Probability and Statistics for SIC

http://stat.epfl.ch

Course material

Probability constitutes roughly the first 60% of the course, and a good book is

- ☐ Ross, S. M. (2007) *Initiation aux Probabilités*. PPUR: Lausanne.
- ☐ Ross, S. M. (2012) *A First Course in Probability*, 9th edition. Pearson: Essex.

Statistics comprises roughly the last 40% of the course. Possible books are

- ☐ Davison, A. C. (2003). *Statistical Models*. Cambridge University Press. Sections 2.1, 2.2; 3.1, 3.2; 4.1–4.5; 7.3.1; 11.1.1, 11.2.1.
- ☐ Morgenthaler, S. (2007) *Introduction à la Statistique*. PPUR: Lausanne.
- ☐ Wild, C. & Seber, G. A. F. (2000). *Chance Encounters: A First Course in Data Analysis and Statistics*. John Wiley & Sons: New York.
- ☐ Helbling, J.-M. & Nuesch, P. (2009). *Probabilités et Statistique* (polycopie).

There are many excellent introductory books on both topics, look in the Rolex Learning Centre.

1.2 Preliminary ideas

Sets

Definition 1. *A* **set** A *is a unordered collection of objects,* $x_1, \ldots, x_n, \ldots$*:*

$$A = \{x_1, \ldots, x_n, \ldots\}\,.$$

We write $x \in A$ *to say that '*x *is an element of* A*', or '*x *belongs to* A*'. The collection of all possible objects in a given context is called the* **universe** Ω*.*
An **ordered set** *is written* $A = (1, 2, \ldots)$*. Thus* $\{1, 2\} = \{2, 1\}$*, but* $(1, 2) \neq (2, 1)$*.*

Examples:

$$\begin{aligned}
C_H &= \{\text{Geneva}, \text{Vaud}, \ldots, \text{Grisons}\} \quad \text{set of Swiss cantons} \\
\{0, 1\} &= \text{finite set made up of the elements 0 and 1} \\
\mathbb{N} &= \{1, 2, \ldots\}, \quad \text{positive integers, countable set} \\
\mathbb{Z} &= \{\ldots, -1, 0, 1, 2, \ldots\}, \quad \text{integers, countable set} \\
\mathbb{R} &= \text{real numbers, uncountable set} \\
\emptyset &= \{\,\} \quad \text{empty set, has no elements}
\end{aligned}$$

Subsets

Definition 2. *A set A is a* **subset** *of a set B if $x \in A$ implies that $x \in B$: we write $A \subset B$.*

- If $A \subset B$ and $B \subset A$, then every element of A is contained within B and vice versa, thus $A = B$: both sets contain exactly the same elements.
- Note that $\emptyset \subset A$ for every set A. Thus,

$$\emptyset \subset \{1,2,3\} \subset \mathbb{N} \subset \mathbb{Z} \subset \mathbb{Q} \subset \mathbb{R} \subset \mathbb{C}, \quad \emptyset \subset \mathbb{I} \subset \mathbb{C}$$

- **Venn diagrams** are useful for grasping the existing elementary relations between sets, but they can be deceptive (not all relations can be represented).

Cardinal of a set

Definition 3. *A finite set A has a finite number of elements, and this number is called its* **cardinal**:

$$\text{card } A, \quad \#A, \quad |A|.$$

- Evidently $|\emptyset| = 0$ and $|\{0,1\}| = 2$
- **Exercise**: Show that if A and B are finite and $A \subset B$, then $|A| \leq |B|$.

Boolean operations

Definition 4. *Let $A, B \subset \Omega$. Then we can define*

- *the* **union** *and the* **intersection** *of A and B to be*

$$A \cup B = \{x \in \Omega : x \in A \text{ or } x \in B\}, \; A \cap B = \{x \in \Omega : x \in A \text{ and } x \in B\};$$

- *the* **complement** *of A in Ω to be $A^c = \{x \in \Omega : x \notin A\}$.*

Evidently $A \cap B \subset A \cup B$,and if the sets are finite, then

$$|A| + |B| = |A \cap B| + |A \cup B|, \quad |A| + |A^c| = |\Omega|.$$

We can also define the **difference between A and B** to be

$$A \setminus B = A \cap B^c = \{x \in \Omega : x \in A \text{ and } x \notin B\},$$

(note that $A \setminus B \neq B \setminus A$), and the **symmetric difference**

$$A \triangle B = (A \setminus B) \cup (B \setminus A).$$

Boolean operations

If $\{A_j\}_{j=1}^{\infty}$ is an infinite set of the subsets of Ω, then

$$\bigcup_{j=1}^{\infty} A_j = A_1 \cup A_2 \cup \cdots : \quad \text{those } x \in \Omega \text{ that belong to } \textbf{at least} \text{ one } A_j;$$

$$\bigcap_{j=1}^{\infty} A_j = A_1 \cap A_2 \cap \cdots : \quad \text{those } x \in \Omega \text{ that belong to } \textbf{every } A_j.$$

The following results are easy to show (e.g., using Venn diagrams):

- $(A^c)^c = A, \quad (A \cup B)^c = A^c \cap B^c, \quad (A \cap B)^c = A^c \cup B^c;$
- $A \cap (B \cup C) = (A \cap B) \cup (A \cap C), \quad A \cup (B \cap C) = (A \cup B) \cap (A \cup C);$
- $(\bigcup_{j=1}^{\infty} A_j)^c = \bigcap_{j=1}^{\infty} A_j^c, \quad (\bigcap_{j=1}^{\infty} A_j)^c = \bigcup_{j=1}^{\infty} A_j^c.$

Partition

Definition 5. *A* **partition** *of Ω is a collection of nonempty subsets $A_1, \ldots, A_n$ in Ω such that*

- *the A_j are* **exhaustive**, *i.e., $A_1 \cup \cdots \cup A_n = \Omega$, and*
- *the A_j are* **disjoint**, *i.e., $A_i \cap A_j = \emptyset$, for $i \neq j$.*

A partition can also be composed of an infinity of sets $\{A_j\}_{j=1}^{\infty}$.

Example 6. *Let $A_j = [j, j+1)$, for $j = \ldots, -1, 0, 1, \ldots$. Do the A_j partition $\Omega = \mathbb{R}$?*

Example 7. *Let A_j be the set of integers that can be divided by j, for $j = 1, 2, \ldots$. Do the A_j partition $\Omega = \mathbb{N}$?*

Note to Example 6

Obviously, $A_j \cap A_i = \emptyset$ if $i \neq j$. Moreover any real number x lies in $A_{\lfloor x \rfloor}$, where $\lfloor x \rfloor$ is the largest integer less than or equal to x. Therefore these sets partition $\mathbb{R}$.

Note to Example 7

Note that $6 \in A_2 \cap A_3$, so these sets do not partition $\mathbb{N}$.

Cartesian product

Definition 8. *The* **Cartesian product** *of two sets A, B is the set of* **ordered** *pairs*

$$A \times B = \{(a, b) : a \in A, b \in B\}.$$

In the same way

$$A_1 \times \cdots \times A_n = \{(a_1, \ldots, a_n) : a_1 \in A_1, \ldots, a_n \in A_n\}.$$

If $A_1 = \cdots = A_n = A$, then we write $A_1 \times \cdots \times A_n = A^n$.

As the pairs are ordered, $A \times B \neq B \times A$ unless $A = B$.
If $A_1, \ldots, A_n$ are all finite, then

$$|A_1 \times \cdots \times A_n| = |A_1| \times \cdots \times |A_n|.$$

Example 9. *Let $A = \{a, b\}, B = \{1, 2, 3\}$. Describe $A \times B$.*

Note to Example 9

$\{(a, 1), (a, 2), \ldots, (b, 3)\}$.

1.3 Combinatorics

Combinatorics: Reminders

Combinatorics is the mathematics of counting. Two basic principles:

- **multiplication:** if I have m hats and n scarves, there are $m \times n$ different ways of wearing both a hat and a scarf;
- **addition:** if I have m red hats and n blue hats, then I have $m + n$ hats in total.

In mathematical terms: if $A_1, \ldots, A_k$ are sets, then

$$|A_1 \times \cdots \times A_k| = |A_1| \times \cdots \times |A_k|, \quad \text{(multiplication)},$$

and if the A_j are disjoint, then

$$|A_1 \cup \cdots \cup A_k| = |A_1| + \cdots + |A_k|, \quad \text{(addition)}.$$

Permutations: Ordered selection

Definition 10. *A* **permutation** *of n distinct objects is an ordered set of those objects.*

Theorem 11. *Given n distinct objects, the number of different permutations (without repetition) of length $r \leq n$ is*

$$n\ (n-1)\ (n-2)\ \cdots\ (n-r+1) = \frac{n!}{(n-r)!}.$$

Thus there are $n!$ permutations of length n.

Theorem 12. *Given $n = \sum_{i=1}^{r} n_i$ objects of r different types, where n_i is the number of objects of type i that are indistinguishable from one another, the number of permutations (without repetition) of the n objects is*

$$\frac{n!}{n_1!\ n_2!\ \cdots\ n_r!}.$$

Example

Example 13. *A class of 20 students choose a committee of size 4 to organise a 'voyage d'études'. In how many different ways can they pick the committee if:*
(a) there are 4 distinct roles (president, secretary, treasurer, travel agent)?
(b) there is one president, one treasurer, and two travel agents?
(c) there are two treasurers and two travel agents?
(d) their roles are indistinguishable?

Note to Example 13

(a) First choose the president, then the secretary, etc., giving $20 \times 19 \times 18 \times 17 = 116280$.
This is the number of permutations of length 4 in a group of size 20.
(b) $20 \times 19 \times 18 \times 17/2! = 58140$.
(c) $20 \times 19 \times 18 \times 17/(2!2!) = 29070$.
(d) The first could have been chosen in 20 ways, the second in 19, etc. But the final group of four could be elected in $4!$ orders, so the number of ways is $20 \times 19 \times 18 \times 17/4! = 4845$.

Multinomial and binomial coefficients

Definition 14. *Let $n_1, \ldots, n_r$ be integers in $0, 1, \ldots, n$, having total $n_1 + \cdots + n_r = n$. Then*

$$\binom{n}{n_1, n_2, \ldots, n_r} = \frac{n!}{n_1!\ n_2!\ \cdots\ n_r!},$$

is called the **multinomial coefficient**.
The most common case arises when $r = 2$:

$$\binom{n}{k} = \frac{n!}{k!(n-k)!} \quad \left(= C_n^k \text{ in certain books} \right)$$

is called the **binomial coefficient**.

Combinations: non ordered selection

Theorem 15. *The number of ways of choosing a set of r objects from a set of n distinct objects without repetition is*

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

Theorem 16. *The number of ways of distributing n distinct objects into r distinct groups of size $n_1, \ldots, n_r$, where $n_1 + \cdots + n_r = n$, is*

$$\frac{n!}{n_1!\, n_2!\, \cdots\, n_r!}.$$

Properties of binomial coefficients

Theorem 17. *If $n, m \in \{1, 2, 3 \ldots\}$ and $r \in \{0, \ldots, n\}$, then:*

$$\begin{aligned}
\binom{n}{r} &= \binom{n}{n-r}; \\
\binom{n+1}{r} &= \binom{n}{r-1} + \binom{n}{r}, \quad \textit{(Pascal's triangle)}; \\
\sum_{j=0}^{r} \binom{m}{j}\binom{n}{r-j} &= \binom{m+n}{r}, \quad \textit{(Vandermonde's formula)}; \\
(a+b)^n &= \sum_{r=0}^{n} \binom{n}{r} a^r b^{n-r}, \quad \textit{(Newton's binomial theorem)}; \\
(1-x)^{-n} &= \sum_{j=0}^{\infty} \binom{n+j-1}{j} x^j, \ |x| < 1, \quad \textit{(negative binomial series)}; \\
\lim_{n\to\infty} n^{-r}\binom{n}{r} &= \frac{1}{r!}, \quad r \in \mathbb{N}.
\end{aligned}$$

Note to Theorem 17

- ☐ The numbers of ways of choosing r objects from n is the same as the number of ways of choosing $n - r$ objects from n.
- ☐ To choose r objects from $n + 1$, we first designate one of the $n + 1$. Then if that object is in the sample, we must choose $r - 1$ from among the other n, and if not, we must choose r from the n, which gives the result.
- ☐ Suppose I have n blue hats and m red hats. Then the number of ways I can choose r hats from all my hats equals the number of ways I can choose j red hats and $r - j$ blue hats, summed over the possible choices of j.
- ☐ The binomial results are standard.
- ☐ For the last part, with r fixed, we have

$$n^{-r}\binom{n}{r} = \frac{n(n-1)\cdots(n-r+1)}{n^r r!} \to \frac{1}{r!}, \quad n \to \infty.$$

Partitions of integers

Theorem 18. *(a) The number of distinct vectors* $(n_1, \ldots, n_r)$ *of positive integers,* $n_1, \ldots, n_r > 0$*, satisfying* $n_1 + \cdots + n_r = n$*, is*

$$\binom{n-1}{r-1}.$$

(b) The number of distinct vectors $(n_1, \ldots, n_r)$ *of non-negative integers,* $n_1, \ldots, n_r \geq 0$*, satisfying* $n_1 + \cdots + n_r = n$*, is*

$$\binom{n+r-1}{n}.$$

Example 19. *How many different ways are there to put 6 identical balls in 3 boxes, in such a way that each box contains at least one ball?*

Example 20. *How many different ways are there to put 6 identical balls into 3 boxes?*

Note to Theorem 18

(a) Line up the n balls, and note that there are $n-1$ spaces between them. You must choose $r-1$ out of these $n-1$ spaces to place these separators, giving the stated formula.
(b) Line up the n balls and the $r-1$ separators. Any distinct configurations of these $n+r-1$ objects will correspond to a different partition, so the number of these partitions is the number of ways the balls and separators can be ordered, and this is the stated formula.

Note to Example 19

We have a total of $n = 6$ balls and $r = 3$ groups, each of which must have at least one member, so the number is

$$\binom{6-1}{3-1} = \frac{5!}{3!2!} = 10.$$

Note to Example 20

Now there is the possibility of empty boxes, so the total number is

$$\binom{6+3-1}{6} = \frac{8!}{6!2!} = 28.$$

Thus there are 18 ways to get at least one empty box.

Reminder: Some series

Theorem 21. *(a) A* geometric series *is of the form* $a, a\theta, a\theta^2, \ldots$*; we have*

$$\sum_{i=0}^{n} a\theta^i = \begin{cases} a\frac{1-\theta^{n+1}}{1-\theta}, & \theta \neq 1, \\ a(n+1), & \theta = 1. \end{cases}$$

If $|\theta| < 1$*, then* $\sum_{i=0}^{\infty} \theta^i = 1/(1-\theta)$*, and*

$$\sum_{i=0}^{\infty} \frac{i!}{r!(i-r)!}\theta^{i-r} = \frac{1}{(1-\theta)^{r+1}}, \quad r = 1, 2, \ldots.$$

The exponential series

$$\exp(x) = \sum_{n=0}^{\infty} \frac{x^n}{n!}$$

converges absolutely for all $x \in \mathbb{C}$.

Small lexicon

Mathematics	English	Français
$\Omega, A, B \ldots$	set	ensemble
$A \cup B$	union	union
$A \cap B$	intersection	intersection
A^c	complement of A (in Ω)	complémentaire de A (en Ω)
$A \setminus B$	difference	différence
$A \Delta B$	symmetric difference	différence symétrique
$A \times B$	Cartesian product	produit cartésien
$\lvert A \rvert$	cardinality	cardinal
	$\{A_j\}_{j=1}^n$ pairwise disjoint	$\{A_j\}_{j=1}^n$ disjoint deux à deux
	partition	partition
	permutation	permutation
	combination	combinaison
$\binom{n}{r}$	binomial coefficient	coefficient binomial (C_n^r)
$\binom{n}{n_1,\ldots,n_r}$	multinomial coefficient	coefficient multinomial
	indistinguishable	indifférentiable
	colour-blind	daltonien (ienne)

2 Probability

Small probabilistic lexicon

Mathematics	English	Français
	one fair die (several fair dice)	dé juste/équilibré (plusieurs dés justes/équilibrés)
	random experiment	expérience aléatoire
Ω	sample space	ensemble fondamental
ω	outcome, elementary event	épreuve, événement élémentaire
$A, B, \ldots$	event	événement
$\mathcal{F}$	event space	espace des événements
	sigma-algebra	tribu
P	probability distribution/probability function	loi de probabilité
$(\Omega, \mathcal{F}, \mathrm{P})$	probability space	espace de probabilité
	inclusion-exclusion formula	formule d'inclusion-exclusion
$\mathrm{P}(A \mid B)$	probability of A given B	probabilité de A sachant B
	independence	indépendance
	(mutually) independent events	événements (mutuellement) indépendants
	pairwise independent events	événements indépendants deux à deux
	conditionally independent events	événements conditionellement indépendants

2.1 Probability Spaces

The Card Players

Paul Cézanne, 1894–95, Musée d'Orsay, Paris

Motivation: Game of dice

We throw two fair dice, one red and one green.

- ☐ (a) What is the set of possible results?
- ☐ (b) Which results give a total of 6?
- ☐ (c) Which results give a total of 12?
- ☐ (d) Which results give an odd total?
- ☐ (e) What are the probabilities of the events (b), (c), (d)?

Calculation of probabilities

- ☐ We can try to calculate the probabilities of events such as (b), (c), (d) by throwing the dice numerous times and letting

$$\text{probability of an event} = \frac{\#\text{ of times event takes place}}{\#\text{ experiments carried out}}.$$

  This is an empirical rather than a mathematical answer, to be reached only after a lot of work (how many times should we roll the dice?), and it will yield different answers each time—**not satisfactory!**

- ☐ For simple examples, we often use symmetry to calculate probabilities. This isn't possible for more complicated cases—we construct mathematical models, based on notions of
  - **random experiments**
  - **probability spaces**.

Random experiment

Definition 22. *A* **random experiment** *is an 'experiment' whose result is (or can be defined as) random.*

Example 23. *I toss a coin.*

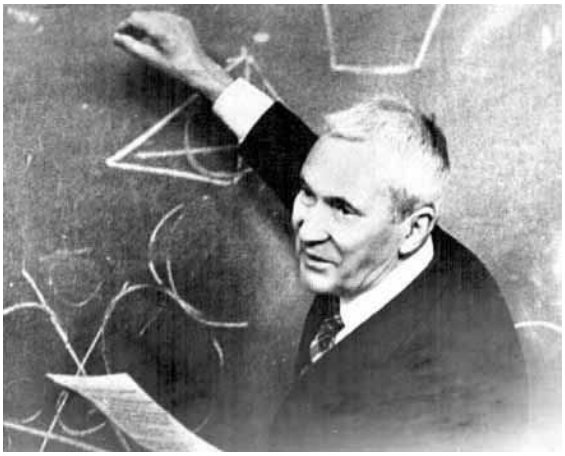
Example 24. *I roll 2 fair dice, one red and one green.*

Example 25. *The number of emails I receive today.*

Example 26. *The waiting time until the end of this lecture.*

Example 27. *The weather here tomorrow at midday.*

Andrey Nikolaevich Kolmogorov (1903–1987)



Grundbegriffe der Wahrscheinlichkeitsrechnung (1933)

(Source: http://en.academic.ru/dic.nsf/enwiki/54484)

Probability space $(\Omega, \mathcal{F}, \mathrm{P})$

A random experiment is modelled by a **probability space**.

Definition 28. *A* **probability space** $(\Omega, \mathcal{F}, \mathrm{P})$ *is a mathematical object associated with a random experiment, comprising:*

- ☐ *a set* Ω*, the* **sample space (universe)***, which contains all the possible* **results (outcomes, elementary events)** ω *of the experiment;*
- ☐ *a collection* $\mathcal{F}$ *of subsets of* Ω*. These subsets are called* **events***, and* $\mathcal{F}$ *is called the* **event space***;*
- ☐ *a function* $\mathrm{P} : \mathcal{F} \mapsto [0, 1]$ *called a* **probability distribution***, which associates a probability* $\mathrm{P}(A) \in [0, 1]$ *to each* $A \in \mathcal{F}$*.*

Sample space

- ☐ The sample space Ω is the space composed of elements representing all the possible results of a random experiment. Each element $\omega \in \Omega$ is associated with a different result.
- ☐ Ω is analogous to the universal set. It can be finite, countable or uncountable.
- ☐ Ω is nonempty. (If $\Omega = \emptyset$, then nothing interesting can happen.)

Example 29. *Describe the sample spaces for Examples 23–27.*

For simple examples with finite Ω, we often choose Ω so that each $\omega \in \Omega$ is equiprobable:

$$\mathrm{P}(\omega) = \frac{1}{|\Omega|}, \quad \text{for every } \omega \in \Omega.$$

Then $\mathrm{P}(A) = |A|/|\Omega|$, for every $A \subset \Omega$.

Note to Example 29

Example 23: Here we can write $\Omega = \{\omega_1, \omega_2\}$, where ω_1 and ω_2 represent Tail and Head respectively.
Example 24: $\Omega = \{\omega_1, \ldots, \omega_{36}\}$, representing all 36 different possibilities.
Example 25: $\Omega = \{\omega_j : j = 0, 1, \ldots, \}$, representing any non-negative number.
Example 26: $\Omega = \{\omega : \omega \in [0, 45]$ minutes$\}$, an uncountable set.
Example 27: Ω? We have to decide what we count as weather outcomes, so this is not so easy.
In general discussion we use ω as an element of Ω, but in examples it is usually easier to write H or T or (r, g) or similar.

Event space

$\mathcal{F}$ is a set of subsets of Ω which represents the events of interest.

Example 30 (Example 24, continued). *Give the events*

A	*the red die shows a 4*,
B	*the total is odd*,
C	*the green die shows a 2*,
$A \cap B$	*the red die shows a 4 and the total is odd.*

Calculate their probabilities.

Note to Example 30

First we set up the probability space Ω. If we write $(2, 4)$ to mean that the red shows 2 and the green shows 4, we have

$$\Omega = \{(r, g) : r, g = 1, \ldots, 6\},$$

giving

$$\begin{aligned} A &= \{(4, g), g = 1, \ldots, 6\}, \\ B &= \{(1,2), (1,4), (1,6), (2,1), (2,3), (2,5), \ldots, (6,1), (6,3), (6,5)\}, \\ C &= \{(r, 2), r = 1, \ldots, 6\}, \\ A \cap B &= \{(4,1), (4,3), (4,5)\}. \end{aligned}$$

By symmetry if the two dice are fair, then $|\Omega| = 36$, $|A| = |C| = 6$, $|B| = 18$, and $|A \cap B| = 3$, so the probabilities are

$$\mathrm{P}(A) = \mathrm{P}(C) = 6/36 = 1/6, \quad \mathrm{P}(B) = 18/36 = 1/2, \quad \mathrm{P}(A \cap B) = 3/36 = 1/12.$$

Event space $\mathcal{F}$, II

Definition 31. *An* **event space** $\mathcal{F}$ *is a set of the subsets of* Ω *such that:*

$(\mathcal{F}1)$ $\mathcal{F}$ *is nonempty;*

$(\mathcal{F}2)$ *if* $A \in \mathcal{F}$ *then* $A^c \in \mathcal{F}$*;*

$(\mathcal{F}3)$ *if* $\{A_i\}_{i=1}^{\infty}$ *are all elements of* $\mathcal{F}$*, then* $\bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$*.*

$\mathcal{F}$ *is also called a* **sigma-algebra** *(en français,* **une tribu***).*

Let $A, B, C, \{A_i\}_{i=1}^{\infty}$ be elements of $\mathcal{F}$. Then the preceding axioms imply that

(a) $\bigcup_{i=1}^{n} A_i \in \mathcal{F}$,

(b) $\Omega \in \mathcal{F}, \quad \emptyset \in \mathcal{F}$,

(c) $A \cap B \in \mathcal{F}, \quad A \setminus B \in \mathcal{F}, \quad A \,\Delta\, B \in \mathcal{F}$,

(d) $\bigcap_{i=1}^{n} A_i \in \mathcal{F}$.

Use of these axioms

To prove (a)–(d), we argue as follows:

(a) Take $A_{n+1} = A_{n+2} = \cdots = A_n$, and apply $(\mathcal{F}3)$.

(b) If $\mathcal{F}$ is non-empty, then it has an element A, and by $(\mathcal{F}2)$ $A^c \in \mathcal{F}$, so $A \cup A^c = \Omega \in \mathcal{F}$. Also, $\Omega^c = \emptyset \in \mathcal{F}$.

(c) Note that $A \cap B = (A^c \cup B^c)^c$, and sets operated on by union and complement remain in $\mathcal{F}$. Likewise for the differences.

(d) We write $\bigcap_{i=1}^{n} A_i = ((\bigcap_{i=1}^{n} A_i)^c)^c = (\bigcup_{i=1}^{n} A_i^c)^c \in \mathcal{F}$.

Event space $\mathcal{F}$, III

- ☐ If Ω is countable, we often take $\mathcal{F}$ to be the set of all the subsets of Ω. This is the biggest (and richest) event space for Ω.
- ☐ We can define different event spaces for the same sample space.

Example 32. *Give the event space for Example 23.*

Example 33. *I roll two fair dice, one red and one green.*
(a) What is my event space $\mathcal{F}_1$*?*
(b) I only tell my friend the total. What is his event space $\mathcal{F}_2$*?*
(c) My friend looks at the dice himself, but he is colour-blind. What then is his event space $\mathcal{F}_3$*?*

Note to Example 32

We can write $\Omega = \{H, T\}$, and then have two choices:

$$\mathcal{F}_1 = \{\{H, T\}, \emptyset\}, \quad \mathcal{F}_2 = \{\{H, T\}, \emptyset, \{H\}, \{T\}\}.$$

Either of these satisfies the axioms (check this) and hence is a valid event space. Only the second, however, is interesting. In the first the only non-null event is $\{H, T\}$, which corresponds to 'the experiment was performed and a head or a tail was observed'.

Note to Example 33

(a) Since we see an outcome of the form (r, g), we can reply to any question about the outcomes; thus we take $\mathcal{F}_1$ to be the set of all possible subsets of $\Omega\{(r, g) : r, g = 1, \ldots, 6\}$. The ordered pair (r, g) corresponds to the event $A_{r,g} = \{(r, g)\}$ ('the experiment was performed and the outcome was (r, g)'), and the 2^{36} distinct elements $B_j$of $\mathcal{F}_1$ can be constructed by taking all possible unions and intersections of the $A_{r,g}$. (Note that the intersection of any two or more disjoint events here will give $\emptyset$, and the union of all of them gives Ω.) This means that $\mathcal{F}_1$ is the power set of $\{A_{r,g} : r, g = 1, \ldots, 6\}$, and $|\mathcal{F}_1| = 2^{36}$.
(b) If I tell him only that the 'total is t' for $t = 2, \ldots, 12$, then he can reply to any question about the total, but nothing else. So his event space $\mathcal{F}_2$ is based on the events $T_2, \ldots, T_{12}$, where

$$T_2 = \{(1, 1)\}, \quad T_3 = \{(1, 2), (2, 1)\}, \quad T_4 = \{(1, 3), (2, 2), (3, 1)\}, \ldots, T_{12} = \{(6, 6)\}.$$

His event space therefore comprises all the possible unions and intersections of these 11 events, and therefore $|\mathcal{F}_2| = 2^{11}$.
(c) Since he is colour-blind, he cannot tell the difference between $(1, 2)$ and $(2, 1)$, etc.. Thus $\mathcal{F}_3$ is made up of all possible unions and intersections of the sets

$$\{(1, 1)\}, \{(2, 2)\}, \ldots, \{(6, 6)\}, \{(1, 2), (2, 1)\}, \{(1, 3), (3, 1)\}, \ldots, \{(5, 6), (6, 5)\}.$$

There are 6 + 15 such sets, so $|\mathcal{F}_3| = 2^{21}$, and obviously $\mathcal{F}_2 \subset \mathcal{F}_3 \subset \mathcal{F}_1$.
In cases (b) and (c) the event spaces have less information than in (a): they represent a coarsening of $\mathcal{F}_1$, so that fewer questions can be answered.

Event space $\mathcal{F}$, III

- ☐ Usually the event space is clear from the context, but it is important to write out Ω and $\mathcal{F}$ explicitly, in order to avoid confusion.
- ☐ This can also be useful when so-called 'paradoxes' appear (generally due to an unclear or erroneous mathematical formulation of the problem).
- ☐ **It is essential to give Ω and $\mathcal{F}$ when doing exercices, tests and exams.**

Examples

Example 34. *A woman planning her future family considers the following possibilities (we suppose that the chances of having a boy or a girl are the same each time) :*

(a) have three children;

(b) keep giving birth until the first girl is born or until three children are born, stop when one of the two situations arises.

(c) keep giving birth until there are one of each gender or until there are three children, stop when one of the two situations arises.

Let B_i be the event 'i boys are born', A the event 'there are more girls than boys'. Calculate $\mathrm{P}(B_1)$ and $\mathrm{P}(A)$ for (a)–(c).

(In fact, the ratio of boys/girls at birth is $\sim 105/100$.)

Example 35 (Birthdays). *n people are in a room. What is the probability that they all have a different birthday?*

Note to Example 34

We learn from this example that:

- ☐ changing the protocol or stopping rule can change the observable outcomes and hence the sample space;
- ☐ the outcomes need not have the same probabilities under different stopping rules;
- ☐ in some cases it is possible to compute probabilities for outcomes in one sample space by comparing it to another sample space.

(a) We can write the sample space under this stopping rule as

$$\Omega_1 = \{BBB, BBG, BGB, BGG, GBB, GBG, GGB, GGG\},$$

where B denotes a boy, G denotes a girl and the ordering is important. These events all have probability 1/8, by symmetry. Then $B_1 = \{BGG, GBG, GGB\}$ and $A = B_1 \cup \{GGG\}$ have probabilities 3/8 and 1/2 respectively. The latter is obvious also by symmetry.
(b) Under this stopping rule the sample space is

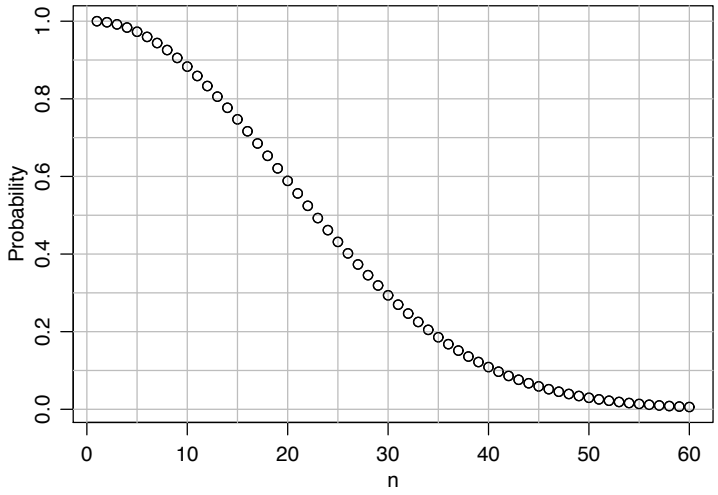
$$\Omega_2 = \{BBB, BBG, BG, G\},$$

but these are not equi-probable; for example $B_1 = \{BG\}$ here corresponds to the event $\{BGG, BGB\}$ in Ω_1 and so has probability 1/4, and $A = \{G\}$ here corresponds to the event $\{GBB, GBG, GGB, GGG\}$ in Ω_1, and so has probability 1/2.
(c) Under this stopping rule the sample space is

$$\Omega_3 = \{BBB, GGG, BBG, GGB, GB, BG\},$$

noting that BG here corresponds to $\{BGG, BGB\}$ in Ω_1, and likewise GB here corresponds to $\{GBG, GBB\}$ in Ω_1. In this case the event $B_1 = \{GB, BG, GGB\}$ in Ω_3 corresponds to $\{GBB, GBG, BGG, BGB, GGB\}$ in Ω_1 and hence has probability 5/8, and in Ω_3 the event $A = \{GGG, GGB\}$ has probability 1/4.

Birthdays



Note to Example 35

The sample space can be written $\Omega = \{1, \ldots, 365\}^n$, and each of these possibilities has probability 365^{-n}. We seek the probability of the event

$$A = \{(i_1, \ldots, i_n) : i_1 \neq i_2 \neq \cdots \neq i_n\}.$$

There are $365 \times 364 \times \cdots (365 - n + 1) = 365!/(365 - n)!$ ways this can happen, so the overall probability is $365!/\{(365 - n)!365^n\}$, which is shown in the graph.

Galileo Galilei (1564–1642)

(Source: Wikipedia, portrait by Ottavio Leoni)

Il Saggiatore, 1623

(Source: Wikipedia)

Il Saggiatore, 1623

La filosofia è scritta in questo grandissimo libro che continuamente ci sta aperto innanzi a gli occhi (io dico l'universo), ma non si può intendere se prima non s'impara a intender la lingua, e conoscer i caratteri, ne' quali è scritto. Egli è scritto in lingua matematica, e i caratteri son triangoli, cerchi, ed altre figure geometriche, senza i quali mezi è impossibile a intenderne umanamente parola; senza questi è un aggirarsi vanamente per un oscuro laberinto.

The book of the Universe cannot be understood unless one first learns to comprehend the language and to understand the alphabet in which it is composed. It is written in the language of mathematics, and its characters are triangles, circles, and other geometric figures, without which it is humanly impossible to understand a single word of it; without these, one wanders about in a dark labyrinth.

Three dice problem

Three fair dice are rolled. Let T_i be the event 'the total is i', for $i = 3, \ldots, 18$. Which is most likely, T_9 or T_{10}?
T_9 occurs if the dice have the following outcomes

$$9 = 6+2+1 = 5+3+1 = 5+2+2 = 4+4+1 = 4+3+2 = 3+3+3.$$

T_{10} occurs if the dice have the following outcomes

$$10 = 6+3+1 = 6+2+2 = 5+4+1 = 5+3+2 = 4+4+2 = 4+3+3.$$

Thus they are equiprobable.

True or false?

Note to the three dice problem

We take $\Omega = \{(r,s,t) : r,s,t = 1,\ldots,6\}$, for a total of $6^3 = 216$ equiprobable outcomes.
Now T_9 occurs if we have $r+s+t=9$, but the outcomes listed are not equiprobable, because $\{1,2,6\}$ and $\{1,3,5\}$ can each arise in 3! ways, while $\{2,2,5\}$ can arise in just 3 ways. Adding up the numbers of outcomes gives $|T_9| = 25$, $|T_{10}| = 27$, so the latter is more probable.

Probability function P

Definition 36. *A* **probability distribution** P *assigns a probability to each element of the event space* $\mathcal{F}$*, with the following properties:*

$(P1)$ *if* $A \in \mathcal{F}$*, then* $0 \le \mathrm{P}(A) \le 1$*;*

$(P2)$ $\mathrm{P}(\Omega) = 1$*;*

$(P3)$ *if* $\{A_i\}_{i=1}^{\infty}$ *are pairwise disjoint (i.e.,* $A_i \cap A_j = \emptyset$*,* $i \neq j$*), then*

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

Properties of P

Theorem 37. *Let* $A, B, \{A_i\}_{i=1}^{\infty}$ *be events of the probability space* $(\Omega, \mathcal{F}, \mathrm{P})$*. Then*

(a) $\mathrm{P}(\emptyset) = 0$*;*

(b) $\mathrm{P}(A^c) = 1 - \mathrm{P}(A)$*;*

(c) $\mathrm{P}(A \cup B) = \mathrm{P}(A) + \mathrm{P}(B) - \mathrm{P}(A \cap B)$*. If* $A \cap B = \emptyset$*, then* $\mathrm{P}(A \cup B) = \mathrm{P}(A) + \mathrm{P}(B)$*;*

(d) if $A \subset B$*, then* $\mathrm{P}(A) \le \mathrm{P}(B)$*, and* $\mathrm{P}(B \setminus A) = \mathrm{P}(B) - \mathrm{P}(A)$*;*

(e) $\mathrm{P}\left(\bigcup_{i=1}^{\infty} A_i\right) \le \sum_{i=1}^{\infty} \mathrm{P}(A_i)$ *(Boole's inequality);*

(f) if $A_1 \subset A_2 \subset \cdots$*, then* $\lim_{n\to\infty} \mathrm{P}(A_n) = \mathrm{P}\left(\bigcup_{i=1}^{\infty} A_i\right)$*;*

(g) if $A_1 \supset A_2 \supset \cdots$*, then* $\lim_{n\to\infty} \mathrm{P}(A_n) = \mathrm{P}\left(\bigcap_{i=1}^{\infty} A_i\right)$*.*

Note to Theorem 37

(a) Since $\emptyset \cap A = \emptyset$ for any $A \in \mathcal{F}$, we can apply $(P3)$ to a finite number of sets, just by adding an infinite number of $\emptyset$s. In particular, $\Omega = \Omega \cup \emptyset \cup \emptyset \cup \cdots$, and these are pairwise disjoint, so

$$1 = \mathrm{P}(\Omega) = \mathrm{P}(\Omega) + \mathrm{P}(\emptyset) + \mathrm{P}(\emptyset) + \cdots,$$

so since $\mathrm{P}(\emptyset) \geq 0$, we must have $\mathrm{P}(\emptyset) = 0$.
Further, if we have a finite collection $A_1, \ldots, A_n$ of pairwise disjoint events, then we can complement them with $A_{n+1} = A_{n+2} = \cdots = \emptyset$, which gives $A_i \cap A_j = \emptyset$ for any $i \neq j$ and all $i, j \in \mathbb{N}$, and then

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

so (P3) also holds for any finite number of disjoint events.
(b) Follows from the finite version of (P3) (in (a)) by setting $A_1 = A$, $A_2 = A^c$, and noting that $1 = \mathrm{P}(\Omega) = \mathrm{P}(A \cup A^c) = \mathrm{P}(A) + \mathrm{P}(A^c)$.
(c) Follows from $(P3)$ by writing $A \cup B = (A \cap B^c) \cup (A \cap B) \cup (A^c \cap B)$, which are pairwise disjoint, and noting that this gives

$$\mathrm{P}(A) = \mathrm{P}(A \cap B) + \mathrm{P}(A \cap B^c), \quad \mathrm{P}(B) = \mathrm{P}(A \cap B) + \mathrm{P}(A^c \cap B),$$

and then

$$\mathrm{P}(A \cup B) = \mathrm{P}(A \cap B^c) + \mathrm{P}(A \cap B) + \mathrm{P}(A^c \cap B) = \{\mathrm{P}(A) - \mathrm{P}(A \cap B)\} + \mathrm{P}(A \cap B) + \{\mathrm{P}(B) - \mathrm{P}(A \cap B)\},$$

giving the required result.
(d) Follows by writing $B = A \cup (B \cap A^c)$, and noting that $B \backslash A = B \cap A^c$.
(e) Iteration: for $k \in \mathbb{N}$, we write $B_{k-1} = \bigcup_{i=k}^{\infty} A_i = A_k \cup \bigcup_{i=k+1}^{\infty} A_i = A_k \cup B_k$, say, and note that (c) gives $\mathrm{P}(B_{k-1}) = \mathrm{P}(A_k \cup B_k) \leq \mathrm{P}(A_k) + \mathrm{P}(B_k)$, resulting in

$$\mathrm{P}\left(\bigcup_{i=1}^{\infty} A_i\right) \leq \mathrm{P}(A_1) + \mathrm{P}(B_1) \leq \sum_{i=1}^{k} \mathrm{P}(A_i) + \mathrm{P}(B_k) \leq \sum_{i=1}^{\infty} \mathrm{P}(A_i)$$

as required.
(f) Now $A_i \subset A_{i+1}$ for every i, so $(A_{i+1} \backslash A_i) \cap (A_{j+1} \backslash A_j) = \emptyset$ when $i \neq j$ (draw picture), and $A_n = \bigcup_{i=1}^{n} (A_i \backslash A_{i-1})$, where we've set $A_0 = \emptyset$. Note that $\mathrm{P}(A_{i+1} \backslash A_i) = \mathrm{P}(A_{i+1}) - \mathrm{P}(A_i)$. Thus by $(P3)$ we have

$$\begin{aligned}
\mathrm{P}(\bigcup_{i=1}^{\infty} A_i) &= \mathrm{P}(A_1) + \sum_{i=2}^{\infty} \mathrm{P}(A_i \backslash A_{i-1}) \\
&= \mathrm{P}(A_1) + \sum_{i=2}^{\infty} \{\mathrm{P}(A_i) - \mathrm{P}(A_{i-1})\}, \\
&= \lim_{n \to \infty} \left[\mathrm{P}(A_1) + \sum_{i=2}^{n} \{\mathrm{P}(A_i) - \mathrm{P}(A_{i-1})\}\right], \\
&= \lim_{n \to \infty} \mathrm{P}(A_n).
\end{aligned}$$

(g) Like (f).

