the scope of the probablity in equation (26.2). Thus, we have

$$\mathrm{P}\left(\frac{1}{\chi^2_{\alpha/2}(n-1)}\sum_{i=1}^n (X_i - \bar{X}_n)^2 \le \sigma^2 \le \frac{1}{\chi^2_{1-\alpha/2}(n-1)}\sum_{i=1}^n (X_i - \bar{X}_n)^2\right) = 1-\alpha.$$

Taking square roots, the same probability can be described as

$$\mathrm{P}\left(\sqrt{\frac{1}{\chi^2_{\alpha/2}(n-1)}\sum_{i=1}^n (X_i - \bar{X}_n)^2} \le \sigma \le \sqrt{\frac{1}{\chi^2_{1-\alpha/2}(n-1)}\sum_{i=1}^n (X_i - \bar{X}_n)^2}\right) = 1-\alpha.$$

[26.1] In fact, it is clear that its density function is zero outside the interval $(0,\infty)$, since a sum of squares cannot be negative.

From here, it is easy t9 write a level $1-\alpha$ confidence interval for σ^2 in terms of observed values. Similarly, one can wite such a confidence interval for σ. Writing out the latter, a level $1-\alpha$ confidence interval for σ is

$$\left(\sqrt{\frac{1}{\chi^2_{\alpha/2}(n-1)}\sum_{i=1}^{n}(x_i-\bar{x}_n)^2},\ \sqrt{\frac{1}{\chi^2_{1-\alpha/2}(n-1)}\sum_{i=1}^{n}(x_i-\bar{x}_n)^2}\right).$$

26.1 Reading

Confidence interval for the variance of the normal is discussed in [2, p. 235].

27 Confidence interval using normal approximation

If n is a large integer, and X_1, X_2, X_3, …, X_n, are independent, identically distributed random variables with expectation θ and variance σ^2, then the sum $U=\sum_{i=1}^{n}X_i$ is approximately normally distributed, according to the Central Limit Theorem. Since the expectation of a sum of random variables equals the sum of expectations, and if these variables are also (pairwise)[27.1] indendent, this is true also for the variances. Hence, the expectation of U is $n\theta$ and its variance is $n\sigma^2$, Hence, writing, as usual,

$$\bar{X}_n=\frac{U}{n}=\frac{1}{n}\sum_{i=1}^{n}X_i,$$

we have $\mathrm{E}(\bar{X}_n)=\mathrm{E}(U)/n=n\theta/n=\theta$ and $\mathrm{V}(\bar{X}_n)=n\sigma^2/n^2=\sigma^2/n$. Thus the variable

$$Y=\frac{\bar{X}_n-\theta}{\sigma/\sqrt{n}}$$

is a variable with expectation 0 and variance 1, and so this variable approximately has disrtibution $\mathcal{N}(0,1)$, the standard normal distribution. Thus, $\mathrm{P}(Y\le x)\approx\Phi(x)$. Hence, one can use normal variable methods to get an approximate answer. If σ^2 is known, then one used the same confidence interval that we have for the normal distribution when σ^2 is known; this is given in (24.4), except tha equation now would only be approximate, since the variable Y is only approximately normally distributed. To restate this result in terms an observed sample, an approximate level $1-\alpha$ confidence interval for θ is

$$\left(\bar{x}_n-\lambda_{\alpha/2}\,\frac{\sigma}{\sqrt{n}},\bar{x}_n+\lambda_{\alpha/2}\,\frac{\sigma}{\sqrt{n}}\right),$$

If σ^2 is not known, we need to estimate it. For this, we use formulas (22.5) and (22.9), or the correspondig lower case versions. For example, we put

$$s=\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x}_n)^2};$$

s of course depends on n, but this dependence is not indicated.

$$\left(\bar{x}_n-\lambda_{\alpha/2}\,\frac{s}{\sqrt{n}},\bar{x}_n+\lambda_{\alpha/2}\,\frac{s}{\sqrt{n}}\right).$$

[27.1] That is, for the variance to be the sum of the variances, we only need pairwise independence. For the Central Limit Theorem, we still need full independence.

As can be seen, we are using the normal distribution, rather than the Student t-distribution. The reason for this that this should be only used for large n, certainly only for $n \geq 30$. Perhaps one should use the Student t-distribution, in that it gives a larger confidence interval, which means one is less certain about the value of θ. But how much one can trust this slightly larger confidence interval is in question. The test for the mean of the normal normal distribution if σ^2 is not known relies on the fact that the test variable in equation (25.2) has the Student t-distribution – but this is only true if the sample variables X_i are normally distributed. This is not assumed to be the case at present, so using the t-distribution is not justified. The branch of statistics we are discussing is called *parametric statistics*, created in the 1920s.[27.2]

With the enormous increase in computational power since the 1920s, many of these parametric methods can be replaced with resampling methods, and in fact these methods can be applied in many situations where the classical parametric methods could not be used. For an early, eminently readable account of these methods see Efron [3]. For a more recent monograph on the subject, see [4] by Bradley Efron and Trevor Hastie. In Section 33, we will be discussion bootstrapping, a useful resampling method invented by Bradley Efron.

27.1 Confidence interval for proportion

We want to find a confidence interval of the probability of an event in a repeated experiment, if performing the experiment n times, independently, the event occurs k times. A typical example would be that in a blood test of 500 persons, 342 carry antibodies for a certain virus. Give a level 95% confidence interval for the proportion of persons that carries antibodies for the virus.

If we denote X for the number of persons carrying antibodies for the virus, then X has a binomial distribution $\mathrm{Bin}(n, \theta)$, where n is the number of persons tested ($n = 500$ in the example), and θ is the probability that a randomly selected person carries antibodies for the virus.

If U_i is the random variable that is 1 if the i person carries antibodies, and it is 0 if the person does not carry antibodies then

$$X = \sum_{i=1}^{n} U_i.$$

Here the variables U_1, U_2, U_3, …, U_n, are independent, and their sum is approximated by a normal variable. So there is a parallel with the example discussed above. In the above example, the distribution of X_i were not known, here the distribution is known except for the value of the parameter θ: $U_i = 0$ or 1. and $\mathrm{P}(U_i = 1) = \theta$, and $\mathrm{P}(U_i = 0) = 1 - \theta$. We have

$$\mathrm{E}(U_i) = \theta.$$

Since $U_i^2 = U_i$, we also have $\mathrm{E}(U_i^2) = \theta$, we have

$$\mathrm{V}(U_i) = E(U_i^2) - \big(\mathrm{E}(U_i)\big)^2 = \theta - \theta^2 = \theta(1 - \theta).$$

Since expectations add up always when adding random variables, and variances add up when these variables are pairwise independent, we can see that $\mathrm{E}(X) = n\theta$ and $V(X) = n\theta(1 - \theta)$. Hence, the distrubution of the random variable

$$Y = \frac{X - n\theta}{\sqrt{n\theta(1 - \theta)}}$$

[27.2] Parametric statistics is based on the assumption that the data satisfy a certain known probabiity distribution. The problem with the name is that one cannot contrast it with the term nonparametric statistics, since the latter term is used in several different senses that is not meant as a replacement for the parametric methods.

is approximately standard normal $\mathcal{N}(0,1)$. Thus, analooously to equation (24.2), we have

$$\mathrm{P}\left(-\lambda_{\alpha/2} \le \frac{X-n\theta}{\sqrt{n\theta(1-\theta)}} \le \lambda_{\alpha/2}\right) \approx 1-\alpha. \tag{27.1}$$

Multiplying the inequality within the scope of the probability, we obtain

$$-\lambda_{\alpha/2}\sqrt{n\theta(1-\theta)} \le X-n\theta \le \lambda_{\alpha/2}\sqrt{n\theta(1-\theta)}$$

Multiplying this inequality by -1, the inequalities will turn around:

$$\lambda_{\alpha/2}\sqrt{n\theta(1-\theta)} \ge -X+n\theta \ge -\lambda_{\alpha/2}\sqrt{n\theta(1-\theta)}.$$

By switching the sides of the inequality, this can be written as

$$-\lambda_{\alpha/2}\sqrt{n\theta(1-\theta)} \le -X+n\theta \le \lambda_{\alpha/2}\sqrt{n\theta(1-\theta)}.$$

Adding X to this inequatily, we obtain

$$X-\lambda_{\alpha/2}\sqrt{n\theta(1-\theta)} \le n\theta \le X+\lambda_{\alpha/2}\sqrt{n\theta(1-\theta)}.$$

Dividing this inequality by n, we obtain

$$\frac{X}{n}-\lambda_{\alpha/2}\sqrt{\frac{\theta(1-\theta)}{n}} \le \theta \le \frac{X}{n}+\lambda_{\alpha/2}\sqrt{\frac{\theta(1-\theta)}{n}}.$$

The trouble here is that θ occurs in all three members of the inequality. Writing $\Theta = X/n$, note that $\theta \approx \Theta$ (note that here θ is an unknown parameter, while Θ is a random variable), replacing both X/n and θ by Θ on the sides (but not in the middle), we obtain

$$\Theta-\lambda_{\alpha/2}\sqrt{\frac{\Theta(1-\Theta)}{n}} \lessapprox \theta \lessapprox \Theta+\lambda_{\alpha/2}\sqrt{\frac{\Theta(1-\Theta)}{n}}. \tag{27.2}$$

The reason we replaced $\le$ with $\lessapprox$ is that the latter ineqquality is only approximately equivalent to the inequalities before. But an important question arises: what justifies approximating θ by Θ on the sides but not the in the middle. The point is that on the sides, the n in the denominator limits the error committed.

If we write

$$f(x) = \sqrt{\frac{x(1-x)}{n}},$$

then

$$f'(x) = \frac{1}{2\sqrt{x(1-x)/n}}\cdot\frac{1-2x}{n} = \frac{1}{\sqrt{n}}\cdot\frac{1-2x}{2\sqrt{x(1-x)}}.$$

Thus, we have

$$\begin{aligned}\sqrt{\frac{\theta(1-\theta)}{n}}-\sqrt{\frac{\Theta(1-\Theta)}{n}} &= f(\theta)-f(\Theta) \approx f'(\Theta)(\theta-\Theta)\\ &= \frac{1}{\sqrt{n}}\cdot\frac{1-2\Theta}{2\sqrt{\Theta(1-\Theta)}}(\theta-\Theta).\end{aligned} \tag{27.3}$$

This also shows that the error is particularly bad when Θ is close to 0 or 1, and it is small when Θ is away from 0 and 1, and when n is large. As for what happens when Θ is close to 0 or 1, we need to remember that in this case the the normal approximation to the binomial distribution is not very accurate, either. In these cases, the estimation of θ and the limit of the accuracy of this estimate is better handled by resampling methods mentioned right before the beginning of the present subsection, on p. 70.

We will rewrite equation (27.2) in terms of observed values. Writing $\hat{\theta} = x/n$, we can write the level $1 - \alpha$ confindence interval for θ as

$$\left(\hat{\theta} - \lambda_{\alpha/2}\sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}},\ \hat{\theta} + \lambda_{\alpha/2}\sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}}\right). \tag{27.4}$$

In order to estimate the error committed by replacing θ by $\hat{\theta}$ on the sides, we will use formula (27.3) (with $\hat{\theta}$ replacing Θ when discussing observed values), where we use the length of the half of the length of the confidence interval in (27.4) to estimate $|\theta - \hat{\theta}|$.[27.3] That is, the absolute value of the error in each of the endpoints of this interval is $\lessapprox$

$$\frac{1}{\sqrt{n}} \cdot \frac{|1-2\hat{\theta}|}{2\sqrt{\hat{\theta}(1-\hat{\theta})}}\,\lambda_{\alpha/2}\sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}}. \tag{27.5}$$

As for the example given at the beginning of this subsection on p. 70, if out of 500 persons 342 test positive for antibodies, then find an approximate 95% confidence interval for the proportion of the population carrying antibodies antibodies for the virus. In this example, $\alpha = .05$, $n = 500$, $x = 342$, so $\hat{\theta} = x/n = .684$. The confidence interval according to formula (27.4) is

$$(0.643, 0.725).$$

According to formula (27.5), the error in the endpoints is approximately ± 0.00144.

27.2 Reading

Confidence interval using the normal approximation is discussed in [2, pp. 241–242]; see also the discussion starting at the bottom of p. 244 and condinuing to the middle of p. 245 in [2], though we believe our discussion is simpler.

28 Hypothesis testing

Consider the following situation. A drug researcher wants to find out whether a compound is useful in curing a certain disease. She sets up two hypotheses. The null hypothesis is H_0, saying that the drug has no beneficial effect. The alternative hypothesis H_1 is that the drug has a beneficial effect. She wants to decide whether the available evidence is enough to reject the null hypothesis, in which case the drug warrants further investigation. In this situation, assuming that the drug is not harmful, she will never be able to prove that the drug has no beneficial effect; after all, the effect of the drug might be so tiny that it is nearly impossible to observe it. On the other hand, if the drug is really useful, rejecting the null hypothesis should be quite feasible. The point we are emphasizing

[27.3] Indeed, $\hat{\theta}$ is the center of this interval, and it is likely that θ is also in this interval.

here is that the situation between H_0 and H_1 is not symmetric. That is, the choice is not between accepting and rejecting H_0. The choice is between rejecting and not rejecting H_0. That is, to decide whether the data show that H_0 is very unlikely to be true, or that it is possible that H_0 is true. A simple enlightening example for hypothesis testing is described in [2, Example 1, pp. 253–254]. The example discusses a person claiming to have extrasensory perception (ESP). In an experiment, he is tested whether he can tell whether a fair coin will come up head or tail in twelve consecutive tosses. Setting up H_0 as the person has no extrasensory perception, that is, predicting the result of each toss is correct with probability 1/2, the hyppthesis H_0 will be rejected if the person guesses correctly in about 10 of the 12 tosses. See loc. cit.[28.1] The discussion of the example continues through several pages, illustrating the theoretical discussion.

28.1 Testing for the mean of a normal variable with known variance

Instead of introducting the general theory of hypothesis testing, we will jump right in the middle of things, and illustrate what hypothesis testing on a simple example. Let X be a normal variable $\mathcal{N}(\theta, \sigma^2)$, where σ^2 is known, we want to test whether $\theta = \theta_0$ for a certain number θ_0. That is, our null hypothesis is $H_0 : \theta = \theta_0$, and our alternative hypothesis is $H_1 : \theta \ne \theta_0$. More subtle choices for the alternative hypothesis will be discussed later. We want to decide whether to reject the null hypothesis, and we want to choose a *significance level* (to be explained next) α for the test; a common value of α is .05 = 5%. We make an observation x of the random variable X, and depending on this observation, we do or do not reject H_0. We want to make sure that if H_0 is true then the chance of rejecting H_0 is not more than α. How to do this was already worked out on account of confidence intervals, so the discussion here will be quite simple. Assuming H_0, that is, that $\theta = \theta_0$, we have $\mathrm{E}(X) = \theta_0$. Thus, according to equation (24.2) we have

$$\mathrm{P}\left(-\lambda_{\alpha/2} \le \frac{X - \theta_0}{\sigma} \le \lambda_{\alpha/2}\right) = 1 - \alpha. \tag{28.1}$$

Multiplying the inequality in the scope of $\mathrm{P}(\cdot)$ by σ, this can also be written as

$$\mathrm{P}(-\lambda_{\alpha/2}\,\sigma \le X - \theta_0 \le +\lambda_{\alpha/2}\,\sigma) = 1 - \alpha.$$

Using absolute values, this inequality is equivalent to

$$\mathrm{P}(|X - \theta_0| \le \lambda_{\alpha/2}\,\sigma) = 1 - \alpha.$$

As for the complementary event, we have

$$\mathrm{P}(|X - \theta_0| > \lambda_{\alpha/2}\,\sigma) = \alpha. \tag{28.2}$$

Thus, this is how the test will be performed. If the observed value x of X satisfies the inequality in the scope of $\mathrm{P}(\cdot)$, then we reject H_0. That is, we reject H_0 if

$$|x - \theta_0| > \lambda_{\alpha/2}\,\sigma$$

at significance level α. In other word, we call the set

$$C = \{t : |t - \theta_0| > \lambda_{\alpha/2}\,\sigma\} \tag{28.3}$$

the *critical region*. We reject H_0 if $x \in C$, i.e., accept H_1, and we do not reject H_0, i.e., do not accept H_1, if $x \notin C$.

[28.1] Latin, short for loco citato, meaning "in the place cited."

28.2 The probablity value (p-value) of a test

Instead of specifying a level of signicicance of a test in advance as in Subsection 28.1 by setting up a critical region as in formula (28.3), one can proceed in a sightly different way, as follows:

As before, we are given a random variable X with distribution $\mathcal{N}(\theta, \sigma^2)$, where σ^2 is known. We are testing the null hypothesis $H_0 : \theta = \theta_0$ against the alternative hypothesis $H_1 : \theta \ne \theta_0$. We make an observation x of X. In contrast to the critical region C given in (28.3), we ask the question, what would be the significance of a test that used the critical region

$$C(x) = \{t : |t - \theta_0| > |x - \theta_0|\}. \tag{28.4}$$

That is, what would be the significance of a test that required an observation as extreme as x for rejecting H_0. This significance level is called the *p-value* of the test. For the type of the test discussed here the p-value is

$$p = \mathrm{P}(|X - \theta_0| > |x - \theta_0|), \tag{28.5}$$

in analogy with equation (28.2).