Continuity of P

Reminder: A function f is continuous at x if for every sequence $\{x_n\}$ such that

$$\lim_{n\to\infty} x_n = x, \text{ we have } \lim_{n\to\infty} f(x_n) = f(x).$$

Parts (f) and (g) of Theorem 37 can be extended to show that for all sequences of sets for which

$$\lim_{n\to\infty} A_n = A, \text{ we have } \lim_{n\to\infty} \mathrm{P}(A_n) = \mathrm{P}(A).$$

Hence P is called a **continuous set function**.

Inclusion-exclusion formulae

If $A_1, \ldots, A_n$ are events of $(\Omega, \mathcal{F}, P)$, then

$$\begin{aligned}
\mathrm{P}(A_1 \cup A_2) &= \mathrm{P}(A_1) + \mathrm{P}(A_2) - \mathrm{P}(A_1 \cap A_2) \\
\mathrm{P}(A_1 \cup A_2 \cup A_3) &= \mathrm{P}(A_1) + \mathrm{P}(A_2) + \mathrm{P}(A_3) \\
&\quad -\mathrm{P}(A_1 \cap A_2) - \mathrm{P}(A_1 \cap A_3) - \mathrm{P}(A_2 \cap A_3) \\
&\quad +\mathrm{P}(A_1 \cap A_2 \cap A_3) \\
&\vdots \\
\mathrm{P}\left(\bigcup_{i=1}^{n} A_i\right) &= \sum_{r=1}^{n} (-1)^{r+1} \sum_{1 \le i_1 < \cdots < i_r \le n} \mathrm{P}(A_{i_1} \cap \cdots \cap A_{i_r}).
\end{aligned}$$

The number of terms in the general formula is

$$\binom{n}{1} + \binom{n}{2} + \binom{n}{3} + \cdots + \binom{n}{n-1} + \binom{n}{n} = 2^n - 1.$$

Note to inclusion-exclusion formulae

We saw the first equality as part (c) of Theorem 37.
For the second, write $B = A_2 \cup A_3$, and note that

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

which is what we want, since the last term is $\mathrm{P}(A_1 \cap A_2 \cap A_3)$. The general formula follows by iteration of this argument.

Note to inclusion-exclusion formulae: II

For example, with $n = 4$, we have

$$\begin{aligned}\mathrm{P}(A_1 \cup A_2 \cup A_3 \cup A_4) &= \mathrm{P}(A_1) + \mathrm{P}(A_2) + \mathrm{P}(A_3) + \mathrm{P}(A_4) \\ &\quad - \{\mathrm{P}(A_1 \cap A_2) + \mathrm{P}(A_1 \cap A_3) + \mathrm{P}(A_1 \cap A_4) \\ &\qquad + \mathrm{P}(A_2 \cap A_3) + \mathrm{P}(A_2 \cap A_4) + \mathrm{P}(A_3 \cap A_4)\} \\ &\quad + \{\mathrm{P}(A_1 \cap A_2 \cap A_3) + \mathrm{P}(A_1 \cap A_2 \cap A_4) \\ &\qquad + \mathrm{P}\{(A_1 \cap A_3 \cap A_4) + \mathrm{P}(A_2 \cap A_3 \cap A_4)\} \\ &\quad - \mathrm{P}(A_1 \cap A_2 \cap A_3 \cap A_4)\end{aligned}$$

where there are 4, 6, 4, 1 terms in the terms having 1, 2, 3, and 4 events, respectively.

Example 38. *What is the probability of getting at least one 6 when I roll three fair dice?*

Example 39. *An urn contains 1000 lottery tickets numbered from 1 to 1000. One ticket is drawn at random. Before the draw a fairground showman offers to pay \$3 to whoever will give him \$2, if the number on the ticket is divisible by 2, 3, or 5. Would you give him your \$2 before the draw? (You lose your money if the ticket is not divisible by 2, 3, or 5.)*

Note to Example 38

Let A_i be the event there is a 6 on die i; we want $\mathrm{P}(A_1 \cup A_2 \cup A_3)$. Now by symmetry $\mathrm{P}(A_i) = 1/6$, $\mathrm{P}(A_i \cap A_j) = 1/36$, and $\mathrm{P}(A_1 \cap A_2 \cap A_3) = 1/216$. Therefore the second inclusion-exclusion formula gives

$$\mathrm{P}(A_1 \cup A_2 \cup A_3) = \frac{3}{6} - \frac{3}{36} + \frac{1}{216} = \frac{91}{216}.$$

Note to Example 39

Here we can write $\Omega = \{1, \ldots, 1000\}$, and let D_i be the event that the number is divisible by i. We want

$$\begin{aligned}\mathrm{P}(D_2 \cup D_3 \cup D_5) &= \mathrm{P}(D_2) + \mathrm{P}(D_3) + \mathrm{P}(D_5) - \mathrm{P}(D_2 \cap D_3) - \mathrm{P}(D_2 \cap D_5) - \mathrm{P}(D_3 \cap D_5) \\ &\quad + \mathrm{P}(D_2 \cap D_3 \cap D_5) \\ &= \mathrm{P}(D_2) + \mathrm{P}(D_3) + \mathrm{P}(D_5) - \mathrm{P}(D_6) - \mathrm{P}(D_{10}) - \mathrm{P}(D_{15}) + \mathrm{P}(D_{30}) \\ &= \frac{500 + 333 + 200 - 166 - 100 - 66 + 33}{1000} = \frac{367}{500} \doteq 0.734.\end{aligned}$$

So with probability 0.734 you gain 3-2=1 and with probability 0.266 you lose 2: the average gain is $1 \times 0.7334 + (-2) \times 0.266 = 0.201$: you will win on average if you play. The 'return on investment' is $0.201/2 \approx 0.1$, or 10%, which is excellent compared to a bank.

2.2 Conditional Probability

Conditional probability

Definition 40. *Let A, B be events of the probability space $(\Omega, \mathcal{F}, \mathrm{P})$, such that $\mathrm{P}(B) > 0$. Then the* **conditional probability of A given B** *is*

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

If $\mathrm{P}(B) = 0$, we adopt the convention $\mathrm{P}(A \cap B) = \mathrm{P}(A \mid B)\mathrm{P}(B)$, so both sides are equal to zero. Thus

$$\mathrm{P}(A) = \mathrm{P}(A \cap B) + \mathrm{P}(A \cap B^c) = \mathrm{P}(A \mid B)\mathrm{P}(B) + \mathrm{P}(A \mid B^c)\mathrm{P}(B^c)$$

even if $\mathrm{P}(B) = 0$ or $\mathrm{P}(B^c) = 0$.

Example 41. *We roll two fair dice, one red and one green. Let A and B be the events 'the total exceeds 8', and 'we get 6 on the red die'. If we know that B has occurred, how does $\mathrm{P}(A)$ change?*

Note to Example 41

We first draw a square containing pairs $\{(r, g) : r, g = 1, \ldots, 6\}$ to display the totals of the two dice. By inspection, and since all the individual outcomes have probability $1/36$, we have $\mathrm{P}(A) = (1 + 2 + 3 + 4)/36 = 5/18$, $\mathrm{P}(B) = 6/36 = 1/6$, and thus by definition the conditional probability is $\mathrm{P}(A \mid B) = \mathrm{P}(A \cap B)/\mathrm{P}(B) = (4/36)/(1/6) = 2/3$.
Thus including the information that B has occurred changes the probability of A: conditioning can be interpreted as inserting information into the calculation of probabilities, resulting in a new probability space, as we see in the next theorem.

Conditional probability distributions

Theorem 42. *Let $(\Omega, \mathcal{F}, \mathrm{P})$ be a probability space, and let $B \in \mathcal{F}$ such that $\mathrm{P}(B) > 0$ and $Q(A) = \mathrm{P}(A \mid B)$. Then $(\Omega, \mathcal{F}, Q)$ is a probability space. In particular,*

- ☐ *if $A \in \mathcal{F}$, then $0 \leq Q(A) \leq 1$;*
- ☐ *$Q(\Omega) = 1$;*
- ☐ *if $\{A_i\}_{i=1}^{\infty}$ are pairwise disjoint, then*

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

Thus conditioning on different events allows us to construct lots of different probability distributions, starting with a single probability distribution.

Note to Theorem 42

We just need to check the axioms. If $A \in \mathcal{F}$, then

$$Q(A) = \mathrm{P}(A \mid B) = \mathrm{P}(A \cap B)/\mathrm{P}(B) \in [0, 1],$$

because $A \cap B \subset B$ and therefore $\mathrm{P}(A \cap B) \leq \mathrm{P}(B)$. Likewise

$$Q(\Omega) = \mathrm{P}(\Omega \cap B)/\mathrm{P}(B) = \mathrm{P}(B)/\mathrm{P}(B) = 1,$$

and finally,

$$Q\left(\bigcup_{i=1}^{\infty} A_i\right) = \frac{\mathrm{P}(\bigcup_{i=1}^{\infty} A_i \cap B)}{\mathrm{P}(B)} = \frac{\mathrm{P}\{\bigcup_{i=1}^{\infty}(A_i \cap B)\}}{\mathrm{P}(B)} = \frac{\sum_{i=1}^{\infty} \mathrm{P}(A_i \cap B)}{\mathrm{P}(B)} = \sum_{i=1}^{\infty} Q(A_i),$$

using the properties of $\mathrm{P}(\cdot)$ and the fact that if $A_1, A_2, \ldots$ are pairwise disjoint, then so too are the $A_1 \cap B, A_2 \cap B, \ldots$.

Thomas Bayes (1702–1761)

Essay towards solving a problem in the doctrine of chances. (1763/4) Philosophical Transactions of the Royal Society of London.
(Source: Wikipedia)

Bayes' theorem

Theorem 43 (Law of total probability). *Let $\{B_i\}_{i=1}^{\infty}$ be pairwise disjoint events (i.e. $B_i \cap B_j = \emptyset$, $i \neq j$) of the probability space $(\Omega, \mathcal{F}, \mathrm{P})$, and let A be an event satisfying $A \subset \bigcup_{i=1}^{\infty} B_i$. Then*

$$\mathrm{P}(A) = \sum_{i=1}^{\infty} \mathrm{P}(A \cap B_i) = \sum_{i=1}^{\infty} \mathrm{P}(A \mid B_i)\mathrm{P}(B_i).$$

Theorem 44 (Bayes). *Suppose that the conditions above are satisfied, and that $\mathrm{P}(A) > 0$. Then*

$$\mathrm{P}(B_j \mid A) = \frac{\mathrm{P}(A \mid B_j)\mathrm{P}(B_j)}{\sum_{i=1}^{\infty} \mathrm{P}(A \mid B_i)\mathrm{P}(B_i)}, \quad j \in \mathbb{N}.$$

These results are also true if the number of B_i is finite, and if the B_i partition Ω.

Note to Theorems 43 and 44

Since the B_i are disjoint, then so are their subsets $A \cap B_i$. Thus

$$\mathrm{P}(A) = \mathrm{P}\left\{A \cap \bigcup_{i=1}^{\infty} B_i\right\} = \mathrm{P}\left\{\bigcup_{i=1}^{\infty}(A \cap B_i)\right\} = \sum_{i=1}^{\infty} \mathrm{P}(A \cap B_i) = \sum_{i=1}^{\infty} \mathrm{P}(A \mid B_i)\mathrm{P}(B_i).$$

For Bayes' theorem, we note that

$$\mathrm{P}(B_j \mid A) = \frac{\mathrm{P}(A \cap B_j)}{\mathrm{P}(A)} = \frac{\mathrm{P}(A \mid B_j)\mathrm{P}(B_j)}{\mathrm{P}(A)} = \frac{\mathrm{P}(A \mid B_j)\mathrm{P}(B_j)}{\sum_{i=1}^{\infty} \mathrm{P}(A \mid B_i)\mathrm{P}(B_i)}$$

using the theorem of total probability, Theorem 43.

Example

Example 45. *You suspect that the man in front of you at the security check at the airport is a terrorist. Knowing that one person out of 10^6 is a terrorist, and that a terrorist is detected by the security check with a probability of 0.9999, but that the alarm goes off when an ordinary person goes through with a probability of 10^{-5}, what is the probability that he is a terrorist, given that the alarm goes off when he passes through security?*

Note to Example 45

Let A and T respectively denote the events 'the alarm sounds' and 'he is a terrorist'. Then we seek

$$\mathrm{P}(T \mid A) = \frac{\mathrm{P}(A \mid T)\mathrm{P}(T)}{\mathrm{P}(A \mid T)\mathrm{P}(T) + \mathrm{P}(A \mid T^c)\mathrm{P}(T^c)} = \frac{0.9999 \times 10^{-6}}{0.9999 \times 10^{-6} + 10^{-5} \times (1 - 10^{-6})} \doteq 0.0909.$$

Thus the odds are around 10:1 that he is not a terrorist.
We would have to decrease the false alarm probability of 10^{-5} to 10^{-6} to have probability 0.5 that he is a terrorist.

Multiple conditioning

Theorem 46 ('Prediction decomposition'). *Let $A_1, \ldots, A_n$ be events in a probability space. Then*

$$\begin{aligned} \mathrm{P}(A_1 \cap A_2) &= \mathrm{P}(A_2 \mid A_1)\mathrm{P}(A_1), \\ \mathrm{P}(A_1 \cap A_2 \cap A_3) &= \mathrm{P}(A_3 \mid A_1 \cap A_2)\mathrm{P}(A_2 \mid A_1)\mathrm{P}(A_1), \\ &\vdots \\ \mathrm{P}(A_1 \cap \cdots \cap A_n) &= \prod_{i=2}^{n} \mathrm{P}(A_i \mid A_1 \cap \cdots \cap A_{i-1}) \times \mathrm{P}(A_1). \end{aligned}$$

Note to Theorem 46

Just iterate. For example, if we let $B = A_1 \cap A_2$ and note that $\mathrm{P}(B) = \mathrm{P}(A_2 \mid A_1)\mathrm{P}(A_1)$ by the definition of conditional probability, then

$$\mathrm{P}(A_1 \cap A_2 \cap A_3) = \mathrm{P}(A_3 \cap B) = \mathrm{P}(A_3 \mid B)\mathrm{P}(B) = \mathrm{P}(A_3 \mid A_1 \cap A_2)\mathrm{P}(A_2 \mid A_1)\mathrm{P}(A_1),$$

on using the definition of conditional probability, twice. For the general case, just extend this idea, by setting

$$\begin{aligned} \mathrm{P}(A_1 \cap \cdots \cap A_n) &= \mathrm{P}(A_n \mid A_1, \ldots, A_{n-1})\mathrm{P}(A_1, \ldots, A_{n-1}) \\ &= \mathrm{P}(A_n \mid A_1, \ldots, A_{n-1})\mathrm{P}(A_{n-1} \mid A_1, \ldots, A_{n-2})\mathrm{P}(A_1, \ldots, A_{n-2}) \\ &\vdots \\ &= \prod_{i=2}^{n} \mathrm{P}(A_i \mid A_1 \cap \cdots \cap A_{i-1}) \times \mathrm{P}(A_1), \end{aligned}$$

as required.

Example

Example 47. *n men go to a dinner. Each leaves his hat in the cloakroom. When they leave, having thoroughly sampled the local wine, they choose their hats randomly.*
(a) What is the probability that no one chooses his own hat?
(b) What is the probability that exactly r men choose their own hats?
(c) What happens when n is very big?

Note to Example 47

- ☐ This is an example of many types of matching problem, going back to Montmort (1708).
- ☐ The sample space here is the permutations of the numbers $\{1, \ldots, n\}$, of size $n!$.
- ☐ Let A_i denote the event that the ith hat is on the ith head, and note that $\mathrm{P}(A_i) = 1/n$,

$$\mathrm{P}(A_i \cap A_j) = \mathrm{P}(A_i \mid A_j)\mathrm{P}(A_j) = \frac{1}{n-1} \times \frac{1}{n}, \ldots, \mathrm{P}(A_1 \cap \cdots \cap A_r) = \frac{(n-r)!}{n!},$$

using the prediction decomposition. Thus the probability that at least r out of n hats are on the right heads is $(n-r)!/n!$. Let $p_n(k)$ denote the probability that exactly k out of n men get the right hat.

- ☐ (a) We want to compute

$$\mathrm{P}(A_1^c \cap \cdots \cap A_n^c) = 1 - \mathrm{P}(A_1 \cup \cdots \cup A_n),$$

so we use the inclusion-exclusion formula to compute $p_n(0) = 1 - \mathrm{P}(A_1 \cup \cdots \cup A_n)$:

$$\begin{aligned} 1 - \mathrm{P}(A_1 \cup \cdots \cup A_n) &= 1 - \left\{ \sum_{r=1}^{n} (-1)^{r+1} \sum_{1 \le i_1 < \cdots < i_r \le n} \mathrm{P}(A_{i_1} \cap \cdots \cap A_{i_r}) \right\} \\ &= 1 - \left\{ n \times n^{-1} - \binom{n}{2} \times \frac{(n-2)!}{n!} + \cdots + (-1)^{n+1} \times \binom{n}{n} \times \frac{(n-n)!}{n!} \right\} \\ &= 1 - \sum_{i=1}^{n} (-1)^{i+1}/i! = \sum_{i=0}^{n} (-1)^i/i! \to e^{-1}, \quad n \to \infty. \end{aligned}$$

- ☐ (b) The probability that men $1, \ldots, r$ have the right hats and no-one else does is

$$\begin{aligned} \mathrm{P}(A_1 \cap \cdots \cap A_r \cap A_{r+1}^c \cap \cdots \cap A_n^c) &= \mathrm{P}(A_1 \cap \cdots \cap A_r) \times \mathrm{P}(A_{r+1}^c \cap \cdots \cap A_n^c \mid A_1 \cap \cdots \cap A_r) \\ &= \frac{(n-r)!}{n!} \times \sum_{i=0}^{n-r} (-1)^i/i!, \end{aligned}$$

but since there are $\binom{n}{r}$ distinct ways of choosing r from n, the total probability is

$$p_n(r) = \frac{n!}{r!(n-r)!} \times \frac{(n-r)!}{n!} \times \sum_{i=0}^{n-r} (-1)^i/i! = \frac{1}{r!} \times \sum_{i=0}^{n-r} (-1)^i/i! \to \frac{1}{r!} e^{-1}, \quad n \to \infty.$$

- ☐ (c) See above.

2.3 Independence slide 70

Independent events

Intuitively, saying that 'A and B are independent' means that the occurrence of one of the two does not affect the occurrence of the other. That is to say that, $\mathrm{P}(A \mid B) = \mathrm{P}(A)$, so the knowledge that B has occurred leaves $\mathrm{P}(A)$ unchanged.

Example 48. *A family has two children.*
(a) We know that the first child is a boy. What is the probability that the second child is a boy?
(b) We know that one of the two children is a boy. What is the probability that the other child is also a boy?

Note to Example 48

The sample space can be written as $\Omega = \{BB, BG, GB, GG\}$, in an obvious notation, and the events that 'the ith child is a boy' are $B_1 = \{BB, BG\}$ and $B_2 = \{BB, GB\}$. Then

- ☐ (a) $\mathrm{P}(B_2 \mid B_1) = \mathrm{P}(B_1 \cap B_2)/\mathrm{P}(B_2) = \mathrm{P}(\{BB\})/\mathrm{P}(B_1) = 1/4 \div 1/2 = 1/2 = \mathrm{P}(B_2)$. Thus B_2 and B_1 are independent.
- ☐ (b) the event 'at least one child is a boy' is $C = B_1 \cup B_2 = \{BB, BG, GB\}$, and the event 'two boys' is $D = \{BB\}$, so now we seek $\mathrm{P}(D \mid C2) = 1/4 \div 3/4 = 1/3 \neq \mathrm{P}(D)$. Thus D and C are not independent.

Note also the importance of precise language: in (a) we know that a specific child is a boy, and in (b) we are told only that one of the two children is a boy. These different pieces of information change the probabilities, because the conditioning event is not the same.

Independence

Definition 49. *Let $(\Omega, \mathcal{F}, \mathrm{P})$ be a probability space. Two events $A, B \in \mathcal{F}$ are* **independent** *(we write $A \perp\!\!\!\perp B$) iff*

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

In compliance with our intuition, this implies that

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

and by symmetry $\mathrm{P}(B \mid A) = \mathrm{P}(B)$.

Example 50. *A pack of cards is well shuffled and one card is packed at random. Are the events A 'the card is an ace', and H 'the card is a heart' independent? What can we say about the events A and K 'the card is a king'?*

Note to Example 50

The sample space Ω consists of the 52 cards, which are equiprobable. $\mathrm{P}(A) = 4/52 = 1/13$ and $\mathrm{P}(H) = 13/52 = 1/4$, and $\mathrm{P}(A \cap H) = 1/52 = \mathrm{P}(A)\mathrm{P}(H)$, so A and H are independent. However $\mathrm{P}(A \cap K) = 0 \neq \mathrm{P}(A)\mathrm{P}(K)$, so these are not independent.

Types of independence

Definition 51. *(a) The events $A_1, \ldots, A_n$ are* **(mutually) independent** *if for all sets of indices $F \subset \{1, \ldots, n\}$,*

$$\mathrm{P}\left(\bigcap_{i \in F} A_i\right) = \prod_{i \in F} \mathrm{P}(A_i).$$

(b) The events $A_1, \ldots, A_n$ are **pairwise independent** *if*

$$\mathrm{P}(A_i \cap A_j) = \mathrm{P}(A_i)\ \mathrm{P}(A_j), \quad 1 \leq i < j \leq n.$$

(c) The events $A_1, \ldots, A_n$ are **conditionally independent given B** *if for all sets of indices $F \subset \{1, \ldots, n\}$,*

$$\mathrm{P}\left(\bigcap_{i \in F} A_i \mid B\right) = \prod_{i \in F} \mathrm{P}(A_i \mid B).$$

A few remarks

- ☐ Mutual independence entails pairwise independence, but the converse is only true when $n = 2$.
- ☐ Mutual independence neither implies nor is implied by conditional independence.
- ☐ Independence is a key idea that greatly simplifies probability calculations. In practice, it is essential to verify whether events are independent, because undetected dependence can greatly modify the probabilities.

Example 52. *A family has two children. Show that the events 'the first born is a boy', 'the second child is a boy', and 'there is exactly one boy' are pairwise independent but not mutually independent.*

Note to Example 52

The sample space is $\Omega = \{BB, BG, GB, GG\}$, so $\mathrm{P}(B_1) = 1/2$, $\mathrm{P}(B_2) = 1/2$, $\mathrm{P}(1B) = 1/2$, using an obvious notation.
Also $\mathrm{P}(B_1 \cap B_2) = \mathrm{P}(B_1 \cap 1B) = \mathrm{P}(B_2 \cap 1B) = 1/4$, but $\mathrm{P}(B_1 \cap B_2 \cap 1B) = 0$, while the product of all three probabilities is $1/8$.

Example 53. *In any given year, the probability that a male driver has an accident and claims on his insurance is μ, independently of other years. The probability for a female driver is $\lambda < \mu$. An insurer has the same number of male drivers and female drivers, and picks one of them at random.*
(a) Give the probability that he (or she) makes a claim this year.
(b) Give the probability that he (or she) makes claims in two consecutive years.
(c) If the company randomly selects a person that made a claim, give the probability that (s)he makes a claim the following year.
(d) Show that the knowledge that a claim was made in one year increases the probability that a claim is made in the following year.

Note to Example 53

Let A_r denote the event that the selected driver has accidents in r successive years, and M denote the event that (s)he is male.
(a) Here the law of total probability gives

$$\mathrm{P}(A_1) = \mathrm{P}(A_1 \mid M)\mathrm{P}(M) + \mathrm{P}(A_1 \mid M^c)\mathrm{P}(M^c) = \mu \times \tfrac{1}{2} + \lambda \times \tfrac{1}{2} = (\mu + \lambda)/2.$$

(b) Independence of accidents from year to year, *for each driver individually*, gives

$$\mathrm{P}(A_2) = \mathrm{P}(A_2 \mid M)\mathrm{P}(M) + \mathrm{P}(A_2 \mid M^c)\mathrm{P}(M^c) = \mu^2 \times \tfrac{1}{2} + \lambda^2 \times \tfrac{1}{2} = (\mu^2 + \lambda^2)/2.$$

(c) Now we want

$$\mathrm{P}(A_2 \mid A_1) = \mathrm{P}(A_2 \cap A_1)/\mathrm{P}(A_1) = \mathrm{P}(A_2)/\mathrm{P}(A_1) = (\lambda^2 + \mu^2)/(\lambda + \mu).$$

(d) Note that $(\lambda^2 + \mu^2)/(\lambda + \mu) > (\lambda + \mu)/2$, because

$$2(\lambda^2 + \mu^2) - (\lambda + \mu)^2 = \lambda^2 + \mu^2 - 2\lambda\mu = (\lambda - \mu)^2 > 0.$$

Thus they would only be equal if $\lambda = \mu$, i.e. with no difference between the sexes.

Series-Parallel Systems

An electric system has components labelled $1, \ldots, n$, which fail independently of each another. Let F_i be the event 'the ith component is faulty', with $\mathrm{P}(F_i) = p_i$. The event S, 'the system fails' occurs if current cannot pass from one end of the system to the other. If the components are arranged in parallel, then

$$\mathrm{P_P}(S) = \mathrm{P}(F_1 \cap \cdots \cap F_n) = \prod_{i=1}^{n} p_i.$$

If the components are arranged in series, then

$$\mathrm{P_S}(S) = \mathrm{P}(F_1 \cup \cdots \cup F_n) = 1 - \prod_{i=1}^{n} (1 - p_i).$$

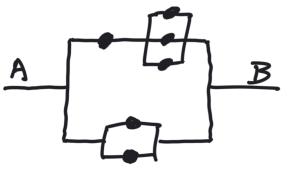
If there exist upper and lower bounds p_+ and p_- such that

$$1 > p_+ > p_i > p_- > 0, \quad i = 1, \ldots, n,$$

and $n \to \infty$, then $\mathrm{P_P}(S) \to 0$, $\mathrm{P_S}(S) \to 1$.

Reliability

Example 54 (Chernobyl). *A nuclear power station depends on a security system whose components are arranged according to:*



The components fail independently with probability p, and the system fails if current cannot pass from A to B.
(a) What is the probability that the system fails?
(b) The components are made in batches, which can be good or bad. For a good batch, $p = 10^{-6}$, whereas for a bad batch $p = 10^{-2}$. The probability that a batch is good is 0.99. What is the probability that the system fails (i) if the components come from different batches? (ii) if all the components come from the same batch?

Note to Example 54

The two parallel systems in the upper right and lower branches have respective probabilities p^3 and $p_l = p^2$ of failing, so the overall probability of failure for the top branch, which is a series system, is $p_u = 1 - (1-p)(1-p^3)$. The upper and lower branches are in parallel, so the probability that they both fail is $p_u \times p_l = p^2\{1 - (1-p)(1-p^3)\} = f(p)$, say.
Such computations can be used recursively to compute failure probabilities for very large systems.
The probability of failure of a component selected randomly from the two sorts of batches is

$$q = 10^{-6} \times 0.99 + 10^{-2} \times 0.1 = 0.00010099,$$

so the probability of failure in case (i) is $f(q) = 1.029995 \times 10^{-12}$, whereas in (ii) it is

$$0.99 f(10^{-6}) + 0.01 f(10^{-2}) = 1.000099 \times 10^{-8},$$

roughly 10^4 times larger than in (i).

2.4 Edifying Examples

Death and the Ladies

(Source: La Danse Macabre des Femmes, Project Gutenberg)

Female smokers

Survival after 20 years for 1314 women in the town of Whickham, England (Appleton et al., 1996, *The American Statistician*). The columns contain: number of dead women after 20 years/number of surviving women at the start of the study (%).

Age (years)	Smokers	Non-smokers
Total	139/582 (24)	230/732 (31)
18–24	2/55 (4)	1/62 (2)
25–34	3/124 (2)	5/157 (3)
35–44	14/109 (13)	7/121 (6)
45–54	27/130 (21)	12/78 (15)
55–64	51/115 (44)	40/121 (33)
65–74	29/36 (81)	101/129 (78)
75+	13/13 (100)	64/64 (100)

According to the totals, there is a beneficial effect of smoking:

$$24\% < 31\%!$$

Simpson's paradox

Define the events 'dead after 20 years', D, 'smoker', S, and 'in age category a at the start', $A = a$. For almost every a we have

$$\mathrm{P}(D \mid S, A = a) > \mathrm{P}(D \mid S^c, A = a),$$

but

$$\mathrm{P}(D \mid S) < \mathrm{P}(D \mid S^c).$$

Note that

$$\begin{aligned} \mathrm{P}(D \mid S) &= \sum_a \mathrm{P}(D \mid S, A = a)\mathrm{P}(A = a), \\ \mathrm{P}(D \mid S^c) &= \sum_a \mathrm{P}(D \mid S^c, A = a)\mathrm{P}(A = a), \end{aligned}$$

so if the probabilities $\mathrm{P}(D \mid S, A = a)$ and $\mathrm{P}(D \mid S^c, A = a)$ vary a lot with a, weighting them with the $\mathrm{P}(A = a)$ can reverse the order of the inequalities.

This is an example of **Simpson's paradox**: 'forgetting' conditioning can change the conclusion of a study.

The tragic story of Sally Clark

An English solicitor, whose first son died of **Sudden Infant Death Syndrome (SIDS)** a few weeks after his birth in 1996. Following the death of her second son in the same manner, she was arrested in 1998 and accused of double murder. Her trial was controversial, as a very eminent paediatrician, Professor Sir Roy Meadow, testified that the probability that two children should die of SIDS in a family such as that of Sally Clark was of 1 in 73 million, a number he obtained as $1/8500^2$, where $1/8500$ was the estimated probability of a single death due to SIDS.

She was convicted in November 1999, then released in January 2003, because it turned out some pathological evidence suggesting her innocence had not been diclosed to her lawyer. As a result of her case, the Attorney-General ordered a review of hundreds of other cases, and two other women in the same situation were released from jail.

She died of alcoholism in March 2007.

The rates of SIDS

Table 3.58 SIDS rates for different factors based on the data from the CESDI SUDI study

	*SIDS rate per 1000 live births**	*SIDS incidence in this group**
Overall rate in the study population	0.768	1 in 1 303
Rate for groups with different factors		
Anybody smokes in the household	1.357	1 in 737
Nobody smokes in the household	0.199	1 in 5 041
No waged income in household	2.057	1 in 486
At least one waged income in household	0.479	1 in 2 088
Mother < 27 years and parity >1	1.762	1 in 567
Mother > 26 years or parity = 1	0.531	1 in 1 882
None of these factors	0.117	1 in 8 543
One of these factors	0.619	1 in 1 616
Two of these factors	1.678	1 in 596
All three of these factors	4.674	1 in 214

* Based on the number of live births in each study region from 1993 to 1995 inclusive (OPCS)

Data on the rates of infantile deaths, (CESMA SUDI report, `http://cemach.interface-test.com/Publications/CESDI-SUDI-Report-(1).aspx`)

Sally Clark: Four tragic errors

- ☐ Estimated probabilities
- ☐ 'Ecological fallacy'
- ☐ Independence? Really?
- ☐ 'Prosecutors' fallacy'

Note on Sally Clark story

- **Estimated probabilities**: How were the probabilities obtained? What is their accuracy? There are very few SIDS deaths, and the number 1/8543 may be based on as few as 4 SIDS deaths. Using standard methods, the estimated probability could be from 0.04 to 0.32 deaths/1000 live births, so (for example), the figure of 1/73 million could be much larger.
- **Ecological fallacy**: Even if we accept the argument above, the SUDI study conflates a lot of different types of families and cases: there is no reason to suppose that the marginal probability of 1/8500 applies to any particular individual (think of Simpson's paradox, which we just met).
- **Independence?** If there is a genetic or environmental factor leading to SIDS, then the probability of two deaths might be much higher than claimed. Just suppose that a genetic factor G is present in 0.1% of families, and leads to a probability of death of 1/10 for each child, and that conditional on G or G^c, deaths are independent. Then we might have

$$\begin{aligned} \mathrm{P}(\text{ two deaths }) &= \mathrm{P}(\text{ two deaths } \mid G)\mathrm{P}(G) + \mathrm{P}(\text{ two deaths } \mid G^c)\mathrm{P}(G^c) \\ &= (1/10)^2 \times 0.001 + (1/8500)^2 \times 0.999 \doteq 0.0001 = 1/10^4 \gg 1/(73 \times 10^6). \end{aligned}$$

- **Prosecutors' Fallacy**: The probability calculated was $\mathrm{P}(\text{ two deaths } \mid \text{ innocent })$, whereas what is wanted is $\mathrm{P}(\text{ innocent } \mid \text{ two deaths })$. To get the latter we need to apply Bayes' theorem. Let E denote the evidence observed (two deaths), and C denote culpability. Then we have

$$\mathrm{P}(C^c \mid E) = \frac{\mathrm{P}(E \mid C^c)\mathrm{P}(C^c)}{\mathrm{P}(E \mid C^c)\mathrm{P}(C^c) + \mathrm{P}(E \mid C)\mathrm{P}(C)},$$

and we see that in order to compute the required probability, we have to have some estimates of $\mathrm{P}(C)$. Suppose that $\mathrm{P}(C) = 10^{-6}$ and that $\mathrm{P}(E \mid C) = 1$, as murdering two of your own children is probably quite rare. Then even using the probabilities above, Bayes's theorem would give that

$$\mathrm{P}(C^c \mid E) \doteq 0.014 \approx 14/10^3,$$

which, though small, is nothing like as small as $1/(73 \times 10^6)$. Thus even accepting the 'squaring of probabilities', the case for the prosecution is not nearly as strong as the original argument suggested.

3 Random Variables

Small probabilistic lexicon

Mathematics	English	Français
	one fair die (several fair dice)	un dé juste/équilibré (plusieurs dés justes/équilibrés)
	random experiment	expérience aléatoire
Ω	sample space	ensemble fondamental
ω	outcome, elementary event	épreuve, événement élémentaire
$A, B, \ldots$	event	événement
$\mathcal{F}$	event space	l'espace des événements
	sigma-algebra	tribu
P	probability distribution/probability function	loi de probabilité
$(\Omega, \mathcal{F}, \mathrm{P})$	probability space	espace de probabilité
	inclusion-exclusion formula	formule d'inclusion-exclusion
$\mathrm{P}(A \mid B)$	probability of A given B	probabilité de A sachant B
	independence	indépendance
	(mutually) independent events	événements (mutuellement) indépendants
	pairwise independent events	événements indépendants deux à deux
	conditionally independent events	événements conditionellement indépendants
$X, Y, Z, W, \ldots$	random variable	variable aléatoire
$F_X(x)$	(cumulative) distribution function	fonction de répartition
$f_X(x)$	(probability) density/mass function (PDF)	fonction de densité/masse (fm)
$\mathrm{E}(X)$	expectation/mean of X	espérance de X
$\mathrm{var}(X)$	variance of X	la variance de X
$\mathrm{var}(X)^{1/2}$	standard deviation of X	deviation standard (ou écart-type, mais ...) de X
$f_X(x \mid B)$	conditional density/mass function	fonction de densité/masse conditionnelle

3.1 Basic Ideas

Random variables

We usually need to consider random numerical quantities.

Example 55. *We roll two fair dice, one red and one green. Let X be the total of the sides facing up. Find all possible values of X, and the corresponding probabilities.*

Definition 56. *Let $(\Omega, \mathcal{F}, \mathrm{P})$ be a probability space. A* **random variable (rv)** *$X : \Omega \mapsto \mathbb{R}$ is a function from the sample space Ω taking values in the real numbers $\mathbb{R}$.*

Definition 57. *The set of values taken by X,*

$$D_X = \{x \in \mathbb{R} : \exists \omega \in \Omega \text{ such that } X(\omega) = x\}$$

is called the **support** *of X. If D_X is countable, then X is a* **discrete random variable**.

The random variable X associates probabilities to subsets S included in $\mathbb{R}$, given by

$$\mathrm{P}(X \in S) = \mathrm{P}(\{w \in \Omega : X(w) \in S\}).$$

In particular, we set $A_x = \{\omega \in \Omega : X(\omega) = x\}$. Note that we must have $A_x \in \mathcal{F}$ for every $x \in \mathbb{R}$, in order to calculate $\mathrm{P}(X = x)$.

Note to Example 55

Draw a grid. X takes values in $D_X = \{2, \ldots, 12\}$, and so is clearly a discrete random variable. By symmetry the 36 points in Ω are equally likely, so, for example,

$$\mathrm{P}(X = 3) = \mathrm{P}(\{(1,2),(2,1)\}) = \frac{2}{36}.$$

Thus the probabilities for $\{2, 3, 4 \ldots, 12\}$ are respectively

$$1/36,\ 2/36,\ 3/36,\ 4/36,\ 5/36,\ 6/36,\ 5/36,\ 4/36,\ 3/36,\ 2/36,\ 1/36.$$

Examples

Example 58. *We toss a coin repeatedly and independently. Let X be the random variable representing the number of throws until we first get heads. Calculate*

$$\mathrm{P}(X = 3),\ \mathrm{P}(X = 15),\ \mathrm{P}(X \leq 3.5),\ \mathrm{P}(X > 1.7),\ \mathrm{P}(1.7 \leq X \leq 3.5).$$

Example 59. *A natural set Ω when I am playing darts is the wall on which the dart board is hanging. The dart lands on a point $\omega \in \Omega \subset \mathbb{R}^2$. My score is $X(\omega) \in D_X = \{0, 1, \ldots, 60\}$.*

Note to Example 58

X takes values in $\{1, 2, 3, \ldots\} = \mathbb{N}$, and so is clearly a discrete random variable, with countable support.
Let $p = \mathrm{P}(F)$; then the event $X = 3$ corresponds to two failures, each with probability $1 - p$, followed by a success, with probability p, giving $\mathrm{P}(X = 3) = (1 - p)^2 p$ by independence of the successive trials. Likewise $\mathrm{P}(X = 15) = (1 - p)^{14} p$, and

$$\begin{aligned} \mathrm{P}(X \leq 3.5) &= \mathrm{P}(X \leq 3) + \mathrm{P}(3 < X \leq 3.5) \\ &= p + (1 - p)p + (1 - p)^2 p \\ &= 1 - \mathrm{P}(X > 3) \\ &= 1 - (1 - p)^3, \end{aligned}$$

and similarly

$$\begin{aligned} \mathrm{P}(1.7 \leq X \leq 3.5) &= \mathrm{P}(X = 2) + \mathrm{P}(X = 3) \\ &= (1 - p)p + (1 - p)^2 p \\ &= p(1 - p)(1 + 1 - p) \\ &= p(1 - p)(2 - p). \end{aligned}$$

Note to Example 59

Here an infinite $\Omega \subset \mathbb{R}^2$ is mapped onto the finite set $\{0, \ldots, 60\}$. Even though the underlying Ω is uncountable, the support of X is countable.

Jacob Bernoulli (1654–1705)

Ars Conjectandi, Basel (1713)
(Source: http://www-history.mcs.st-and.ac.uk/PictDisplay/Bernoulli_Jacob.html)

Bernoulli random variables

Definition 60. *A random variable that takes only the values 0 and 1 is called an* **indicator variable**, *or a* **Bernoulli random variable**, *or a* **Bernoulli trial**.

Typically the values 0/1 correspond to false/true, failure/success, bad/good, ...

Example 61. *Suppose that n identical coins are tossed independently, let H_i be the event 'we get heads for the ith coin', and let $I_i = I(H_i)$ be the indicator of this event. Then*

$$\mathrm{P}(I_i = 1) = \mathrm{P}(H_i) = p, \quad \mathrm{P}(I_i = 0) = \mathrm{P}(H_i^c) = 1 - p,$$

where p is the probability of obtaining heads.

- ☐ *If $n = 3$ and $X = I_1 + I_2 + I_3$, describe Ω, D_X and the sets A_x.*
- ☐ *What do*

$$X = I_1 + \cdots + I_n, \quad Y = I_1(1 - I_2)(1 - I_3), \quad Z = \sum_{j=2}^{n} I_{j-1}(1 - I_j)$$

represent?