28.2.1 The connection between p-value and significance

. The connection is quite simple. If in a type of test, an observation has p-value p. then for any $\alpha > p$, in a test of significance level α, the observation would result in the rejection of the null hypothesis H_0.

28.2.2 Tables, computers, and p-values

Statisticians relied on tables since the 1920s to construct confidence intervals and tests of given significance levels. This was feasible if one used a limited number of confidence and significance levels (.95, .99, .999, ... for confidence levels and .05, .01, .001, ... for significance levels), because there was only a limited amount of data that the tables needed to contain. This is different for p-values; for this, one needs the whole distribution function of distributions that have a degree of freedom parameter; using a whole page for each degree of freedom, say, under 120, the statistical tables would have to be collected in a book of considerable size. Fortunately, with computers, this is no longer a problem, and the needed values can be calculated in no time at all. Thus, computers introduced an era of greater reliance on p-values than significance. Talking in terms of levels of significance is still important when communicating with the public.

28.3 One-sided tests

The tests considered above in this section were two sided test. In some investigations, one needs to consider one-sided tests. In this case, the null hypothesis may be $H_0 : \theta = \theta_0$, with the alternative hypothesis being $H_1 : \theta > \theta_0$.[28.2] The discussion here is analgous; what differs here is that only one tail of the distribution needs to be cut off. Thus, to construct a test of significance α, instead of (28.1), we put

$$\mathrm{P}\left(\frac{X - \theta_0}{\sigma} \le \lambda_\alpha\right) = 1 - \alpha. \tag{28.6}$$

[28.2] Or else $H_1 : \theta < \theta_0$. In case of the normal distribution, this does not warrant a separate discussion in view of the symmetry of the normal distribution. Some other distributions, such as the χ^2 distribution, used to test for the variance of a normal distribution are not symmetric.

or else

$$(28.7) \qquad \mathrm{P}(X > \theta_0 + \lambda_\alpha\, \sigma) = \alpha.$$

In this case, the critical region is

$$(28.8) \qquad C = \{t : t > \theta_0 + \lambda_\alpha\, \sigma\} = (\theta_0 + \lambda_\alpha\, \sigma, +\infty).$$

We reject H_0 if $x \in C$, i.e., accept H_1, and we do not reject H_0, i.e., do not accept H_1, if $x \notin C$.

28.3.1 The probablity value (p-value) of a one-sided test

As before, given the test with the critical region in (28.8), instead of setting up a critical region in advance, we make a single observation x of the variable X, and set up a critical region that gives the maximum significance so tis observation does not result in the rejection of H_0. Similarly to formula (28.5), in this case we have the p-value

$$(28.9) \qquad p = \mathrm{P}(X > x).$$

28.4 Reading

The introduction to hypothesis testing in [2, Chapter 14, starting on p. 252], especially the ESP example is interesting reading; the example continues through a discussion of the theory of hypothesis testing. In our discussion we focused on the relationship of confidence intervals and hypothesis testing, discussed in [2, §14.4, p. 261–262].

29 Confidence intervals converted into tests

As we saw in Section 28, confidence intervals can easily be converted into tests (though we do not want to give a general description, since there may be exceptions), the discussion of statistical tests will be fairly simple. We are focusing on two-sided tests, just as we focused on two-sided confidence intervals, a discussion of one-sided tests can easily be added.

29.1 A test for the mean of the normal distribution when the variance is known

Give a random sample from a normal distribution $\mathcal{N}(\theta, \sigma^2)$, that is given identically distributed independent random variables X_1, X_2, X_3, ..., X_n with distribution $\mathcal{N}(\theta, \sigma^2)$, assuming that σ^2 is known, we would like to test with null hypothesis $H_0 : \theta = \theta_0$ and alternative hypothesis $H_1 : \theta \neq \theta_0$. For this, we will consider the variable

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

It is known that $\bar{X}_n$ is also normally distributed. As we discussed in Subsection 24.2, we have $E(\bar{X}_n) = \theta$ and $\mathrm{V}(\bar{X}_n) = \sigma^2/n$, that is, the standard deviation of $\bar{X}_n$ is $\sigma/\sqrt{n}$. The rest of the discussion is identical to the case of a single normal variable with expectation of θ and variance σ^2/n, as given in Subsecion 28.1. Similarly to formula (28.2), we have

$$(29.1) \qquad \mathrm{P}\left(|\bar{X}_n - \theta_0| > \lambda_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = \alpha.$$

The critical region is the set

$$C = \left\{ t : |t - \theta_0| > \lambda_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right\}. \tag{29.2}$$

We reject H_0 if $\bar{x}_n \in C$, i.e., accept H_1, and we do not reject H_0, i.e., do not accept H_1, if $\bar{x}_n \notin C$. If the observed values of the random variables X_1, X_2, X_3, …, X_n are x_1, x_2, x_3, …, x_n, then, writing

$$\bar{x}_n = \sum_{i=1}^{n} x_i,$$

the p-value of these observations in the analogous test is

$$p = \mathrm{P}\big(|\bar{X}_n - \theta_0| > |\bar{x}_n - \theta_0|\big). \tag{29.3}$$

29.2 A test for the mean of the normal distribution when the variance is not known

The construction of a confidence interval was discussed in Section 25. Using the notation introduced there, the variablr T given in (25.2) with θ_0 replacing θ; has Student t-distribution with degree of freedom $n-1$. Assuming the null hypothesis $H_0 : \theta = \theta_0$, similarly to equation (25.5), we have

$$\mathrm{P}\left(-t_{\alpha/2}(n-1) \le \frac{\bar{X}_n - \theta_0}{S/\sqrt{n}} \le t_{\alpha/2}(n-1) \right) = 1 - \alpha. \tag{29.4}$$

This can also be written as

$$\mathrm{P}\left(\left| \frac{\bar{X}_n - \theta_0}{S/\sqrt{n}} \right| \le t_{\alpha/2}(n-1) \right) = 1 - \alpha,$$

or else

$$\mathrm{P}\left(\left| \frac{\bar{X}_n - \theta_0}{S/\sqrt{n}} \right| > t_{\alpha/2}(n-1) \right) = \alpha,$$

That is, given a an observed sample x_1, x_2, x_3, …, x_n, and calculating the quantities $\bar{x}_n$ and s as their upper case variants, with X_i replaced with x_i. the critical region for the test is

$$C = \left\{ (u, v) : \left| \frac{u - \theta_0}{v/\sqrt{n}} \right| > t_{\alpha/2}(n-1) \right\}.$$

We reject H_0 if $(\bar{x}_n, s) \in C$, i.e., accept H_1; we do not reject H_0, i.e., do not accept H_1, if $(\bar{x}_n, s) \notin C$.[29.1] If we use p-values, the p-value of the analogous test is

$$p = \mathrm{P}\left(\left| \frac{\bar{X}_n - \theta_0}{S/\sqrt{n}} \right| > \left| \frac{\bar{x}_n - \theta_0}{s/\sqrt{n}} \right| \right).$$

[29.1] We wrote the critical region this way, to indicate that the critical region does not need to be a subset of $\mathbb{R}$. The same test could have been described with the critical region

$$C' = \left\{ u : |u| > t_{\alpha/2}(n-1) \right\}.$$

Writing

$$t = \frac{\bar{x}_n - \theta_0}{s/\sqrt{n}},$$

We reject H_0 if $t \in C'$, i.e., accept H_1; we do not reject H_0, i.e., do not accept H_1, if $t \notin C$.

The fact that the random variable

$$\frac{\bar{X}_n - \theta_0}{S/\sqrt{n}}$$

has Student t-distribution with degree of freedom $n-1$ makes it difficult to find this p-value in tables. The statistical tables given as [2, Table 7, p. 326] is certainly not suitable for this. On the other hand, the software Maxima can do the calculation in no time at all.

29.3 Reading

The discussion in [2, §14.5, pp. 262–264].

30 The power of a test

Assume we are given a random variable X that depends on the parameter θ, and we want to test the hypothesis $H_0 : \theta = \theta_0$ against the alternative hypothesis $H_1 : \theta \ne \theta_0$. In case $\theta \ne \theta_0$, H_0 needs to be rejected. Realistically, though, this is unlikely to happen if θ is close to θ_0 even if $\theta \ne \theta_0$. Recall that the significance of the test was the probability that H_0 is rejected even if H_0 is true. The power function of the test in question is defined as

$$h(\theta) \overset{def}{=} \mathrm{P}(H_0 \text{ is rejected if } \theta \text{ is the correct value of the parameter}). \tag{30.1}$$

30.1 The case of a single normal variable

We will consider the following example, discussed in [2, Example 2, p. 256]. Assume we are given a normal distribution $\mathcal{N}(\theta, \sigma^2)$, where we want to test the hypothesis $H_0 : \theta = \theta_0$ against the alternative hypothesis $H_1 : \theta \ne \theta_0$, assuming that σ^2 is known. The critical region for this test is given in (28.3) is

$$\begin{aligned} C &= \{t : t < \theta_0 - \lambda_{\alpha/2}\,\sigma \quad \text{or} \quad t > \theta_0 + \lambda_{\alpha/2}\,\sigma\} \\ &= (-\infty, \theta_0 - \lambda_{\alpha/2}\,\sigma) \cup (\theta_0 + \lambda_{\alpha/2}\,\sigma, +\infty). \end{aligned} \tag{30.2}$$

Now, if X has distribution $\mathcal{N}(\theta, \sigma^2)$ distribution, then $(X-\theta)/\sigma$ has standard normal distribution, so

$$\begin{aligned} h(\theta) &= \mathrm{P}(X \in C) = \mathrm{P}(X < \theta_0 - \lambda_{\alpha/2}\,\sigma) + \mathrm{P}(X > \theta_0 + \lambda_{\alpha/2}\,\sigma) \\ &= \mathrm{P}\left(\frac{X-\theta}{\sigma} < \frac{\theta_0-\theta}{\sigma} - \lambda_{\alpha/2}\right) + \mathrm{P}\left(\frac{X-\theta}{\sigma} > \frac{\theta_0-\theta}{\sigma} + \lambda_{\alpha/2}\right) \\ &= \Phi\left(\frac{\theta_0-\theta}{\sigma} - \lambda_{\alpha/2}\right) + \left(1 - \Phi\left(\frac{\theta_0-\theta}{\sigma} + \lambda_{\alpha/2}\right)\right). \end{aligned} \tag{30.3}$$

For $\theta_0 = 2$, $\sigma = .04$, and $\alpha = .05$, i.e., for a 5% significance test, this gives

$$C = \{t : t < 1.9216 \quad \text{or} \quad t > 2.0784\} = (-\infty, 1.9216) \cup (2.0784, +\infty).$$

Further, $h(2) = .05$, which is not surprising, since we are testing for $H_0 : \theta = 2$ at significance level 5%, and so $h(2)$ should equal to the siginince level of the test. We have $h(2.01) \approx .05719$, $h(2.05) \approx .23952$, $h(2.1) \approx .7054$, $h(2.2) \approx .9988$. This power function is graphed in Figure 30.1.

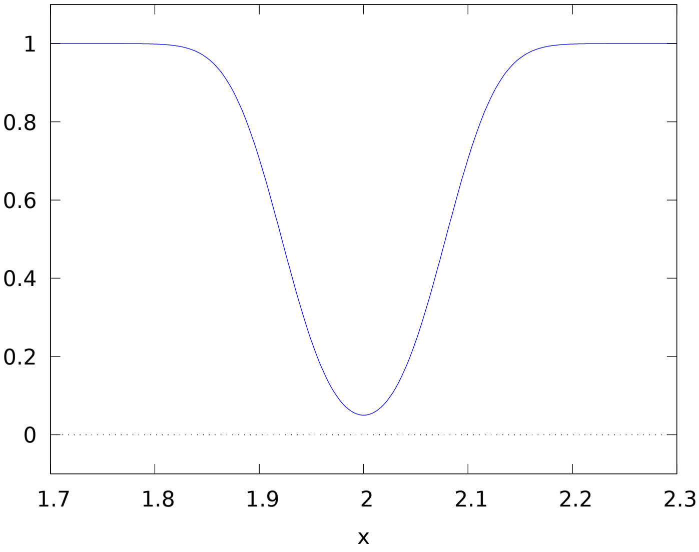

Figure 30.1: Power of test $H_0 : \theta = 2$, $\sigma = .04$

30.2 Power and sample size

Above, we discussed the case of a single normal variable. We will continue consideration of the example, discussed in [2, Example 2, p. 256] for larger sample size.

Assume that we are given a random sample of size n. That is, we are given random variables idependent, identically distributed random variables X_1, X_2, $\dots$, X_n, each having normal distribution $\mathcal{N}(\theta, \sigma^2)$, where σ^2 is known but θ is not. We want to test the hypothesis $H_0 : \theta = \theta_0$ against the alternative hypothesis $H_1 : \theta \ne \theta_0$. In case $\theta \ne \theta_0$, H_0 needs to be rejected. As we saw above, the random variable

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$$

has distribution $\mathcal{N}(\theta, \sigma^2/n)$. The considerations of Subsection 30.1 can be adapted to the present situation by replacing X by $\bar{X}_n$ and σ by $\sigma/\sqrt{n}$. Thus formula (30.2) is replaced by

$$\begin{aligned} C &= \left\{ t : t < \theta_0 - \lambda_{\alpha/2} \frac{\sigma}{\sqrt{n}} \quad \text{or} \quad t > \theta_0 + \lambda_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right\} \\ &= \left(-\infty, \theta_0 - \lambda_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right) \cup \left(\theta_0 + \lambda_{\alpha/2} \frac{\sigma}{\sqrt{n}}), +\infty \right), \end{aligned} \tag{30.4}$$

and H_0 is rejected if the test variable belongs to this set, i.e., if $\bar{X}_n \in C$, and it is not rejected if $\bar{X}_n \notin C$.

Now, if $\bar{X}_n$ has distribution $\mathcal{N}(\theta, \sigma^2/n)$ distribution, then $(X_n - \theta)/(\sigma/\sqrt{n})$ has standard normal distribution, so

$$
\begin{aligned}
h(\theta) = \mathrm{P}(\bar{X}_n \in C) &= \mathrm{P}\left(X < \theta_0 - \lambda_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) + \mathrm{P}\left(X > \theta_0 + \lambda_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) \\
&= \mathrm{P}\left(\frac{X-\theta}{\sigma/\sqrt{n}} < \frac{\theta_0-\theta}{\sigma} - \lambda_{\alpha/2}\right) + \mathrm{P}\left(\frac{X-\theta}{\sigma/\sqrt{n}} > \frac{\theta_0-\theta}{\sigma} + \lambda_{\alpha/2}\right) \qquad (30.5) \\
&= \Phi\left(\frac{\theta_0-\theta}{\sigma/\sqrt{n}} - \lambda_{\alpha/2}\right) + \left(1 - \Phi\left(\frac{\theta_0-\theta}{\sigma/\sqrt{n}} + \lambda_{\alpha/2}\right)\right).
\end{aligned}
$$

For $\theta_0 = 2$, $\sigma = .04$, $n = 100$, and $\alpha = .05$, i.e., for a 5% significance test with sample size 100, this gives

$$C = \{t : t < 1.9922 \quad \text{or} \quad t > 2.0078\} = (-\infty, 1.9922) \cup (2.0078, +\infty).$$

Further, $h(2) = .05$, which is not surprising, since we are testing for $H_0 : \theta = 2$ at significance level 5%, and so $h(2)$ should equal to the significance level of the test. We have $h(2.01) \approx .7054$, $h(2.02) \approx .9988$. The power functions for $n = 1$ (in blue), $n = 10$ (in red), and $n = 100$ (in green) are graphed together in Figure 30.2.

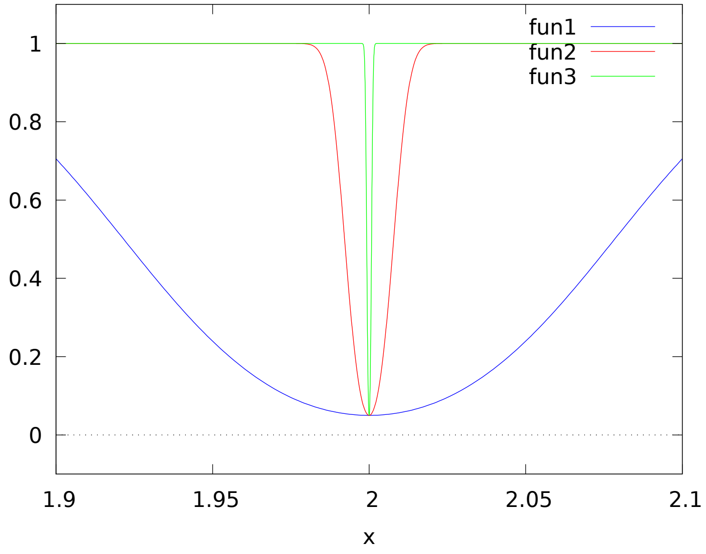

Figure 30.2: Power of test $H_0 : \theta = 2$, $\sigma = .04$

31 Types of errors

Suppose that in a statistical test, you have a null hypothesis H_0 and an alternative hypothesis H_1, and one has to choose which one is likely to be true. This kind of testing occurs all the time in practice; for example, one needs to decide if a given person has a disease or not. Making an incorrect decision can happen in two ways. *Type I error*, or *error of the first kind*, is made if H_0 is rejected when H_0 is true, and *type II error*, or *error of the second kind*, is made if H_0 is not rejected when

H_1 is true. The probability of a type I error is customarily denoted by α and the probability of a type II error is usually denoted by β; α is also called the significance of the test.