Note to Example 61

- ☐ If $n = 3$, then we can write the sample space as $\Omega = \{TTT, TTH, THT, HTT, THH, HTH, HHT, HHH\}$. Clearly $D_X = \{0, 1, 2, 3\}$, and

$$A_0 = \{TTT\}, \quad A_1 = \{TTH, THT, HTT\}, \quad A_2 = \{THH, HTH, HHT\}, \quad A_3 = \{HHH\}.$$

- ☐ X is the total number of heads in the first n tosses, $Y = 1$ if and only if the sequence starts HTT, and Z counts the number of times a 1 is followed by a 0 in the sequence of n tosses.

Mass functions

A random variable X associates probabilities to subsets of $\mathbb{R}$. In particular when X is discrete, we have

$$A_x = \{\omega \in \Omega : X(\omega) = x\},$$

and we can define:

Definition 62. *The* **probability mass function (PMF)** *of a discrete random variable X is*

$$f_X(x) = \mathrm{P}(X = x) = \mathrm{P}(A_x), \quad x \in \mathbb{R}.$$

It has two key properties :

(i) $f_X(x) \geq 0$, and it is only positive for $x \in D_X$, where D_X is the image of the function X, i.e., the **support** *of f_X;*

(ii) the total probability $\sum_{\{i:x_i \in D_X\}} f_X(x_i) = 1$.

When there is no risk of confusion, we write $f_X \equiv f$ and $D_X \equiv D$.

Binomial random variable

Example 63 (Example 61 continued). *Give the PMFs and supports of I_i, of Y and of X.*

Definition 64. *A* **binomial** *random variable X has PMF*

$$f(x) = \binom{n}{x} p^x (1-p)^{n-x}, \quad x = 0, 1, \ldots, n, \quad n \in \mathbb{N}, 0 \leq p \leq 1.$$

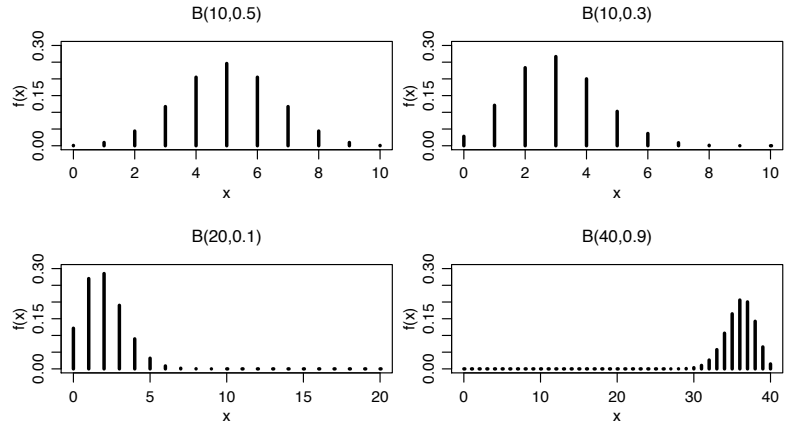
We write $X \sim B(n, p)$, and call n the **denominator** *and p the* **probability of success**. *With $n = 1$, this is a Bernoulli variable.*

Remark: we use $\sim$ to mean 'has the distribution'.
The binomial model is used when we are considering the number of 'successes' of a trial which is independently repeated a fixed number of times, and where each trial has the same probability of success.

Note to Example 63

- ☐ I_i takes values 0 and 1, with probabilities $\mathrm{P}(I_i = 1) = p$, and $\mathrm{P}(I_i = 0) = 1 - p$.
- ☐ Y is also binary with $\mathrm{P}(Y = 1) = p(1-p)^2$, $\mathrm{P}(Y = 0) = 1 - p(1-p)^2$.
- ☐ X takes values $0, 1, \ldots, n$, with binomial probabilities (see below).

Binomial probability mass functions



Examples

Example 65. *A multiple choice test contains 20 questions. For each question you must choose the correct answer amongst 5 possible answers. A pass is obtained with 10 correct answers. A student picks his answers at random.*

- ☐ *Give the distribution for his number of correct answers.*
- ☐ *What is the probability that he will pass the test?*

Note to Example 65

Since $n = 20$ and $p = 1/5 = 0.2$, the number of correct replies is $X \sim B(20, 0.2)$. The probability of success is

$$\mathrm{P}(X \geq 10) = \sum_{x=10}^{20} \binom{20}{x} 0.2^x (1-0.2)^{20-x} \doteq 0.0026$$

after a painful calculation, or, better, using R,

```
> 1-pbinom(q=9, size=20, prob=0.2)
[1] 0.002594827
> pbinom(q=9, size=20, prob=0.2, lower.tail=FALSE)
[1] 0.002594827
```


Geometric distribution

Definition 66. *A* **geometric** *random variable X has PMF*

$$f_X(x) = p(1-p)^{x-1}, \quad x = 1, 2, \ldots, \quad 0 \leq p \leq 1.$$

We write $X \sim \mathrm{Geom}(p)$, and we call p the **success probability**.

This models the waiting time until a first event, in a series of independent trials having the same success probability.

Example 67. *To start a board game, m players each throw a die in turn. The first to get six begins. Give the probabilities that the 3rd player will begin on his first throw of the die, that he will begin, and of waiting for at least 6 throws of the die before the start of the game.*

Theorem 68 (**Lack of memory**). *If $X \sim \mathrm{Geom}(p)$, then*

$$\mathrm{P}(X > n + m \mid X > m) = \mathrm{P}(X > n).$$

This is also sometimes called **memorylessness**.

Note to Example 67

In this case $D_X = \mathbb{N}$.
Here $p = 1/6$, so the probability that the third person starts on his first throw of the die is $(5/6)^2 \times 1/6 = 0.116$. He starts if the first six appears on throw $3, m+3, 2m+3, \ldots$ and this equals

$$\sum_{i=0}^{\infty} \mathrm{P}(X = 3 + im) = \sum_{i=0}^{\infty} p(1-p)^{3+im-1} = p(1-p)^2 \sum_{i=0}^{\infty} (1-p)^{im} = \frac{p(1-p)^2}{1-(1-p)^m},$$

where $p = 1/6$.
The probability of waiting for at least 6 tosses is $(1-p)^6 = 0.335$.

Note to Theorem 68

Since $\mathrm{P}(X > n) = (1-p)^n$, we seek

$$\mathrm{P}(X > n+m \mid X > m) = (1-p)^{m+n}/(1-p)^m = (1-p)^n = \mathrm{P}(X > n).$$

Thus we see that there is a 'lack of memory': knowing that $X > m$ does not change the probability that we have to wait at least another n trials before seeing the event.

Negative binomial distribution

Definition 69. *A* **negative binomial** *random variable X with parameters n and p has PMF*

$$f_X(x) = \binom{x-1}{n-1} p^n (1-p)^{x-n}, \quad x = n, n+1, n+2, \ldots, \quad 0 \le p \le 1.$$

We write $X \sim \mathrm{NegBin}(n,p)$. *When* $n = 1$, $X \sim \mathrm{Geom}(p)$.

It models the waiting time until the nth success in a series of independent trials having the same success probability.

Example 70. *Give the probability of seeing 2 heads before 5 tails in repeated tosses of a coin.*

Note to Example 70

This is the probability that $X \le 6$, where X is the waiting time for $n = 2$ heads. It is

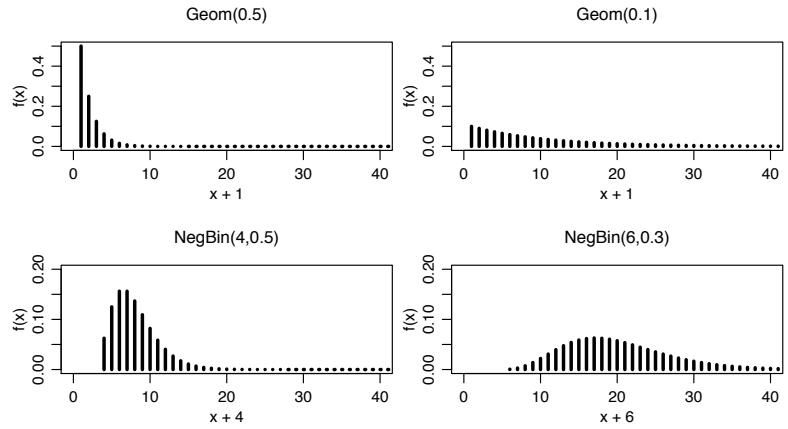
$$\binom{2-1}{2-1} p^2 (1-p)^{2-2} + \binom{3-1}{2-1} p^2 (1-p)^{3-2}$$
$$+ \binom{4-1}{2-1} p^2 (1-p)^{4-2} + \binom{5-1}{2-1} p^2 (1-p)^{5-2} + \binom{6-1}{2-1} p^2 (1-p)^{6-2}.$$

If we assume that the coin is fair, so $p = 0.5$, `R` gives

```
pnbinom(q=4, size=2, prob=0.5)
[1] 0.890625
```

where note that $q = x - n$ in the parametrization used in `R`.

Geometric and negative binomial PMFs



Negative binomial distribution: alternative version

We sometimes write the geometric and negative binomial variables in a more general form, setting $Y = X - n$, and then the probability mass function is

$$f_Y(y) = \frac{\Gamma(y+\alpha)}{\Gamma(\alpha)y!}p^\alpha(1-p)^y, \quad y = 0, 1, 2, \ldots, \quad 0 \le p \le 1, \alpha > 0,$$

where

$$\Gamma(\alpha) = \int_0^\infty u^{\alpha-1}e^{-u}\,du, \quad \alpha > 0$$

is the **Gamma function**. The principal properties of $\Gamma(\alpha)$ are:

$$\begin{aligned} \Gamma(1) &= 1; \\ \Gamma(\alpha+1) &= \alpha\Gamma(\alpha), \quad \alpha > 0; \\ \Gamma(n) &= (n-1)!, \quad n = 1, 2, 3, \ldots; \\ \Gamma(\tfrac{1}{2}) &= \sqrt{\pi}. \end{aligned}$$

They will be useful later.

Hypergeometric distribution

Definition 71. *We draw a sample of m balls without replacement from an urn containing w white balls and b black balls. Let X be the number of white balls drawn. Then*

$$\mathrm{P}(X = x) = \frac{\binom{w}{x}\binom{b}{m-x}}{\binom{w+b}{m}}, \quad x = \max(0, m-b), \ldots, \min(w, m),$$

and the distribution of X is **hypergeometric**. *We write* $X \sim \mathrm{HyperGeom}(w, b; m)$.

Example 72. *I leave for a camping trip in Ireland with six tins of food, two of which contain fruit. It pours with rain, and the labels come off the tins. If I pick three of the six tins at random, find the distribution of the number of tins of fruit among the three I have chosen.*

Note to Example 72

White balls correspond to fruit tins, black balls to others, so $w = 2$, $b = 4$, and I take $m = 3$. Therefore the number of fruit tins X drawn has probability

$$\mathrm{P}(X = x) = \frac{\binom{2}{x}\binom{4}{3-x}}{\binom{6}{3}}, \quad x = 0, \ldots, 2,$$

and some calculation gives $\mathrm{P}(X = 0) = 1/5$, $\mathrm{P}(X = 1) = 3/5$, $\mathrm{P}(X = 2) = 1/5$.

Capture-recapture

Example 73. *In order to estimate the number of fish N in a lake, we first catch r fish, mark them, and let them go. After having waited long enough for the fish population to become well-mixed, we catch another sample of size s.*

- ☐ *Find the distribution of the number of marked fish, M, in this sample.*
- ☐ *Show that the value of N which maximises $\mathrm{P}(M = m)$ is $\lfloor rs/m \rfloor$, and calculate the best estimation of N when $s = 50$, $r = 40$, and $m = 4$.*

The basic idea behind this example is used to estimate the sizes of populations of endangered species, the number of drug addicts or of illegal immigrants in human populations, etc. One practical problem often encountered is that certain individuals become harder to recapture, whereas others enjoy it; thus the probabilities of recapture are heterogeneous, unlike in the example above.

Note to Example 73

The total number is N, of which r are marked and $N - r$ unmarked. The distribution of M is

$$\mathrm{P}_N(M = m) = \frac{\binom{r}{m}\binom{N-r}{s-m}}{\binom{N}{s}}, \quad m = \max(0, s + r - N), \ldots, \min(r, s),$$

(work out the limits carefully).
For the second part, we seek to maximise this probability with respect to N. Now compare the probabilities for N and $N - 1$ and take ratios, giving

$$\frac{\mathrm{P}_N(M = m)}{\mathrm{P}_{N-1}(M = m)} = \frac{\binom{r}{m}\binom{N-r}{s-m}}{\binom{N}{s}} \Big/ \frac{\binom{r}{m}\binom{N-1-r}{s-m}}{\binom{N-1}{s}} = \frac{(N-r)(N-s)}{N(N+m-r-s)} > 1$$

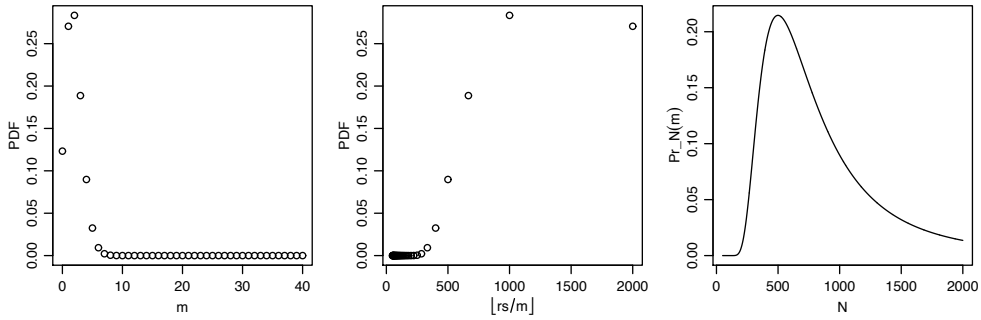
provided that (after a little algebra) $rs/m > N$. Hence the largest value of N for which this ratio increases is $\widehat{N} = \lfloor rs/m \rfloor$, which therefore maximises the probability, because we can write

$$\mathrm{P}_N(M = m) = \frac{\mathrm{P}_N(M = m)}{\mathrm{P}_{N-1}(M = m)} \times \cdots \times \frac{\mathrm{P}_{N_{\min}+1}(M = m)}{\mathrm{P}_{N_{\min}}(M = m)} \mathrm{P}_{N_{\min}}(M = m),$$

where the latter probability is for the smallest value of N for which the probability that $M = m$ is positive.
In the example given, $\widehat{N} = \lfloor 50 \times 40/4 \rfloor = 500$.
The behaviour of such estimators can be very poor.

Hypergeometric PMFs

Probability mass functions of M (left) and of $\lfloor rs/M \rfloor$ (centre) in Example 73, when $r = 40$, $s = 50$ and $N = 1000$, without $\lfloor rs/M \rfloor = +\infty$, which corresponds to $M = 0$, and $\mathrm{P}_N(M = m)$ as a function of N (right):



Discrete uniform distribution

Definition 74. *A* **discrete uniform** *random variable X has PMF*

$$f_X(x) = \frac{1}{b-a+1}, \quad x = a, a+1, \ldots, b, \quad a < b, \quad a, b \in \mathbb{Z}.$$

We write $U \sim \mathrm{DU}(a, b)$.

This definition generalizes the outcome of a die throw, which corresponds to the $\mathrm{DU}(1, 6)$ distribution.

Siméon-Denis Poisson (1781–1840)

'Life is good for only two things, discovering mathematics and teaching mathematics.'
(Source: http://www-history.mcs.st-and.ac.uk/PictDisplay/Poisson.html)

Poisson distribution

Definition 75. *A* Poisson *random variable* X *has the PMF*

$$f_X(x) = \frac{\lambda^x}{x!} e^{-\lambda}, \quad x = 0, 1, \ldots, \quad \lambda > 0.$$

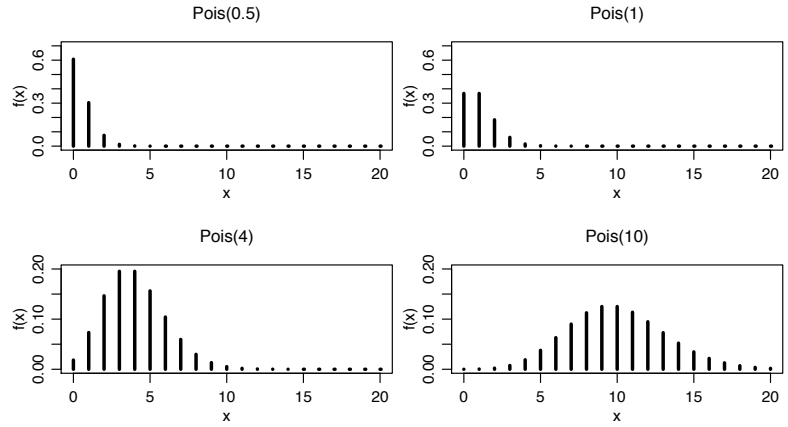
We write $X \sim \text{Pois}(\lambda)$.

- Since $\lambda^x / x! > 0$ for any $\lambda > 0$ and $x \in \{0, 1, \ldots\}$, and

$$e^{-\lambda} = \frac{1}{e^{\lambda}} = \frac{1}{\sum_{x=0}^{\infty} \frac{\lambda^x}{x!}} > 0,$$

  we see that $f_X(x) > 0$ and $\sum_{x=0}^{\infty} f_X(x) = 1$, so this is a probability distribution.
- The Poisson distribution appears everywhere in probability and statistics, often as a model for counts, or for a number of rare events.
- It also provides approximations to probabilities, for example for random permutations (Example 47, random hats) or the binomial distribution (later).

Poisson probability mass functions



Cumulative distribution function

Definition 76. *The* **cumulative distribution function (CDF)** *of a random variable X is*

$$F_X(x) = \mathrm{P}(X \leq x), \quad x \in \mathbb{R}.$$

If X is discrete, we can write

$$F_X(x) = \sum_{\{x_i \in D_X : x_i \leq x\}} \mathrm{P}(X = x_i),$$

which is a step function with jumps at the points of the support D_X of $f_X(x)$.

When there is no risk of confusion, we write $F \equiv F_X$.

Example 77. *Give the support and the probability mass and cumulative distribution functions of a Bernoulli random variable.*

Example 78. *Give the cumulative distribution function of a geometric random variable.*

Note to Example 77

The support is $D = \{0, 1\}$, and the CDF is

$$F(x) = \begin{cases} 0, & x < 0, \\ 1 - p, & 0 \leq x < 1, \\ 1, & x \geq 1. \end{cases}$$

Draw a picture, showing a step function with a jump of $1 - p$ at $x = 0$ and of p at $x = 1$.

Note to Example 78

The support is $D = \mathbb{N}$, and for $x \geq 1$ we have

$$\mathrm{P}(X \leq x) = \sum_{r=1}^{\lfloor x \rfloor} p(1-p)^{r-1},$$

so we need to sum a geometric series with common ratio $1 - p$, giving

$$\mathrm{P}(X \leq x) = \frac{p\{1 - (1-p)^{\lfloor x \rfloor}\}}{1 - (1-p)} = 1 - (1-p)^{\lfloor x \rfloor}.$$

Thus

$$\mathrm{P}(X \leq x) = \begin{cases} 0, & x < 1, \\ 1 - (1-p)^{\lfloor x \rfloor}, & x \geq 1. \end{cases}$$

Properties of a cumulative distribution function

Theorem 79. *Let $(\Omega, \mathcal{F}, \mathrm{P})$ be a probability space and $X : \Omega \mapsto \mathbb{R}$ a random variable. Its cumulative distribution function F_X satisfies:*

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

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

(c) F_X *is non-decreasing, so* $F_X(x) \leq F_X(y)$ *for* $x \leq y$;

(d) F_X *is continuous on the right, thus*

$$\lim_{t \downarrow 0} F_X(x + t) = F_X(x), \quad x \in \mathbb{R};$$

(e) $\mathrm{P}(X > x) = 1 - F_X(x)$;

(f) if $x < y$, *then* $\mathrm{P}(x < X \leq y) = F_X(y) - F_X(x)$.

Note to Theorem 79

(a) If not, there must be a blob of mass at $-\infty$, which is not allowed, as $X \in \mathbb{R}$.
(b) Ditto, for $+\infty$.
(c) If $y \geq x$, then $F(y) = F(x) + \mathrm{P}(x < X \leq y)$, so the difference is always non-negative.
(d) Now $F(x + t) = \mathrm{P}(X \leq x) + \mathrm{P}(x < X \leq x + t)$, and the second term here tends to zero, because any point in the interval $(x, x + t]$ at which there is positive probability must lie to the right of x.
(e) We have $\mathrm{P}(X > x) = 1 - \mathrm{P}(X \leq x) = 1 - F_X(x)$.
(f) We have $\mathrm{P}(x < X \leq y) = \mathrm{P}(X \leq y) - \mathrm{P}(X \leq x) = F_X(y) - F_X(x)$.

Remarks

- We can obtain the probability mass function of a discrete random variable from the cumulative distribution function using
$$f(x) = F(x) - \lim_{y \uparrow x} F(y).$$
In many cases X only takes integer values, $D_X \subset \mathbb{Z}$, and so $f(x) = F(x) - F(x - 1)$ for $x \in \mathbb{Z}$.
- From now on we will mostly ignore the implicit probability space $(\Omega, \mathcal{F}, \mathrm{P})$ when dealing with a random variable X. We will rather think in terms of X, $F_X(x)$, and $f_X(x)$. We can legitimise this 'oversight' mathematically.
- We can specify the distribution of a random variable in an equivalent way by saying (for example):
  - X follows a Poisson distribution with parameter λ; or
  - $X \sim \mathrm{Pois}(\lambda)$; or
  - by giving the probability mass function of X; or
  - by giving the cumulative distribution function of X.

Transformations of discrete random variables

Real-valued functions of random variables are random variables themselves, so they possess probability mass and cumulative distribution functions.

Theorem 80. *If X is a random variable and $Y = g(X)$, then*

$$f_Y(y) = \sum_{x:g(x)=y} f_X(x).$$

Example 81. *Calculate the PMF of $Y = I(X \geq 1)$ when $X \sim \text{Pois}(\lambda)$.*

Example 82. *Let Y be the remainder of the division by four of the total of two independent dice throws. Calculate the PMF of Y.*

Note to Theorem 80

We have

$$f_Y(y) = \text{P}(Y = y) = \sum_{x:g(x)=y} \text{P}(X = x) = \sum_{x:g(x)=y} f_X(x).$$

Note to Example 81

Here $Y = I(X \geq 1)$ takes values 0 and 1, and

$$f_Y(0) = \text{P}(Y = 0) = \text{P}(X = 0) = e^{-\lambda}, \quad f_Y(1) = \text{P}(Y = 1) = \sum_{x=1}^{\infty} \text{P}(X = x) = \sum_{x=1}^{\infty} \frac{\lambda^x}{x!} e^{-\lambda} = 1 - e^{-\lambda}.$$

Note to Example 82

Y has support $0, 1, 2, 3$, and mass function given by

$$\begin{aligned} f_Y(0) &= \text{P}(Y = 0) = \text{P}(X \in \{4, 8, 12\}) = (3 + 5 + 1)/36 = 9/36, \\ f_Y(1) &= \text{P}(Y = 1) = \text{P}(X \in \{5, 9\}) = (4 + 4)/36 = 8/36, \\ f_Y(2) &= \text{P}(Y = 2) = \text{P}(X \in \{2, 6, 10\}) = (1 + 5 + 3)/36 = 9/36, \\ f_Y(3) &= \text{P}(Y = 3) = \text{P}(X \in \{3, 7, 11\}) = (2 + 6 + 2)/36 = 10/36, \end{aligned}$$

which fortunately adds to 36/36.

3.2 Expectation

Expectation

Definition 83. *Let X be a discrete random variable for which $\sum_{x\in D_X}|x|f_X(x)<\infty$, where D_X is the support of f_X. The* **expectation** *(or* **expected value** *or* **mean***) of X is*

$$\mathrm{E}(X)=\sum_{x\in D_X} x\mathrm{P}(X=x)=\sum_{x\in D_X} x f_X(x).$$

- If $\mathrm{E}(|X|)=\sum_{x\in D_X}|x|f_X(x)$ is not finite, then $\mathrm{E}(X)$ is not well defined.
- $\mathrm{E}(X)$ is also sometimes called the "average of X". We will limit the use of the word "average" to empirical quantities.
- The expectation is analogous in mechanics to the notion of **centre of gravity** of an object whose mass is distributed according to f_X.

Example 84. *Calculate the expectation of a Bernoulli random variable with probability p.*

Example 85. *Calculate the expectation of $X\sim B(n,p)$.*

Example 86. *Calculate the expectation of the random variables with PMFs*

$$f_X(x)=\frac{4}{x(x+1)(x+2)},\quad f_Y(x)=\frac{1}{x(x+1)},\quad x=1,2,\ldots.$$

Note to Example 84

First we note that if the support of X is finite, then $\mathrm{E}(|X|)<\max_{x\in D_X}|x|<\infty$.
If I is Bernoulli with probability p, then $E(I)=0\times(1-p)+1\times p=p$.

Note to Example 85

Here $D_X=\{0,1,\ldots,n\}$ is finite, so $E(|X|)<\infty$.
We get

$$\begin{aligned}
E(X) &= \sum_{x=0}^{n} x\binom{n}{x}p^x(1-p)^{n-x}\\
&= \sum_{x=1}^{n}\frac{n\times(n-1)!}{(x-1)!\{n-1-(x-1)\}!}p\times p^{x-1}(1-p)^{(n-1)-(x-1)}\\
&= np\sum_{y=0}^{n-1}\binom{n-1}{y}p^y(1-p)^{n-1-y}=np,
\end{aligned}$$

where we have set $y=x-1$. This agrees with the previous example, since X can be viewed as a sum $I_1+\cdots+I_n$.

Note to Example 86

Note that f_Y sums to unity: since the series is absolutely convergent we can re-organise the brackets in the sums, giving

$$\sum_{x=1}^{\infty} \frac{1}{x(x+1)} = \sum_{x=1}^{\infty} \left(\frac{1}{x} - \frac{1}{x+1} \right) = \frac{1}{1} + \sum_{x=1}^{\infty} \left(\frac{1}{x+1} - \frac{1}{x+1} \right) = 1,$$

after cancelling terms.
A similar argument works for f_X, since

$$\begin{aligned} \sum_{x=1}^{\infty} \frac{4}{x(x+1)(x+2)} &= 2 \sum_{x=1}^{\infty} \left(\frac{1}{x(x+1)} - \frac{1}{(x+1)(x+2)} \right) \\ &= 2 \left\{ \frac{1}{1 \times 2} + \sum_{x=1}^{\infty} \left(\frac{1}{(x+1)(x+2)} - \frac{1}{(x+1)(x+2)} \right) \right\} = 1. \end{aligned}$$

Now since the sum below is absolutely convergent, we have

$$\mathrm{E}(X) = 4 \sum_{x=1}^{\infty} \frac{1}{(x+1)(x+2)} = 4 \sum_{x=1}^{\infty} \left(\frac{1}{x+1} - \frac{1}{x+2} \right) = 4 \left\{ \frac{1}{1+1} + \sum_{x=1}^{\infty} \left(\frac{1}{x+2} - \frac{1}{x+2} \right) \right\} = 2.$$

However,

$$\mathrm{E}(Y) = \sum_{x=1}^{\infty} \frac{1}{x+1} = +\infty.$$

Thus it is relatively easy to construct random variables whose expectations are infinite: existence of an expected value is not guaranteed.

Expected value of a function

Theorem 87. *Let X be a random variable with mass function f, and let g be a real-valued function of X. Then*

$$\mathrm{E}\{g(X)\} = \sum_{x \in D_X} g(x) f(x),$$

when $\sum_{x \in D_X} |g(x)| f(x) < \infty$.

Example 88. *Let $X \sim \mathrm{Pois}(\lambda)$. Calculate the expectations of*

$$X, \quad X(X-1), \quad X(X-1) \cdots (X-r+1).$$

Note to Theorem 87

Write $Y = g(X)$, and note that for any y in the support D_Y of Y, we have

$$f_Y(y) = \mathrm{P}(Y = y) = \mathrm{P}\{g(X) = y\} = \sum_{\{x \in D_X : g(x) = y\}} \mathrm{P}(X = x) = \sum_{\{x \in D_X : g(x) = y\}} f_X(x).$$

Therefore

$$\mathrm{E}(Y) = \sum_{y \in D_Y} y f_Y(y) = \sum_{y \in D_Y} y \sum_{\{x \in D_X : g(x) = y\}} f_X(x) = \sum_{y \in D_Y} \sum_{x : g(x) = y} g(x) f_X(x) = \sum_{x \in D_X} g(x) f_X(x),$$

as required.

Note to Example 88

Note that

$$\mathrm{E}\{X(X-1)\cdots(X-r+1)\} = \sum_{x=0}^{\infty} x(x-1)\cdots(x-r+1)\frac{\lambda^x}{x!}e^{-\lambda} = \lambda^r \sum_{x-r=0}^{\infty} \frac{\lambda^{x-r}}{(x-r)!}e^{-\lambda} = \lambda^r,$$

which yields $\mathrm{E}(X) = \lambda$ and $\mathrm{E}\{X(X-1)\} = \lambda^2$.

Properties of the expected value

Theorem 89. *Let X be a random variable with a finite expected value $\mathrm{E}(X)$, and let $a, b \in \mathbb{R}$ be constants. Then*

(a) $\mathrm{E}(\cdot)$ is a linear operator, i.e., $\mathrm{E}(aX + b) = a\mathrm{E}(X) + b$;

(b) if $g(X)$ and $h(X)$ have finite expected values, then

$$\mathrm{E}\{g(X) + h(X)\} = \mathrm{E}\{g(X)\} + \mathrm{E}\{h(X)\};$$

(c) if $\mathrm{P}(X = b) = 1$, then $\mathrm{E}(X) = b$;

(d) if $\mathrm{P}(a < X \leq b) = 1$, then $a < \mathrm{E}(X) \leq b$;

(e) $\{\mathrm{E}(X)\}^2 \leq \mathrm{E}(X^2)$.

Remark: Linearity of the expected value, (a) and (b), and fact (c), are very useful in calculations.

Note to Theorem 89

(a) We need to show absolute convergence:

$$\sum_x |ax+b| f(x) \le \sum_x (|a||x|+|b|) f(x) = |a| \sum_x |x| f(x) + |b| \sum_x f(x) < \infty,$$

and after that we just apply linearity of the summation.
(b) Follows using same argument as in (a), after noting that $|g(x)+h(x)| \le |g(x)| + |h(x)|$.
(c) Here $f(b) = \mathrm{P}(X=b) = 1$, so $\mathrm{E}(X) = bf(b) = b$ by definition.
(d) Now $f(x) = 0$ for $x \notin (a,b]$, so $\mathrm{E}(X) = \sum_x xf(x) \le \sum_x bf(x) = b$ and similarly $\mathrm{E}(X) > a$.
(e) For any real a, linearity of the expectation gives

$$0 \le \mathrm{E}\left\{(X-a)^2\right\} = \mathrm{E}\left\{X^2 - 2aX + a^2\right\} = \mathrm{E}(X^2) - 2a\mathrm{E}(X) + a^2,$$

and setting $a = \mathrm{E}(X)$ and simplifying the right-hand side to $\mathrm{E}(X^2) - \mathrm{E}(X)^2$ yields the result.

Moments of a distribution

Definition 90. *If X has a PMF $f(x)$ such that $\sum_x |x|^r f(x) < \infty$, then*

(a) the rth **moment** *of X is $\mathrm{E}(X^r)$;*
(b) the rth **central moment** *of X is $\mathrm{E}[\{X - \mathrm{E}(X)\}^r]$;*
(c) the **variance** *of X is $\mathrm{var}(X) = \mathrm{E}[\{X - \mathrm{E}(X)\}^2]$ (the second central moment);*
(d) the **standard deviation** *of X is defined as $\sqrt{\mathrm{var}(X)}$ (non-negative);*
(e) the rth **factorial moment** *of X is $\mathrm{E}\{X(X-1)\cdots(X-r+1)\}$.*

Remarks:

- $\mathrm{E}(X)$ and $\mathrm{var}(X)$ are the most important moments: they represent the 'average value' $\mathrm{E}(X)$ of X, and the 'average squared distance' of X from its mean, $\mathrm{E}(X)$.
- The variance is analogous to the **moment of inertia** in mechanics: it measures the scatter of X around its mean, $\mathrm{E}(X)$, with small variance corresponding to small scatter, and conversely.
- The expectation and standard deviation have the same units (kg, m, ...) as X.

Example 91. *Calculate the expectation and variance of the score when we roll a die.*

Note to Example 91

Now X takes values $1, \ldots, 6$ with equal probabilities $1/6$. Obviously $\mathrm{E}(|X|) < \infty$, and $\mathrm{E}(X) = (1 + \cdots + 6)/6 = 21/6 = 7/2$. The variance is

$$\mathrm{E}[\{X - \mathrm{E}(X)\}^2] = \sum_{x=1}^{6} \frac{1}{6}(x - 7/2)^2 = \tfrac{2}{6} \times \tfrac{1}{4} \times (1 + 9 + 25) = 35/12.$$

Properties of the variance

Theorem 92. *Let X be a random variable whose variance exists, and let a, b be constants. Then*

$$\begin{aligned} \mathrm{var}(X) &= \mathrm{E}(X^2) - \mathrm{E}(X)^2 = \mathrm{E}\{X(X-1)\} + \mathrm{E}(X) - \mathrm{E}(X)^2; \\ \mathrm{var}(aX+b) &= a^2\mathrm{var}(X); \\ \mathrm{var}(X) = 0 &\Rightarrow X \text{ is constant with probability 1.} \end{aligned}$$

- The first of these formulae expresses the variance in terms of either the ordinary moments, or the factorial moments. Usually the first is more useful, but occasionally the second can be used.
- The second formula shows that the variance does not change if X is shifted by a fixed quantity b, but the dispersion is increased by the square of a multiplier a.
- The third shows that the variance is appropriately named: if X has zero variance, then it does not vary.

Example 93. *Calculate the variance of a Poisson random variable.*

Note to Theorem 92

(a) Just expand, use linearity of E, and simplify.
(b) Ditto.
(c) If we write $\mathrm{E}(X) = \mu$ and

$$\mathrm{var}(X) = \mathrm{E}[\{X - \mathrm{E}(X)\}^2] = \mathrm{E}[\{X-\mu\}^2] = \sum_x f(x)(x-\mu)^2 = 0,$$

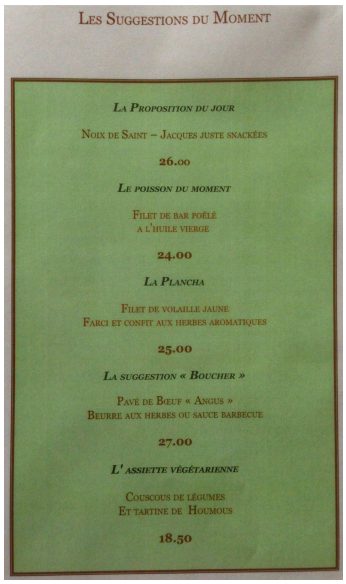
then for each $x \in D_X$, either $x = \mu$ or $f(x) = 0$. Suppose that $f(a), f(b) > 0$ and $a \neq b$. Then if $\mathrm{var}(X) = 0$, we must have $a = \mu = b$, which is a contradiction. Therefore $f(x) > 0$ for a unique value of x, and then we must have $f(x) = 1$, so $\mathrm{P}(X = x) = 1$ and $(x-\mu)^2 = 0$; thus $\mathrm{P}(X = \mu) = f_X(\mu) = 1$.

Note to Example 93

By recalling Example 88, we find

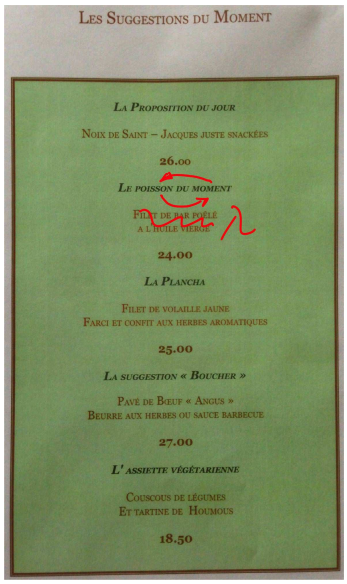
$$\mathrm{var}(X) = \mathrm{E}\{X(X-1)\} + \mathrm{E}(X) - \mathrm{E}(X)^2 = \lambda^2 + \lambda - \lambda^2 = \lambda.$$

Poisson du moment



(Source: Copernic)

Moment du Poisson



(Source: Copernic)

Properties of the variance II

Theorem 94. *If X takes its values in $\{0, 1, \ldots\}$, $r \geq 2$, and $\mathrm{E}(X) < \infty$, then*

$$\begin{aligned} \mathrm{E}(X) &= \sum_{x=1}^{\infty} \mathrm{P}(X \geq x), \\ \mathrm{E}\{X(X-1)\cdots(X-r+1)\} &= r\sum_{x=r}^{\infty}(x-1)\cdots(x-r+1)\mathrm{P}(X \geq x). \end{aligned}$$

Example 95. *Let $X \sim \mathrm{Geom}(p)$. Calculate $\mathrm{E}(X)$ and $\mathrm{var}(X)$.*

Example 96. *Each packet of a certain product has equal chances of containing one of n different types of tokens, independently of each other packet. What is the expected number of packets you will need to buy in order to get at least one of each type of token?*

Note to Theorem 94

☐ The first part of this is

$$\mathrm{E}(X) = \sum_{x=1}^{\infty} xf(x) = \sum_{x=1}^{\infty} \mathrm{P}(X = x)\sum_{r=1}^{x} 1 = \sum_{x=1}^{\infty} \mathrm{P}(X \geq x),$$

as follows on changing the order of summation, noting that since all the terms are positive, this is a legal operation.

☐ The second part is proved in the same way, first writing

$$r(x-1)\cdots(x-r+1) = r!\frac{(x-1)!}{(r-1)!(x-r)!} = r!\binom{x-1}{r-1}.$$

Then we write

$$r\sum_{x=r}^{\infty}(x-1)\cdots(x-r+1)\mathrm{P}(X \geq x) = \sum_{x=r}^{\infty} r!\binom{x-1}{r-1}\sum_{y=x}^{\infty} f_X(y) = \sum_{y=r}^{\infty} f_X(y) r!\sum_{x=r}^{y}\binom{x-1}{r-1},$$

and use Pascal's triangle (Theorem 17) to find that

$$r!\sum_{x=r}^{y}\binom{x-1}{r-1} = r!\sum_{x=r}^{y}\left\{\binom{x}{r} - \binom{x-1}{r}\right\} = r!\binom{y}{r} = y(y-1)\cdots(y-r+1)$$

after cancellations. As required, this gives

$$\sum_{y=r}^{\infty} f_X(y)y(y-1)\cdots(y-r+1) = \mathrm{E}\{X(X-1)\cdots(X-r+1)\}.$$

Note to Example 95

In this case $X \in \{1, 2, \ldots\}$, and Theorem 94 yields

$$\mathrm{E}(X) = \sum_{x=1}^{\infty} (1-p)^{x-1} = \frac{1}{1-(1-p)} = 1/p \geq 1.$$

For the variance, note that the second part of Theorem 94, with $r = 2$, gives

$$\begin{aligned} \mathrm{E}\{X(X-1)\} &= 2\sum_{x=2}^{\infty} (x-1)(1-p)^{x-1} \\ &= 2(1-p)\frac{d}{dp}\left\{-\sum_{x=1}^{\infty}(1-p)^{x-1}\right\} \\ &= 2(1-p)\frac{d}{dp}(-1/p) = 2(1-p)/p^2. \end{aligned}$$

Hence the variance is

$$\mathrm{var}(X) = \mathrm{E}\{X(X-1)\} + \mathrm{E}(X) - \mathrm{E}(X)^2 = 2(1-p)/p^2 + 1/p - 1/p^2 = (1-p)/p^2.$$

This gets smaller as $p \to 1$, and larger as $p \to 0$, as expected.