In our description of statistical testing, the alternative hypothesis was invariably taken to be the negation of the null hypothesis, such as in examples when we had $H_0 : \theta = \theta_0$ and $H_1 : \theta \neq \theta_0$. this was the approach taken by Ronald Fisher. However, in the example discussed above, at the end of Subsection 30.2, even for sample size $n = 1,000,000$, we have $h(2.000,1) = .050,007,159$. Given that the significance of the test is $h(2) = .05$, it does not seem reasonable to distinguish between $H_0 : \theta = 2$ and $\theta = 2.000,1$, when in fact the alternative hypothesis $H_1 : \theta \neq 2$ is true. In a practical situation, one wants an alternative hypothesis H_1 where a reasonable test can tell whether accepting H_0 or H_1 is the more reasonable choice.

31.1 How large a sample is needed?

In statistics, a *population* is a set of things from which samples may be drawn. To continue the example above, assume we are given a population with distribution $\mathcal{N}(\theta, .04^2)$. We want to test the null hypothesis $H_0 : \theta = 2$ against the alternative hypothesis $H_1 : |\theta - 2| \geq .01$. What this really means that we do not care if $\theta = 2$ does not exactly hold, because exact equality is not testable, but we want to ensure that $|\theta - 2| < .01$ holds. We take a sample of size n, that is, we take independent random variables X_1, X_2, ..., X_n, each with distribution $\mathcal{N}(\theta, .04^2)$. The question is, how large n should be that we have $\alpha \leq .05$ for the type I error, and $\beta \leq .05$ for the type II error. The restriction on α does not actually influence the size of n; the optimal size of n is determined by the restriction on β, and then the critical region needs to be chosen appropriately for the given n. Writing $H_1 = \{x : |x - 2| \geq .01$, the power function $h(\theta)$ has the smallest value for $\theta \in H_1$ is closest to 2. We then have $\beta = 1 - h(\theta)$ for this value of θ. That is,

$$1 - \beta = \min\{h(\theta) : \theta \in H_1\} = h(1.99) = h(2.01);$$

the equality here holds for reasons of symmetry.[31.1] Using formula (30.5), we can evaluate $h(2.01)$ for various values of n. Using binary search on a computer, one can fairly quickly determine that for $n = 207$, $h(2.01) = .949175$ and for $n = 208$, $h(2.01) = .950076$.[31.2] That is, a sample size of $n = 208$ ensures that $\beta = 1 - h(2.01) = .049924 < .05$. For this value of n, a the critical region of $.05 = 5\%$ significance is $(1.9946, 2.0054)$.

32 Maxima examples

The computer algebra program Maxima can be used to handle many of the problems discussed in this course.

32.1 Binomial distribution

Problem 32.1. Let X be a a random variable with $\mathrm{Bin}(n, p)$ distribution. In particular, consider the following

[31.1] In the general situation, $\min\{h(\theta) : \theta \in H_1\}$ may not exist, and one should take infimum (greatest lower bound, instead of minimum; see [6].

[31.2] In a binary search, if one finds that for $h(2.01) > .95$ for $n = 150$ and $h(2.01) < .95$ for $n = 100$, you try the middle fo the interval, $n = 125$, to see if the optimail value of n should be sought between 100 and 125 or between 125 and 150, and then picks n in the middle of the appropriate interval, and so on. If one wants to do binary search often, it is not hard to program it; if one wants to work out a single example, it is much faster to do the binary search manually.

(1) Assume $n = 600$ and $p = 3/5$. What is the probability that $X \leq 370$.

(2) Assume $n = 2600$ and $p = .0015$. What is the probability that $X = 5$.

Solution. The probability function of the random variable with binomial distribution $\mathrm{Bin}(n, p)$ can be described as

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

Hence, we have

$$\mathrm{P}(X \leq k) = \sum_{i=0}^{k} \binom{n}{i} p^k (1-p)^{n-i}. \tag{32.1}$$

Since this sum is difficult to evaluate without a computer, we used approximations to evaluate this sum. We will comment on the appropriate approximations below; first we show how to use computers to evaluate this sum. The following Maxima program will work:

```
 1  kill(all)$
 2  binprob(n, p, k) := float(binomial(n,k)*p^k*(1-p)^(n-k));
 3  bindist(n, p, k) :=
 4  do(
 5    sum : 0,
 6      for i : 0 step 1 thru k do(
 7        term : binprob(n, p, i),
 8        sum : float(sum+term)), return(sum)
 9  );
10  answer1 : bindist(600,3/5,370);
11  answer2 : binprob(2600, .0015, 5);
```

The numbers on the left sere line numbers, and are not part of the program. The first line kills all earlier variables; this is probably superfluous if the command is run from the Linux commandline. For example, if the file containing the above lines is called `binomdist.max`, then it can be run from the Linux command line with the command

```
$ maxima -b binomdist.max > binomdist.out
```

Here the $ sign at the beginning of the line is the prompt printed by the computer (meaning that the computer invites you to type something), and it should not be typed. However, the program can be run in different ways, such as when running Maxima interactively, and then it is important to delete the variables specified by earlier commands. In the command, the part "`> binomdist.out`" will redirect the output of the program into the file `binomdist.out` rather than print it to the screen (where it may not be readable it if does not fit into the viewing window). Line 2 in the program doed the formula for $\mathrm{P}(X = k)$. Here `bindist` is defined as a function of n, p, and k. The assignment symbol `:=` defines a function; `float` indicates that the result is interpreted as a floating point number. The predefined function `binomial(n,k)` describes the binomial coefficient $\binom{n}{k}$. The command `do(...)` on lines 4–11 groups several commands, seaparated by commas, into a single command. Usually, commands are terminated with a semicolon, but inside a `do` command, they are terminated by commas. Instead of a semicolon, one can also terminate a command by the dollar sign; this prevents the command from reporting its execution in the output. On line 11, in the command `sum : 0`, the colon is an assignment symbol, setting the value of a variable (rather than defining a function). On lines 6–8, the loop adds up the values probablities in the binomial distribution as described in formula (32.1). On lines 10 and 11, the answers to questions (1) and (2) of the problem are printed.

The output of the above problem is as follows:

```
 1 Maxima 5.38.1 http://maxima.sourceforge.net
 2 using Lisp GNU Common Lisp (GCL) GCL 2.6.12
 3 Distributed under the GNU Public License. See the file
 4 COPYING.
 5 Dedicated to the memory of William Schelter.
 6 The function bug_report() provides bug reporting
 7 information.
 8 (%i1) batch("binomdist.max")
 9
10 read and interpret file:
11 #p/home/mate/courses/elemprob/notes.dir/maxima/binomdist.max
12 (%i2) kill(all)
13 (%i1) binprob(n,p,k):=float(binomial(n,k)*p^k*(1-p)^(n-k))
14                                                      k
15   n - k
16 (%o1)      binprob(n, p, k) := float(binomial(n, k) p  (1 -
17 p)     )
18 (%i2) bindist(n,p,k):=do
19               (sum:0,
20                for i from 0 thru k do
21
22 (term:binprob(n,p,i),sum:float(sum+term)),return(sum))
23 (%o2) bindist(n, p, k) := do (sum : 0,
24 for i from 0 thru k do (term : binprob(n, p, i), sum :
25 float(sum + term)),
26 return(sum))
27 (%i3) answer1:bindist(600,3/5,370)
28 (%o3)                         0.8090292621686978
29 (%i4) answer2:binprob(2600,0.0015,5)
30 (%o4)                         0.1523035592368988
31 (%o4)                            binomdist.max
```

If the program is run from the command line with the command described above, the first line of the output is always a blank line; this has been deleted. Discountitn this blank line, the first seven lines of the output identifies the version of Maxima and related information. After this, the symbols `(%i1)`, `(%i2)`, `(%i3)`, ..., indentify input lines (usually just repeating what was in the input file, and the symbols `(%o1)`, `(%o2)`, `(%o3)` ..., identify output lines. After the `kill(all)` command listed in the output on line 12, no output line is given; this is because in the program file above, the first line was terminated by the dollar symbol `$` and not by a semicolon. Much of the output just repeats the lines of the input. The answers are given on lines 17–31. These show that the answer to question (1) is 0.8090292621686978, and the answer to question (2) is 0.1523035592368988.

The following program gives a solution to question (1) using the normal approximation for the binomial distribution:

```
kill(all);
func : 1/sqrt(2*%pi)*exp(-t^2/2);
normdist(x) := float(1/2+romberg(func,t,0,x));
n : 600;
```

```
5 p : 3/5;
6 x : 370;
7 y : (x-n*p+1/2)/sqrt(n*p*(1-p));
8 normdist(y);
```

Rather than usin Maxima resources for the normal distribution, the density function of the standard normal distribution is given on line 2; here `%pi` denotes the constant π. On line 2 `romberg` refers to a powerful method of numerical integration method invented by Werner Romberg. On line 6, the binomial variable x is given the value 370, and the correspoinding standard normal variable is calculated. In the numerator, 1/2 is added as continuity correction.

The output file is given next:

```
Maxima 5.38.1 http://maxima.sourceforge.net
using Lisp GNU Common Lisp (GCL) GCL 2.6.12
Distributed under the GNU Public License. See the file
COPYING.
Dedicated to the memory of William Schelter.
The function bug_report() provides bug reporting
information.
(%i1) batch("normal.max")

read and interpret file:
#p/home/mate/courses/elemprob/notes.dir/maxima/normal.max
(%i2) kill(all)
(%o0)                                done
(%i1) func:(1/sqrt(2*%pi))*exp((-t^2)/2)
                                         2
                                        t
                                      - --
                                        2
                                    %e
(%o1)                          -----------------
                               sqrt(2) sqrt(%pi)
(%i2) normdist(x):=float(1/2+romberg(func,t,0,x))
                                    1
(%o2)          normdist(x) := float(- + romberg(func, t, 0,
x))
                                    2
(%i3) n:600
(%o3)                                 600
(%i4) p:3/5
                                       3
(%o4)                                  -
                                       5
(%i5) x:370
(%o5)                                 370
(%i6) y:(x-n*p+1/2)/sqrt(n*p*(1-p))
                                       7
(%o6)                                  -
                                       8
(%i7) normdist(y)
(%o7)                         0.8092130401048883
(%o7)                             normal.max
```

Note that on line 13, the output line `done` appears; this is because in the program file, the first line was terminated by a semicolon, rather than a dollar sign $. The answer 0.8092130401048883 is printed on line 40; note that it agrees with the answer given above up to three decimals.

As for question (2), an approximate answwer can be given using the Poisson approximation to the binomial distribution. The program file doing this calculation is the following:

```
1  kill(all)$
2  po(z,k) := z^k/k!*exp(-z);
3  n : 2600;
4  p : .0015;
5  k : 5;
6  z : n*p;
7  po(z,k);
```

The probability function P($Z = k$) with of a random variable Z with a Poisson distribution Po(z) is defined on line 2. The binomial variable in the quesion is approximated by the variable Z with $z = np = 2600 \cdot .0015$. The output file is as follows:

```
 1  Maxima 5.38.1 http://maxima.sourceforge.net
 2  using Lisp GNU Common Lisp (GCL) GCL 2.6.12
 3  Distributed under the GNU Public License. See the file
 4  COPYING.
 5  Dedicated to the memory of William Schelter.
 6  The function bug_report() provides bug reporting
 7  information.
 8  (%i1) batch("poisson.max")
 9
10  read and interpret file:
11  #p/home/mate/courses/elemprob/notes.dir/maxima/poisson.max
12  (%i2) kill(all)
13  (%i1) po(z,k):=(z^k/k!)*exp(-z)
14                                           k
15                                          z
16  (%o1)                      po(z, k) := (--) exp(- z)
17                                          k!
18  (%i2) n:2600
19  (%o2)                                2600
20  (%i3) p:0.0015
21  (%o3)                               0.0015
22  (%i4) k:5
23  (%o4)                                  5
24  (%i5) z:n*p
25  (%o5)                                 3.9
26  (%i6) po(z,k)
27  (%o6)                         0.1521925205355527
28  (%o6)                             poisson.max
```

The value of z is calculated to be 3.9 on line 25, and the answer to the question is given on line 27 as 0.1521925205355527. Note that the first three decimals of this answer agree with the answer given above.

33 Bootstrapping

Assume we are given two random samples, X_1, X_2, …, X_m, and Y_1, Y_2, …, Y_n, where we assume that all the X_i are identically distributed, and similarly, all the Y_j are identically distributed, but the distribution of the X_i and that of the Y_i are not known to be the same. We want to test whether $\mathrm{E}(X_i) = \mathrm{E}(Y_j)$. This kind of question is very important in medical research. The article [3, p. 450] by Bradley Efron describes such a situation where out of 16 mice, 7 received a certain inoculation expected to increase a certain kind of cell count. The question was whether the inoculation was effective. Before computers, statisticians would assume that the distributions of both populations are normal, but the means and the variances of the two populations may be different. There is a variant of the Student t-test, called the Welch's t-test, that then can test if the means are the same; however, the assumptions that the cell count of the mice is normally distributed is questionable. Besides, the theory of this test is beyond the scope of the present notes.

A computer-intensive methods that can be used in this situation, and many other situations where no satisfactory theory is available, are resampling methods, creating an empirical distribution from the available data, and check the location of the sample parameter in this empirical distribution. The bootstrap method adapted the situation described above can be carried out as follows. Assume we are given observed data x_1, x_2, …, x_m, and y_1, y_2, …, y_n, we first calculate the means

$$\bar{x} = \frac{1}{m}\sum_{i=1}^{m} x_i \quad \text{and} \quad \bar{y} = \frac{1}{n}\sum_{j=1}^{n} y_j.$$

Then we calculate the value $\bar{x} - \bar{y}$; this will be our *test statistic.* Then we pool the data in a single list x_1, x_2, …, x_{m+n} where $x_{m+j} = y_j$ for j with $1 \le j \le n$. Finally, we randomly pick u_1, u_2, …, u_m, and v_1, v_2, …, v_n *with replacement* from the pooled data, and calculate the means

$$\bar{u} = \frac{1}{m}\sum_{i=1}^{m} u_i \quad \text{and} \quad \bar{v} = \frac{1}{n}\sum_{j=1}^{n} v_j.$$

and calculate the *data point* $\bar{u} - \bar{v}$. We repeat this a large number of times; these data points form our empirical distribution. We locate where our test statistic lies in this empirical distribution to evaluate the result of this test. An example for the method is given in the following program:

```
kill(all)$
linel : 60$
numer : true$
load(distrib)$
st : make_random_state(63088571546)$
set_random_state(st);
size1 : 10;
size2 : 20;
n : 1000000;
x : random_normal(12,3,size1);
y : random_normal(11,5,size2);
do(
xbar : 0,
for i : 1 step 1 thru size1 do(
   xbar : xbar+x[i]), return(xbar)
);
xbar : xbar/size1;
```

```
18  do(
19  ybar : 0,
20  for i : 1 step 1 thru size2 do(
21     ybar : ybar+y[i]), return(ybar)
22  );
23  ybar : ybar/size2;
24  bardiff : ybar-xbar;
25  x : append(x,y)$
26  count : 0;
27  do(
28     for i : 1 step 1 thru n do(
29
30  do(
31  xbar : 0,
32  for i : 1 step 1 thru size1 do(
33     xbar : xbar+x[random(size1+size2)+1]), return(xbar)
34  ),
35  xbar : xbar/size1,
36  do(
37  ybar : 0,
38  for i : 1 step 1 thru size2 do(
39     ybar : ybar+x[random(size1+size2)+1]), return(ybar)
40  ),
41  ybar : ybar/size2,
42  barindiff : ybar-xbar,
43  count : if barindiff <= bardiff then count+1 else count
44     ),
45     return(count/n)
46  );
```

We will explain the lines in this program. On line 2, the command `linel : 60` specifies the maximum length of output lines in the program (longer lines will be broken up). This is important only for the present manuscript, to that the listing of the output can be nicely displayed in the manuscript. On line 3 the program is told to do numerical (i.e., mostly floating point) calculations by the command `numer : true$`, as we mentioned before, the dollar sign $ is to separate commands (also called statements), suppressing output (unlike a semicolon would). On line 4, a package with probability distributions and related items is loaded. On line 5 a *random seed* (see below) is created and assigned to the variable `st`. This is set as the random seed to be used in the calculations that follows.

There are many computer algorithms that rely on random features or random numbers. The random numbers that are used are not really numbers, since they are calculated from a given number called seed; such numbers are called pseudo-random numbers. Truly random numbers would require a separate unit in the computer that generate such numbers, and it is difficult to incorporate such a device in a computer that can reliably do this. Another, more important point is that pseudorandom numbers in repeated calculations will always be the same, a point that is important for testing programs. In later runs of the program, the seed may be taken from unpredictable values when running the program, such as the time on the clock, the process number of the latest process running, etc. Often, random programs play an important role in computer security, and in such cases it is important that the seed cannot be guessed. In fact, an early ransomware attack on Linux failed because it was possible to recover the encryption key from the seed the random number generator used; see the article Linux Ransomware Debut Fails on Predictable Encryption Key.