Note to Example 96

This can be represented as $X_1 + X_2 + \cdots + X_n$, where X_1 is the number of packets to the first token, then X_2 is the number of packets to the next different token (i.e., not the first), etc. Thus the X_r are independent geometric variables with probabilities $p = n/n, (n-1)/n, \ldots, 1/n$. Hence the expectation is $n(1 + 1/2 + 1/3 + \cdots + 1/n) \sim n \log n$, which $\to \infty$ as $n \to \infty$.

Conditional probability distributions

Definition 97. *Let $(\Omega, \mathcal{F}, \mathrm{P})$ be a probability space, on which we define a random variable X, and let $B \in \mathcal{F}$ with $\mathrm{P}(B) > 0$. Then the* conditional probability mass function *of X given B is*

$$f_X(x \mid B) = \mathrm{P}(X = x \mid B) = \mathrm{P}(A_x \cap B)/\mathrm{P}(B),$$

where $A_x = \{\omega \in \Omega : X(\omega) = x\}$.

Theorem 98. *The function $f_X(x \mid B)$ satisfies*

$$f_X(x \mid B) \geq 0, \quad \sum_x f_X(x \mid B) = 1,$$

and is thus a well-defined mass function.

Often B is an event of form $X \in \mathcal{B}$, for some $\mathcal{B} \subset \mathbb{R}$, and then

$$f_X(x \mid B) = \frac{\mathrm{P}(X = x, X \in \mathcal{B})}{\mathrm{P}(X \in \mathcal{B})} = \frac{\mathrm{P}(X \in \mathcal{B} \mid X = x)\mathrm{P}(X = x)}{\mathrm{P}(X \in \mathcal{B})} = \frac{I(x \in \mathcal{B})}{\mathrm{P}(X \in \mathcal{B})} f_X(x),$$

so $f_X(x \mid B) = 0$ $(x \notin \mathcal{B})$ and $f_X(x \mid B) \propto f_X(x)$ $(x \in \mathcal{B})$, rescaled to have unit probability.

Example 99. *Calculate the conditional PMFs of $X \sim \mathrm{Geom}(p)$, (a) given that $X > n$, (b) given that $X \leq n$.*

Note to Theorem 98

We need to check the two properties of a distribution function.

- ☐ Non-negativity is obvious because the $f_X(x \mid B) = \mathrm{P}(X = x \mid B)$ are conditional probabilities.
- ☐ Now $A_x \cap A_y = \emptyset$ if $x \neq y$, and $\bigcup_{x \in \mathbb{R}} A_x = \Omega$. Hence the A_x partition $\mathbb{R}$, and thus

$$\sum_x f_X(x \mid B) = \sum_x \mathrm{P}(A_x \cap B)/\mathrm{P}(B) = \mathrm{P}(B)/\mathrm{P}(B) = 1.$$

Note to Example 99

(a) The event $B_1 = \{X > n\}$ has probability $(1-p)^n$, so the new mass function is

$$f_X(x \mid B_1) = \frac{\mathrm{P}(X = x \cap X > n)}{\mathrm{P}(X > n)} = \frac{f_X(x) I(x > n)}{\mathrm{P}(X > n)} = p(1-p)^{x-n-1}, \quad x = n+1, n+2, \ldots.$$

This implies that conditional on $X > n$, $X - n$ has the same distribution as did X originally.
(b) The event $B_2 = B_1^c = \{X \leq n\}$ has probability $1 - (1-p)^n$, so the new mass function is

$$f_X(x \mid B_2) = \frac{\mathrm{P}(X = x \cap X \leq n)}{\mathrm{P}(X \leq n)} = \frac{f_X(x) I(x \leq n)}{1 - (1-p)^n} = \frac{p(1-p)^{x-1}}{1 - (1-p)^n}, \quad x = 1, \ldots, n.$$

Conditional expected value

Definition 100. *Suppose that* $\sum_x |g(x)| f_X(x \mid B) < \infty$. *Then the conditional expected value of* $g(X)$ *given* B *is*

$$\mathrm{E}\{g(X) \mid B\} = \sum_x g(x) f_X(x \mid B).$$

Theorem 101. *Let* X *be a random variable with expected value* $\mathrm{E}(X)$ *and let* B *be an event with* $\mathrm{P}(B), \mathrm{P}(B^c) > 0$. *Then*

$$\mathrm{E}(X) = \mathrm{E}(X \mid B)\mathrm{P}(B) + \mathrm{E}(X \mid B^c)\mathrm{P}(B^c).$$

More generally, when $\{B_i\}_{i=1}^{\infty}$ *is a partition of* Ω, $\mathrm{P}(B_i) > 0$ *for all* i, *and the sum is absolutely convergent,*

$$\mathrm{E}(X) = \sum_{i=1}^{\infty} \mathrm{E}(X \mid B_i)\mathrm{P}(B_i).$$

Note to Theorem 101

We prove the second part, of which the first is a special case. The total probability theorem, Theorem 43, gives

$$f(x) = \mathrm{P}(X = x) = \sum_{i=1}^{\infty} \mathrm{P}(X = x \mid B_i)\mathrm{P}(B_i) = \sum_{i=1}^{\infty} f(x \mid B_i)\mathrm{P}(B_i),$$

and this gives

$$\mathrm{E}(X) = \sum_x x f(x) = \sum_x x \sum_{i=1}^{\infty} f(x \mid B_i)\mathrm{P}(B_i) = \sum_{i=1}^{\infty} \left\{ \sum_x x f(x \mid B_i) \right\} \mathrm{P}(B_i) = \sum_{i=1}^{\infty} \mathrm{E}(X \mid B_i)\mathrm{P}(B_i),$$

as required. The first part follows on setting $B_1 = B$, $B_2 = B^c$, $B_3 = B_4 = \cdots = \emptyset$.

Example

Example 102. *Calculate the expected values for the distributions in Example 99.*

Note to Example 99

(a) Since

$$f_X(x \mid B_1) = p(1-p)^{x-n-1}, \quad x = n+1, n+2, \ldots,$$

we have

$$\mathrm{E}(X \mid B_1) = \sum_{x=n+1}^{\infty} xp(1-p)^{x-n-1} = \sum_{y=1}^{\infty} (n+y)p(1-p)^{y-1}$$

where we have set $y = x - n$, and hence

$$\mathrm{E}(X \mid B_1) = n\sum_{y=1}^{\infty} p(1-p)^{y-1} + \sum_{y=1}^{\infty} yp(1-p)^{y-1} = n + 1/p,$$

since the first sum equals unity and the second is the expectation of a Geom(p) variable.
(b) We can tackle this directly using the expression

$$\mathrm{E}(X \mid B_2) = \sum_{x=1}^{n} x\frac{p(1-p)^{x-1}}{1-(1-p)^n}$$

or indirectly by writing $B = B_1$ and $B^c = B_2$, which are complementary events, and using Theorem 101:

$$\mathrm{E}(X) = \mathrm{E}(X \mid B_1)\mathrm{P}(B_1) + \mathrm{E}(X \mid B_2)\mathrm{P}(B_2),$$

giving

$$1/p = (n+1/p)(1-p)^n + \mathrm{E}(X \mid B_2)\{1-(1-p)^n\},$$

and a little algebra yields

$$\mathrm{E}(X \mid B_2) = \frac{1/p - (n+1/p)(1-p)^n}{1-(1-p)^n}.$$

3.4 Notions of Convergence

Convergence of distributions

We often want to approximate one distribution by another. The mathematical basis for doing so is the convergence of distributions.

Definition 103. *Let $\{X_n\}$, X be random variables whose cumulative distribution functions are $\{F_n\}$, F. Then we say that the random variables $\{X_n\}$* **converge in distribution** *(or* **converge in law***) to X, if, for all $x \in \mathbb{R}$ where F is continuous,*

$$F_n(x) \to F(x), \quad n \to \infty.$$

We write $X_n \xrightarrow{D} X$.

If $D_X \subset \mathbb{Z}$, then $F_n(x) \to F(x)$ if $f_n(x) \to f(x)$ for all x, $n \to \infty$.

Law of small numbers

Recall from Theorem 17 that $n^{-r}\binom{n}{r} \to 1/r!$ for all $r \in \mathbb{N}$, when $n \to \infty$.

Theorem 104 (Law of small numbers). *Let $X_n \sim B(n, p_n)$, and suppose that $np_n \to \lambda > 0$ when $n \to \infty$. Then $X_n \xrightarrow{D} X$, where $X \sim \mathrm{Pois}(\lambda)$.*

Theorem 104 can be used to approximate binomial probabilities for large n and small p by Poisson probabilities.

Example 105. *In Example 47 we saw that the probability of having exactly r fixed points in a random permutation of n objects is*

$$\frac{1}{r!}\sum_{k=0}^{n-r}\frac{(-1)^k}{k!} \to \frac{e^{-1}}{r!}, \quad r = 0, 1, \ldots, \quad n \to \infty,$$

Thus the number of fixed points has a limiting $\mathrm{Pois}(1)$ distribution.

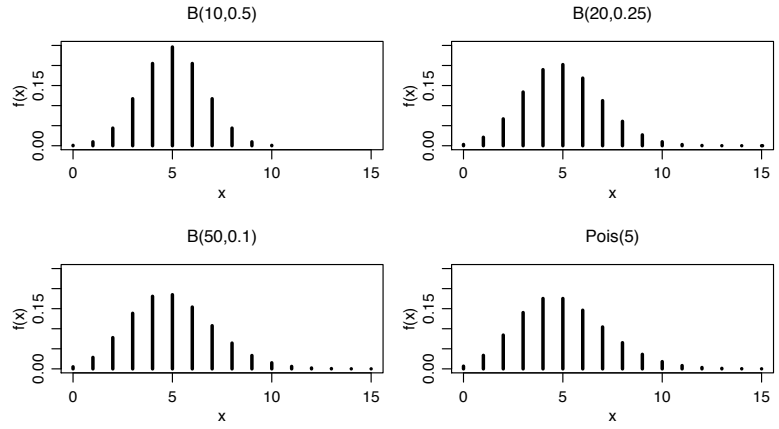
Note to Theorem 104

For any fixed r we have

$$\binom{n}{r}p_n^r(1-p_n)^{n-r} = n^{-r}\binom{n}{r} \times (np_n)^r(1-np_n/n)^{n-r} \to \frac{1}{r!}\lambda^r e^{-\lambda}, \quad n \to \infty,$$

which is the required Poisson mass function; call this limiting Poisson random variable X. This convergence implies that $\mathrm{P}(X_n \le x) \to \mathrm{P}(X \le x)$ for any fixed real x, since $\mathrm{P}(X_n \le x)$ is just then a finite sum of probabilities, each of which is converging to the limiting Poisson probability.

Law of small numbers



Mass functions of three binomial distributions and the Poisson distribution, all with expectation 5.

Numerical comparison

Example 106 (Binomial and Poisson distributions). *Compare* $\mathrm{P}(X \leq 3)$ *for* $X \sim B(20, p)$, *with* $p = 0.05, 0.1, 0.2, 0.5$ *with the results from a Poisson approximation,* $\mathrm{P}(X' \leq 3)$, *with* $X' \sim \mathrm{Pois}(np)$, *using the functions* `pbinom` *and* `ppois` *in the software* `R` *— see*

`http://www.r-project.org/`

Thus for example we have:

```
> pbinom(3,size=20,prob=0.05)  # Binomial prob, Pr(X <= 3)
[1] 0.9840985
> ppois(3,lambda=20*0.05)      # Poisson approx, Pr(X' <= 3)
[1] 0.9810118
```


People versus Collins

Example 107. *In 1964 a handbag was stolen in Los Angeles by a young woman with blond hair in a pony tail. The thief disappeared, but soon afterwards she was spotted in a yellow car with a bearded black man with a moustache. The police then arrested a woman called Janet Collins, who matched the description, and had a black bearded friend with a moustache, who drove a yellow car.*

Due to a lack of evidence and of reliable witnesses, the prosecutor tried to convince the jury that Collins and her friend were the only pair in Los Angeles who could have committed the crime. He found a probability of $p = 1/(12 \times 10^6)$ *that a couple picked at random should fit the description, and they were convicted.*

In a higher court it was argued that the number of couples X *fitting the description must follow a Poisson distribution with* $\lambda = np$, *where* n *is the size of the population to which the couple belong. To be certain that the couple were guilty,* $\mathrm{P}(X > 1 \mid X \geq 1)$ *must be very small. But with* $n = 10^6$, 2×10^6, 5×10^6, 10×10^6, *these probabilities are* 0.041, 0.081, 0.194, 0.359*: it was therefore very far from certain that they were guilty. They were finally cleared.*

Note to Example 107

Here the law of small numbers applies, so

$$\mathrm{P}(X > 1 \mid X \geq 1) = 1 - \mathrm{P}(X = 1 \mid X \geq 1) = 1 - \lambda e^{-\lambda}/(1 - e^{-\lambda}) = 1 - \lambda/(e^{\lambda} - 1),$$

with Poisson parameter $\lambda = np = 1/12$, 1/6, 5/12 and 1 respectively. Calculation gives the required numbers. In fact here X has a truncated Poisson distribution.

Example

Example 108. *Let X_N be a hypergeometric variable, then*

$$\mathrm{P}(X_N = x) = \frac{\binom{m}{x}\binom{N-m}{n-x}}{\binom{N}{n}}, \quad x = \max(0, m+n-N), \ldots, \min(m, n).$$

This is the distribution of the number of white balls obtained when we take a random sample of size n without replacement from an urn containing m white balls and $N-m$ black balls. Show that when $N, m \to \infty$ in such a way that $m/N \to p$, where $0 < p < 1$,

$$\mathrm{P}(X_N = x) \to \binom{n}{x} p^x (1-p)^{n-x}, \quad i = 0, \ldots, n.$$

Hence the limiting distribution of X_N is $B(n, p)$.

Note to Example 108

We apply the last part of Theorem 17, writing

$$\begin{aligned} \frac{\binom{m}{x}\binom{N-m}{n-x}}{\binom{N}{n}} &= \frac{m^x (N-m)^{n-x}}{N^n} \times \frac{m^{-x}\binom{m}{x}(N-m)^{-(n-x)}\binom{N-m}{n-x}}{N^{-n}\binom{N}{n}} \\ &\to p^x (1-p)^{n-x} \times \frac{n!}{x!(n-x)!}, \quad N \to \infty, \end{aligned}$$

under the terms of the theorem, as required.

Which distribution?

We have encountered several distributions: Bernoulli, binomial, geometric, negative binomial, hypergeometric, Poisson—how to choose? Here is a little algorithm to help your reasoning:

Is X based on independent trials (0/1) with a same probability p, or on draws from a finite population, with replacement?

- ☐ If **Yes**, is the total number of trials n fixed, so $X \in \{0, \ldots, n\}$?
  - If **Yes**: use the **binomial** distribution, $X \sim B(n, p)$ (and thus the **Bernoulli** distribution if $n = 1$).
    - ▷ If $n \approx \infty$ or $n \gg np$, we can use the **Poisson** distribution, $X \sim \mathrm{Pois}(np)$.
  - If **No**, then $X \in \{n, n+1, \ldots\}$, and we use the **geometric** (if X is the number of trials until one success) or **negative binomial** (if X is the number of trials until the last of several successes) distributions.
- ☐ If **No**, then if the draw is independent but without replacement from a finite population, then $X \sim$ **hypergeometric** distribution.

There are many more distributions, and we may choose a distribution on empirical grounds. The following map comes from Leemis and McQueston (2008, American Statistician) ...

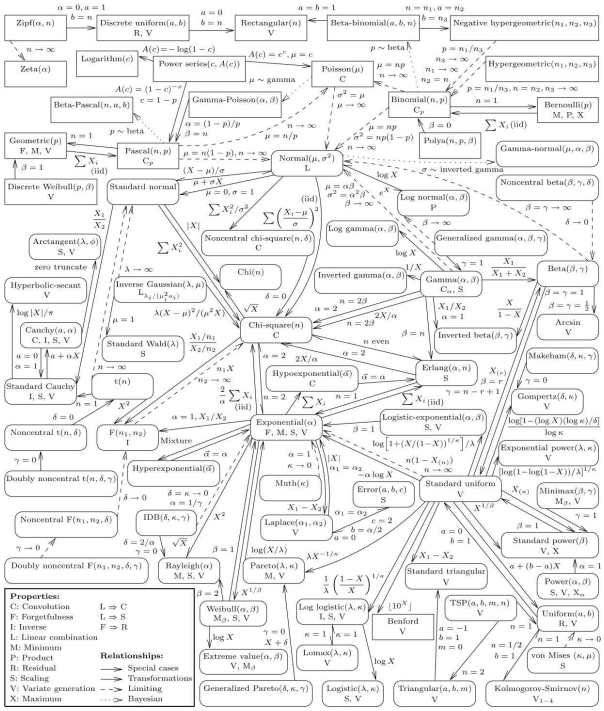


Figure 1. Univariate distribution relationships.

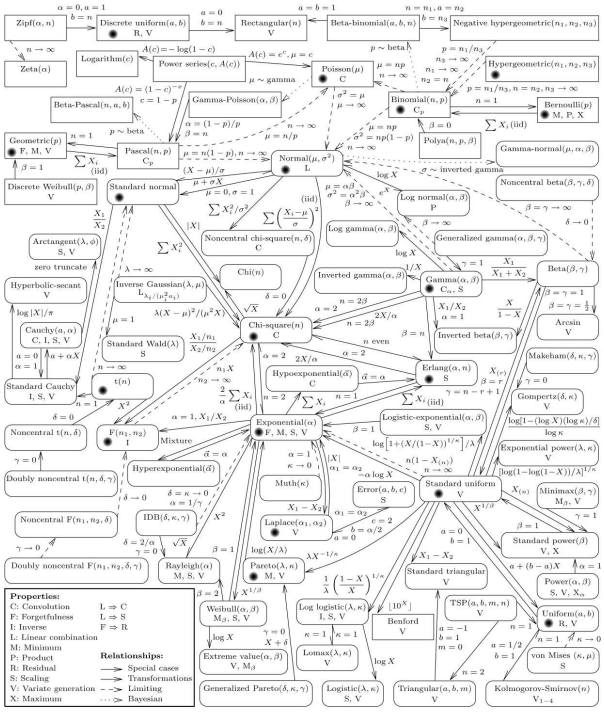


Figure 1. Univariate distribution relationships.

4 Continuous Random Variables

slide 135

4.1 Basic Ideas

slide 136

Continuous random variables

In many situations, we must work with continuous variables:

- ☐ the time until the end of the lecture $\in (0, 45)$ min;
- ☐ the pair (height, weight) $\in (0, \infty)^2$.

Until now we supposed that the support

$$D_X = \{x \in \mathbb{R} : X(\omega) = x, \omega \in \Omega\}$$

of X is countable, so X is a discrete random variable. We suppose now that D_X is not countable, which implies also that Ω itself is not countable.

Definition 109 (Reminder). *Let $(\Omega, \mathcal{F}, \mathrm{P})$ be a probability space. The* **cumulative distribution function** *of a rv X defined on $(\Omega, \mathcal{F}, \mathrm{P})$ is*

$$F(x) = \mathrm{P}(X \leq x) = \mathrm{P}(\mathcal{B}_x), \quad x \in \mathbb{R},$$

where $\mathcal{B}_x = \{\omega : X(\omega) \leq x\} \subset \Omega$.

Probability density functions

Definition 110. *A random variable X is* **continuous** *if there exists a function $f(x)$, called the* **probability density function (or density) (PDF)** *of X, such that*

$$\mathrm{P}(X \leq x) = F(x) = \int_{-\infty}^{x} f(u)\, du, \quad x \in \mathbb{R}.$$

The properties of F imply that (i) $f(x) \geq 0$, and (ii) $\int_{-\infty}^{\infty} f(x)\, dx = 1$.

Remarks:

- ☐ Evidently,
$$f(x) = \frac{dF(x)}{dx}.$$
- ☐ Since $\mathrm{P}(x < X \leq y) = \int_x^y f(u)\, du$ for $x < y$, for all $x \in \mathbb{R}$,
$$\mathrm{P}(X = x) = \lim_{y \downarrow x} \mathrm{P}(x < X \leq y) = \lim_{y \downarrow x} \int_x^y f(u)\, du = \int_x^x f(u)\, du = 0.$$
- ☐ If X is discrete, then its PMF $f(x)$ is often also called its density function.

Motivation

We study continuous random variables for several reasons:

- ☐ they appear in simple but powerful models—for example, the **exponential** distribution often represents the waiting time in a process where events occur completely at random;
- ☐ they give simple but very useful approximations for complex problems—for example, the **normal** distribution appears as an approximation for the distribution of an average, under fairly general conditions;
- ☐ they are the basis for modelling complex problems either in probability or in statistics—for example, the **Pareto** distribution is often a good approximation for heavy-tailed data, in finance and for the internet.

We will discuss a few well-known distributions, but there are plenty more (see map at the end of Chapter 3) ....

Basic distributions

Definition 111 (Uniform distribution). *The random variable U having density*

$$f(u) = \begin{cases} \frac{1}{b-a}, & a \le u \le b, \\ 0, & \textit{otherwise}, \end{cases} \qquad a < b,$$

is called a **uniform random variable**. *We write* $U \sim U(a, b)$.

Definition 112 (Exponential distribution). *The random variable X having density*

$$f(x) = \begin{cases} \lambda e^{-\lambda x}, & x > 0, \\ 0, & \textit{otherwise}, \end{cases}$$

is called an **exponential random variable** *with parameter $\lambda > 0$. We write* $X \sim \exp(\lambda)$.

In practice random variables are almost always either discrete or continuous, with exceptions such as daily rain totals.

Example 113. *Find the cumulative distribution functions of the uniform and exponential distributions, and establish the* **lack of memory** *(or* **memorylessness***) property of X:*

$$\mathrm{P}(X > x + t \mid X > t) = \mathrm{P}(X > x), \quad t, x > 0.$$

Note to Example 113

Integration of the uniform density gives

$$F(u) = \begin{cases} 0, & u \le a, \\ (u-a)/(b-a), & a < u \le b, \\ 1, & u > b. \end{cases}$$

Sketch the density and the CDF.
Integration of the exponential density gives

$$F(x) = \begin{cases} 0, & x \le 0, \\ 1 - \exp(-\lambda x), & x > 0. \end{cases}$$

Draw the density and the CDF.
For the lack of memory of the exponential distribution, note that

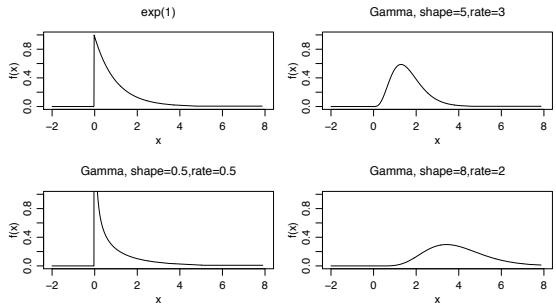
$$\mathrm{P}(X > x + t \mid X > t) = \frac{\mathrm{P}(X > x+t)}{\mathrm{P}(X > t)} = \frac{\exp\{-\lambda(x+t)\}}{\exp(-\lambda t)} = \exp(-\lambda x), \quad x > 0.$$

Gamma distribution

Definition 114 (Gamma distribution). *The random variable X having density*

$$f(x) = \begin{cases} \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x}, & x > 0, \\ 0, & \textit{otherwise}, \end{cases}$$

is called a **gamma random variable** *with parameters $\alpha, \lambda > 0$; we write $X \sim \mathrm{Gamma}(\alpha, \lambda)$. Here α is called the* **shape parameter** *and λ is called the* **rate**, *with λ^{-1} the* **scale parameter**. *By letting $\alpha = 1$ we get the exponential density, and when $\alpha = 2, 3, \ldots$ we get the* **Erlang** *density. Slide 99 gives the properties of $\Gamma(\cdot)$.*



Laplace distribution

Definition 115 (Laplace). *The random variable X having density*

$$f(x) = \frac{\lambda}{2} e^{-\lambda|x-\eta|}, \quad x \in \mathbb{R}, \quad \eta \in \mathbb{R}, \lambda > 0,$$

is called a **Laplace random variable** *(or sometimes a* **double exponential***) random variable.*

(Source: http://www-history.mcs.st-and.ac.uk/PictDisplay/Laplace.html)
Pierre-Simon Laplace (1749–1827): **Théorie Analytique des Probabilités** (1814)
According to Napoleon Bonaparte: 'Laplace did not consider any question from the right angle: he sought subtleties everywhere, conceived only problems, and brought the spirit of "infinitesimals" into the administration.'

Pareto distribution

Definition 116 (Pareto). *The random variable X with cumulative distribution function*

$$F(x) = \begin{cases} 0, & x < \beta, \\ 1 - \left(\frac{\beta}{x}\right)^{\alpha}, & x \geq \beta, \end{cases}, \quad \alpha, \beta > 0,$$

is called a **Pareto random variable**.

Vilfredo Pareto (1848–1923): Professor at Lausanne University, father of economic science.
(Source: http://www.gametheory.net/dictionary/People/VilfredoPareto.html)

Example 117. *Find the cumulative distribution function of the Laplace distribution, and the probability density function of the Pareto distribution.*

Note to Example 117

For the Laplace distribution, integration of the density gives

$$F(x) = \begin{cases} \frac{1}{2}e^{-\lambda|x-\eta|}, & x \leq \eta, \\ 1 - \frac{1}{2}e^{-\lambda|x-\eta|}, & x > \eta. \end{cases}$$

Note that $F(\eta) = 1/2$, so η is the **median** of the distribution.
Sketch the density and the CDF.
For the Pareto density, just differentiate with respect to x to obtain the density function,

$$f(x) = \begin{cases} 0, & x < \beta, \\ \frac{\alpha\beta^{\alpha}}{x^{\alpha+1}}, & x \geq \beta. \end{cases}$$

Moments

Definition 118. *Let $g(x)$ be a real-valued function, and X a continuous random variable of density $f(x)$. Then if $\mathrm{E}\{|g(X)|\} < \infty$, we define the* **expectation** *of $g(X)$ to be*

$$\mathrm{E}\{g(X)\} = \int_{-\infty}^{\infty} g(x)f(x)\,dx.$$

In particular the **expectation** *and the* **variance** *of X are*

$$\begin{aligned} \mathrm{E}(X) &= \int_{-\infty}^{\infty} xf(x)\,dx, \\ \mathrm{var}(X) &= \int_{-\infty}^{\infty} \{x - \mathrm{E}(X)\}^2 f(x)\,dx = \mathrm{E}(X^2) - \mathrm{E}(X)^2. \end{aligned}$$

Example 119. *Calculate the expectation and the variance of the following distributions: (a) $U(a, b)$; (b) gamma; (c) Pareto.*

Note to Example 119

(a) Note that we need to compute $\mathrm{E}(U^r)$ for $r = 1, 2$, and this is $\frac{1}{r+1}(b^{r+1} - a^{r+1})/(b-a)$. Hence $\mathrm{E}(X) = \frac{1}{2}(b^2 - a^2)/(b-a) = (b+a)/2$, as expected. For the variance, note that

$$\mathrm{E}(X^2) - \mathrm{E}(X)^2 = \tfrac{1}{3}\frac{b^3 - a^3}{b-a} - (b+a)^2/4 = \tfrac{1}{3}(b^2 + ab + a^2) - (b^2 + 2ab + a^2)/4 = (b-a)^2/12.$$

(b) In this case

$$\begin{aligned} \mathrm{E}(X^r) &= \int_0^\infty x^r \times \lambda^\alpha x^{\alpha-1}\Gamma(\alpha)^{-1}\exp(-\lambda x)\, dx \\ &= \lambda^{-r}\Gamma(\alpha)^{-1}\int_0^\infty u^{r+\alpha-1}e^{-u}\, du \\ &= \lambda^{-r}\Gamma(r+\alpha)/\Gamma(\alpha). \end{aligned}$$

Properties of the gamma function (slide 99) give

$$\mathrm{E}(X) = \alpha/\lambda, \quad \mathrm{E}(X^2) = \alpha(\alpha+1)/\lambda^2, \quad \mathrm{var}(X) = \mathrm{E}(X^2) - \mathrm{E}(X)^2 = \alpha/\lambda^2.$$

(c) The expectation is

$$\mathrm{E}(X^r) = \int_\beta^\infty \alpha\beta^\alpha x^{r-\alpha-1}\, dx = \alpha\beta^r/(\alpha - r)$$

provided that $\alpha > r$. If $\alpha \le r$ then the moment does not exist. In particular, $\mathrm{E}(X) < \infty$ only if $\alpha > 1$.

Conditional densities

We can also calculate conditional cumulative distribution and density functions: for reasonable subsets $\mathcal{A} \subset \mathbb{R}$ we have

$$F_X(x \mid X \in \mathcal{A}) = \mathrm{P}(X \le x \mid X \in \mathcal{A}) = \frac{\mathrm{P}(X \le x \cap X \in \mathcal{A})}{\mathrm{P}(X \in \mathcal{A})} = \frac{\int_{\mathcal{A}_x} f(y)\, dy}{\mathrm{P}(X \in \mathcal{A})},$$

where $\mathcal{A}_x = \{y : y \le x, y \in \mathcal{A}\}$, and

$$f_X(x \mid X \in \mathcal{A}) = \begin{cases} \frac{f_X(x)}{\mathrm{P}(X \in \mathcal{A})}, & x \in \mathcal{A}, \\ 0, & \text{otherwise.} \end{cases}$$

With $I(X \in \mathcal{A})$ the indicator variable of the event $X \in \mathcal{A}$, we can write

$$\mathrm{E}\{g(X) \mid X \in \mathcal{A}\} = \frac{\mathrm{E}\{g(X)\ I(X \in \mathcal{A})\}}{\mathrm{P}(X \in \mathcal{A})},$$

Example 120. *Let $X \sim \exp(\lambda)$. Find the density and the cumulative distribution function of X, given that $X > 3$.*

Note to Example 120

With $\mathcal{A} = (3, \infty)$, we have $\mathrm{P}(X \in \mathcal{A}) = \exp(-3\lambda)$. Hence

$$F_X(x \mid X \in \mathcal{A}) = \begin{cases} 0, & x < 3, \\ \frac{\exp(-3\lambda) - \exp(-\lambda x)}{\exp(-3\lambda)}, & x \geq 3, \end{cases}$$

and the formula here reduces to $1 - \exp\{-(x-3)\lambda\}$, $x > 3$. This is just the exponential density, shifted along to $x = 3$. There is a close relation to the lack of memory property.

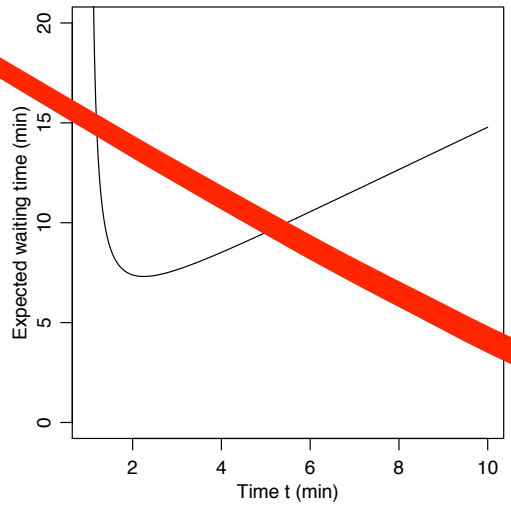
Example

Example 121. *To get a visa for a foreign country, you call its consulate every morning at 10 am. On any given day the civil servant is only there to answer telephone calls with probability 1/2, and when he does answer, he lets the phone ring for a random amount of time T (min) whose distribution is*

$$F_T(t) = \begin{cases} 0, & t \leq 1, \\ 1 - t^{-1}, & t > 1. \end{cases}$$

(a) If you call one morning and don't hang up, what is the probability that you will listen to the ringing tone for at least s minutes?
(b) You decide to call once every day, but to hang up if there has been no answer after s^* minutes. Find the value of s^* which minimises your time spent listening to the ringing tone.

Waiting time in Example 121



X discrete or continuous?

	Discrete	Continuous
Support D_X	countable	contains an interval $(x_-, x_+) \subset \mathbb{R}$
f_X	mass function dimensionless $0 \le f_X(x) \le 1$ $\sum_{x\in\mathbb{R}} f_X(x) = 1$	density function units $[x]^{-1}$ $0 \le f_X(x)$ $\int_{-\infty}^{\infty} f_X(x)\,dx = 1$
$F_X(a) = \mathrm{P}(X \le a)$	$\sum_{x\le a} f_X(x)$	$\int_{-\infty}^{a} f_X(x)\,dx$
$\mathrm{P}(X \in \mathcal{A})$	$\sum_{x\in\mathcal{A}} f_X(x)$	$\int_{\mathcal{A}} f_X(x)\,dx$
$\mathrm{P}(a < X \le b)$	$\sum_{\{x:a<x\le b\}} f_X(x)$	$\int_a^b f_X(x)\,dx$
$\mathrm{P}(X = a)$	$f_X(a) \ge 0$	$\int_a^a f_X(x)\,dx = \mathbf{0}$
$\mathrm{E}\{g(X)\}$ (if well defined)	$\sum_{x\in\mathbb{R}} g(x) f_X(x)$	$\int_{-\infty}^{\infty} g(x) f_X(x)\,dx$

4.2 Further Ideas

Quantiles

Definition 122. *Let $0 < p < 1$. We define the p* **quantile** *of the cumulative distribution function $F(x)$ to be*

$$x_p = \inf\{x : F(x) \ge p\}.$$

For most continuous random variables, x_p is unique and equals $x_p = F^{-1}(p)$, where F^{-1} is the inverse function F; then x_p is the value for which $\mathrm{P}(X \le x_p) = p$. In particular, we call the 0.5 quantile the **median** *of F.*

Example 123. *Let $X \sim \exp(\lambda)$. Show that $x_p = -\lambda^{-1}\log(1-p)$.*

Example 124. *Find the p quantile of the Pareto distribution.*

The infimum is needed when there are jumps in the distribution function, or when it is flat over some interval. Here is an example:

Example 125. *Compute $x_{0.5}$ and $x_{0.9}$ for a Bernoulli random variable with $p = 1/2$.*

Note to Example 123

We have to solve $F(x_p) = 1 - \exp(-\lambda x_p) = p$, which gives the required result.

Note to Example 124

We have to solve $F(x_p) = 1 - (\beta/x_p)^{\alpha} = p$, which gives $x_p = \beta(1-p)^{-1/\alpha}$.

Note to Example 125

Recall that in this case

$$F(x) = \begin{cases} 0, & x < 0, \\ 1/2, & 0 \le x < 1, \\ 1, & x \ge 1. \end{cases}$$

There is no value of x such that $F(x) = 0.9$, but $F(x) \ge 0.9$ for every $x \ge 1$, so

$$x_{0.9} = \inf\{x : F(x) \ge 0.9\} = \inf\{x : x \ge 1\} = 1.$$

Likewise

$$x_{0.5} = \inf\{x : F(x) \ge 0.5\} = \inf\{x : x \ge 0\} = 0.$$

Transformations

We often consider $Y = g(X)$, where g is a known function, and we want to calculate F_Y and f_Y given F_X and f_X.

Example 126. *Let $Y = -\log(1-U)$, where $U \sim U(0,1)$. Calculate $F_Y(y)$ and discuss. Calculate also the density and cumulative distribution function of $W = -\log U$. Explain.*

Example 127. *Let $Y = \lceil X \rceil$, where $X \sim \exp(\lambda)$ (thus Y is the smallest integer greater than X). Calculate $F_Y(y)$ and $f_Y(y)$.*

Note to Example 126

Note first that since $0 < U < 1$, $1 - U > 0$ and taking the log is OK, and we get $Y = -\log(1-U) > 0$. Hence

$$\mathrm{P}(Y \le y) = \mathrm{P}\{-\log(1-U) \le y\} = \mathrm{P}\{U \le 1 - \exp(-y)\} = 1 - \exp(-y), \quad y > 0$$

which is the exponential density; note that the transformation here is monotone. Thus Y has an exponential distribution.
For $W = -\log U$, we have

$$\begin{aligned} \mathrm{P}(W \le w) &= \mathrm{P}\{-\log(U) \le w\} \\ &= \mathrm{P}\{\log U \ge -w\} \\ &= \mathrm{P}(U \ge e^{-w}) \\ &= 1 - \mathrm{P}(U < e^{-w}) = 1 - e^{-w}, \quad w > 0, \end{aligned}$$

where the $<$ can become an $\le$ because there is no probability at individual points in $\mathbb{R}$.
Hence W also has an exponential distribution. This is obvious, because if $U \sim U(0,1)$, then $1 - U \sim U(0,1)$ also.

Note to Example 127

$Y = r$ iff $r - 1 < X \leq r$, so for $r = 1, 2, \ldots,$ we have

$$\mathrm{P}(Y = r) = \int_{r-1}^{r} f_X(x)\, dx = \int_{r-1}^{r} \lambda e^{-\lambda x}\, dx = (e^{-\lambda(r-1)} - e^{-\lambda r}) = (e^{-\lambda})^{r-1}(1 - e^{-\lambda}).$$

This is the geometric distribution with probability $p = 1 - e^{-\lambda}$.

General transformation

We can formalise the previous discussion in the following way:

Definition 128. *Let $g : \mathbb{R} \mapsto \mathbb{R}$ be a function and $\mathcal{B} \subset \mathbb{R}$ any subset of $\mathbb{R}$. Then $g^{-1}(\mathcal{B}) \subset \mathbb{R}$ is the set for which $g\{g^{-1}(\mathcal{B})\} = \mathcal{B}$.*

Theorem 129. *Let $Y = g(X)$ be a random variable and $\mathcal{B}_y = (-\infty, y]$. Then*

$$F_Y(y) = \mathrm{P}(Y \leq y) = \begin{cases} \int_{g^{-1}(\mathcal{B}_y)} f_X(x)\, dx, & X \text{ continuous}, \\ \sum_{x \in g^{-1}(\mathcal{B}_y)} f_X(x), & X \text{ discrete}, \end{cases}$$

where $g^{-1}(\mathcal{B}_y) = \{x \in \mathbb{R} : g(x) \leq y\}$. When g is monotone increasing or decreasing and has differentiable inverse g^{-1}, then

$$f_Y(y) = \left| \frac{dg^{-1}(y)}{dy} \right| f_X\{g^{-1}(y)\}, \quad y \in \mathbb{R}.$$

Example 130. *If $X \sim \exp(\lambda)$ and $Y = \exp(X)$, find F_Y and f_Y.*

Example 131. *Find the distribution and density functions of $Y = \cos(X)$, where $X \sim \exp(1)$.*

Note to Theorem 129

We have

$$\mathrm{P}(Y \in \mathcal{B}) = \mathrm{P}\{g(X) \in \mathcal{B}\} = \mathrm{P}\{X \in g^{-1}(\mathcal{B})\},$$

because $X \in g^{-1}(\mathcal{B})$ if and only if $g(X) \in g\{g^{-1}(\mathcal{B})\} = \mathcal{B}$.
To find $F_Y(y)$ we take $\mathcal{B}_y = (-\infty, y]$, giving

$$F_Y(y) = \mathrm{P}(Y \leq y) = \mathrm{P}\{g(X) \in \mathcal{B}_y\} = \mathrm{P}\{X \in g^{-1}(\mathcal{B}_y)\},$$

which is the formula in the theorem.
When g is monotone increasing with (monotone increasing) inverse g^{-1}, we have $g^{-1}\{(-\infty, y]\} = (-\infty, g^{-1}(y)]$, and hence

$$F_Y(y) = \mathrm{P}\{Y \in \mathcal{B}_y\} = \mathrm{P}\{X \in g^{-1}(\mathcal{B}_y)\} = \mathrm{P}\{X \leq g^{-1}(y)\} = F_X\{g^{-1}(y)\}, \quad y \in \mathbb{R}.$$

In the case of a continuous random variable X, differentiation gives

$$f_Y(y) = \frac{dg^{-1}(y)}{dy} f_X\{g^{-1}(y)\}, \quad y \in \mathbb{R}.$$

When g is monotone decreasing with (monotone decreasing) inverse g^{-1}, we have $g^{-1}\{(-\infty, y]\} = [g^{-1}(y), \infty)$, and hence

$$F_Y(y) = \mathrm{P}\{Y \in \mathcal{B}_y\} = \mathrm{P}\{X \in g^{-1}(\mathcal{B}_y)\} = \mathrm{P}\{X \geq g^{-1}(y)\}, \quad y \in \mathbb{R}.$$

In the case of a continuous density, $F_Y(y) = \mathrm{P}\{X \geq g^{-1}(y)\} = 1 - F_X\{g^{-1}(y)\}$ and differentiation gives

$$f_Y(y) = -\frac{dg^{-1}(y)}{dy} f_X\{g^{-1}(y)\}, \quad y \in \mathbb{R};$$

note that $-dg^{-1}(y)/dy \geq 0$, because $g^{-1}(y)$ is monotone decreasing.
Thus in both cases we can write

$$f_Y(y) = \left| \frac{dg^{-1}(y)}{dy} \right| f_X\{g^{-1}(y)\}, \quad y \in \mathbb{R}.$$

Note to Example 130

Note first that since X only puts probability on $\mathbb{R}_+$, $Y \in (1, \infty)$.
In terms of the theorem, let $\mathcal{B}_y = (-\infty, y]$, and note that $g(x) = e^x$ is monotone increasing, with $g^{-1}(y) = \log y$, so

$$\mathrm{P}(Y \leq y) = \mathrm{P}(Y \in \mathcal{B}) = \mathrm{P}\{g(X) \in \mathcal{B}\} = \mathrm{P}\{X \in g^{-1}(\mathcal{B})\} = \mathrm{P}\{X \in (-\infty, \log y]\} = F_X(\log y),$$

so

$$\mathrm{P}(Y \leq y) = 1 - \exp\{-\lambda \log y\} = 1 - y^{-\lambda}, \quad y > 1.$$

Hence Y has the Pareto distribution with $\beta = 1$, $\alpha = \lambda$, and

$$f_Y(y) = \begin{cases} 0, & y \leq 1, \\ \lambda y^{-\lambda - 1}, & y > 1. \end{cases}$$

To get the density directly, we note that $dg^{-1}(y)/dy = 1/y$, and

$$f_Y(y) = \left| \frac{dg^{-1}(y)}{dy} \right| f_X\{g^{-1}(y)\} = |y^{-1}| \times \lambda e^{-\lambda \log y} = \lambda y^{-\lambda - 1}, \quad y > 1,$$

and $f_Y(y) = 0$ for $y \leq 1$, because if $y < 1$, then $\log y < 0$, and $f_X(x) = 0$ for $x < 0$.

Note to Example 131

- Here $Y = g(X) = \cos(X)$ takes values only in the range $-1 \le y \le 1$, so if $y < -1$, $\mathcal{B}_y = \emptyset$, and if $y \ge 1$, $\mathcal{B}_y = \mathbb{R}$, thus giving

$$F_Y(y) = \begin{cases} 0, & y < -1 \\ 1, & y \ge 1. \end{cases}$$

- A sketch of the function $\cos x$ for $x \ge 0$ shows that in the range $0 < x < 2\pi$, and for $-1 < y < 1$, the event $\cos(X) \le y$ is equivalent to the event $\cos^{-1}(y) \le X \le 2\pi - \cos^{-1}(y)$. Since the cosine function is periodic, the set $\mathcal{B}_y$ is an infinite union of disjoint intervals. In fact

$$\cos(X) \le y \quad \Leftrightarrow \quad X \in g^{-1}(B_y) = \bigcup_{j=0}^{\infty} \{x : 2\pi j + \cos^{-1}(y) \le x \le 2\pi(j+1) - \cos^{-1}(y)\},$$

and therefore

$$\begin{aligned} \mathrm{P}(Y \le y) &= \mathrm{P}\{X \in g^{-1}(\mathcal{B})\} \\ &= \sum_{j=0}^{\infty} \mathrm{P}\left\{2\pi j + \cos^{-1}(y) \le X \le 2\pi(j+1) - \cos^{-1}(y)\right\} \\ &= \sum_{j=0}^{\infty} \left(\exp[-\lambda\{2\pi j + \cos^{-1}(y)\}] - \exp[-\lambda\{2\pi(j+1) - \cos^{-1}(y)\}]\right) \\ &= \frac{\exp\{-\lambda\cos^{-1}(y)\} - \exp\{\lambda\cos^{-1}(y) - 2\pi\lambda\}}{1 - \exp(-2\pi\lambda)}, \end{aligned}$$

where we noticed that the summation is proportional to a geometric series.

- Note that if $y = 1$, then $\cos^{-1}(y) = 0$, and so $\mathrm{P}(Y \le 1) = 1$, and if $y = -1$, then $\cos^{-1}(y) = \pi$, and then $\mathrm{P}(Y \le -1) = 0$, as required. Here we used values of $\cos^{-1}(y)$ in the range $[0, \pi]$.
- The density function is found by differentiation: since $\cos\{\cos^{-1}(y)\} = y$, we have

$$\frac{d\cos^{-1}(y)}{dy} = -\frac{1}{\sin\{\cos^{-1}(y)\}},$$

and this gives

$$f_Y(y) = \frac{\lambda}{\sin\{\cos^{-1}(y)\}} \times \frac{\exp\{-\lambda\cos^{-1}(y)\} + \exp\{\lambda\cos^{-1}(y) - 2\pi\lambda\}}{1 - \exp(-2\pi\lambda)}, \quad y \in (-1, 1).$$

Normal distribution

Definition 132. *A random variable X having density*

$$f(x) = \frac{1}{(2\pi)^{1/2}\sigma} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}, \quad x \in \mathbb{R}, \quad \mu \in \mathbb{R}, \sigma > 0,$$

is a **normal random variable** *with expectation μ and variance σ^2: we write $X \sim \mathcal{N}(\mu, \sigma^2)$. (The standard deviation of X is $\sqrt{\sigma^2} = \sigma > 0$.)*
When $\mu = 0$, $\sigma^2 = 1$, the corresponding random variable Z is **standard normal**, *$Z \sim \mathcal{N}(0, 1)$, with density*

$$\phi(z) = (2\pi)^{-1/2} e^{-z^2/2}, \quad z \in \mathbb{R}.$$

Then

$$F_Z(x) = \mathrm{P}(Z \le x) = \Phi(x) = \int_{-\infty}^{x} \phi(z)\, dz = \frac{1}{(2\pi)^{1/2}} \int_{-\infty}^{x} e^{-z^2/2}\, dz.$$

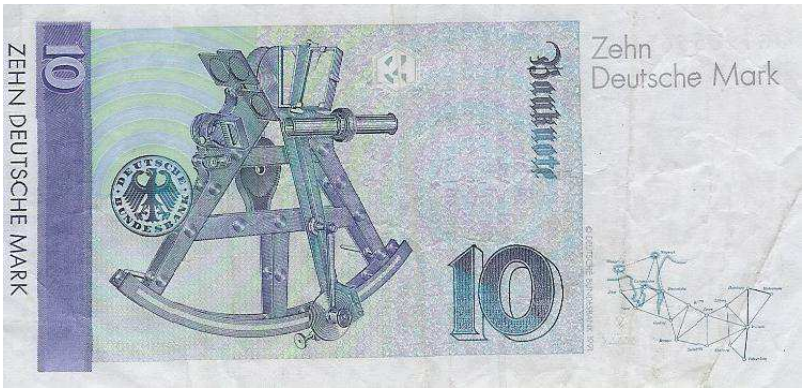
This integral is given in the Formulaire.

Note that $f(x) = \sigma^{-1}\phi\{(x-\mu)/\sigma\}$ for $x \in \mathbb{R}$.

Johann Carl Friedrich Gauss (1777–1855)

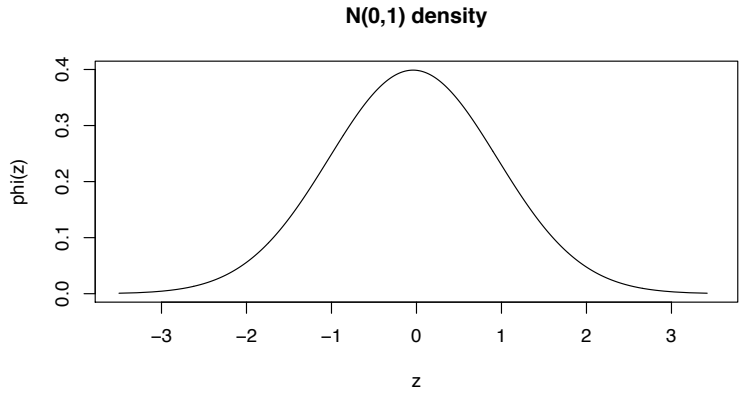
The normal distribution is often called the **Gaussian distribution**. Gauss used it for the combination of astronomical and topographical measures.

Johann Carl Friedrich Gauss (1777–1855)



The normal distribution is often called the **Gaussian distribution**. Gauss used it for the combination of astronomical and topographical measures.

Standard normal density



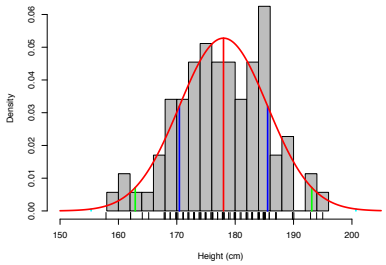
The famous bell curve:

$$\phi(z) = (2\pi)^{-1/2} e^{-z^2/2}, \quad z \in \mathbb{R}.$$

Interpretation of $\mathcal{N}(\mu, \sigma^2)$

- ☐ The density function is centred at μ, which is the most likely value and also the median;
- ☐ the standard deviation σ is a measure of the spread of the values around μ:
  - 68% of the probability lies in the interval $\mu \pm \sigma$;
  - 95% of the probability lies in the interval $\mu \pm 2\sigma$;
  - 99.7% of the probability lies in the interval $\mu \pm 3\sigma$.

Example 133. *The average height for a class of students was 178 cm, with standard deviation 7.6 cm. If this is representative of the population, then 68% have heights in the interval* 178 ± 7.6 *cm (blue lines), 95% in the interval* $178 \pm 2 \times 7.6$ *cm (green lines), and 99.7% in the interval* $178 \pm 3 \times 7.6$ *cm (cyan lines, almost invisible).*



Properties

Theorem 134. *The density* $\phi(z)$*, the cumulative distribution function* $\Phi(z)$*, and the quantiles* z_p *of* $Z \sim \mathcal{N}(0,1)$ *satisfy, for all* $z \in \mathbb{R}$*:*

(a) the density is symmetric with respect to $z = 0$*, i.e.,* $\phi(z) = \phi(-z)$*;*

(b) $\mathrm{P}(Z \leq z) = \Phi(z) = 1 - \Phi(-z) = 1 - \mathrm{P}(Z \geq z)$*;*

(c) the standard normal quantiles z_p *satisfy* $z_p = -z_{1-p}$*, for all* $0 < p < 1$*;*

(d) $z^r \phi(z) \to 0$ *when* $z \to \pm\infty$*, for all* $r > 0$*. This imples that the moments* $\mathrm{E}(Z^r)$ *exist for all* $r \in \mathbb{N}$*;*

(e) we have

$$\phi'(z) = -z\phi(z), \quad \phi''(z) = (z^2 - 1)\phi(z), \quad \phi'''(z) = -(z^3 - 3z)\phi(z), \quad \ldots$$

This implies that $\mathrm{E}(Z) = 0$*,* $\mathrm{var}(Z) = 1$*,* $\mathrm{E}(Z^3) = 0$*, etc.*

(f) If $X \sim \mathcal{N}(\mu, \sigma^2)$*, then* $Z = (X - \mu)/\sigma \sim \mathcal{N}(0,1)$*.*

Note that if $X \sim \mathcal{N}(\mu, \sigma^2)$, then we can write $X = \mu + \sigma Z$, where $Z \sim \mathcal{N}(0,1)$.

Theorem 134

(a) Obvious by substitution:

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

(b) Obvious by the symmetry of $\phi(z)$, as

$$\Phi(z) = \int_{-\infty}^{z} \phi(x)\, dx = \int_{-z}^{\infty} \phi(x)\, dx = 1 - \Phi(-z),$$

which implies that

$$\mathrm{P}(Z \leq z) = \Phi(z) = 1 - \Phi(-z) = 1 - \mathrm{P}(Z \leq -z) = 1 - \Phi(-z).$$

(c) Again obvious by symmetry, using (b): $p = \Phi(z) = 1 - \Phi(-z)$ implies that $z_p = -z_{1-p}$.
(d) This is just a fact from analysis, since for any $r \geq 0$, we have

$$z^r \phi(z) \propto \frac{z^r}{\sum_{i=0}^{\infty} z^{2i}/i!} < \frac{z^r}{z^{2(r+1)}} \to 0, \quad z \to \infty,$$

and by symmetry the same will be true when $z \to -\infty$.
(e) Differentiate $\phi(z)$ repeatedly, and then note that

$$\mathrm{E}(Z) = \int z\phi(z)\, dz = [-\phi(z)]_{-\infty}^{\infty} = 0, \quad \mathrm{E}(Z^2 - 1) = \left[\phi'(z)\right]_{-\infty}^{\infty} = 0,$$

etc. by (d). Hence $\mathrm{E}(Z) = 0$, $\mathrm{E}(Z^2) = 1$, etc.
(f) This is just a change of variable in the density function.

Values of the function $\Phi(z)$

z	0	1	2	3	4	5	6	7	8	9
0.0	.50000	.50399	.50798	.51197	.51595	.51994	.52392	.52790	.53188	.53586
0.1	.53983	.54380	.54776	.55172	.55567	.55962	.56356	.56750	.57142	.57535
0.2	.57926	.58317	.58706	.59095	.59483	.59871	.60257	.60642	.61026	.61409
0.3	.61791	.62172	.62552	.62930	.63307	.63683	.64058	.64431	.64803	.65173
0.4	.65542	.65910	.66276	.66640	.67003	.67364	.67724	.68082	.68439	.68793
0.5	.69146	.69497	.69847	.70194	.70540	.70884	.71226	.71566	.71904	.72240
0.6	.72575	.72907	.73237	.73565	.73891	.74215	.74537	.74857	.75175	.75490
0.7	.75804	.76115	.76424	.76730	.77035	.77337	.77637	.77935	.78230	.78524
0.8	.78814	.79103	.79389	.79673	.79955	.80234	.80511	.80785	.81057	.81327
0.9	.81594	.81859	.82121	.82381	.82639	.82894	.83147	.83398	.83646	.83891
1.0	.84134	.84375	.84614	.84850	.85083	.85314	.85543	.85769	.85993	.86214
1.1	.86433	.86650	.86864	.87076	.87286	.87493	.87698	.87900	.88100	.88298
1.2	.88493	.88686	.88877	.89065	.89251	.89435	.89617	.89796	.89973	.90147
1.3	.90320	.90490	.90658	.90824	.90988	.91149	.91309	.91466	.91621	.91774
1.4	.91924	.92073	.92220	.92364	.92507	.92647	.92786	.92922	.93056	.93189
1.5	.93319	.93448	.93574	.93699	.93822	.93943	.94062	.94179	.94295	.94408
1.6	.94520	.94630	.94738	.94845	.94950	.95053	.95154	.95254	.95352	.95449
1.7	.95543	.95637	.95728	.95818	.95907	.95994	.96080	.96164	.96246	.96327
1.8	.96407	.96485	.96562	.96638	.96712	.96784	.96856	.96926	.96995	.97062
1.9	.97128	.97193	.97257	.97320	.97381	.97441	.97500	.97558	.97615	.97670
2.0	.97725	.97778	.97831	.97882	.97932	.97982	.98030	.98077	.98124	.98169

Remark: A more detailed table can be found in the *Formulaire*. You may also use the function `pnorm` in the software R: $\Phi(z) =$ `pnorm(z)`.

Example 135. *Calculate*

$$\mathrm{P}(Z \leq 0.53), \quad \mathrm{P}(Z \leq -1.86), \quad \mathrm{P}(-1.86 < Z < 0.53), \quad z_{0.95}, \quad z_{0.025}, \quad z_{0.5}.$$

Note to Example 135

In R we use `pnorm` for Φ and `qnorm` for Φ^{-1}:

```
> pnorm(0.53)
[1] 0.701944
> pnorm(-1.86)
[1] 0.03144276
> pnorm(0.53)- pnorm(-1.86)
[1] 0.6705013
> qnorm(0.95)
[1] 1.644854
> qnorm(0.025)
[1] -1.959964
> qnorm(0.5)
[1] 0
```

Examples and calculations

Example 136. *The duration in minutes of a maths lecture is $\mathcal{N}(47, 4)$, but should be 45. Give the probability that (a) the lecture finishes early, (b) the lecture finishes at least 5 minutes late.*

Example 137. *Show that the expectation and variance of $X \sim \mathcal{N}(\mu, \sigma^2)$ are μ and σ^2, and find the p quantile of X.*

Example 138. *Calculate the cumulative distribution function and the density of $Y = |Z|$ and $W = Z^2$, where $Z \sim \mathcal{N}(0, 1)$.*

Note to Example 136

(a) Note that we can write $X = \mu + \sigma Z$, where $Z \sim \mathcal{N}(0, 1)$. We have $X \sim \mathcal{N}(47, 4)$, and we seek $\mathrm{P}(X < 45) = \mathrm{P}\{(X - 47)/2 < (45 - 47)/2\} = \mathrm{P}(Z < -1) = 1 - 0.84134 \doteq 0.16$.
(b) $\mathrm{P}(X > 50) = \mathrm{P}\{(X - 47)/2 > (50 - 47)/2\} = \mathrm{P}(Z > 1.5) = 1 - 0.93319 \doteq 0.067$.

Note to Example 137

Since we can write $X = \mu + \sigma Z$, and $\mathrm{E}(Z) = 0$ and $\mathrm{E}(Z^2) = \mathrm{var}(Z) = 1$ by Theorem 134(e), we just apply the properties of mean and variance from Theorems 89 and 92.

Note to Example 138

- For Y, note that if $y > 0$, then $\mathrm{P}(Y \le y) = \mathrm{P}(-y \le Z \le y) = \Phi(y) - \Phi(-y)$, and differentiate to obtain $2\phi(y)$, for $y > 0$ and zero otherwise.
  Alternatively, in the terms of Theorem 129, we have $g(x) = |x|$ and therefore $g^{-1}(\mathcal{B}_y) = g^{-1}\{(-\infty, y]\} = (-y, y)$, provided that $y \ge 0$, and $g^{-1}(\mathcal{B}_y) = \emptyset$ if $y < 0$. Therefore

$$\mathrm{P}(Y \le y) = \int_{g^{-1}(\mathcal{B}_y)} \phi(x)\, dx = \int_{-y}^{y} \phi(x)\, dx = \Phi(y) - \Phi(-y), \quad y > 0,$$

  as before.
- For W, the same argument gives $\mathrm{P}(W \le w) = \mathrm{P}(-\sqrt{w} \le Z \le \sqrt{w}) = \Phi(\sqrt{w}) - \Phi(-\sqrt{w})$, for $w > 0$. Then differentiate to obtain the density.
  In this case $g(x) = x^2$ and $g^{-1}(\mathcal{B}_w) = g^{-1}\{(-\infty, w]\} = (-\sqrt{w}, \sqrt{w})$ for $w \ge 0$ and $g^{-1}(\mathcal{B}_w) = \emptyset$ for $w < 0$. This gives the previous result, by a slightly more laborious route.

Normal approximation to the binomial distribution

The normal distribution is a central to probability, partly because it can be used to approximate probabilities of other distributions. One of the basic results is:

Theorem 139 (de Moivre–Laplace). *Let $X_n \sim B(n,p)$, where $0 < p < 1$, let*

$$\mu_n = \mathrm{E}(X_n) = np, \quad \sigma_n^2 = \mathrm{var}(X_n) = np(1-p),$$

and let $Z \sim \mathcal{N}(0,1)$. When $n \to \infty$,

$$\mathrm{P}\left(\frac{X_n - \mu_n}{\sigma_n} \leq z\right) \to \Phi(z),\ z \in \mathbb{R}; \quad \textit{i.e.,} \quad \frac{X_n - \mu_n}{\sigma_n} \xrightarrow{D} Z.$$

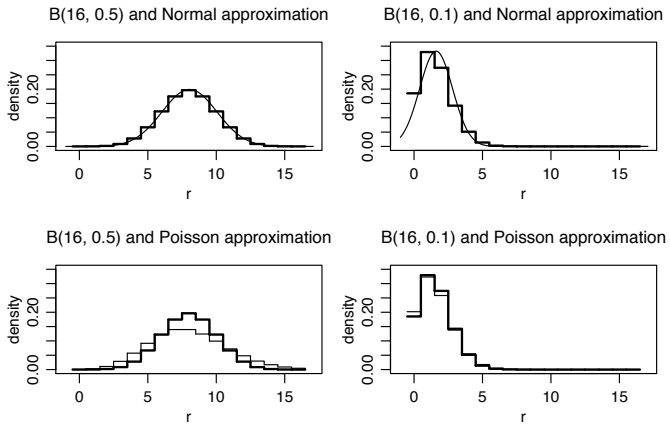
This gives us an approximation of the probability that $X_n \leq r$:

$$\mathrm{P}(X_n \leq r) = \mathrm{P}\left(\frac{X_n - \mu_n}{\sigma_n} \leq \frac{r - \mu_n}{\sigma_n}\right) \doteq \Phi\left(\frac{r - \mu_n}{\sigma_n}\right),$$

which corresponds to $X_n \dot{\sim} \mathcal{N}\{np, np(1-p)\}$.
In practice the approximation is bad when $\min\{np, n(1-p)\} < 5$.

Normal and Poisson approximations to the binomial

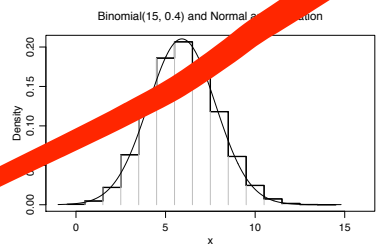
We have already encountered the Poisson approximation to the binomial distribution, valid for large n and small p. The normal approximation is valid for large n and $\min\{np, n(1-p)\} \geq 5$. Left: a case where the normal approximation is valid. Right: a case where the Poisson approximation is valid.



Continuity correction

A better approximation to $\mathrm{P}(X_n \leq r)$ is given by replacing r by $r + \frac{1}{2}$; the $\frac{1}{2}$ is [illegible]d the **continuity correction**. This gives

$$\mathrm{P}(X_n \leq r) \doteq \Phi\left(\frac{r + \frac{1}{2} - np}{\sqrt{np(1-p)}}\right)$$



Example 140. *Let $X \sim B(15, 0.4)$. Calculate the exact and approximate values of $\mathrm{P}(X \leq r)$ for $r = 1, 8, 10$, with and without the continuity correction. Comment.*

Note to Example 140

The following R code shows how to do this, but fir[illegible]o some of it on the board using the normal table.

NumeRical Results

```
pbinom(c(1,8,10),15,prob=0.4)
[1] 0.005172035 0.9[illegible]4952592 0.990652339

pnorm(c(1,8,10),mean=15*0.4,sd=sqrt(15*0.4*0.6))
[1] 0.0042039[illegible]97 0.854079727 0.982492509

pnorm(c(1,8,10)+0.5,mean=15*0.4,sd=sqrt(15*0.4*0.6))
[1] 0.008853033 0.906183835 0.991146967
```


Example

Example 141. *The total number of students in a class is 100.*
(a) Each student goes independently to a maths lecture with probability 0.6. What is the size of the smallest classroom suited for the number of students who go to class, with a probability of 0.95?
(b) There are 14 lectures per semester, and the students decide to go to each lecture independently. What is now the size of the smallest classroom necessary?

Note to Example 141

(a) The number of students present X is $B(100, 0.6)$, so the mean is $100 \times 0.6 = 60$ and the variance is $100 \times 0.6 \times 0.4 = 24$. We seek x such that

$$0.95 = \mathrm{P}(X \le x) = \mathrm{P}\left\{\frac{X-60}{\sqrt{24}} \le \frac{x-60}{\sqrt{24}}\right\} \doteq \Phi\left\{\frac{x-60}{\sqrt{24}}\right\},$$

and this implies that $(x-60)/\sqrt{24} = \Phi^{-1}(0.95) = 1.65$, and thus $x = 60 + \sqrt{24} \times 1.65 = 68.08$. Better have a room for 69.

(b) Now we want to solve the equation

$$0.95 = \mathrm{P}(X \le x)^{14} = \mathrm{P}\left\{\frac{X-60}{\sqrt{24}} \le \frac{x-60}{\sqrt{24}}\right\}^{14} \doteq \Phi\left\{\frac{x-60}{\sqrt{24}}\right\}^{14},$$

and this implies that $(x-60)/\sqrt{24} = \Phi^{-1}(0.95^{1/14}) = 2.68$, and thus $x = 60 + \sqrt{24} \times 2.68 = 73.14$. Better have a room for 74.

4.4 Q-Q Plots

Quantile-quantile (Q-Q) plots

One way of comparing a sample $X_1, \ldots, X_n$ with a theoretical distribution F:

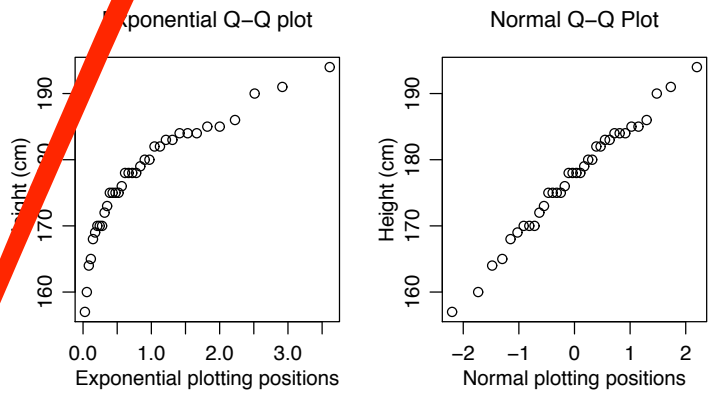
- ☐ we order the X_j, giving
$$X_{(1)} \le X_{(2)} \le \cdots \le X_{(n)},$$
then we plot the graph against $F^{-1}\{1/(n+1)\}, F^{-1}\{2/(n+1)\} \ldots, F^{-1}\{n/(n+1)\}$.
- ☐ **The idea**: in an ideal case $U_1, \ldots, U_n \sim U(0,1)$ should cut the interval $(0,1)$ into $n+1$ sub-intervals of width $1/(n+1)$, so we should plot the graph of the $U_{(j)}$ against $1/(n+1)$, …, $n/(n+1)$, and thus the $X_{(j)} \stackrel{D}{=} F^{-1}(U_{(j)})$ against the $F^{-1}\{j/(n+1)\}$;
- ☐ the closer the graph is to a straight line, the more the data resemble a sample from F;
- ☐ we often take a standard version of F (e.g., $\exp(1)$ or $\mathcal{N}(0,1)$), and then the $F^{-1}\{j/(n+1)\}$ are called the **plotting positions** of F—then the slope gives an estimation of the dispersion parameter of the distribution, and the value at the origin gives an estimation of the position parameter;
- ☐ for the distributions $\exp(1)$ and $\mathcal{N}(0,1)$ we have respectively
$$F^{-1}\left(\frac{j}{n+1}\right) = -\log\left(1 - \frac{j}{n+1}\right), \quad F^{-1}\left(\frac{j}{n+1}\right) = \Phi^{-1}\left(\frac{j}{n+1}\right);$$
- ☐ it is difficult to draw strong conclusions from such a graph for small n, as the variability is then large—we have a tendency to over-interpret patterns in the plot.

Note to the following graphs

- ☐ First graph: the normal graph is close to a straight line, whereas the exponential one is not. Suggests that the normal would be a reasonable model for these data. Derive the formula for the exponential plotting positions, using the quantile formula for the exponential distribution.
- ☐ Second graph: Here we compare the real data (top centre) with simulated data. The fact that it is hard to tell which is which (you need to remember the shape of the first graph, or to note that tied observations are impossible with simulations) suggests that the heights can be considered to be normal.
- ☐ The lower left is gamma: there is clearer nonlinearity than with the other panels—but it is hard to be sure with this sample size.
- ☐ The lower middle is obviously not normal; the sample size is big, however.

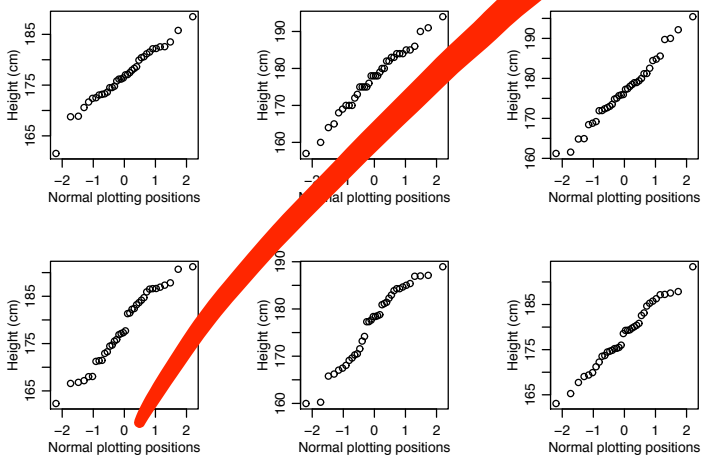
Heights of students

Q-Q plots for the heights of $n = 36$ students in SSC, for the exponential and normal distributions.



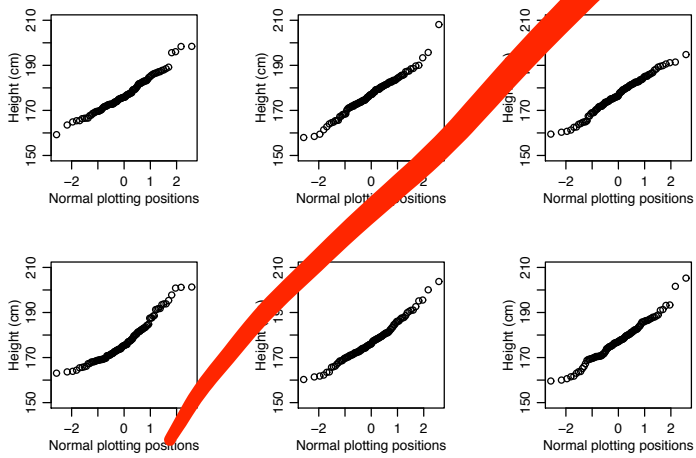
$n = 36$: Which sample is not normal?

There are five samples of simulated normal variables, and some real data.



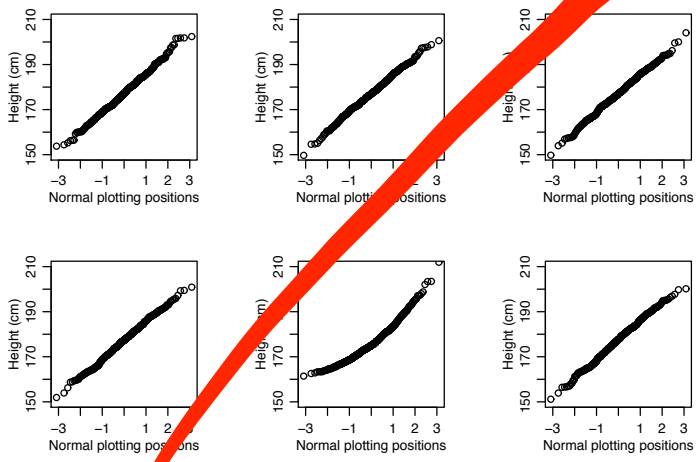
$n = 100$: **Which sample is not normal?**

There are five samples of simulated normal variables, and one simulated gamma sample.



$n = 500$: **Which sample is not normal?**

There are five samples of simulated normal variables, and one simulated gamma sample.



Which density?

- **Uniform** variables lie in a finite interval, and give equal probability to each part of the interval;
- **exponential** and **gamma** variables lie in $(0, \infty)$, and are often used to model waiting times and other positive quantities,
  - the gamma has two parameters and is more flexible, but the exponential is simpler and has some elegant properties;
- **Pareto** variables lie in the interval (β, ∞), so are not appropriate for arbitrary positive quantities (which could be smaller than β), but are often used to model financial losses over some threshold β;
- **normal** variables lie in $\mathbb{R}$ and are used to model quantities that arise (or might arise) through averaging of many small effects (e.g., height and weight, which are influenced by many genetic factors), or where measurements are subject to error;
- **Laplace** variables lie in $\mathbb{R}$; the Laplace distribution can be used in place of the normal in situations where outliers might be present.

5. Several Random Variables

slide 174

Lexicon

Mathematics	English	Français
$\mathrm{E}(X)$	expected value/expectation of X	espérance de X
$\mathrm{E}(X^r)$	rth moment of X	$r$ième moment de X
$\mathrm{var}(X)$	variance of X	variance de X
$M_X(t)$	moment generating function of X, or the Laplace transform of $f_X(x)$	fonction génératrice des moments ou transformée de Laplace de $f_X(x)$
$f_{X,Y}(x,y)$	joint density/mass function	densité/fonction de masse conjointe
$F_{X,Y}(x,y)$	joint (cumulative) distribution function	fonction de répartition conjointe
$f_{X\mid Y}(x \mid y)$	conditional density function	densité conditionnelle
$f_{X,Y}(x,y) = f_X(x)f_Y(y)$	X, Y independent	X, Y independantes
$X_1, \ldots, X_n \overset{\text{iid}}{\sim} F$	random sample from F	échantillon aléatoire
$\mathrm{E}(X^r Y^s)$	joint moment	moment conjoint
$\mathrm{cov}(X,Y)$	covariance of X and Y	covariance de X et Y
$\mathrm{corr}(X,Y)$	correlation of X and Y	correlation de X et Y
$\mathrm{E}(X \mid Y = y)$	conditional expectation of X	espérance conditionnelle de X
$\mathrm{var}(X \mid Y = y)$	conditional variance of X	variance conditionnelle de X
$X_{(r)}$	rth order statistic	$r$ième statistique d'ordre

5.1 Basic Notions

slide 176

Motivation

Often we have to consider the way in which several variables vary simultaneously. Some examples:

Example 142. *The distribution of (height, weight) of a student picked at random from the class.*

Example 143 (Hats, continuation of Example 47)**.** *Three men with hats permute them in a random way. Let I_1 be the indicator of the event in which man 1 has his hat, etc. Find the joint distribution of (I_1, I_2, I_3).*

Our previous definitions generalise in a natural way to this situation.

Note to Example 143

The possibilities each have probability 1/6, and with the notation that I_j indicates that the jth hat is on the right head, are

1	2	3	
1	2	3	$(I_1, I_2, I_3) = (1, 1, 1)$
1	3	2	$(I_1, I_2, I_3) = (1, 0, 0)$
2	1	3	$(I_1, I_2, I_3) = (0, 0, 1)$
2	3	1	$(I_1, I_2, I_3) = (0, 0, 0)$
3	1	2	$(I_1, I_2, I_3) = (0, 0, 0)$
3	2	1	$(I_1, I_2, I_3) = (0, 1, 0)$

from which we can compute anything we like.

Discrete random variables

Definition 144. *Let (X, Y) be a discrete random variable: the set*

$$D = \{(x, y) \in \mathbb{R}^2 : \mathrm{P}\{(X, Y) = (x, y)\} > 0\}$$

is countable. The **(joint) probability mass function** *of (X, Y) is*

$$f_{X,Y}(x, y) = \mathrm{P}\{(X, Y) = (x, y)\}, \quad (x, y) \in \mathbb{R}^2,$$

and the **(joint) cumulative distribution function** *of (X, Y) is*

$$F_{X,Y}(x, y) = \mathrm{P}(X \leq x, Y \leq y), \quad (x, y) \in \mathbb{R}^2.$$

Example 145 (Hats, Continuation of Example 143). *Find the joint distribution of $(X, Y) = (I_1, I_2 + I_3)$.*

Note to Example 145

The lines of the table in the previous example all have probabilities $1/6$, so, for example, we have

$$f(0, 0) = \mathrm{P}(X = 0, Y = 0) = \mathrm{P}\{\text{configuration } (2, 3, 1)\} + \mathrm{P}\{\text{configuration } (3, 1, 2)\} = 2/6.$$

In a similar manner, we obtain:

x	y	$f(x, y)$
0	0	$2/6$
0	1	$2/6$
0	2	$0/6$
1	0	$1/6$
1	1	$0/6$
1	2	$1/6$

Continuous random variables

Definition 146. *The random variable (X, Y) is said to be* **(jointly) continuous** *if there exists a function $f_{X,Y}(x, y)$, called the* **(joint) density** *of (X, Y), such that*

$$\mathrm{P}\{(X, Y) \in A\} = \int\int_{(u,v)\in A} f_{X,Y}(u, v)\, dudv, \quad \mathcal{A} \subset \mathbb{R}^2.$$

By letting $\mathcal{A} = \{(u, v) : u \leq x, v \leq y\}$, we see that the **(joint) cumulative distribution function** *of (X, Y) can be written*

$$F_{X,Y}(x, y) = \mathrm{P}(X \leq x, Y \leq y) = \int_{-\infty}^{x}\int_{-\infty}^{y} f_{X,Y}(u, v)\, dudv, \quad (x, y) \in \mathbb{R}^2,$$

and this implies that

$$f_{X,Y}(x, y) = \frac{\partial^2}{\partial x \partial y} F_{X,Y}(x, y).$$

Example

Example 147. *Calculate the joint cumulative distribution function and $\mathrm{P}(X \leq 1, Y \leq 2)$ when*

$$f_{X,Y}(x, y) \propto \begin{cases} e^{-x-y}, & y > x > 0, \\ 0, & \text{otherwise.} \end{cases}$$

We can write $f(x, y) = ce^{-x-y} I(y > x) I(x > 0)$, where $I(\mathcal{A})$ is the indicator function of the set $\mathcal{A}$.

Note to Example 147

Note that if $\min(x,y) \le 0$, then $F(x,y) = 0$, and consider the integral for $y > x$ (sketch):

$$\begin{aligned} F(x,y) &= \int_{-\infty}^{x} du \int_{-\infty}^{y} dv\, f(u,v) \\ &= c\int_0^x e^{-u}\,du \int_u^y e^{-v}\,dv \\ &= c\int_0^x du\, e^{-u}\left[e^{-v}\right]_y^u \\ &= c\int_0^x du\, e^{-u}\left[e^{-u} - e^{-y}\right] \\ &= c\int_0^x du\,(e^{-2u} - e^{-u-y}) \\ &= c\left[e^{-u-y} - \tfrac{1}{2}e^{-2u}\right]_0^x \\ &= \tfrac{1}{2}c\left[1 - e^{-2x} - 2e^{-y} + 2e^{-y-x}\right]. \end{aligned}$$

On setting $x = y = +\infty$, we get $\frac{1}{2}c = 1$, and this implies that $c = 2$.
Now for $y \le x$, consideration of areas shows that we should take the above formula with $y = x$, so

$$F(x,y) = \begin{cases} 1 - e^{-2x} + 2e^{-x-y} - 2e^{-y}, & y > x > 0, \\ 1 + e^{-2y} - 2e^{-y}, & x \ge y > 0, \\ 0, & \text{otherwise.} \end{cases}$$

Thus $F(1,2) = 1 - e^{-2} + 2e^{-3} - 2e^{-2} = 1 - 3e^{-2} + 2e^{-3}$.

Exponential families

Definition 148. *Let $(X_1,\ldots,X_n)$ be a discrete or continuous random variable with mass/density function of the form*

$$f(x_1,\ldots,x_n) = \exp\left\{\sum_{i=1}^{p} s_i(x)\theta_i - \kappa(\theta_1,\ldots,\theta_p) + c(x_1,\ldots,x_n)\right\}, \quad (x_1,\ldots,x_n) \in D \subset \mathbb{R}^n,$$

where $(\theta_1,\ldots,\theta_p) \in \Theta \subset \mathbb{R}^p$. This is called an **exponential family** *distribution—not to be confused with the exponential distribution.*

Example 149. *Show that the (a) Poisson and (b) gamma distributions are exponential families.*

Example 150 (Random graph model). *(a) Suppose that we have $d \ge 3$ nodes, and links appear between nodes i and j ($i \ne j$) independently with probability p. Let $X_{i,j}$ be the indicator that there is a link between i and j. Show that the joint mass function of $X_{1,2},\ldots,X_{d-1,d}$ is an exponential family. (b) If $s_1(x) = \sum_{i<j} x_{i,j}$ and $s_2(x) = \sum_{i<j<k} x_{i,j}x_{j,k}x_{k,i}$, discuss the properties of data from an exponential family with mass function*

$$f(x_{1,2},\ldots,x_{d-1,d}) = \exp\left\{s_1(x)\theta + s_2(x)\beta - \kappa(\theta,\beta) + c(x_{1,2},\ldots,x_{d-1,d})\right\}, \quad \theta,\beta \in \mathbb{R},$$

as θ and β vary.