On lines 7 and 8, the sizes of the two samples are given, and the samples are generated by the commands on lines 10 and 11. The first command generate a list of size `size1` of random values from a normal distribution with mean 12 and variance 3; the second command works similarly. Going back to line 9, `n` will denote the number of random re-sample points that will be generated. On lines 13 through 23 the averages of the lists tt x and `y` are determined, and on line 23 their difference, the test statistic to be used, is determined. Note that on lines 18 and 22 `do` introduces a compound statement; inside a compound statements, the individual statements must be separated by commas, and not semicolons or the dollar symbols $.

On line 25, the list `x` is extended by attaching the list tt y at the end. On line 26, the variable `count` will count the number of times the randomly generated data points are exceeded by the test statistic. On lines 27–44 the random data points are generated, ond the number for which the the test statistic exceeds the value of the generated data points is counted. Recall that in the big compound statement encompassing these lines, the individual commands must all be separated by commas.

We are going to give more detain in this command (or statement). On lines 33 and 39 a random member of the expanded list `x` is taken. Here the command `random(size1+size2)` generates a random integer between 0 and the argument minus 1, inclusive, i.e., between 0 and *size1+size2-1*, inclusive.[33.1] However, the members a list (same as an array in Maxima) are numbered between 1 and the size of the list (in, so 1 needs to be added to find a random member of the list `x`.[33.2] The variable names `xbar` and `ybar` are re-used in these commands, since their old meanings are no longer needed. On line 43, 1 is added to the variable `count` is the test statistic exceeds the value of the random data point generated. On line 45, the value of the ratio of the "hits" to the total number of "tries" is calculated. This is the proportion of the data point that are exceeded by the test statistic.

Next, the output of the program is given.

```
Maxima 5.38.1 http://maxima.sourceforge.net
using Lisp GNU Common Lisp (GCL) GCL 2.6.12
Distributed under the GNU Public License. See the file
COPYING.
Dedicated to the memory of William Schelter.
The function bug_report() provides bug reporting
information.
(%i1) batch("bootstrap.max")

read and interpret file: #p/home/mate/courses/elemprob
/notes.dir/maxima/bootstrap.max
(%i2) kill(all)
(%i1) linel:60
(%i2) numer:true
(%i3) load(distrib)
(%i4) st:make_random_state(63088571546)
(%i5) set_random_state(st)
(%o5)                           done
(%i6) size1:10
(%o6)                            10
(%i7) size2:20
(%o7)                            20
```

[33.1] The command `random` behaves this way if the argument is an integer; it generates a real if the argument is a floating point number.

[33.2] Many programming languages number the members of the arrays between 0 and the size of the array minus 1.

```
23  (%i8) n:1000000
24  (%o8)                     1000000
25  (%i9) x:random_normal(12,3,size1)
26  (%o9) [12.64574680588857, 13.97966559399711,
27  13.28417370283619, 3.841194836990418, 12.92512172990199,
28  16.82840523437749, 10.1622143562734 8, 19.8113697127616,
29  11.67454720345394, 10.6041752980218]
30  (%i10) y:random_normal(11,5,size2)
31  (%o10) [14.09697526729592, 17.00558847647043,
32  6.929914408687217, 14.99987194697551, 5.394536281479344,
33  4.57483038703998, 20.47900143486329, 4.437983820671345,
34  10.04598103959843, 1.158480883500161, 13.14915633518043,
35  13.85346089584671, 10.39256504632421, 8.054827805217816,
36  11.57127973699703, 13.53056022281369, 13.90649381039518,
37  8.974917606328738, 14.61315494011777, 5.609666371340634]
38  (%i11) do (xbar:0,for i thru size1 do xbar:xbar+x[i],
39             return(xbar))
40  (%o11)               125.7566144745026
41  (%i12) xbar:xbar/size1
42  (%o12)               12.57566144745026
43  (%i13) do (ybar:0,for i thru size2 do ybar:ybar+y[i],
44             return(ybar))
45  (%o13)               212.7792467171439
46  (%i14) ybar:ybar/size2
47  (%o14)               10.63896233585719
48  (%i15) bardiff:ybar-xbar
49  (%o15)                - 1.936699111593062
50  (%i16) x:append(x,y)
51  (%i17) count:0
52  (%o17)                       0
53  (%i18) do (for i thru n do
54                 (do (xbar:0,
55                      for i thru size1 do
56                          xbar:xbar+x[random(size1+size2)+1],
57                      return(xbar)),xbar:xbar/size1,
58                  do (ybar:0,
59                      for i thru size2 do
60                          ybar:ybar+x[random(size1+size2)+1],
61                      return(ybar)),ybar:ybar/size2,
62                  barindiff:ybar-xbar,
63                  count:if barindiff <= bardiff then count+1
64                            else count),return(count/n))
65  (%o18)                    0.140866
66  (%o18)                 bootstrap.max
```

There is not much to say about the output, since most of the output lines just echo the input lines in the program. On line 65, the proportion of the data points exceeded by the test statistic is printed. It is printed as a floating point number because of the directive `numer : true$` on line 3 of the program.

We ran this program several times with the same input data `x` and `y` without specifying the random seed to see how much the random seed influences the result. To this end, lines 5 and 6 need to be removed, and lines 10 and 11 defining `x` and `y` as a random list by their values calculated by the given random seeds, since if the computer chooses its own seed to generate these lists, they will be different, and the comparison as to how the statistical test in the program works with different random seeds will not be valid. That is, lines 10 and 11 of the program will need to be replaced by

```
x : [12.64574680588857, 13.97966559399711,
13.28417370283619, 3.841194836990418, 12.92512172990199,
16.82840523437749, 10.16221435627348, 19.8113697127616,
11.67454720345394, 10.6041752980218];
y : [14.09697526729592, 17.00558847647043,
6.929914408687217, 14.99987194697551, 5.394536281479344,
4.57483038703998, 20.47900143486329, 4.437983820671345,
10.04598103959843, 1.158480883500161, 13.14915633518043,
13.85346089584671, 10.39256504632421, 8.054827805217816,
11.57127973699703, 13.53056022281369, 13.90649381039518,
8.974917606328738, 14.61315494011777, 5.609666371340634];
```

Here the values for x and y are taken from lines 26–37 of the output above.

The program with one million random data points runs on my computer in less than a minute and a half even with the lowest priority. The various values of the proportion calculated without the program specifying the seed were 0.140655, 0.140541, 0.141095, 0.140749, 0.141391. With $100,000$ random data points, the program runs for about 22 seconds, and the results are 0.14123, 0.13908, 0.14012, 0.14039, 0.14042. With $10,000$ data points, it runs for about 11 seconds, and the results are 0.1386, 0.1392, 0.1446, 0.1394, 0.1423, 0.1459. With 1000 data points, the program runs for less than a second, and the results are 0.15, 0.156, 0.133, 0.154, 0.13.

34 The Weierstrass approximation theorem: a proof using probability theory

The Weierstrass approximation theorem says that function continuous on a closed interval can be approximated arbitrarily well on that interval by polynomials. Later, we will precisely formulate this result, and give the proof found by Sergei Bernstein (see [1]). The proof uses Chebyshev's inequality (21.1) applied to a random variable $X(n,x)$ having binomial distribution $\mathrm{Bin}(n,x)$ for integers $n>0$ and reals x with $0\le x\le 1$. In our discussion in Section 15 we required $0<x<1$ (we used p instead of x) but adding the cases $x=0$ and $x=1$ causes no difficulty, even though considering these cases was not important in the context of probability theory, since $X(n,0)=0$ and $X(n,1)=n$ with probability 1. The important relations $\mathrm{E}\big(X(n,x)\big)=nx$, $\mathrm{V}\big(X(n,x)\big)=nx(1-x)$, and $\mathrm{P}\big(X(n,x)=k\big)=\binom{n}{k}x^k(1-x)^{n-k}$ for k with $0\le k\le n$, remain valid also in cases $x=0$ and $x=1$, the last one with the special stipulation that we take $x^0=1$ in case $x=0$ and $(1-x)^0=1$ in case $x=1$.[34.1] We will adopt this stipulation throughout this section.

34.1 Requirements from elementary analysis

The topic of elementary analysis is the same as basic calculus, but with rigorous proofs.[34.2] We will use two important result in elementary analysis. The first one is asserts that continuous functions on a closed intervals are bounded:

[34.1] 0^0 is usually undefined, for good reason, but such a stipulation is always adopted when discussing power series, for example.