Note to Example 149

☐ (a) We write

$$f(x;\lambda) = \lambda^x \exp(-\lambda)/x! = \exp(x\log\lambda - \lambda - \log x!), \quad \lambda > 0, x \in \{0,1,\ldots\},$$

which is of the required form with $n = p = 1$, $s(x) = x$, $\theta = \log\lambda \in \Theta = \mathbb{R}$, $\kappa(\theta) = \exp(\theta)$, and $c(x) = -\log x!$.

☐ (b) We write

$$\begin{aligned} f(x;\lambda,\alpha) &= \lambda^\alpha x^{\alpha-1}\exp(-\lambda x)/\Gamma(\alpha) \\ &= \exp\left\{\alpha\log x - \lambda x + \alpha\log\lambda - \log\Gamma(\alpha) - \log x\right\}, \quad \lambda,\alpha > 0, x > 0, \end{aligned}$$

which is of the required form with $n = 1$, $p = 2$, $\theta_1 = \alpha$, $\theta_2 = -\lambda$, so $\Theta = \mathbb{R}_+ \times \mathbb{R}_-$, $s_1(x) = \log x$, $s_2(x) = x$, so $D = \mathbb{R}\times\mathbb{R}_+$ and $\kappa(\theta) = \log\Gamma(\theta_1) - \theta_1\log(-\theta_2)$, $c(x) = -\log x$.

Note to Example 150

☐ (a) Since the $X_{i,j}$ are Bernoulli variables, we can write $f(x_{i,j}) = p^{x_{i,j}}(1-p)^{1-x_{i,j}}$, where $x_{i,j} \in \{0,1\}$, and $0 < p < 1$. Since they are independent, their joint mass function is

$$\begin{aligned} f(x_{1,2},\ldots,x_{d-1,d}) &= \prod_{i<j} p^{x_{i,j}}(1-p)^{1-x_{i,j}} \\ &= \exp\left\{\sum_{i<j} x_{i,j}\log p + \sum_{i<j}(1-x_{i,j})\log(1-p)\right\} \\ &= \exp\left[\sum_{i<j} x_{i,j}\log\{p/(1-p)\} + \frac{d(d-1)}{2}\log(1-p)\right], \end{aligned}$$

which is of the given form with $n = d(d-1)/2$, $p = 1$, $s(x) = \sum_{i<j} x_{i,j}$, $c(x_{1,2},\ldots,x_{d-1,d}) \equiv 0$, $\theta = \log\{p/(1-p)\} \in \Theta = \mathbb{R}$, and $\kappa(\theta) = d(d-1)\log(1+e^\theta)/2$ (check this).
Note that $p = 1/2$ corresponds to $\theta = 0$, which corresponds to links appearing independently with probability 0.5, whereas setting $\theta \ll 0$ will give a very sparse graph, with very few links.

☐ (b) Here $s_1(x)$ counts how many links there are, and $s_2(x)$ counts how many triangles there are. Increasing β therefore gives more probability to graphs with lots of triangles, whereas decreasing β makes triangles less likely. So, taking $\theta \ll 0$ and $\beta \gg 0$ will tend to give a graph with a few links, but mostly in triangles. Note that the normalising constant is very complex, as it is

$$\kappa(\theta,\beta) = \log\sum\exp\left\{s_1(x)\theta + s_2(x)\beta\right\},$$

where the sum is over all 2^n possible values of $(x_{1,2},\ldots,x_{d-1,d})$.

☐ Exponential families are very useful in practice, because

- many standard distributions can be written as exponential families,
- we can construct new ones to model things of interest to us,
- they have a unified probabilistic and statistical theory, with many nice properties.

Marginal and conditional distributions

Definition 151. *The* **marginal probability mass/density function** *of* X *is*

$$f_X(x) = \begin{cases} \sum_y f_{X,Y}(x,y), & \text{discrete case,} \\ \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dy, & \text{continuous case,} \end{cases} \qquad x \in \mathbb{R}.$$

The **conditional probability mass/density function** *of* Y *given* X *is*

$$f_{Y|X}(y \mid x) = \frac{f_{X,Y}(x,y)}{f_X(x)}, \quad y \in \mathbb{R},$$

provided $f_X(x) > 0$. *If* (X,Y) *is discrete,*

$$f_X(x) = \mathrm{P}(X = x), \quad f_{Y|X}(y \mid x) = \mathrm{P}(Y = y \mid X = x).$$

- ☐ The conditional density $f_{Y|X}(y \mid x)$ is undefined if $f_X(x) = 0$. (Why?)
- ☐ Analogous definitions exist for $f_Y(y), f_{X|Y}(x \mid y)$, and for the conditional cumulative distribution functions $F_{X|Y}(x \mid y), F_{Y|X}(y \mid x)$.

Examples

Example 152. *Calculate the conditional PMF of* Y *given* X, *and the marginal PMFs of Example 145.*

Example 153. *Calculate the marginal and conditional densities for Example 147.*

Example 154. *Every day I receive a number of emails whose distribution is Poisson, with parameter* $\mu = 100$. *Each is a spam independently with probability* $p = 0.9$. *Find the distribution of the number of good emails which I receive. Given that I have received 15 good ones, find the distribution of the total number that I received.*

Note to Example 152

The joint mass function can be represented as

x	y	$f(x,y)$
0	0	2/6
0	1	2/6
1	0	1/6
1	2	1/6

so

$$\begin{aligned} f_X(0) &= f(0,0)+f(0,1)=2/3, \quad f_X(1)=f(1,0)+f(1,2)=1/3, \\ f_Y(0) &= f(0,0)+f(1,0)=1/2, \quad f_Y(1)=f(0,1)=1/3, \quad f_Y(2)=f(1,2)=1/6, \end{aligned}$$

and from which we can compute the required conditional distribution.
For example, we have

$$f_{Y|X}(y \mid x=0) = \frac{f_{X,Y}(0,y)}{f_X(0)} = \begin{cases} \frac{2/6}{2/3}=\frac{1}{2}, & y \in \{0,1\}, \\ 0, & \text{otherwise}, \end{cases}$$

and so we obtain

x	y	$f(y \mid x)$
0	0	1/2
0	1	1/2
1	0	1/2
1	2	1/2

Note to Example 153

- The only interesting cases are when $x, y > 0$. In this case the marginal density of X is

$$f_X(x) = 2\int_{y=x}^{\infty} e^{-x-y}\,dy = 2e^{-2x}, \quad x>0,$$

  and obviously this integrates to unity. The marginal density of Y is

$$f_Y(y) = 2\int_{x=0}^{y} e^{-x-y}\,dx = 2e^{-y}(1-e^{-y}), \quad y>0,$$

  and its integral is $2(1-1/2)=1$, so this is also a valid density function.
- For the conditional densities we have

$$f(y \mid x) = 2e^{-x-y}/(2e^{-2x}) = e^{x-y}, \quad y>x,$$

  and

$$f(x \mid y) = 2e^{-x-y}/\{2e^{-y}(1-e^{-y})\} = e^{-x}/(1-e^{-y}), \quad 0<x<y.$$

  It is easy to check that both conditional densities integrate to unity. Compare to Example 120.

Note to Example 154

Let N denote the total number of emails, and G the number of good ones. Then conditional on $N = n$, $G \sim B(n, p)$, so

$$f_{G,N}(g,n) = f_{G|N}(g \mid n) f_N(n) = \frac{n!}{g!(n-g)!}(1-p)^g p^{n-g} \times \frac{\mu^n}{n!} e^{-\mu}, \quad n \in \{0,1,2,\ldots\}, g \in \{0,1,\ldots,n\}$$

where $\mu > 0$ and $0 < p < 1$. Thus the number of good emails G has density

$$\begin{aligned} f_G(g) &= \sum_{n=g}^{\infty} f_{G,N}(g,n) \\ &= \frac{e^{-\mu}\mu^g(1-p)^g}{g!} \times \sum_{n=g}^{\infty} \frac{1}{(n-g)!}\mu^{n-g}p^{n-g} \\ &= \frac{e^{-\mu}\mu^g(1-p)^g}{g!} \times \sum_{r=0}^{\infty} \frac{1}{r!}(\mu p)^r, \quad \text{where } r = n-g, \\ &= \frac{e^{-\mu}\mu^g(1-p)^g}{g!} \times e^{\mu p} = \frac{\{\mu(1-p)\}^g}{g!} e^{-\mu(1-p)}, \quad g \in \{0,1,\ldots\}, \end{aligned}$$

which is the Poisson mass function with parameter $\mu(1-p)$.
Finally, given that $G = g$,

$$f_{N|G}(n \mid g) = \frac{f_{G,N}(g,n)}{f_G(g)} = \frac{\frac{n!}{g!(n-g)!}(1-p)^g p^{n-g} \times \frac{\mu^n}{n!}e^{-\mu}}{e^{-\mu(1-p)}\mu^g(1-p)^g/g!} = e^{-p\mu}\frac{(p\mu)^{n-g}}{(n-g)!}, \quad n = g, g+1, \ldots,$$

which is a Poisson distribution with mean μp, shifted to start at $n = g$. Thus the number of spams $S = N - G$ has a Poisson distribution, with mean μp.

Multivariate random variables

Definition 155. *Let $X_1, \ldots, X_n$ be rvs defined on the same probability space. Their* **joint cumulative distribution function** *is*

$$F_{X_1,\ldots,X_n}(x_1,\ldots,x_n) = \mathrm{P}(X_1 \le x_1, \ldots, X_n \le x_n)$$

and their **joint density/mass probability function** *is*

$$f_{X_1,\ldots,X_n}(x_1,\ldots,x_n) = \begin{cases} \mathrm{P}(X_1 = x_1, \ldots, X_n = x_n), & \text{discrete case,} \\ \frac{\partial^n F_{X_1,\ldots,X_n}(x_1,\ldots,x_n)}{\partial x_1 \cdots \partial x_n}, & \text{continuous case.} \end{cases}$$

We analogously define the conditional and marginal densities, the cumulative distribution functions, etc., by replacing (X, Y) by $X = X_{\mathcal{A}}$, $Y = X_{\mathcal{B}}$, where $\mathcal{A}, \mathcal{B} \subset \{1,\ldots,n\}$ and $\mathcal{A} \cap \mathcal{B} = \emptyset$. So for example, if $n = 4$, we can consider the marginal distribution of (X_1, X_2) and its conditional distribution given (X_3, X_4).

Subsequently everything can be generalised to n variables, but for ease of notation we will mostly limit ourselves to the bivariate case.

Multinomial distribution

Definition 156. *The random variable* $(X_1, \ldots, X_k)$ *has the* **multinomial distribution of denominator** m **and probabilities** $(p_1, \ldots, p_k)$ *if its mass function is*

$$f(x_1, \ldots, x_k) = \frac{m!}{x_1! \times \cdots \times x_k!} p_1^{x_1} p_2^{x_2} \cdots p_k^{x_k}, \quad x_1, \ldots, x_k \in \{0, \ldots, m\}, \sum_{j=1}^{k} x_j = m,$$

where $m \in \mathbb{N}$ *and* $p_1 \ldots, p_k \in [0, 1]$, *with* $p_1 + \cdots + p_k = 1$.

This distribution appears as the distribution of the number of individuals in the categories $\{1, \ldots, k\}$ when m independent individuals fall into the classes with probabilities $\{p_1, \ldots, p_k\}$. It generalises the binomial distribution to $k > 2$ categories.

Example 157 (Vote). n *students vote for three candidates for the presidency of their syndicate. Let* X_1, X_2, X_3 *be the number of corresponding votes, and suppose that the* n *students vote independently with probabilities* $p_1 = 0.45$, $p_2 = 0.4$, *and* $p_3 = 0.15$. *Find the joint distribution of* X_1, X_2, X_3, *calculate the marginal distribution of* X_3, *and the conditional distribution of* X_1 *given* $X_3 = x_3$.

Note to Example 157

☐ This is a multinomial distribution with $k = 3$, denominator n, and the given probabilities. The joint density is therefore

$$f(x_1, x_2, x_3) = \frac{n!}{x_1!x_2!x_3!} p_1^{x_1} p_2^{x_2} p_3^{x_3}, \quad x_1, x_2, x_3 \in \{0, \ldots, n\}, \sum_{j=1}^{3} x_j = n.$$

☐ The marginal distribution of X_3 is the number of votes for the third candidate. If we say that a vote for him is a success, and a vote for one of the other two is a failure, we see that $X_3 \sim B(n, p_3)$: X_3 is binomial with denominator n and probability 0.15.
Alternatively we can compute the marginal density for $x_3 = 0, \ldots, n$ using Definition 151 with $X = X_3$ and $Y = (X_1, X_2)$ as

$$\begin{aligned}
\mathrm{P}(X_3 = x_3) &= \sum_{\{(x_1,x_2):x_1+x_2=n-x_3\}} \frac{n!}{x_1!x_2!x_3!} p_1^{x_1} p_2^{x_2} p_3^{x_3} \\
&= \frac{n!}{x_3!(x_1+x_2)!} p_3^{x_3} \sum_{\{(x_1,x_2):x_1+x_2=n-x_3\}} \frac{(x_1+x_2)!}{x_1!x_2!} p_1^{x_1} p_2^{x_2} \\
&= \frac{n!}{x_3!(x_1+x_2)!} p_3^{x_3} (p_1+p_2)^{n-x_3} \\
&= \frac{n!}{(n-x_3)!x_3!} p_3^{x_3} (1-p_3)^{n-x_3},
\end{aligned}$$

using Newton's binomial formula (Theorem 17) and the fact that $p_1 + p_2 = 1 - p_3$. Thus again we see that $X_3 \sim B(n, p_3)$.

☐ If we now take the ratio of the joint density of (X_1, X_2, X_3) to the marginal density of X_3, we obtain the conditional density

$$\begin{aligned}
f_{X_1,X_2|X_3}(x_1, x_2 \mid x_3) &= \frac{f_{X_1,X_2,X_3}(x_1, x_2, x_3)}{f_{X_3}(x_3)} \\
&= \frac{\frac{n!}{x_1!x_2!x_3!} p_1^{x_1} p_2^{x_2} p_3^{x_3}}{\frac{n!}{x_3!(x_1+x_2)!} (p_1+p_2)^{x_1+x_2} p_3^{x_3}} \\
&= \frac{(x_1+x_2)!}{x_1!x_2!} \pi_1^{x_1} \pi_2^{x_2}, \quad 0 \le x_1 \le x+1+x_2,
\end{aligned}$$

where $\pi_1 = p_1/(p_1 + p_2)$, $\pi_2 = 1 - \pi_1$. This density is binomial with denominator $n - x_3 = x_1 + x_2$ and probability $\pi_1 = p_1/(1 - p_3)$. Note that $X_2 = n - x_3 - X_1$, so although the conditional mass function here has two arguments X_1, X_2, in reality it is of dimension 1.

☐ We conclude that, conditional on knowing the vote for one candidate, $X_3 = x_3$, the split of votes for the other two candidates has a binomial distribution. If we regard a vote for candidate 1 as a 'success', then $X_1 \sim B(n - x_3, \pi_1)$, where $n - x_3$ is the number of votes not for candidate 3, and π_1 is the conditional probability of voting for candidate 1, given that a voter has not chosen candidate 3.

Independence

Definition 158. *Random variables X, Y defined on the same probability space are* **independent** *if*

$$\mathrm{P}(X \in \mathcal{A}, Y \in \mathcal{B}) = \mathrm{P}(X \in \mathcal{A})\mathrm{P}(Y \in \mathcal{B}), \quad \mathcal{A}, \mathcal{B} \subset \mathbb{R}.$$

By letting $\mathcal{A} = (-\infty, x]$ and $\mathcal{B} = (-\infty, y]$, we find that

$$F_{X,Y}(x, y) = \cdots = F_X(x)F_Y(y), \quad x, y \in \mathbb{R},$$

implying the equivalent condition

$$f_{X,Y}(x, y) = f_X(x)f_Y(y), \quad x, y \in \mathbb{R}, \tag{2}$$

which will be our criterion of independence. This condition concerns the **functions** *$f_{X,Y}(x, y)$, $f_X(x)$, $f_Y(y)$: X, Y are independent iff (2) remains true* **for all** *$x, y \in \mathbb{R}$.*

If X, Y are independent, then for all x such that $f_X(x) > 0$,

$$f_{Y|X}(y \mid x) = \frac{f_{X,Y}(x, y)}{f_X(x)} = \frac{f_X(x)f_Y(y)}{f_X(x)} = f_Y(y), \quad y \in \mathbb{R}.$$

Thus knowing that $X = x$ does not affect the density of Y: this is an obvious meaning of "independence". By symmetry $f_{X|Y}(x \mid y) = f_X(x)$ for all y such that $f_Y(y) > 0$.

Examples

Example 159. *Are (X, Y) independent in (a) Example 145? (b) Example 147? (c) when*

$$f_{X,Y}(x, y) \propto \begin{cases} e^{-3x-2y}, & x, y > 0, \\ 0, & \textit{sinon.} \end{cases}$$

If X and Y are independent, then in particular the support of (X, Y) must be of the form $S_X \times S_Y \subset \mathbb{R}^2$.

Definition 160. *A* **random sample of size** *n from a distribution F of density f is a set of n independent random variables which all have a distribution F. Equivalently we say that $X_1, \ldots, X_n$ are* **independent and identically distributed (iid)** *with distribution F, or with density f, and write $X_1, \ldots, X_n \overset{\text{iid}}{\sim} F$ or $X_1, \ldots, X_n \overset{\text{iid}}{\sim} f$.*

By independence, the joint density of $X_1, \ldots, X_n \overset{\text{iid}}{\sim} f$ is

$$f_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = \prod_{j=1}^{n} f_X(x_j).$$

Example 161. *If $X_1, X_2, X_3 \overset{\text{iid}}{\sim} \exp(\lambda)$, give their joint density.*

Note to Example 159

(a) Since

$$f_X(0)f_Y(2) = \frac{2}{3} \times \frac{1}{6} \neq f_{X,Y}(0,2) = 0,$$

X and Y are dependent. This is obvious, because if I have the wrong hat (i.e., $X = 0$), then it is impossible that both other persons have the correct hats (i.e., $Y = 2$ is impossible).
Finding a single pair (x, y) giving $f_{X,Y}(x, y) \neq f_X(x)f_Y(y)$ is enough to show dependence, while to show independence it must be true that $f_{X,Y}(x, y) = f_X(x)f_Y(y)$ for every possible (x, y).
(b) In this case

$$f_{X,Y}(x,y) = \begin{cases} 2e^{-x-y}, & y > x > 0, \\ 0, & \text{otherwise.} \end{cases}$$

and we previously saw that

$$f_X(x) = 2\exp(-2x)I(x > 0), \quad f_Y(y) = 2\exp(-y)\{1 - \exp(-y)\}I(y > 0),$$

so obviously the joint density is not the product of the marginals. This is equally obvious on looking at the conditional densities.
In this case, the dependence is clear without any computations, as the support of (X, Y) cannot be the product of sets $I_{\mathcal{A}}(x)I_{\mathcal{B}}(y)$, but it would have to be if they were independent.
(c) The density factorizes and the support is a Cartesian product, so they are independent.

Note to Example 161

The variables are independent, so

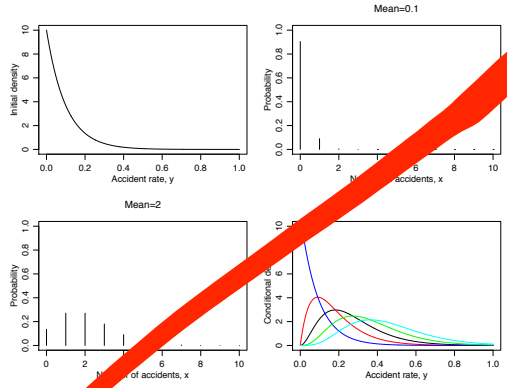
$$f(x_1, x_2, x_3) = f(x_1)f(x_2)f(x_3) = \lambda^3 \exp\{-\lambda(x_1 + x_2 + x_3)\}, \quad x_1, x_2, x_3 > 0, \quad \lambda > 0.$$

Mixed distributions

We sometimes encounter distributions with X discrete and Y continuous, or vice versa.

Example 162. *A big insurance company observes that the distribution of the number of insurance claims X in one year for its clients does not follow a Poisson distribution. However, a claim is a rare event, and so it seems reasonable that the distribution of small numbers should be applied. To model X, we suppose that for each client, the number of claims X in one year follows a Poisson distribution $\mathrm{Pois}(y)$, but that $Y \sim \mathrm{Gamma}(\alpha, \lambda)$: the mean number of claims for a client with $Y = y$ is then $\mathrm{E}(X \mid Y = y) = y$, since certain clients are more likely to make a claim than others.*
Find the joint distribution of (X, Y), the marginal distribution of X, and the conditional distribution of Y given $X = x$.

Insurance and learning



The graph shows how the knowledge of the number of accidents changes the distribution of the rate of accidents y for an insured party. Top left: the original density $f_Y(y)$. Top right: the conditional mass function $f_{X|Y}(x \mid y = 0.1)$ for a good driver. Bottom left: the conditional mass function $f_{X|Y}(x \mid y = 2)$ for a bad driver. Bottom right: the conditional densities $f_{Y|X}(y \mid x)$ with $x = 0$ (blue), 1 (red), 2 (black), 3 (green), 4 (cyan) (in order of decreasing maximal density).

5.2 Dependence slide 190

Joint moments

Definition 163. *Let X, Y be random variables of density $f_{X,Y}(x, y)$. Then if $\mathrm{E}\{|g(X,Y)|\} < \infty$, we can define the* **expectation** *of $g(X, Y)$ to be*

$$\mathrm{E}\{g(X,Y)\} = \begin{cases} \sum_{x,y} g(x,y) f_{X,Y}(x,y), & \textit{discrete case,} \\ \iint g(x,y) f_{X,Y}(x,y)\, dxdy, & \textit{continuous case.} \end{cases}$$

In particular we define the **joint moments** *and the* **joint central moments** *by*

$$\mathrm{E}(X^r Y^s), \quad \mathrm{E}\left[\{X - \mathrm{E}(X)\}^r \{Y - \mathrm{E}(Y)\}^s\right], \quad r, s \in \mathbb{N}.$$

The most important of these is the **covariance** *of X and Y,*

$$\mathrm{cov}(X,Y) = \mathrm{E}\left[\{X - \mathrm{E}(X)\}\{Y - \mathrm{E}(Y)\}\right] = \mathrm{E}(XY) - \mathrm{E}(X)\mathrm{E}(Y).$$

Properties of covariance

Theorem 164. *Let X, Y, Z be random variables and $a, b, c, d \in \mathbb{R}$ constants. The covariance satisfies:*

$$\begin{aligned}
\operatorname{cov}(X, X) &= \operatorname{var}(X); \\
\operatorname{cov}(a, X) &= 0; \\
\operatorname{cov}(X, Y) &= \operatorname{cov}(Y, X), \quad \textit{(symmetry)}; \\
\operatorname{cov}(a + bX + cY, Z) &= b\operatorname{cov}(X, Z) + c\operatorname{cov}(Y, Z), \quad \textit{(bilinearity)}; \\
\operatorname{cov}(a + bX, c + dY) &= bd\operatorname{cov}(X, Y); \\
\operatorname{var}(a + bX + cY) &= b^2\operatorname{var}(X) + 2bc\operatorname{cov}(X, Y) + c^2\operatorname{var}(Y); \\
\operatorname{cov}(X, Y)^2 &\leq \operatorname{var}(X)\operatorname{var}(Y), \quad \textit{(Cauchy–Schwarz inequality)}.
\end{aligned}$$

Note to Theorem 164

- All of this is mechanical computation. The only part that needs any thought is the last. For any $a \in \mathbb{R}$, we have

$$\operatorname{var}(aX + Y) = a^2\operatorname{var}(X) + 2a\operatorname{cov}(X, Y) + \operatorname{var}(Y) = Aa^2 + Ba + C \geq 0,$$

  and since this quadratic polynomial in a has at most one real root, we have

$$B^2 - 4AC = 4\operatorname{cov}(X, Y)^2 - 4\operatorname{var}(X)\operatorname{var}(Y) \leq 0,$$

  leading to $\operatorname{cov}(X, Y)^2 \leq \operatorname{var}(X)\operatorname{var}(Y)$.
- Equality would mean that there is precisely one real root, so $\operatorname{var}(aX + Y) = 0$ for some a, in which case $aX + Y$ is a constant, c, say, with probability one, and therefore provided $a \neq 0$ there is an exact linear relation $aX + Y = c$ between X and Y.

Independence and covariance

If X and Y are independent and $g(X)$, $h(Y)$ are functions whose expectations exist, then

$$\mathrm{E}\{g(X)h(Y)\} = \cdots = \mathrm{E}\{g(X)\}\mathrm{E}\{h(Y)\}.$$

By letting $g(X) = X - \mathrm{E}(X)$ and $h(Y) = Y - \mathrm{E}(Y)$, we can see that if X and Y are independent, then

$$\operatorname{cov}(X, Y) = \cdots = 0.$$

Thus X, Y indep $\Rightarrow \operatorname{cov}(X, Y) = 0$. However, the converse is false.

Linear combinations of random variables

Definition 165. *The* average *of random variables* $X_1, \ldots, X_n$ *is* $\overline{X} = n^{-1} \sum_{j=1}^n X_j$.

Lemma 166. *Let* $X_1, \ldots, X_n$ *be random variables and* $a, b_1, \ldots, b_n$ *be constants. Then (a)*

$$\begin{aligned} \mathrm{E}(a + b_1 X_1 + \cdots + b_n X_n) &= a + \sum_{j=1}^n b_j \mathrm{E}(X_j), \\ \mathrm{var}(a + b_1 X_1 + \cdots + b_n X_n) &= \sum_{j=1}^n b_j^2 \mathrm{var}(X_j) + \sum_{j \neq k} b_j b_k \,\mathrm{cov}(X_j, X_k). \end{aligned}$$

(b) If $X_1, \ldots, X_n$ *are independent, then* $\mathrm{cov}(X_j, X_k) = 0$, $j \neq k$, *so*

$$\mathrm{var}(a + b_1 X_1 + \cdots + b_n X_n) = \sum_{j=1}^n b_j^2 \mathrm{var}(X_j).$$

(c) If $X_1, \ldots, X_n$ *are independent and all have mean* μ *and variance* σ^2, *then*

$$\mathrm{E}(\overline{X}) = \mu, \quad \mathrm{var}(\overline{X}) = \sigma^2/n.$$

Example 167. *Let* X_1, X_2 *be independent rv's with* $\mathrm{E}(X_1) = 1$, $\mathrm{var}(X_1) = 1$, $\mathrm{E}(X_2) = 2$, $\mathrm{var}(X_2) = 4$, *and* $Y = 16 + 5X_1 - 6X_2$. *Calculate* $\mathrm{E}(Y)$, $\mathrm{var}(Y)$.

Note to Lemma 166

(a) The expectation of $a + b_1 X_1 + \cdots + b_n X_n$ follows easily from the fact that expectation is a linear operator.
The variance of $a + b_1 X_1 + \cdots + b_n X_n$ follows by extending the result on $\mathrm{var}(a + bX + cY)$ from Theorem 164 in an obvious way.
(b) Obvious.
(c) Use (a) and (b) with $a = 0$, $b_1 = \cdots = b_n = 1/n$ and the facts that $\mathrm{E}(X_j) = \mu$, $\mathrm{var}(X_j) = \sigma^2$ and $\mathrm{cov}(X_j, X_k) = 0$ when $j \neq k$, since the variables are independent.

Note to Example 167

Lemma 166 gives

$$\begin{aligned} \mathrm{E}(Y) &= \mathrm{E}(16 + 5X_1 - 6X_2) = 16 + 5\mathrm{E}(X_1) - 6\mathrm{E}(X_2) = 16 + 5 \times 1 - 6 \times 2 = 9, \\ \mathrm{var}(Y) &= \mathrm{var}(16 + 5X_1 - 6X_2) = 5^2 \mathrm{var}(X_1) + (-6)^2 \mathrm{var}(X_2) = 25 \times 1 + 36 \times 4 = 169. \end{aligned}$$

Correlation

Unfortunately the covariance depends on the units of measurement, so we often use the following dimensionless measure of dependence.

Definition 168. *The* correlation *of* X, Y *is*

$$\mathrm{corr}(X,Y) = \frac{\mathrm{cov}(X,Y)}{\{\mathrm{var}(X)\mathrm{var}(Y)\}^{1/2}}.$$

This measures the linear dependence between X and Y.

Example 169. *We can model the heredity of a quantitative genetic characteristic as follows. Let* X *be its value for a parent, and* Y_1 *and* Y_2 *its values for two children.*
Let $Z_1, Z_2, Z_3 \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$ *and take*

$$X = Z_1, \quad Y_1 = \rho Z_1 + (1-\rho^2)^{1/2} Z_2, \quad Y_2 = \rho Z_1 + (1-\rho^2)^{1/2} Z_3, \quad 0 < \rho < 1.$$

Calculate $\mathrm{E}(X)$, $\mathrm{E}(Y_j)$, $\mathrm{corr}(X, Y_j)$ *and* $\mathrm{corr}(Y_1, Y_2)$.

Note to Example 169

- ☐ Easy to use linearity of expectation to see that $\mathrm{E}(X) = \mathrm{E}(Y_j) = 0$ and that $\mathrm{var}(X) = \mathrm{var}(Y_j) = 1$.
- ☐ Since the Zs are independent and therefore are uncorrelated, and using the bilinearity of covariance, we have

$$\begin{aligned}
\mathrm{cov}(X, Y_j) &= \mathrm{cov}\{Z_1, \rho Z_1 + (1-\rho^2)^{1/2} Z_j\} \\
&= \mathrm{cov}(Z_1, \rho Z_1) + \mathrm{cov}\{Z_1, (1-\rho^2)^{1/2} Z_j\} \\
&= \rho \mathrm{cov}(Z_1, Z_1) + (1-\rho^2)^{1/2} \mathrm{cov}(Z_1, Z_j) \\
&= \rho \mathrm{var}(Z_1) + 0 = \rho.
\end{aligned}$$

- ☐ Likewise

$$\begin{aligned}
\mathrm{cov}(Y_1, Y_2) &= \mathrm{cov}\{\rho Z_1 + (1-\rho^2)^{1/2} Z_2, \rho Z_1 + (1-\rho^2)^{1/2} Z_3\} \\
&= \rho^2 \mathrm{cov}(Z_1, Z_1) + \rho(1-\rho^2)^{1/2} \mathrm{cov}(Z_1, Z_3) + \rho(1-\rho^2)^{1/2} \mathrm{cov}(Z_2, Z_1) \\
&\quad + (1-\rho^2)\mathrm{cov}(Z_2, Z_3) \\
&= \rho^2 \mathrm{cov}(Z_1, Z_1) \\
&= \rho^2 \mathrm{var}(Z_1) = \rho^2.
\end{aligned}$$

- ☐ Therefore $\mathrm{corr}(X, Y_j) = \rho > \mathrm{corr}(Y_1, Y_2) = \rho^2$, since $0 < \rho < 1$
  So the correlation between siblings is less than that between a parent and his/her offspring. A similar computation shows that the correlation between cousins will be ρ^4.

Properties of correlation

Theorem 170. *Let X, Y be random variables having correlation $\rho = \mathrm{corr}(X, Y)$. Then:*

(a) $-1 \leq \rho \leq 1$*;*

(b) if $\rho = \pm 1$*, then there exist* $a, b, c \in \mathbb{R}$ *such that*

$$aX + bY + c = 0$$

with probability 1 (X and Y are then linearly dependent);

(c) if X, Y are independent, then $\mathrm{corr}(X, Y) = 0$*;*

(d) the effect of the transformation

$$(X, Y) \mapsto (a + bX, c + dY)$$

is

$$\mathrm{corr}(X, Y) \mapsto \mathrm{sign}(bd)\mathrm{corr}(X, Y).$$

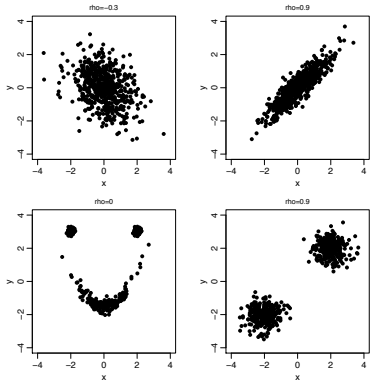
Note to Theorem 170

(a) Just apply the Cauchy–Schwarz inequality.
(b) Equality in the Cauchy–Schwarz inequality arises iff we have $\mathrm{var}(aX + bY + c) = 0$ for some $a, b, c \in \mathbb{R}$, and this can only mean that $aX + bY + c = 0$ with probability 1.
(c), (d) Just computations.

Limitations of correlation

Note that:

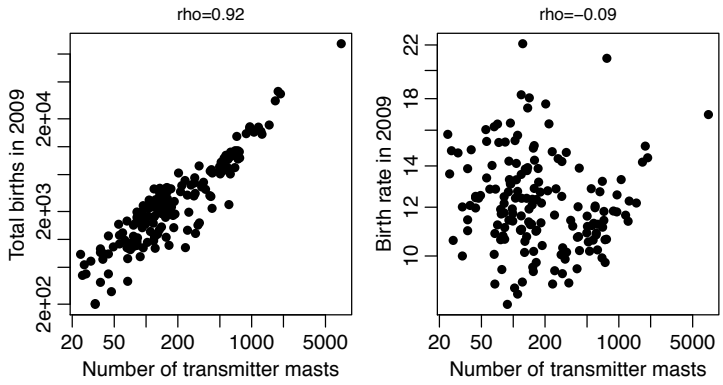
- ☐ correlation measures **linear** dependence, as in the upper panels below;
- ☐ we can have strong nonlinear dependence, but correlation zero, as in the bottom left panel;
- ☐ correlation can be strong but **specious**, as in the bottom right, where two sub-populations, each without correlation, are combined.



Correlation $\neq$ causation

Two variables can be very correlated without one *causing* changes in the other.

- ☐ The left panel shows strong dependence between the number of mobile phone transmitter masts, and the number of births in UK towns. Do masts increase fertility?
- ☐ The right panel shows that this dependence disappears when we allow for population size: more people $\Rightarrow$ more births and more transmitter masts. Adding masts will not lead to more babies.



Conditional expectation

Definition 171. *Let $g(X,Y)$ be a function of a random vector (X,Y). Its* **conditional expectation** *given $X = x$ is*

$$\mathrm{E}\{g(X,Y) \mid X = x\} = \begin{cases} \sum_y g(x,y) f_{Y|X}(y \mid x), & \text{in the discrete case,} \\ \int_{-\infty}^{\infty} g(x,y) f_{Y|X}(y \mid x)\, dy, & \text{in the continuous case,} \end{cases}$$

on the condition that $f_X(x) > 0$ and $\mathrm{E}\{|g(X,Y)| \mid X = x\} < \infty$. Note that the conditional expectation $\mathrm{E}\{g(X,Y) \mid X = x\}$ is a function of x.

Example 172. *Let $Z = XY$, where X and Y are independent, X having a Bernoulli distribution with probability p, and Y having the Poisson distribution with parameter λ.*

- ☐ *Find the density of Z.*
- ☐ *Find $\mathrm{E}(Z \mid X = x)$.*

Example 173. *Calculate the conditional expectation and variance of the total number of emails received in Example 154, given the arrival of g good emails.*

Note to Example 172

The event $Z = 0$ occurs iff we have either $X = 0$ and Y takes any value, or if $X = 1$ and $Y = 0$. Since X and Y are independent, we therefore have

$$\begin{aligned} f_Z(0) &= \sum_{y=0}^{\infty} \mathrm{P}(X = 0, Y = y) + \mathrm{P}(X = 1, Y = 0) \\ &= \sum_{y=0}^{\infty} \mathrm{P}(X = 0)\mathrm{P}(Y = y) + \mathrm{P}(X = 1)\mathrm{P}(Y = 0) \\ &= \mathrm{P}(X = 0) \sum_{y=0}^{\infty} \mathrm{P}(Y = y) + \mathrm{P}(X = 1)\mathrm{P}(Y = 0) \\ &= (1 - p) \times 1 + p \times e^{-\lambda}. \end{aligned}$$

Similarly

$$f_Z(z) = \mathrm{P}(X = 1, Y = z) = \mathrm{P}(X = 1)\mathrm{P}(Y = z) = p \times \lambda^z e^{-\lambda}/z!, \quad z = 1, 2, \ldots.$$

No other values for Z are possible. Clearly the above probabilities are non-negative, and

$$\sum_{z=0}^{\infty} f_Z(z) = (1 - p) + pe^{-\lambda} + \sum_{z=1}^{\infty} p\lambda^z e^{-\lambda}/z! = (1 - p) + p \sum_{z=0}^{\infty} \lambda^z e^{-\lambda}/z! = (1 - p) + p = 1,$$

so

$$f_Z = \begin{cases} (1 - p) + pe^{-\lambda}, & z = 0, \\ p\lambda^z e^{-\lambda}/z!, & z = 1, 2, \ldots, \end{cases}$$

is indeed a density function.
Now

$$\mathrm{E}(Z \mid X = x) = \mathrm{E}(XY \mid X = x) = \mathrm{E}(xY \mid X = x) = x\mathrm{E}(Y \mid X = x) = x\mathrm{E}(Y) = x\lambda,$$

since if we know that $X = x$, then the value x of X is a constant, and since Y and X are independent, $E\{h(Y) \mid X = x\} = x\mathrm{E}\{h(Y)\}$ for any function $h(Y)$. Therefore

$$\mathrm{E}(Z \mid X = 0) = 0, \quad \mathrm{E}(Z \mid X = 1) = \lambda.$$

Note to Example 173

The number of spams $S = N - G$ has a Poisson distribution, with mean $p\mu$. Thus the conditional expectation of N given $G = g$ is

$$\mathrm{E}(N \mid G = g) = \mathrm{E}(S + G \mid G = g) = \mathrm{E}(S + g \mid G = g) = p\mu + g,$$

because conditional on $G = g$, we treat g as a constant, and $S \sim \mathrm{Poiss}(p\mu)$. Likewise

$$\mathrm{var}(N \mid G = g) = \mathrm{var}(S + G \mid G = g) = \mathrm{var}(S + g \mid G = g) = \mathrm{var}(S \mid G = g) = p\mu.$$

Expectation and conditioning

Sometimes it is easier to calculate $\mathrm{E}\{g(X,Y)\}$ in stages.

Theorem 174. *If the required expectations exist, then*

$$\begin{aligned}\mathrm{E}\{g(X,Y)\} &= \mathrm{E}_X\left[\mathrm{E}\{g(X,Y) \mid X=x\}\right], \\ \mathrm{var}\{g(X,Y)\} &= \mathrm{E}_X\left[\mathrm{var}\{g(X,Y) \mid X=x\}\right] + \mathrm{var}_X\left[\mathrm{E}\{g(X,Y) \mid X=x\}\right].\end{aligned}$$

where E_X and var_X represent expectation and variance according to the distribution of X.

Example 175. *$n = 200$ persons pass a busker on a given day. Each one of them decides independently with probability $p = 0.05$ to give him money. The donations are independent, and have expectation $\mu = 2$\$ and variance $\sigma^2 = 1$\$^{2}. Find the expectation and the variance of the amount of money he receives.*

Note to Example 175

☐ Let $X_j = 1$ if the jth person decides to give him money and $X_j = 0$ otherwise, and let Y_j be the amount of money given by the jth person, if money is given. Then we can write his total takings as

$$T = g(X, Y) = Y_1 X_1 + \cdots + Y_n X_n,$$

where $X_1, \ldots, X_n \overset{\text{iid}}{\sim} B(1, p)$ are independent Bernoulli variables and $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} (\mu, \sigma^2)$. We want to compute $\mathrm{E}(T)$ and $\mathrm{var}(T)$.

☐ We first condition on $X_1, \ldots, X_n$, in which case (using an obvious shorthand notation)

$$\begin{aligned}
\mathrm{E}(T \mid X = x) &= \mathrm{E}(Y_1 X_1 + \cdots + Y_n X_n \mid X = x) \\
&= \sum_{j=1}^{n} \mathrm{E}(Y_j X_j \mid X = x) \\
&= \sum_{j=1}^{n} x_j \mathrm{E}(Y_j \mid X = x) = \sum_{j=1}^{n} x_j \mathrm{E}(Y_j) = \sum_{j=1}^{n} x_j \mu = \mu \sum_{j=1}^{n} x_j, \\
\mathrm{var}(T \mid X = x) &= \mathrm{var}(Y_1 X_1 + \cdots + Y_n X_n \mid X = x) \\
&= \sum_{j=1}^{n} \mathrm{var}(Y_j X_j \mid X = x) \quad \text{by independence of the } Y_j \\
&= \sum_{j=1}^{n} x_j^2 \mathrm{var}(Y_j \mid X = x) = \sum_{j=1}^{n} x_j^2 \sigma^2 = \sigma^2 \sum_{j=1}^{n} x_j.
\end{aligned}$$

In these expressions the X_j are treated as fixed quantities x_j and are regarded as constants, since the computations are conditional on $X_j = x_j$. Note that $x_j^2 = x_j$, since $x_j = 0, 1$.

☐ Now we 'uncondition', by replacing the values x_j of the X_j by the corresponding random variables, and in order to calculate the expressions in Theorem 174 we therefore need to compute

$$\mathrm{E}\left(\mu \sum_{j=1}^{n} X_j\right), \quad \mathrm{var}\left(\mu \sum_{j=1}^{n} X_j\right), \quad \mathrm{E}\left(\sigma^2 \sum_{j=1}^{n} X_j\right).$$

We have that $S = \sum_{j=1}^{n} X_j \sim B(n, p)$, so S has mean np and variance $np(1-p)$, and this yields

$$\begin{aligned}
\mathrm{E}(T) &= \mathrm{E}_X\left[\mathrm{E}\{T \mid X = x\}\right] = \mathrm{E}_X\left(\mu S\right) = \mu \mathrm{E}_X(S) = np\mu = 200 \times 0.05 \times 2 = 20, \\
\mathrm{var}(T) &= \mathrm{E}_X\left[\mathrm{var}\{T \mid X = x\}\right] + \mathrm{var}_X\left[\mathrm{E}\{T \mid X = x\}\right] \\
&= \mathrm{E}_X(\sigma^2 S) + \mathrm{var}_X(\mu S) = np\sigma^2 + \mu^2 np(1-p) \\
&= 200 \times 0.05 \times 1 + 2^2 \times 200 \times 0.05 \times 0.95 = 48.
\end{aligned}$$

Definition

Definition 176. *We define the* **moment-generating function** *of a random variable X by*

$$M_X(t) = \mathrm{E}(e^{tX})$$

for $t \in \mathbb{R}$ such that $M_X(t) < \infty$.

- $M_X(t)$ is also called the **Laplace transform** of $f_X(x)$.
- The MGF is useful as a summary of all the properties of X, we can write

$$M_X(t) = \mathrm{E}(e^{tX}) = \mathrm{E}\left(\sum_{r=0}^{\infty} \frac{t^r X^r}{r!}\right) = \sum_{r=0}^{\infty} \frac{t^r}{r!}\mathrm{E}(X^r),$$

  from which we can obtain all the moments $\mathrm{E}(X^r)$ by differentiation.

Example 177. *Calculate $M_X(t)$ when: (a) $X = c$ with probability one; (a) X is an indicator variable; (c) $X \sim B(n,p)$; (d) $X \sim \mathrm{Pois}(\lambda)$; (e) $X \sim \mathcal{N}(\mu, \sigma^2)$.*

Note to Example 177

(a) X is discrete, so $M_X(t) = 1 \times e^{t\times c} = e^{ct}$, valid for $t \in \mathbb{R}$.
(b) Here $M_X(t) = (1-p)e^{t\times 0} + pe^{t\times 1} = 1 - p + pe^t$, valid for $t \in \mathbb{R}$.
(c) Using the binomial theorem we have

$$M_X(t) = \sum_{x=0}^{n} e^{tx}\binom{n}{x}p^x(1-p)^{n-x} = \sum_{x=0}^{n}\binom{n}{x}(pe^t)^x(1-p)^{n-x} = (1-p+pe^t)^n, \quad t \in \mathbb{R}.$$

(d) We have

$$M_X(t) = \sum_{x=0}^{\infty} e^{xt}\frac{\lambda^x}{x!}e^{-\lambda} = \sum_{x=0}^{\infty}\frac{(\lambda e^t)^x}{x!}e^{-\lambda} = \exp(\lambda e^t)e^{-\lambda} = \exp\{\lambda(e^t-1)\}, \quad t \in \mathbb{R},$$

where we have used the exponential series $e^a = \sum_{n=0}^{\infty} a^n/n!$ for any $a \in \mathbb{R}$.
(e) We first consider $Z \sim \mathcal{N}(0,1)$ and compute

$$\mathrm{E}(e^{tZ}) = \int_{-\infty}^{\infty} e^{tz} \times \phi(z)\,dz = \int_{-\infty}^{\infty} e^{tz} \times \frac{1}{\sqrt{2\pi}}e^{-z^2/2}\,dz.$$

The fact that the $\mathcal{N}(\mu,\sigma^2)$ density integrates to 1, i.e.,

$$\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi\sigma^2}}e^{-(x-\mu)^2/(2\sigma^2)}\,dx = 1, \quad \mu \in \mathbb{R}, \sigma > 0$$

implies, on expanding the exponent and re-arranging the result, that

$$\int_{-\infty}^{\infty}\frac{1}{(2\pi)^{1/2}}\exp\left\{-x^2/(2\sigma^2) + x\mu/\sigma^2\right\}\,dx = \sigma\exp\left\{\mu^2/(2\sigma^2)\right\}, \quad \mu \in \mathbb{R}, \sigma \in \mathbb{R}_+.$$

If we take $\sigma = 1, \mu = t$, the left-hand side is the MGF of Z, and the right is $e^{t^2/2}$, valid for any $t \in \mathbb{R}$. (As an aside, note that if we take $\mu = 0, \sigma^2 = 1/(1-2t)$, then the left-hand side is the MGF of Z^2, and the right is $(1-2t)^{-1/2}$, valid only if $t < 1/2$. Thus

$$M_{Z^2}(t) = (1-2t)^{-1/2}, \quad t < 1/2.$$

This is the moment-generating function of a chi-squared random variable with one degree of freedom.) Now note that

$$\begin{aligned}
\mathrm{E}(e^{tX}) &= \mathrm{E}[\exp\{t(\mu+\sigma Z)\}] \\
&= \exp(t\mu)\mathrm{E}[\exp\{(t\sigma)Z\}] \\
&= \exp\{t\mu + (t\sigma)^2/2\} \\
&= \exp(t\mu + t^2\sigma^2/2), \quad t \in \mathbb{R}.
\end{aligned}$$

Important theorems I

Theorem 178. *If $M(t)$ is the MGF of a random variable X, then*

$$\begin{aligned} M_X(0) &= 1; \\ M_{a+bX}(t) &= e^{at} M_X(bt); \\ \mathrm{E}(X^r) &= \left.\frac{\partial^r M_X(t)}{\partial t^r}\right|_{t=0}; \\ \mathrm{E}(X) &= M'_X(0); \\ \mathrm{var}(X) &= M''_X(0) - M'_X(0)^2. \end{aligned}$$

Example 179. *Find the expectation and the variance of $X \sim \exp(\lambda)$.*

Theorem 180 (No proof). *There exists an injection between the cumulative distribution functions $F_X(x)$ and the moment-generating functions $M_X(t)$.*

Theorem 180 is very useful, as it says that if we recognise a MGF, we know to which distribution it corresponds.

Note to Theorem 178

This is just a series of mechanical computations, the last three of which involve differentiation of $M_X(t)$ with respect to t under the integral sign.

Note to Example 179

A simple calculation gives

$$M_X(t) = \int_0^\infty e^{xt} \times \lambda e^{-\lambda x}\, dx = \lambda \int_0^\infty e^{-(\lambda - t)x}\, dx = \frac{\lambda}{\lambda - t}, \quad t < \lambda.$$

Then we just differentiate $M_X(t)$ twice, getting

$$M'_X(t) = \lambda/(\lambda - t)^2, \quad M''_X(t) = 2\lambda/(\lambda - t)^3,$$

and set $t = 0$, using Theorem 178 to get the expectation and variance, λ^{-1} and λ^{-2} respectively.

Linear combinations

Theorem 181. *Let $a, b_1, \ldots, b_n \in \mathbb{R}$ and $X_1, \ldots, X_n$ be independent rv's whose MGFs exist. Then $Y = a + b_1X_1 + \cdots + b_nX_n$ has MGF*

$$M_Y(t) = \cdots = e^{ta} \prod_{j=1}^{n} M_{X_j}(tb_j).$$

In particular, if $X_1, \ldots, X_n$ is a random sample, then $S = X_1 + \cdots + X_n$ has

$$M_S(t) = M_X(t)^n.$$

Example 182. *Let $X_1, X_2 \overset{\text{ind}}{\sim}$ Pois(λ), Pois(μ). Find the distribution of $X_1 + X_2$.*

Example 183. *Let $X_1, \ldots, X_n$ be independent with $X_j \sim \mathcal{N}(\mu_j, \sigma_j^2)$. Show that*

$$Y = a + b_1X_1 + \cdots + b_nX_n \ \sim \ N(a + b_1\mu_1 + \cdots + b_n\mu_n, b_1^2\sigma_1^2 + \cdots + b_n^2\sigma_n^2) :$$

thus a linear combination of normal rv's is normal.

Note to Theorem 181

This simple calculation uses the fact that independence of $X_1, \ldots, X_n$ implies that the expectation of a product is the product of the expectations.

Note to Example 182

Theorem 181 implies that

$$M_{X_1+X_2}(t) = M_{X_1}(t)M_{X_2}(t) = \exp\{(\lambda_1 + \lambda_2)(e^t - 1)\}, \quad t \in \mathbb{R},$$

so by Theorem 180 and Theorem 177(d) we recognise that $X_1 + X_2$ is a Poisson variable with parameter $\lambda_1 + \lambda_2$.

Note to Example 183

Since the X_j are independent and their MGFs are $M_{X_j}(t) = \exp(t\mu_j + t^2\sigma_j^2/2)$, we can first use Theorem 181 to see that

$$\begin{aligned} M_Y(t) &= e^{ta} \prod_{j=1}^{n} M_{X_j}(tb_j) \\ &= \exp(ta) \prod_{j=1}^{n} \exp(tb_j\mu_j + t^2b_j^2\sigma_j^2/2) \\ &= \exp[t(a + b_1\mu_1 + \cdots + b_n\mu_n) + (t^2/2)(\sigma_1^2b_1^2 + \cdots + \sigma_n^2b_n^2)], \end{aligned}$$

and then Theorem 180 to obtain

$$Y \sim \mathcal{N}(a + b_1\mu_1 + \cdots + b_n\mu_n, b_1^2\sigma_1^2 + \cdots + b_n^2\sigma_n^2).$$

Important theorems II

Definition 184 ($\xrightarrow{D}$, Reminder). *Let $\{X_n\}$, X be random variables whose cumulative distribution functions are $\{F_n\}$, F. Then we say that the random variables $\{X_n\}$* **converge in distribution** *to X, if, for all $x \in \mathbb{R}$ where F is continuous,*

$$F_n(x) \to F(x), \quad n \to \infty.$$

We then write $X_n \xrightarrow{D} X$.

Theorem 185 (Continuity, no proof). *Let $\{X_n\}$, X be random variables with distribution functions $\{F_n\}$, F, whose MGFs $M_n(t)$, $M(t)$ exist for $0 \leq |t| < b$. Then if $M_n(t) \to M(t)$ for $|t| \leq a < b$ when $n \to \infty$, then $X_n \xrightarrow{D} X$, i.e., $F_n(x) \to F(x)$ at each $x \in \mathbb{R}$ where F is continuous.*

Example 186 (Law of small numbers, II). *Let $X_n \sim B(n, p_n)$ and $X \sim \text{Pois}(\lambda)$. Show that if $n \to \infty, p_n \to 0$ in such a way that $np_n \to \lambda$, then*

$$X_n \xrightarrow{D} X.$$

Note to Example 186

The results from Example 177 give $M_{X_n}(t) = (1 - p_n + p_n e^t)^n$ for $X_n \sim B(n, p_n)$ and $M_X(t) = \exp\{\lambda(e^t - 1)\}$ for $X \sim \text{Pois}(\lambda)$, both valid for $t \in \mathbb{R}$.
We use the fact that if $a \in \mathbb{R}$, then $(1 + a/n)^n \to e^a$ as $n \to \infty$.
If $n \to \infty$ and $np_n \to \lambda$, we can write

$$M_{X_n}(t) = (1 - p_n + p_n e^t)^n = \left\{1 + \frac{np_n(e^t - 1)}{n}\right\}^n \to \exp\{\lambda(e^t - 1)\} = M_X(t), \quad t \in \mathbb{R},$$

and this is true for any $t \in \mathbb{R}$. Hence the hypothesis of the theorem is clearly satisfied, and thus $X_n \xrightarrow{D} X$.

Mean vector and covariance matrix

Definition 187. *Let $X = (X_1, \ldots, X_p)^{\mathrm{T}}$ be a $p \times 1$ vector of random variables. Then*

$$\mathrm{E}(X)_{p\times 1} = \begin{pmatrix} \mathrm{E}(X_1) \\ \vdots \\ \mathrm{E}(X_p) \end{pmatrix},$$

$$\mathrm{var}(X)_{p\times p} = \begin{pmatrix} \mathrm{var}(X_1) & \mathrm{cov}(X_1, X_2) & \cdots & \mathrm{cov}(X_1, X_p) \\ \mathrm{cov}(X_1, X_2) & \mathrm{var}(X_2) & \cdots & \mathrm{cov}(X_2, X_p) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{cov}(X_1, X_p) & \mathrm{cov}(X_2, X_p) & \cdots & \mathrm{var}(X_p) \end{pmatrix},$$

are called the **expectation (mean vector)** *and the* **(co)-variance matrix** *of X.*

The matrix $\mathrm{var}(X)$ is positive semi-definite, since

$$\mathrm{var}\left(\sum_{j=1}^{p} a_j X_j\right) = a^{\mathrm{T}} \mathrm{var}(X) a \geq 0$$

for all vectors $a = (a_1, \ldots, a_p)^{\mathrm{T}} \in \mathbb{R}^p$.

Moment-generating function: multivariate case

Definition 188. *The* **moment-generating function (MGF)** *of a random vector $X_{p\times 1} = (X_1, \ldots, X_p)^{\mathrm{T}}$ is*

$$M_X(t) = \mathrm{E}(e^{t^{\mathrm{T}} X}) = \mathrm{E}(e^{\sum_{r=1}^{p} t_r X_r}), \quad t \in \mathcal{T},$$

where $\mathcal{T} = \{t \in \mathbb{R}^p : M_X(t) < \infty\}$. Let the rth and (r,s)th elements of the **mean vector** *$\mathrm{E}(X)_{p\times 1}$ and of the* **covariance matrix** *$\mathrm{var}(X)_{p\times p}$ be the quantities $\mathrm{E}(X_r)$ and $\mathrm{cov}(X_r, X_s)$.*

The MGF has the following properties:

- $0 \in \mathcal{T}$, thus $M_X(0) = 1$;
- we have

$$\mathrm{E}(X)_{p\times 1} = M'_X(0) = \left.\frac{\partial M_X(t)}{\partial t}\right|_{t=0}, \quad \mathrm{var}(X)_{p\times p} = \left.\frac{\partial^2 M_X(t)}{\partial t \partial t^{\mathrm{T}}}\right|_{t=0} - M'_X(0) M'_X(0)^{\mathrm{T}};$$

- if $\mathcal{A}, \mathcal{B} \subset \{1, \ldots, p\}$ and $\mathcal{A} \cap \mathcal{B} = \emptyset$, and we write $X_{\mathcal{A}}$ for the subvector of X containing $\{X_j : j \in \mathcal{A}\}$, etc., then $X_{\mathcal{A}}$ and $X_{\mathcal{B}}$ are independent iff

$$M_X(t) = \mathrm{E}(e^{t_{\mathcal{A}}^{\mathrm{T}} X_{\mathcal{A}} + t_{\mathcal{B}}^{\mathrm{T}} X_{\mathcal{B}}}) = M_{X_{\mathcal{A}}}(t_{\mathcal{A}}) M_{X_{\mathcal{B}}}(t_{\mathcal{B}}), \quad t \in \mathcal{T};$$

- there is an injective mapping between MGFs and probability distributions.

Example

Example 189. *Emails arrive as a Poisson process with rate λ emails per day: the number of emails arriving each day has the Poisson distribution with parameter λ. Each is a spam with probability p. Show that the numbers of good emails and of spams are independent Poisson variables with parameters $(1-p)\lambda$ and $p\lambda$.*

Note to Example 189

Let $N = S + G$ be the total number of spam and good emails, and note that $N \sim \text{Poiss}(\lambda)$, while $S = \sum_{j=1}^{N} I_j$, $G = \sum_{j=1}^{N}(1 - I_j)$, with I_j being the indicator that the jth message is a spam. The joint MGF of S and G is therefore

$$\begin{aligned} \mathrm{E}\left[\exp(t_1 S + t_2 G)\right] &= \mathrm{E}\left[\exp\left\{\sum_{j=1}^{N} t_1 I_j + t_2(1-I_j)\right\}\right] \\ &= \mathrm{E}_N\left(\mathrm{E}\left[\exp\left\{\sum_{j=1}^{N} t_1 I_j + t_2(1-I_j)\right\} \mid N = n\right]\right), \end{aligned}$$

where we have used the iterated expectation formula from Theorem 174. The inner expectation is

$$\mathrm{E}\left[\exp\left\{\sum_{j=1}^{N} t_1 I_j + t_2(1-I_j)\right\} \mid N = n\right] = \prod_{j=1}^{n} \mathrm{E}\left[\exp\left\{t_1 I_j + t_2(1-I_j)\right\}\right]$$

because conditional on $N = n$, the $I_1, \ldots, I_n$ are independent, and because they are Bernoulli variables each with success probability p, we have

$$\mathrm{E}\left[\exp\left\{t_1 I_j + t_2(1-I_j)\right\}\right] = e^{t_1}\mathrm{P}(I_j = 1) + e^{t_2}\mathrm{P}(I_j = 0) = pe^{t_1} + (1-p)e^{t_2}.$$

Therefore

$$\mathrm{E}\left[\exp\left\{\sum_{j=1}^{N} t_1 I_j + t_2(1-I_j)\right\} \mid N = n\right] = \left\{(1-p)e^{t_2} + pe^{t_1}\right\}^n,$$

and on inserting the right-hand side of this into the original expectation, and then treating $N = n$ as random with a Poiss(λ) distribution, we get

$$\begin{aligned} \mathrm{E}\left\{\exp(t_1 S + t_2 G)\right\} &= \mathrm{E}_N\left[\left\{(1-p)e^{t_2} + pe^{t_1}\right\}^N\right] \\ &= \sum_{n=0}^{\infty} \frac{\lambda^n}{n!} e^{-\lambda}\left\{(1-p)e^{t_2} + pe^{t_1}\right\}^n \\ &= \exp\left[-\lambda + \lambda\left\{(1-p)e^{t_2} + pe^{t_1}\right\}\right] \\ &= \exp\left[-\lambda(1-p+p) + \lambda\left\{(1-p)e^{t_2} + pe^{t_1}\right\}\right] \\ &= \exp\left\{-\lambda(1-p) + \lambda(1-p)e^{t_2}\right\} \times \exp\left(-\lambda p + \lambda p e^{t_1}\right) \\ &= \mathrm{E}\left\{\exp(t_2 G)\right\} \times \mathrm{E}\left\{\exp(t_1 S)\right\}, \end{aligned}$$

which is the MGF of two independent Poisson variables G and S with means $(1-p)\lambda$ and $p\lambda$, as required.

Parenthesis: Characteristic function

Many distributions do not have a MGF, since $\mathrm{E}(e^{tX}) < \infty$ only for $t = 0$. In this case, the Laplace transform of the density is not useful. Instead we can use the Fourier transform, leading us to the following definition.

Definition 190. *Let $i = \sqrt{-1}$. The* **characteristic function** *of X is*

$$\varphi_X(t) = \mathrm{E}(e^{itX}), \quad t \in \mathbb{R}.$$

Every random variable has a characteristic function, which possesses the same key properties as the MGF. Characteristic functions are however more complicated to handle, as they require ideas from complex analysis (path integrals, Cauchy's residue theorem, etc.).

Theorem 191 (No proof). *X and Y have the same cumulative distribution function if and only if they have the same characteristic function. If X is continuous and has density f and characteristic function φ then*

$$f(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-itx} \varphi(t)\, dt$$

for all x at which f is differentiable.

Parenthesis: Cumulant-generating function

Definition 192. *The* **cumulant-generating function (CGF)** *of X is $K_X(t) = \log M_X(t)$. The* **cumulants** *κ_r of X are defined by*

$$K_X(t) = \sum_{r=1}^{\infty} \frac{t^r}{r!} \kappa_r, \quad \kappa_r = \left. \frac{d^r K_X(t)}{dt^r} \right|_{t=0}.$$

It is easy to verify that $\mathrm{E}(X) = \kappa_1$ and $\mathrm{var}(X) = \kappa_2$.
The CGF is equivalent to the MGF, and so shares its properties, but it is often easier to work with the CGF.

Example 193. *Calculate the CGF and the cumulants of (a) $X \sim \mathcal{N}(\mu, \sigma^2)$; (b) $Y \sim \mathrm{Pois}(\lambda)$.*

Note to Example 193

We get directly from Example 177(d) that if $X \sim \mathcal{N}(\mu, \sigma^2)$, then (a)

$$K_X(t) = \log M_X(t) = t\mu + t^2\sigma^2/2, \quad t \in \mathbb{R},$$

so we see that $\kappa_1 = \mu$ and $\kappa_2 = \sigma^2$, with all other cumulants zero.
(b) Likewise from Example 177(c),

$$K_Y(t) = \log M_Y(t) = \lambda(e^t - 1), \quad t \in \mathbb{R},$$

so $\kappa_r = \lambda$ for all r.

Cumulants of sums of random variables

Theorem 194. *If $a, b_1, \ldots, b_n$ are constants and $X_1, \ldots, X_n$ are independent random variables, then*

$$K_{a+b_1X_1+\cdots+b_nX_n}(t) = ta + \sum_{j=1}^{n} K_{X_j}(tb_j).$$

If $X_1, \ldots, X_n$ are independent variables having cumulants $\kappa_{j,r}$, then the CGF of $S = X_1 + \cdots + X_n$ is

$$K_S(t) = \sum_{j=1}^{n} K_{X_j}(t) = \sum_{j=1}^{n}\sum_{r=1}^{\infty} \frac{t^r}{r!}\kappa_{j,r} = \sum_{r=1}^{\infty} \frac{t^r}{r!}\sum_{j=1}^{n} \kappa_{j,r} :$$

the rth cumulant of $X_1 + \cdots + X_n$ is the sum of the rth cumulants of the X_j. If the $X_1, \ldots, X_n \overset{\text{iid}}{\sim} F$ and have CGF $K(t)$, then t has CGF $nK(t)$ and has rth cumulant $n\kappa_r$.

Note to Theorem 194

For the first result, just use take logarithms in Theorem 181. For the second, just use the definition of the CGF in terms of the infinite series.
This result is very useful when looking at linear combinations of independent variables.

Multivariate cumulant-generating function

Definition 195. *The* **cumulant-generating function (CGF)** *of a random variable $X_{p\times 1} = (X_1, \ldots, X_p)^{\mathrm{T}}$ is*

$$K_X(t) = \log M_X(t) = \log \mathrm{E}(e^{t^{\mathrm{T}}X}), \quad t \in \mathcal{T},$$

where $\mathcal{T} = \{t \in \mathbb{R}^p : M_X(t) < \infty\}$.

The CGF has the following properties:

- $0 \in \mathcal{T}$, thus $K_X(0) = 0$;
- we have
$$\mathrm{E}(X)_{p\times 1} = K'_X(0) = \left.\frac{\partial K_X(t)}{\partial t}\right|_{t=0}, \quad \mathrm{var}(X)_{p\times p} = \left.\frac{\partial^2 K_X(t)}{\partial t \partial t^{\mathrm{T}}}\right|_{t=0};$$
- if $\mathcal{A}, \mathcal{B} \subset \{1, \ldots, p\}$ and $\mathcal{A} \cap \mathcal{B} = \emptyset$, and we write $X_{\mathcal{A}}$ for the subvector of X containing $\{X_j : j \in \mathcal{A}\}$, etc., then $X_{\mathcal{A}}$ and $X_{\mathcal{B}}$ are independent iff
$$K_X(t) = \log \mathrm{E}(e^{t_{\mathcal{A}}^{\mathrm{T}}X_{\mathcal{A}} + t_{\mathcal{B}}^{\mathrm{T}}X_{\mathcal{B}}}) = K_{X_{\mathcal{A}}}(t_{\mathcal{A}}) + K_{X_{\mathcal{B}}}(t_{\mathcal{B}}), \quad t \in \mathcal{T};$$
- there is an injective mapping between CGFs and probability distributions.

5.4 Multivariate Normal Distribution

Multivariate normal distribution

Definition 196. *The random vector $X = (X_1, \ldots, X_p)^{\mathrm{T}}$ has a* **multivariate normal distribution** *if there exist a $p \times 1$ vector $\mu = (\mu_1, \ldots, \mu_p)^{\mathrm{T}} \in \mathbb{R}^p$ and a $p \times p$ symmetric matrix Ω with elements ω_{jk} such that*

$$u^{\mathrm{T}} X \sim \mathcal{N}(u^{\mathrm{T}}\mu, u^{\mathrm{T}}\Omega u), \quad u \in \mathbb{R}^p;$$

then we write $X \sim \mathcal{N}_p(\mu, \Omega)$.

- ☐ Since $\mathrm{var}(u^{\mathrm{T}}X) = u^{\mathrm{T}}\Omega u \geq 0$ for any $u \in \mathbb{R}^p$, Ω must be positive semi-definite.
- ☐ This definition allows degenerate distributions, for which there exists a u such that $\mathrm{var}(u^{\mathrm{T}}X) = 0$. This gives mathematically clean results but can be avoided in applications by reformulating the problem to avoid degeneracy, effectively working in a space of dimension $m < p$.

Multivariate normal distribution, II

Lemma 197. *(a) We have*

$$\mathrm{E}(X_j) = \mu_j, \quad \mathrm{var}(X_j) = \omega_{jj}, \quad \mathrm{cov}(X_j, X_k) = \omega_{jk}, \quad j \neq k,$$

so μ and Ω are called the **mean vector** *and* **covariance matrix** *of X.*
(b) The moment-generating function of X is $M_X(u) = \exp(u^{\mathrm{T}}\mu + \frac{1}{2}u^{\mathrm{T}}\Omega u)$, for $u \in \mathbb{R}^p$.
(c) If $\mathcal{A}, \mathcal{B} \subset \{1, \ldots, p\}$, and $\mathcal{A} \cap \mathcal{B} = \emptyset$ then

$$X_{\mathcal{A}} \perp\!\!\!\perp X_{\mathcal{B}} \quad \Leftrightarrow \quad \Omega_{\mathcal{A},\mathcal{B}} = 0.$$

(d) If $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$, then $X_{n\times 1} = (X_1, \ldots, X_n)^{\mathrm{T}} \sim \mathcal{N}_n(\mu 1_n, \sigma^2 I_n)$.
(e) Linear combinations of normal variables are normal:

$$a_{r\times 1} + B_{r\times p}X \sim \mathcal{N}_r(a + B\mu, B\Omega B^{\mathrm{T}}).$$

Lemma 198. *The random vector $X \sim \mathcal{N}_p(\mu, \Omega)$ has a density function on $\mathbb{R}^p$ if and only if Ω is positive definite, i.e., Ω has rank p. If so, the density function is*

$$f(x; \mu, \Omega) = \frac{1}{(2\pi)^{p/2}|\Omega|^{1/2}} \exp\left\{-\tfrac{1}{2}(x-\mu)^{\mathrm{T}}\Omega^{-1}(x-\mu)\right\}, \quad x \in \mathbb{R}^p. \tag{3}$$

If not, X is a linear combination of variables that have a density function on $\mathbb{R}^m$, where $m < p$ is the rank of Ω.

Note to Lemma 197

(a) Let e_j denote the p-vector with 1 in the jth place and zeros everywhere else. Then $X_j = e_j^{\mathrm{T}} X \sim N(\mu_j, \omega_{jj})$, giving the mean and variance of X_j.
Now $\mathrm{var}(X_j + X_k) = \mathrm{var}(X_j) + \mathrm{var}(X_k) + 2\mathrm{cov}(X_j, X_k)$, and

$$X_j + X_k = (e_j + e_k)^{\mathrm{T}} X \sim \mathcal{N}(\mu_j + \mu_k, \omega_{jj} + \omega_{kk} + 2\omega_{jk}),$$

which implies that $\mathrm{cov}(X_j, X_k) = \omega_{jk} = \omega_{kj}$.
(b) Since $u^{\mathrm{T}} X \sim \mathcal{N}(u^{\mathrm{T}}\mu, u^{\mathrm{T}}\Omega u)$, its MGF is $M_{u^{\mathrm{T}}X}(t) = \mathrm{E}(e^{tu^{\mathrm{T}}X}) = \exp(tu^{\mathrm{T}}\mu + \frac{1}{2}t^2 u^{\mathrm{T}}\Omega u)$. The MGF of X is $M_X(u) = \mathrm{E}(e^{u^{\mathrm{T}}X}) = M_{u^{\mathrm{T}}X}(1) = \exp(u^{\mathrm{T}}\mu + \frac{1}{2}u^{\mathrm{T}}\Omega u)$, for any $u \in \mathbb{R}^p$, as stated.
(c) Without loss of generality, let $X_{\mathcal{A}} = (X_1, \ldots, X_q)^{\mathrm{T}}$, for $1 \le q < p$, and partition $t^{\mathrm{T}} = (t_{\mathcal{A}}^{\mathrm{T}}, t_{\mathcal{B}}^{\mathrm{T}})$, $\mu^{\mathrm{T}} = (\mu_{\mathcal{A}}^{\mathrm{T}}, \mu_{\mathcal{B}}^{\mathrm{T}})$, etc. Also without loss of generality suppose that $\mathcal{A} \cup \mathcal{B} = \{1, \ldots, n\}$, since otherwise we can just set $t_j = 0$ for $j \notin \mathcal{A} \cup \mathcal{B}$. Then, using matrix algebra, the joint CGF of X can be written as

$$K_X(t) = t^{\mathrm{T}}\mu + \tfrac{1}{2}t^{\mathrm{T}}\Omega t = t_{\mathcal{A}}^{\mathrm{T}}\mu_{\mathcal{A}} + t_{\mathcal{B}}^{\mathrm{T}}\mu_{\mathcal{B}} + \tfrac{1}{2}t_{\mathcal{A}}^{\mathrm{T}}\Omega_{\mathcal{A}\mathcal{A}}t_{\mathcal{A}} + \tfrac{1}{2}t_{\mathcal{B}}^{\mathrm{T}}\Omega_{\mathcal{B}\mathcal{B}}t_{\mathcal{B}} + t_{\mathcal{A}}^{\mathrm{T}}\Omega_{\mathcal{A}\mathcal{B}}t_{\mathcal{B}}.$$

This equals the sum of the CGFs of $X_{\mathcal{A}}$ and $X_{\mathcal{B}}$, i.e.,

$$K_{X_{\mathcal{A}}}(t) + K_{X_{\mathcal{B}}}(t) = t_{\mathcal{A}}^{\mathrm{T}}\mu_{\mathcal{A}} + \tfrac{1}{2}t_{\mathcal{A}}^{\mathrm{T}}\Omega_{\mathcal{A}\mathcal{A}}t_{\mathcal{A}} + t_{\mathcal{B}}^{\mathrm{T}}\mu_{\mathcal{B}} + +\tfrac{1}{2}t_{\mathcal{B}}^{\mathrm{T}}\Omega_{\mathcal{B}\mathcal{B}}t_{\mathcal{B}}$$

if and only if the final term of $K_X(t)$ equals zero for all t, which occurs if and only if $\Omega_{\mathcal{A}\mathcal{B}} = 0$. Hence the elements of the variance matrix corresponding to $\mathrm{cov}(X_r, X_s)$ must equal zero for any $r \in \mathcal{A}$ and $s \notin \mathcal{A}$, as required. Clearly this also holds if $\mathcal{A} \cup \mathcal{B} \neq \{1, \ldots, p\}$.
(d) In this case each of the X_j has mean μ and variance σ^2, and since they are independent, $\mathrm{cov}(X_j, X_k) = 0$ for $j \neq k$. If $u \in \mathbb{R}^n$, then $u^{\mathrm{T}}X$ is a linear combination of normal variables, with mean and variance

$$\sum u_j\mu = u^{\mathrm{T}}\mu 1_n, \quad \sum u_j^2\sigma^2 = u^{\mathrm{T}}\sigma_n^2 u,$$

so $X \sim \mathcal{N}_n(\mu 1_n, \sigma^2 I_n)$, as required.
(e) The MGF of $a + BX$ equals

$$\begin{aligned}
\mathrm{E}\left[\exp\{t^{\mathrm{T}}(a + BX)\}\right] = \mathrm{E}\left[\exp\{t^{\mathrm{T}}a + (B^{\mathrm{T}}t)^{\mathrm{T}}X)\}\right] &= e^{t^{\mathrm{T}}a} M_X(B^{\mathrm{T}}t) \\
&= \exp\{t^{\mathrm{T}}a + (B^{\mathrm{T}}t)^{\mathrm{T}}\mu + \tfrac{1}{2}(B^{\mathrm{T}}t)^{\mathrm{T}}\Omega(B^{\mathrm{T}}t)\} \\
&= \exp\left\{t^{\mathrm{T}}(a + B\mu) + \tfrac{1}{2}t^{\mathrm{T}}(B\Omega B^{\mathrm{T}})t\right\},
\end{aligned}$$

which is the MGF of the $\mathcal{N}_r(a + B\mu, B\Omega B^{\mathrm{T}})$ distribution.

Note I to Lemma 198

- ☐ Since Ω is positive semi-definite, the spectral theorem tells us that we may write $\Omega = A^{\mathrm{T}} D A$, where $D = \mathrm{diag}(d_1, \ldots, d_p)$ contains the eigenvalues of Ω, with $d_1 \geq d_2 \geq \cdots \geq d_p \geq 0$, and A is a $p \times p$ orthogonal matrix, i.e., $A^{\mathrm{T}} A = A A^{\mathrm{T}} = I_p$ and $|A| = 1$. Note that $|\Omega| = |A^{\mathrm{T}} D A| = |A^{\mathrm{T}}| \times |D| \times |A| = |D|$, and that if the inverse exists, $\Omega^{-1} = A^{\mathrm{T}} D^{-1} A$.
- ☐ Now $Y = AX \sim N_p(A\mu, A\Omega A^{\mathrm{T}})$, and $A\Omega A^{\mathrm{T}} = A A^{\mathrm{T}} D A A^{\mathrm{T}} = D$ is diagonal, so $Y_1, \ldots, Y_p$ are independent normal variables with means b_j given by the elements of $A\mu$ and variances d_j.
- ☐ Suppose that $d_p > 0$, so that Ω has rank p. Then all the Y_j have non-degenerate normal densities, and since they are independent, their joint density is

$$f_Y(y) = \prod_{j=1}^{p} (2\pi d_j)^{-1/2} \exp\left\{ -\frac{(y_j - b_j)^2}{2 d_j} \right\} = (2\pi)^{-p/2} |D|^{-1/2} \exp\left\{ -\tfrac{1}{2} (y-b)^{\mathrm{T}} D^{-1} (y-b) \right\}.$$

  Since $Y = AX$ and $A^{-1} = A^{\mathrm{T}}$, we have that $X = A^{\mathrm{T}} Y$, and this transformation has Jacobian $|A^{\mathrm{T}}| = 1$. Since $|D| = |\Omega|$, we can appeal to Theorem 204 and hence write the density of X as

$$f_X(x) = |A^{\mathrm{T}}| f_Y(Ax) = (2\pi)^{-p/2} |\Omega|^{-1/2} \exp\left\{ -\tfrac{1}{2} (Ax - A\mu)^{\mathrm{T}} D^{-1} (Ax - A\mu) \right\}, \quad x \in \mathbb{R}^p,$$

  where $(Ax - A\mu)^{\mathrm{T}} D^{-1} (Ax - A\mu) = (x-\mu)^{\mathrm{T}} A^{\mathrm{T}} D^{-1} A (x-\mu) = (x-\mu)^{\mathrm{T}} \Omega^{-1} (x-\mu)$, giving (3).
- ☐ If Ω has rank $m < p$, then $d_m > 0$ but $d_{m+1} = \cdots = d_p = 0$. In this case only $Y_1, \ldots, Y_m$ have positive variances, and the argument above allows us to construct a joint density for $Y_1, \ldots, Y_m$ on $\mathbb{R}^m$. Since $Y_m = b_m, \ldots, Y_p = b_p$ with probability one, we can write

$$X = A^{\mathrm{T}} Y = A^{\mathrm{T}} (Y_1, \ldots, Y_m, b_{m+1}, \ldots, b_p)^{\mathrm{T}},$$

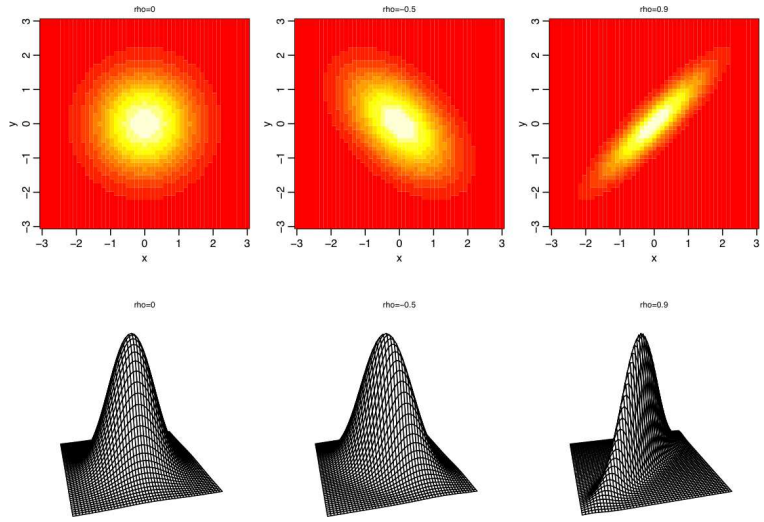
  which confirms that the density of X is positive only on an m-dimensional linear subspace of $\mathbb{R}^p$ generated by the variation of $Y_1, \ldots, Y_m$; it might be said to have only 'm degrees of freedom'.