[34.2] The college course discussing elementary analysis is usually called Advanced Calculus. While the students reading these notes may not have taken Advanced Calculus, the explanations we give here should suffice.

Theorem 34.1. *Given a function f that is continuous on a closed interval $[a, b]$, there is an M such that $|f(x)| \leq M$ for all $x \in [a, b]$.*

Such an M is usually called a *bound* for the absolute value of f on $[a, b]$, and f itself can be described as *bounded.* Another important result we need is the following

Theorem 34.2. *Given a function f that is continuous on a closed interval $[a, b]$, for every $\epsilon > 0$ there is a $\delta > 0$ such that $|f(x) - f(y)| < \epsilon$ whenever $x, y \in [a, b]$ are such that $|x - y| < \delta$.*

The property of f described in this theorem is called *uniform continuity*, and f having this property is called *uniformly continuous.* One can concisely formulate this result by saying that a function that is continuous on a closed interval is also uniformly continuous. The difference of the concept of continuity and uniform continuity may be a little hard to appreciate, so we make a detour to explain the difference in some detail.

34.1.1 Continuity versus uniform continuity

Since the property of uniform continuity mentioned in Theorem 34.2 may not be familiar to some of the readers of these notes, we will make some effort to explain the difference between continuity and uniform continuity. Let S be a set of real numbers,

Definition 34.1. The function $f : S \to \mathbb{R}$ is said to be continuous on S if for every $x \in S$ and for every $\epsilon > 0$ there is a $\delta > 0$ such that for every $y \in S$, if $|x - y| < \delta$ then $|f(x) - f(y)| < \epsilon$.

Note. The above definition says that f is continuous in S at x for every $x \in S$. Using logic notation, the function $f : S \to \mathbb{R}$ is said to be continuous on S if

$$(\forall x \in S)(\forall \epsilon > 0)(\exists \delta > 0)(\forall y \in S)\big(|x - y| < \delta \to |f(x) - f(y)| < \epsilon\big).$$

The first two quantifiers are interchangeable here, since two quantifiers of the same type (i.e., two universal quantifiers, or two existential quantifiers) are interchangeable, so we can write this also as

$$(\forall \epsilon > 0)(\forall x \in S)(\exists \delta > 0)(\forall y \in S)\big(|x - y| < \delta \to |f(x) - f(y))| < \epsilon\big). \tag{34.1}$$

Definition 34.2. The function $f : S \to \mathbb{R}$ is said to be uniformly continuous on S if for every $\epsilon > 0$ there is a $\delta > 0$ such that for every $x \in S$ and for every $y \in S$, if $|x - y| < \delta$ then $|f(x) - f(y)| < \epsilon$.

Note. Using logic notation, the function $f : S \to \mathbb{R}$ is said to be uniformly continuous on S if

$$(\forall \epsilon > 0)(\exists \delta > 0)(\forall x \in S)(\forall y \in S)\big(|x - y| < \delta \to |f(x) - f(y)| < \epsilon\big). \tag{34.2}$$

The formal difference between continuity and uniform continuity is indicated by the different order of the second and third quantifiers in formulas (34.1) and (34.2); aside from this difference, the two formulas are identical. However, these quantifiers are of different type (one is a universal quantifier, the other is an existential quantifier), and so they are not interchangeable. Hence the meanings of these two formulas are different.

To explain the difference in a less formal way, in case of continuity, δ depends on the choice of x as well as on ϵ (and, of course, on the function f itself), whereas in case of uniform continuity, δ depends only on ϵ but not on x (but it does depend on the function f itself).

34.2 Statement and proof of the Weierstrass approximation theorem

We will formulate the result specifically for the interval $[0,1]$.

Theorem 34.3 (Weierstrass approximation theorem)**.** *Let f be a function that is continuous on the interval $[0,1]$, and let $\epsilon > 0$ be a real number. Then there is a polynomial P such that $|f(x) - P(x)| < \epsilon$ for all $x \in [0,1]$.*

More specifically: writing

$$P_n(x) = \sum_{k=0}^{n} \binom{n}{k} f\left(\frac{k}{n}\right) x^k (1-x)^{n-k} \qquad (n > 0),$$

there is an integer $N > 0$ such that $|f(x) - P_n(x)| < \epsilon$ for all $x \in [0,1]$ and for all $n \geq N$.

The statement in the second paragraph is due to Sergei Bernstein.

Proof. Let M be a bound for f on $|f|$ in the interval $[0,1]$ (see Theorem 34.1), that is, M is such that $|f(x)| \leq M$ for all $x \in [0,1]$. Further, let $\delta > 0$ be such that $|f(x) - f(y)| < \epsilon/2$ whenever $x, y \in [0,1]$ and $|x - y| < \delta$.

Let $n > 0$ be an integer. Pick a fixed $x \in [0,1]$, and let k be an integer with $0 \leq k \leq n$. Then we have

$$\left| f(x) - f\left(\frac{k}{n}\right) \right| \leq 2M$$

and

$$\left| f(x) - f\left(\frac{k}{n}\right) \right| \leq \frac{\epsilon}{2} \qquad \text{if} \quad |k - nx| < n\delta$$

according to the above estimates. Writing $X = X(n,x)$ for a random variable with $\mathrm{Bin}(n,x)$ distribution, we have

$$P_n(x) = \sum_{k=0}^{n} f\left(\frac{n}{k}\right) \mathrm{P}(X = k) \qquad \text{and} \qquad f(x) = \sum_{k=0}^{n} f(x)\, \mathrm{P}(X = k).$$

Hence

$$\begin{aligned} |f(x) - P_n(x)| &\leq \sum_{k=0}^{n} \left| f(x) - f\left(\frac{k}{n}\right) \right| \mathrm{P}(X = k) \\ &\leq 2M \sum_{\substack{k: 0 \leq k \leq n \\ |k - nx| \geq n\delta}} \mathrm{P}(X = k) + \frac{\epsilon}{2} \sum_{\substack{k: 0 \leq k \leq n \\ |k - nx| < n\delta}} \mathrm{P}(X = k) \\ &= 2M\, \mathrm{P}(|X - nx| \geq n\delta) + \frac{\epsilon}{2}\, \mathrm{P}(|X - nx| < n\delta), \end{aligned}$$

where the inequality follows form the above estimates for the change of f. We have $\mathrm{E}(X) = nx$ and $\mathrm{V}(X) = nx(1-x) < n$. Hence, by Chebyshev's inequality (21.1) we have

$$\mathrm{P}(|X - nx| \geq n\delta) \leq \frac{n}{n^2\delta^2} = \frac{1}{n\delta^2}.$$

This estimate together with the trivial estimate $P(|X - nx| < n\delta) \leq 1$, gives

$$|f(x) - P_n(x)| \leq \frac{2M}{n\delta^2} + \frac{\epsilon}{2} \leq \frac{2M}{N\delta^2} + \frac{\epsilon}{2}$$

for $n \geq N$. Choosing $N > 4M/(\delta^2\epsilon)$, the right-hand side will be less than ϵ, completing the proof. □

References

[1] S. Bernstein. Démonstration du théorème de Weierstrass fondée sur le calcul des probabilités. *Soobshch. Kharkov. Mat. Obshch. (Comm. of the Kharkhov Math. Soc.)*, 13, 2nd Ser.:909–996, 1913. On the Web as the original paper and as an English translation.

[2] Gunnar Blom. *Probability and Statistics: Theory and Applications.* Springer-Verlag, New York, 1989.

[3] Bradley Efron. Computers and the theory of statistics: Thinking the unthinkable. *SIAM Review*, 21(4):460–480, 1979. JSTOR stable url: `http://www.jstor.org/stable/2030104`.

[4] Bradley Efron and Trevor Hastie. *Computer Age Statistical Inference: Algorithms, Evidence, and Data Science.* Cambridge University Press, Cambridge, United Kingdom–New York, USA, 2016.

[5] Attila Máté. Introduction to numerical analysis with C programs. `http://www.sci.brooklyn.cuny.edu/~mate/nml/numanal.pdf`, August 2013.

[6] Attila Máté. Supplementary notes on Introduction to Analysis. `http://www.sci.brooklyn.cuny.edu/~mate/anl/analysis.pdf`, January 2013.

[7] Attila Máté. The natural exponential function. `http://www.sci.brooklyn.cuny.edu/~mate/misc/exp_x.pdf`, September 2015.

[8] Attila Máté. Aspects of time series, December 2016. `http://www.sci.brooklyn.cuny.edu/~mate/misc/time_series.pdf`.

[9] W. V. Quine. *Mathematical Logic.* Harvard University Press, Cambridge, MA, USA, revised edition, 1991.

[10] Stanislaw Ulam. Zur Masstheorie in der allgemeinen Mengenlehre. *Fundamenta Mathematicae*, 16(1):140–150, 1930. Free access: European Digital Mathematics Library.