Note II to Lemma 198

- Since Ω is symmetric and positive semi-definite, the spectral theorem tells us that we may write $\Omega = ADA^{\mathrm{T}}$, where $D = \mathrm{diag}(d_1, \ldots, d_p)$ contains the (real) eigenvalues of Ω, with $d_1 \geq d_2 \geq \cdots \geq d_p \geq 0$, and A is a $p \times p$ orthogonal matrix, i.e., $A^{\mathrm{T}}A = AA^{\mathrm{T}} = I_p$ and $|A| = 1$. The columns $A_1, \ldots, A_p$ of A are the eigenvectors corresponding to the respective eigenvalues; note that
$$\Omega = ADA^{\mathrm{T}} = \sum_{j=1}^{p} d_j a_j a_j^{\mathrm{T}},$$
that $|\Omega| = |ADA^{\mathrm{T}}| = |A| \times |D| \times |A^{\mathrm{T}}| = |D|$, and that $\Omega^{-1} = AD^{-1}A^{\mathrm{T}}$ if the inverse exists.
- Now let $Y_j \sim \mathcal{N}(0, d_j)$ be independent variables, let $Y = (Y_1, \ldots, Y_p)^{\mathrm{T}}$, and let $u \in \mathbb{R}^p$; note that if $d_j = 0$ then $Y_j = 0$ with probability one. Then
$$u^{\mathrm{T}}X = u^{\mathrm{T}}(\mu + AY) = u^{\mathrm{T}}\mu + \sum_{j=1}^{p} Y_j u^{\mathrm{T}} a_j$$
is a linear combination of normal variables, so it has a normal distribution, with mean $u^{\mathrm{T}}\mu$ and variance
$$\mathrm{var}\left(u^{\mathrm{T}}\mu + \sum_{j=1}^{p} Y_j u^{\mathrm{T}} a_j\right) = \sum_{j=1}^{n} (u^{\mathrm{T}}a_j)^2 \mathrm{var}(Y_j) = u^{\mathrm{T}}\left(\sum_{j=1}^{n} d_j a_j a_j^{\mathrm{T}}\right) u = u^{\mathrm{T}}\Omega u,$$
which implies that $X = \mu + AY \sim N_p(\mu, \Omega)$, according to Definition 196.
- Now $X = \mu + \sum_{j=1}^{p} Y_j a_j$ can be constructed by scaling the eigenvectors a_j of Ω by normally-distributed factors Y_j, so $X - \mu$ lies in the linear space $\mathcal{S} = \mathrm{span}(a_1, \ldots, a_m)$ generated by the eigenvectors a_j for which $d_j > 0$. If $d_p > 0$, then $m = p$ and $\mathcal{S} = \mathbb{R}^p$, but otherwise $\mathcal{S}$ is a proper subspace of $\mathbb{R}^p$ generated by $a_1, \ldots, a_m$, and $d_1 \geq \cdots \geq d_m > 0$ but $d_{m+1} = \cdots = d_p = 0$. In this case X has a density on $\mu + \mathcal{S}$, but places no probability elsewhere.
- For example, suppose that $p = 2$, $a_1 = (1, 0)^{\mathrm{T}}$ and $a_2 = (0, 1)^{\mathrm{T}}$. If $m = 2$, then $d_1, d_2 > 0$, and X can lie anywhere in $\mathbb{R}^2$, whereas if $m = 1$, then $d_1 > 0$ but $d_2 = 0$, and X can only take values in the x-axis, within which its density is $\mathcal{N}(\mu_1, d_1)$. If $m = 0$, then X takes the constant value μ with probability one.
- To compute the density of X, suppose that $m \geq 1$ and note that the non-degenerate part of $Y = A^{\mathrm{T}}(X - \mu)$, $Y_+ = (Y_1, \ldots, Y_m)^{\mathrm{T}}$, say, has joint density
$$f_{Y_+}(y_+) = \prod_{j=1}^{m} (2\pi d_j)^{-1/2} \exp\left\{-y_j^2/(2d_j)\right\} = (2\pi)^{-p/2} |D_+|^{-1/2} \exp\left(-\tfrac{1}{2} y_+^{\mathrm{T}} D_+^{-1} y_+\right),$$
where $D_+ = \mathrm{diag}(d_1, \ldots, d_m)$ and $y_+ = (y_1, \ldots, y_m)^{\mathrm{T}}$.
- If Ω has rank $m < p$, then $d_m > 0$ but $d_{m+1} = \cdots = d_p = 0$. In this case only $Y_1, \ldots, Y_m$ have positive variances, and the argument above allows us to construct a joint density for $Y_1, \ldots, Y_m$ on $\mathbb{R}^m$. Since $Y_m = b_m, \ldots, Y_p = b_p$ with probability one, we can write
$$X = A^{\mathrm{T}}Y = A^{\mathrm{T}}(Y_1, \ldots, Y_m, b_{m+1}, \ldots, b_p)^{\mathrm{T}},$$
which confirms that the density of X is positive only on an m-dimensional linear subspace of $\mathbb{R}^p$ generated by the variation of $Y_1, \ldots, Y_m$; it might be said to have only 'm degrees of freedom'.

Bivariate normal densities

Normal PDF with $p = 2$, $\mu_1 = \mu_2 = 0$, $\omega_{11} = \omega_{22} = 1$, and correlation $\rho = \omega_{12}/(\omega_{11}\omega_{22})^{1/2} = 0$ (left), -0.5 (centre) and 0.9 (right).



Examples

Example 199. *If $X \sim \mathcal{N}(1, 4)$, $Y \sim \mathcal{N}(-1, 9)$, $\mathrm{corr}(X, Y) = -1/6$, and they have a joint normal distribution, give the joint distribution of (X, Y). Hence find the distribution of $W = X + Y$.*

Example 200. *If $X_1, \ldots, X_4 \overset{\text{iid}}{\sim} N(0, \sigma^2)$, find the distribution of $Y = BX$ when*

$$B = \begin{pmatrix} 1 & -1 & -1 & -1 \\ 1 & -1 & 1 & 1 \\ 1 & 1 & -1 & 1 \\ 1 & 1 & 1 & -1 \end{pmatrix}.$$

Note to Example 199

Part (a) of Lemma 197 gives that

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim \mathcal{N}_2 \left\{ \begin{pmatrix} \mathrm{E}(X) \\ \mathrm{E}(Y) \end{pmatrix}, \begin{pmatrix} \mathrm{var}(X) & \mathrm{cov}(X,Y) \\ \mathrm{cov}(X,Y) & \mathrm{var}(Y) \end{pmatrix} \right\},$$

and we know all the elements of the matrices except

$$\mathrm{cov}(X,Y) = \mathrm{corr}(X,Y) \times \sqrt{\mathrm{var}(X)} \times \sqrt{\mathrm{var}(Y)} = -1/6 \times 2 \times 3 = -1.$$

Therefore

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim \mathcal{N}_2 \left\{ \begin{pmatrix} 1 \\ -1 \end{pmatrix}, \begin{pmatrix} 4 & -1 \\ -1 & 9 \end{pmatrix} \right\}.$$

Since W is a linear combination of normal variables, it has a normal distribution, and we can apply Part (e) of Lemma 197 with $r = 1$, $p = 2$, $a = 0$ and $B = (1,1)$ to obtain

$$W = (1,1) \begin{pmatrix} X \\ Y \end{pmatrix} \sim \mathcal{N}_1 \left\{ 0 + (1,1)(1,-1)^{\mathrm{T}}, (1,1) \begin{pmatrix} 4 & -1 \\ -1 & 9 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \end{pmatrix} \right\} = \mathcal{N}(0, 11).$$

Note to Example 200

For this we use parts (d) and (e) of Lemma 197. For (d) we take $\mu = 0_{4\times 1}$ and $\Omega = \sigma^2 I_4$, and for (e) we take $a = 0_{4\times 1}$ and the stated matrix B. Thus

$$Y = a + BX \sim \mathcal{N}_4 \left(a + B\mu, B\Omega B^{\mathrm{T}}\right) \overset{D}{=} \mathcal{N}_4 \left(0, \sigma^2 BB^{\mathrm{T}}\right) \overset{D}{=} \mathcal{N}_4 \left(0, 4\sigma^2 I_4\right),$$

because it is easy to check that $BB^{\mathrm{T}} = 4I_4$. Thus the variables $Y_1, \ldots, Y_4$ have $N(0, 4\sigma^2)$ distributions, and are independent because their covariance matrix is diagonal.

Marginal and conditional distributions

Theorem 201. *Let $X \sim \mathcal{N}_p(\mu_{p\times 1}, \Omega_{p\times p})$, where $|\Omega| > 0$, and let $\mathcal{A}, \mathcal{B} \subset \{1, \ldots, p\}$ with $|\mathcal{A}| = q < p, |\mathcal{B}| = r < p$ and $\mathcal{A} \cap \mathcal{B} = \emptyset$.*
Let $\mu_{\mathcal{A}}$, $\Omega_{\mathcal{A}}$ and $\Omega_{\mathcal{AB}}$ be respectively the $q \times 1$ subvector of μ, $q \times q$ and $q \times r$ submatrices of Ω conformable with $\mathcal{A}$, $\mathcal{A} \times \mathcal{A}$ and $\mathcal{A} \times \mathcal{B}$. Then:
(a) the marginal distribution of $X_{\mathcal{A}}$ is normal,

$$X_{\mathcal{A}} \sim \mathcal{N}_q(\mu_{\mathcal{A}}, \Omega_{\mathcal{A}});$$

(b) the conditional distribution of $X_{\mathcal{A}}$ given $X_{\mathcal{B}} = x_{\mathcal{B}}$ is normal,

$$X_{\mathcal{A}} \mid X_{\mathcal{B}} = x_{\mathcal{B}} \sim \mathcal{N}_q \left\{ \mu_{\mathcal{A}} + \Omega_{\mathcal{AB}} \Omega_{\mathcal{B}}^{-1} (x_{\mathcal{B}} - \mu_{\mathcal{B}}), \Omega_{\mathcal{A}} - \Omega_{\mathcal{AB}} \Omega_{\mathcal{B}}^{-1} \Omega_{\mathcal{BA}} \right\}.$$

This has two important implications:

- ☐ (a) implies that any subvector of X also has a multivariate normal distribution;
- ☐ (b) implies that two components of $X_{\mathcal{A}}$ are conditionally independent given $X_{\mathcal{B}}$ if and only if the corresponding off-diagonal element of $\Omega_{\mathcal{A}} - \Omega_{\mathcal{AB}} \Omega_{\mathcal{B}}^{-1} \Omega_{\mathcal{BA}}$ equals zero.

Proof of Theorem 201

First note that without loss of generality we can permute the elements of X so that the components of $X_{\mathcal{A}}$ appear before those of $X_{\mathcal{B}}$, then writing $X^{\mathrm{T}} = (X_{\mathcal{A}}^{\mathrm{T}}, X_{\mathcal{B}}^{\mathrm{T}})$. Partition the vectors t, μ, and the matrix Ω conformally with X, using obvious notation.
(a) The CGF of X is

$$K_X(t) = t^{\mathrm{T}}\mu + \tfrac{1}{2}t^{\mathrm{T}}\Omega t = \begin{pmatrix} t_{\mathcal{A}} \\ t_{\mathcal{B}} \end{pmatrix}^{\mathrm{T}} \begin{pmatrix} \mu_{\mathcal{A}} \\ \mu_{\mathcal{B}} \end{pmatrix} + \tfrac{1}{2} \begin{pmatrix} t_{\mathcal{A}} \\ t_{\mathcal{B}} \end{pmatrix}^{\mathrm{T}} \begin{pmatrix} \Omega_{\mathcal{A}} & \Omega_{\mathcal{AB}} \\ \Omega_{\mathcal{BA}} & \Omega_{\mathcal{B}} \end{pmatrix} \begin{pmatrix} t_{\mathcal{A}} \\ t_{\mathcal{B}} \end{pmatrix}$$

We obtain the marginal CGF of $X_{\mathcal{A}}$ by setting $t_{\mathcal{B}} = 0$, giving

$$K_X(t) = \begin{pmatrix} t_{\mathcal{A}} \\ 0 \end{pmatrix}^{\mathrm{T}} \begin{pmatrix} \mu_{\mathcal{A}} \\ \mu_{\mathcal{B}} \end{pmatrix} + \tfrac{1}{2} \begin{pmatrix} t_{\mathcal{A}} \\ 0 \end{pmatrix}^{\mathrm{T}} \begin{pmatrix} \Omega_{\mathcal{A}} & \Omega_{\mathcal{AB}} \\ \Omega_{\mathcal{BA}} & \Omega_{\mathcal{B}} \end{pmatrix} \begin{pmatrix} t_{\mathcal{A}} \\ 0 \end{pmatrix} = t_{\mathcal{A}}\mu_{\mathcal{A}} + \tfrac{1}{2}t_{\mathcal{A}}^{\mathrm{T}}\Omega_{\mathcal{A}}t_{\mathcal{A}},$$

which is the CGF of the $\mathcal{N}_q(\mu_{\mathcal{A}}, \Omega_{\mathcal{A}})$ distribution.
(b) Consider $W = X_{\mathcal{A}} - \Omega_{\mathcal{AB}}\Omega_{\mathcal{B}}^{-1}X_{\mathcal{B}}$. This is a linear combination of normals and so is normal, and its mean and variance matrix are

$$\mu_{\mathcal{A}} - \Omega_{\mathcal{AB}}\Omega_{\mathcal{B}}^{-1}\mu_{\mathcal{B}}, \quad \Omega_{\mathcal{A}} - \Omega_{\mathcal{AB}}\Omega_{\mathcal{B}}^{-1}\Omega_{\mathcal{BA}},$$

and as $\mathrm{cov}(X_{\mathcal{B}}, W) = 0$ (check!) and they are jointly normally distributed, $W \perp\!\!\!\perp X_{\mathcal{B}}$. Now

$$X_{\mathcal{A}} = W + \Omega_{\mathcal{AB}}\Omega_{\mathcal{B}}^{-1}X_{\mathcal{B}},$$

and as W and $X_{\mathcal{B}}$ are independent, the distribution of W is unchanged by conditioning on the event $X_{\mathcal{B}} = x_{\mathcal{B}}$. The conditional mean of $X_{\mathcal{A}}$ is therefore

$$\mathrm{E}(X_{\mathcal{A}} \mid X_{\mathcal{B}} = x_{\mathcal{B}}) = \mathrm{E}(W + \Omega_{\mathcal{AB}}\Omega_{\mathcal{B}}^{-1}X_{\mathcal{B}} \mid X_{\mathcal{B}} = x_{\mathcal{B}}) = \mathrm{E}(W) + \Omega_{\mathcal{AB}}\Omega_{\mathcal{B}}^{-1}x_{\mathcal{B}} = \mu_{\mathcal{A}} + \Omega_{\mathcal{AB}}\Omega_{\mathcal{B}}^{-1}(x_{\mathcal{B}} - \mu_{\mathcal{B}})$$

as required. Likewise

$$\mathrm{var}(X_{\mathcal{A}} \mid X_{\mathcal{B}} = x_{\mathcal{B}}) = \mathrm{var}(W + \Omega_{\mathcal{AB}}\Omega_{\mathcal{B}}^{-1}X_{\mathcal{B}} \mid X_{\mathcal{B}} = x_{\mathcal{B}}) = \mathrm{var}(W) = \Omega_{\mathcal{A}} - \Omega_{\mathcal{AB}}\Omega_{\mathcal{B}}^{-1}\Omega_{\mathcal{BA}},$$

because the term in $X_{\mathcal{B}}$ is conditionally constant. This gives the required result.

Example

Example 202. *Let (X_1, X_2) be the pair (height (cm), weight (kg)) for a population of people aged 20. To model this, we take*

$$\mu = \begin{pmatrix} 180 \\ 70 \end{pmatrix}, \quad \Omega = \begin{pmatrix} 225 & 90 \\ 90 & 100 \end{pmatrix}.$$

(a) Find the marginal distributions of X_1 and of X_2, and $\mathrm{corr}(X_1, X_2)$.
(b) Do the marginal distributions determine the joint distribution?
(c) Find the conditional distribution of X_2 given that $X_1 = x_1$, and of X_1 given that $X_2 = x_2$.

Note to Example 202

(a) The marginal distributions are $X_1 \sim \mathcal{N}(180, 225)$ and $X_2 \sim \mathcal{N}(70, 100)$. The correlation is

$$\frac{\omega_{12}}{\sqrt{\omega_{11}\omega_{22}}} = \frac{90}{\sqrt{225 \times 100}} = \frac{90}{150} = 0.6.$$

(b) Clearly not, because they don't determine the correlation.
(c) For this we have

$$X_{\mathcal{A}} \mid X_{\mathcal{B}} = x_{\mathcal{B}} \sim \mathcal{N}_q \left\{ \mu_{\mathcal{A}} + \Omega_{\mathcal{A}\mathcal{B}}\Omega_{\mathcal{B}}^{-1}(x_{\mathcal{B}} - \mu_{\mathcal{B}}), \Omega_{\mathcal{A}} - \Omega_{\mathcal{A}\mathcal{B}}\Omega_{\mathcal{B}}^{-1}\Omega_{\mathcal{B}\mathcal{A}} \right\}.$$

where $X_{\mathcal{A}} = X_2$, $X_{\mathcal{B}} = X_1$, so

$$\begin{aligned} \mu_{\mathcal{A}} + \Omega_{\mathcal{A}\mathcal{B}}\Omega_{\mathcal{B}}^{-1}(x_{\mathcal{B}} - \mu_{\mathcal{B}}) &= \mu_2 + \omega_{21}\omega_{11}^{-1}(x_1 - \mu_1) = 70 + 0.4(x_1 - 180), \\ \Omega_{\mathcal{A}} - \Omega_{\mathcal{A}\mathcal{B}}\Omega_{\mathcal{B}}^{-1}\Omega_{\mathcal{B}\mathcal{A}} &= 100 - 90^2/225 = 64. \end{aligned}$$

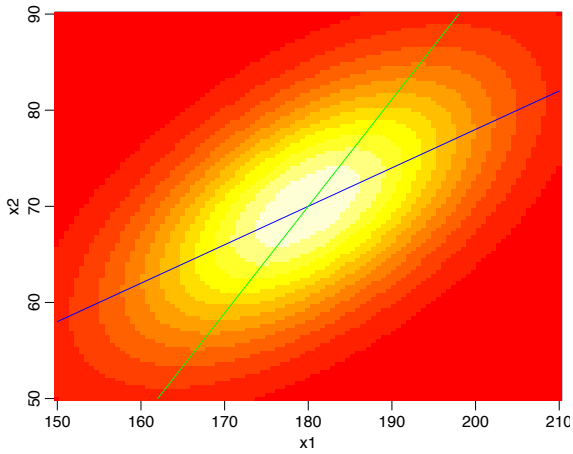
Thus $X_2 \mid X_1 = x_1 \sim \mathcal{N}\{70 + 0.4(x_1 - 180), 64\}$: larger height leads to larger weight, on average. A similar computation gives

$$X_1 \mid X_2 = x_2 \sim \mathcal{N}\{180 + 0.9(x_2 - 70), 144\}.$$

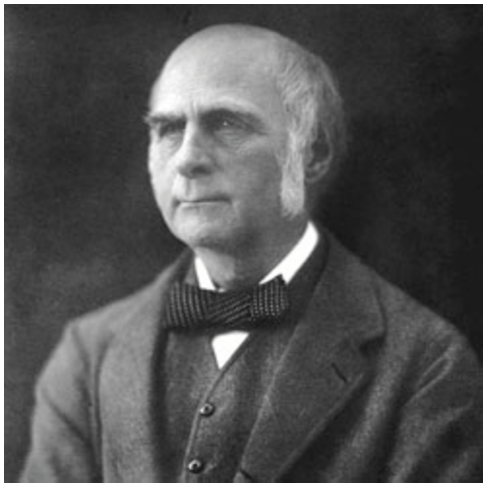
In each case the mean depends linearly on the conditioning variable, and the conditional variance is smaller than the marginal variance, consistent with the idea that conditioning adds information and therefore reduces uncertainty.

Bivariate normal distribution

The normal bivariate density for (X_1, X_2) =(hauteur, poids), as well as the straight lines $\mathrm{E}(X_2 \mid X_1 = x_1) = 70 + 0.4(x_1 - 180)$ (blue) and $\mathrm{E}(X_1 \mid X_2 = x_2) = 180 + 0.9(x_2 - 70)$ (green).



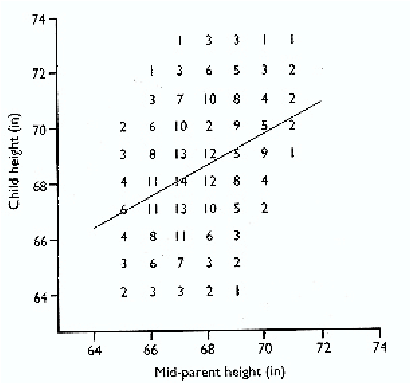
Francis Galton (1822–1911)



(Source: Wikipedia)

Regression to the mean

- ☐ Galton obtained the heights of parents and of their children, and fitted a line.
- ☐ The slope of the line < 1: the children of tall parents are smaller than them, on average, and the children of small parents are larger than them, on average.
- ☐ This effect is called **regression to the mean**, and appears in many contexts. For example, someone with an above-average mark on a midterm test will tend to do worse in the final, on average.



5.5 Transformations

slide 223

Reminder: Transformation of random variables

We often want to calculate the distributions of random variables based on other random variables.

- ☐ Let $Y = g(X)$, where the function g is known. We want to obtain F_Y and f_Y from F_X and f_X.
- ☐ Let $g : \mathbb{R} \mapsto \mathbb{R}$, $\mathcal{B} \subset \mathbb{R}$, and $g^{-1}(\mathcal{B}) \subset \mathbb{R}$ be the set for which $g\{g^{-1}(\mathcal{B})\} = \mathcal{B}$. Then

$$\mathrm{P}(Y \in \mathcal{B}) = \mathrm{P}\{g(X) \in \mathcal{B}\} = \mathrm{P}\{X \in g^{-1}(\mathcal{B})\},$$

  since $X \in g^{-1}(\mathcal{B})$ iff $g(X) = Y \in g\{g^{-1}(\mathcal{B})\} = \mathcal{B}$.
- ☐ To find $F_Y(y)$, we take $\mathcal{B}_y = (-\infty, y]$, giving

$$F_Y(y) = \mathrm{P}(Y \leq y) = \mathrm{P}\{g(X) \in \mathcal{B}_y\} = \mathrm{P}\{X \in g^{-1}(\mathcal{B}_y)\}.$$

- ☐ If the function g is monotonic increasing with (monotonic increasing) inverse g^{-1}, then

$$f_Y(y) = \frac{dF_Y(y)}{dy} = \frac{dF_X\{g^{-1}(y)\}}{dy} = f_X\{g^{-1}(y)\} \times \left|\frac{dg^{-1}(y)}{dy}\right|,$$

  where the $|\cdot|$ ensures that the same formula holds with monotonic decreasing g.

X bivariate

We calculate $\mathrm{P}(Y \in \mathcal{B})$, with $Y \in \mathbb{R}^d$ a function of $X \in \mathbb{R}^2$ and

$$Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_d \end{pmatrix} = \begin{pmatrix} g_1(X_1, X_2) \\ \vdots \\ g_d(X_1, X_2) \end{pmatrix} = g(X).$$

Let $g : \mathbb{R}^2 \mapsto \mathbb{R}^d$ be a known function, $\mathcal{B} \subset \mathbb{R}^d$, and $g^{-1}(\mathcal{B}) \subset \mathbb{R}^2$ be the set for which $g\{g^{-1}(\mathcal{B})\} = \mathcal{B}$. Then

$$\mathrm{P}(Y \in \mathcal{B}) = \mathrm{P}\{g(X) \in \mathcal{B}\} = \mathrm{P}\{X \in g^{-1}(\mathcal{B})\}.$$

Example 203. *If $X_1, X_2 \overset{\text{iid}}{\sim} \exp(\lambda)$, calculate the distribution of $X_1 + X_2$.*

It can be helpful to include indicator functions in formulae for densities of new variables (examples later).

Note to Example 203

We want to compute $\mathrm{P}(Y \le y) = \mathrm{P}(X_1 + X_2 \le y)$, and with $\mathcal{B}_y = (-\infty, y]$ and $g(x_1, x_2) = x_1 + x_2$, we have that

$$g^{-1}(\mathcal{B}_y) = \{(x_1, x_2) \in \mathbb{R}^2 : x_1 + x_2 \le y\}.$$

Thus we want to compute $F_Y(y) = \mathrm{P}(X_1 + X_2 \le y)$. If $y < 0$ this is zero, and otherwise equals

$$\begin{aligned} F_Y(y) = \mathrm{P}(X_1 + X_2 \le y) &= \int_0^y dx_1 \int_0^{y-x_1} dx_2 \; \lambda^2 e^{-\lambda(x_1+x_2)} \\ &= \lambda \int_0^y dx_1 e^{-\lambda x_1} \left[e^{-\lambda x_2} \right]_{y-x_1}^0 \\ &= \lambda \int_0^y dx_1 e^{-\lambda x_1} (1 - e^{-\lambda(y-x_1)}) \\ &= 1 - e^{-\lambda y} - \lambda y e^{-\lambda y}, \quad y \ge 0, \end{aligned}$$

giving

$$F_Y(y) = \begin{cases} 0, & y < 0, \\ 1 - e^{-\lambda y} - \lambda y e^{-\lambda y}, & y \ge 0. \end{cases}$$

Differentiation gives $f_Y(y) = \lambda^2 y e^{-\lambda y}$ for $y > 0$, (the gamma density with shape parameter $\alpha = 2$).

Transformations of joint continuous densities

Theorem 204. *Let $X = (X_1, X_2) \in \mathbb{R}^2$ be a continuous random variable, and let $Y = (Y_1, Y_2)$ with $Y_1 = g_1(X_1, X_2)$ and $Y_2 = g_2(X_1, X_2)$, where:*
(a) the system of equations $y_1 = g_1(x_1, x_2), y_2 = g_2(x_1, x_2)$ can be solved for all (y_1, y_2), giving the solutions $x_1 = h_1(y_1, y_2)$, $x_2 = h_2(y_1, y_2)$; and
(b) g_1 and g_2 are continuously differentiable and have Jacobian

$$J(x_1, x_2) = \begin{vmatrix} \frac{\partial g_1}{\partial x_1} & \frac{\partial g_1}{\partial x_2} \\ \frac{\partial g_2}{\partial x_1} & \frac{\partial g_2}{\partial x_2} \end{vmatrix} = \left| \frac{\partial g_1}{\partial x_1}\frac{\partial g_2}{\partial x_2} - \frac{\partial g_1}{\partial x_2}\frac{\partial g_2}{\partial x_1} \right|,$$

which is positive if $f_{X_1,X_2}(x_1, x_2) > 0$.
Then

$$f_{Y_1,Y_2}(y_1, y_2) = f_{X_1,X_2}(x_1, x_2) \times |J(x_1, x_2)|^{-1}\big|_{x_1=h_1(y_1,y_2), x_2=h_2(y_1,y_2)}.$$

Example 205. *Calculate the joint density of $X_1 + X_2$ and $X_1 - X_2$ when $X_1, X_2 \overset{\text{iid}}{\sim} N(0, 1)$.*

Example 206. *Calculate the joint density of $X_1 + X_2$ and $X_1/(X_1 + X_2)$ when $X_1, X_2 \overset{\text{iid}}{\sim} \exp(\lambda)$.*

Note to Example 205

☐ We already have one way to do this, as we can write

$$\begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} = \begin{pmatrix} X_1 + X_2 \\ X_1 - X_2 \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} = B \begin{pmatrix} X_1 \\ X_2 \end{pmatrix},$$

say, and use results for the multivariate normal distribution in Lemma 197(e).

☐ Using Theorem 204 instead, we need to compute

$$f_{Y_1,Y_2}(y_1, y_2) = f_{X_1,X_2}(x_1, x_2) \times |J(x_1, x_2)|^{-1}\big|_{x_1=h_1(y_1,y_2), x_2=h_2(y_1,y_2)}.$$

First, note that the Jacobian of the transformation $(x_1, x_2) \mapsto (y_1, y_2)$ is

$$J(x_1, x_2) = |B| = |-2| = 2.$$

Now we need to express the density

$$f_{X_1,X_2}(x_1, x_2) = f_{X_1}(x_1) f_{X_2}(x_2) = \frac{1}{\sqrt{2\pi}} e^{-x_1^2/2} \times \frac{1}{\sqrt{2\pi}} e^{-x_2^2/2}, \quad x_1, x_2 \in \mathbb{R},$$

in terms of (y_1, y_2). As $x_1 = (y_1 + y_2)/2$ and $x_2 = (y_1 - y_2)/2$, the exponent may be written in terms of the new variables y_1, y_2 as

$$-\tfrac{1}{2}(x_1^2 + x_2^2) = -\tfrac{1}{2}\left[\{(y_1 + y_2)/2\}^2 + \{(y_1 - y_2)/2\}^2\right] = -\frac{1}{2 \times 4}(2y_1^2 + 2y_2^2) = -\frac{1}{2 \times 2}(y_1^2 + y_2^2),$$

so

$$f_{Y_1,Y_2}(y_1, y_2) = \frac{1}{2} \times \frac{1}{2\pi} \exp\left\{-\frac{1}{2 \times 2}(y_1^2 + y_2^2)\right\}, \quad y_1, y_2 \in \mathbb{R},$$

and we see that Y_1 and Y_2 are mutually independent $\mathcal{N}(0, 2)$ variables.

Note to Example 206

We write

$$f(x_1, x_2) = \lambda^2 \exp\{-\lambda(x_1 + x_2)\} I(x_1 > 0) I(x_2 > 0).$$

With $Y_1 = X_1 + X_2 > 0$ and $Y_2 = X_1/(X_1 + X_2) \in (0, 1)$, we have

$$y_1 = g_1(x_1, x_2) = x_1 + x_2 > 0, \quad y_2 = g_2(x_1, x_2) = x_1/(x_1 + x_2) \in (0, 1),$$

and the corresponding inverse transformation is

$$x_1 = h(y_1, y_2) = y_1 y_2, \quad x_2 = h(y_1, y_2) = y_1(1 - y_2), \quad x_1, x_2 > 0.$$

Clearly these transformations satisfy the conditions of Theorem 204. We can either compute

$$J = \begin{vmatrix} 1 & 1 \\ \frac{x_2}{(x_1+x_2)^2} & -\frac{x_1}{(x_1+x_2)^2} \end{vmatrix} = \left| -\frac{(x_1 + x_2)}{(x_1 + x_2)^2} \right| = 1/y_1 > 0,$$

or (maybe better),

$$J^{-1} = \begin{vmatrix} \frac{\partial h_1}{\partial y_1} & \frac{\partial h_1}{\partial y_2} \\ \frac{\partial h_2}{\partial y_1} & \frac{\partial h_2}{\partial y_2} \end{vmatrix} = y_1 > 0.$$

Thus

$$\begin{aligned} f(y_1, y_2) &= \lambda^2 \exp\{-\lambda(x_1 + x_2)\} I(x_1 > 0) I(x_2 > 0) |J^{-1}|\big|_{x_1 = y_1 y_2, x_2 = y_1(1-y_2)} \\ &= y_1 \lambda^2 \exp(-\lambda y_1) I(y_1 y_2 > 0) I\{y_1(1 - y_2) > 0\}, \\ &= y_1 \lambda^2 \exp(-\lambda y_1) I(y_1 > 0) \times I(0 < y_2 < 1) \\ &= f_{Y_1}(y_1) \times f_{Y_2}(y_2). \end{aligned}$$

Integration over y_2 shows that the marginal density of Y_1 is $y_1 \lambda^2 \exp(-\lambda y_1) I(y_1 > 0)$, and so $Y_1 \sim \text{Gamma}(1, \lambda)$ and $Y_2 \sim U(0, 1)$, independently.

Sums of independent variables

Theorem 207. *If X, Y are independent random variables, then the PDF of their sum $S = X + Y$ is the convolution $f_X * f_Y$ of the PDFs f_X, f_Y:*

$$f_S(s) = f_X * f_Y(s) = \begin{cases} \int_{-\infty}^{\infty} f_X(x) f_Y(s - x)\, dx, & X, Y \text{ continuous}, \\ \sum_x f_X(x) f_Y(s - x), & X, Y \text{ discrete}. \end{cases}$$

Example 208. *Show that the sum of independent exponential and gamma variables has a gamma distribution.*

Note to Theorem 207

Change variables to $W = X$ and $S = X + Y$, so the Jacobian is

$$J = \begin{vmatrix} 1 & 0 \\ 1 & 1 \end{vmatrix} = 1,$$

and note that $x = w$ and $y = s - w$. Thus, since X and Y are independent, an application of Theorem 204 gives

$$f_{W,S}(w, s) = f_{X,Y}(w, s - w) \times |J^{-1}| = f_X(w) f_Y(s - w) \times 1.$$

Therefore the marginal density of S in the continuous case is

$$f_S(s) = \int_{-\infty}^{\infty} f_X(w) f_Y(s - w)\, dw.$$

The computation in the discrete case is similar, but the Jacobian is not needed.

Note to Example 208

Use indicator functions to write the densities as

$$f_X(x) = \frac{\lambda^\alpha x^{\alpha-1}}{\Gamma(\alpha)} e^{-\lambda x} I(x > 0), \quad f_Y(y) = \lambda e^{-\lambda y} I(y > 0), \quad \lambda, \alpha > 0,$$

and use the convolution formula to give that $S = X + Y$ has density

$$f_S(s) = \int_{-\infty}^{\infty} f_X(w) f_Y(s - w)\, dw = \int_{-\infty}^{\infty} \frac{\lambda^\alpha w^{\alpha-1}}{\Gamma(\alpha)} e^{-\lambda w} I(w > 0) \times \lambda e^{-\lambda(s-w)} I(s - w > 0)\, dw.$$

The product of the indicator functions is positive only when $w > 0$ and $s - w > 0$ simultaneously, i.e., when $0 < w < s$, and hence on putting constants outside the integral, we have

$$f_S(s) = \frac{\lambda^{\alpha+1} e^{-\lambda s}}{\Gamma(\alpha)} \int_0^s w^{\alpha-1}\, dw.$$

On noting that the integral equals s^α/α and recalling that $\alpha\Gamma(\alpha) = \Gamma(\alpha + 1)$, we have

$$f_S(s) = \frac{\lambda^{\alpha+1} s^\alpha}{\Gamma(\alpha + 1)} e^{-\lambda s}, \quad s > 0,$$

so $S \sim$ gamma$(\alpha + 1, \lambda)$. In particular, a sum of two exponential variables has a gamma(2,λ) distribution.

Multivariate case

Theorem 204 extends to random vectors with continuous density, $Y = g(X) \in \mathbb{R}^n$, where $X \in \mathbb{R}^n$ is a continuous variable:

$$(X_1, \ldots, X_n) \;\mapsto\; (Y_1 = g_1(X_1, \ldots, X_n), \ldots, Y_n = g_n(X_1, \ldots, X_n)).$$

If the inverse transformation h exists, and has Jacobian

$$J(x_1, \ldots, x_n) = \begin{vmatrix} \frac{\partial g_1}{\partial x_1} & \cdots & \frac{\partial g_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial g_n}{\partial x_1} & \cdots & \frac{\partial g_n}{\partial x_n} \end{vmatrix},$$

then

$$f_{Y_1,\ldots,Y_n}(y_1, \ldots, y_n) = f_{X_1,\ldots,X_n}(x_1, \ldots, x_n)\ |J(x_1, \ldots, x_n)|^{-1},$$

evaluated at $x_1 = h_1(y_1, \ldots, y_n), \ldots, x_n = h_n(y_1, \ldots, y_n)$.

Convolution and sums of random variables

Theorem 209. *If $X_1, \ldots, X_n$ are independent random variables, then the PDF of $S = X_1 + \cdots + X_n$ is the convolution*

$$f_S(s) = f_{X_1} * \cdots * f_{X_n}(s).$$

In fact it is easier to use the MGFs for convolutions, if possible.

Example 210. *Show that if $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \exp(\lambda)$, then $Y = X_1 + \cdots + X_n \sim$ gamma(n, λ).*

Note to Example 210

The MGF of $X \sim \exp(\lambda)$ is $M_X(t) = \lambda/(\lambda - t)$, for $t < \lambda$. Now if Y has the gamma(n, λ) distribution,

$$\begin{aligned} \mathrm{E}(e^{tY}) &= \int_0^\infty e^{ty} \frac{\lambda^n y^{n-1}}{\Gamma(n)} e^{-\lambda y}\, dy \\ &= \frac{\lambda^n}{\Gamma(n)} \int_0^\infty y^{n-1} e^{-(\lambda - t)y}\, dy \\ &= \frac{\lambda^n}{(\lambda - t)^n \Gamma(n)} \int_0^\infty (\lambda - t)^n y^{n-1} e^{-(\lambda - t)y}\, dy \\ &= \left(\frac{\lambda}{\lambda - t}\right)^n \times 1, \end{aligned}$$

provided that $\lambda - t > 0$, or equivalently that $t < \lambda$. The last step just notes that the integral corresponds to the density of the gamma$(n, \lambda - t)$ distribution, and so equals unity.
Now $M_Y(t) = M_X(t)^n = \lambda^n/(\lambda - t)^n$, so Y has the stated gamma distribution, since there is a bijection between MGFs and distributions.

5.6 Order Statistics

Definition

Definition 211. *The* **order statistics** *of the rv's $X_1, \ldots, X_n$ are the ordered values*

$$X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n-1)} \leq X_{(n)}.$$

If the $X_1, \ldots, X_n$ are continuous, then no two of the X_j can be equal, i.e.,

$$X_{(1)} < X_{(2)} < \cdots < X_{(n-1)} < X_{(n)}.$$

In particular, the **minimum** *is $X_{(1)}$, the* **maximum** *is $X_{(n)}$, and the* **median** *is*

$$X_{(m+1)} \quad (n = 2m+1, \text{ odd}), \qquad \tfrac{1}{2}(X_{(m)} + X_{(m+1)}) \quad (n = 2m, \text{ even}).$$

The median is the central value of $X_1, \ldots, X_n$.

Theorem 212. *Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} F$, from a continuous distribution with density f, then:*

$$\begin{aligned} \mathrm{P}(X_{(n)} \leq x) &= F(x)^n; \\ \mathrm{P}(X_{(1)} \leq x) &= 1 - \{1 - F(x)\}^n; \\ f_{X_{(r)}}(x) &= \frac{n!}{(r-1)!(n-r)!} F(x)^{r-1} f(x) \{1 - F(x)\}^{n-r}, \quad r = 1, \ldots, n. \end{aligned}$$

Example 213. *If $X_1, X_2, X_3 \overset{\text{iid}}{\sim} \exp(\lambda)$, give the densities of the $X_{(r)}$.*

Example 214. *Abélard and Héloïse make an appointment to work at the Learning Centre. Both are late independently of each other, arriving at times distributed uniformly up to one hour after the time agreed. Find the distribution and the expectation of the time at which the first one arrives, and give the density of his (or her) waiting time. Find the expected time at which they can start to work.*

Note to Theorem 212

- First, $X_{(n)} \leq x$ if and only if all the $X_i \leq x$, and this has probability $F(x)^n$.
- Likewise $X_{(1)} > x$ if and only if all the $X_i > x$, and this has probability $\{1 - F(x)\}^n$. Thus the required CDF is $\mathrm{P}(X_{(1)} \leq x) = 1 - \{1 - F(x)\}^n$.
- Finally, for the event $X_{(r)} \in [x, x + dx)$, we need to have split the sample into three groups of respective sizes $r - 1$, 1 and $n - r$ (hence the combinatorial coefficient) and probabilities $F(x)$, $f(x)dx$, and $1 - F(x)$. This gives the required formula.

Note to Example 213

We note that in this case $f(x) = \lambda e^{-\lambda x}$ and $F(x) = 1 - \exp(-\lambda x)$, and then just apply the theorem with $n = 3$ and $r = 1, 2, 3$:

$$\begin{aligned} f_{X_{(1)}}(x) &= 3\lambda e^{-\lambda x} \times (e^{-\lambda x})^2, \quad x > 0 \\ f_{X_{(2)}}(x) &= 6(1 - e^{-\lambda x}) \times \lambda e^{-\lambda x} \times e^{-\lambda x}, \quad x > 0 \\ f_{X_{(3)}}(x) &= 3(1 - e^{-\lambda x})^2 \times \lambda e^{-\lambda x}, \quad x > 0. \end{aligned}$$

Note to Example 214

Let $0 < U < V < 1$ denote the ordered arrival times.
U is the minimum of $n = 2$ independent $U(0, 1)$ variables, each with $F(u) = u$ $(0 < u < 1)$, so according to the second line of Theorem 212 U has distribution function $F_U(u) = 1 - (1 - u)^2$ and corresponding density

$$f_U(u) = \frac{dF_U(u)}{du} = \frac{d\{1 - (1 - u)^2\}}{du} = 2(1 - u), \quad 0 < u < 1;$$

consequently $\mathrm{E}(U) = \int_0^1 u \times 2(1 - u)\, du = 1 - 2/3 = 1/3$. To compute the joint density we note that the uniformity of the arrival times implies that

$$\mathrm{P}(V \le v, U \le u) = \mathrm{P}(V \le v) - \mathrm{P}(V \le v, U > u) = v^2 - (v - u)^2, \quad 0 < u < v < 1,$$

because the event $V < v$ occurs if and only if both of them independently arrive before v, and the event $V \le v, U > u$ occurs if and only if they both arrive in the interval (u, v). It follows that the joint density is

$$f(u, v) = \frac{\partial^2 \mathrm{P}(V \le v, U \le u)}{\partial u \partial v} = 2I(0 < u < v < 1).$$

Therefore $w = v - u$ has density

$$\begin{aligned} f(w) = \int_{u=0}^{1} 2I(0 < u < v < 1)\, du &= 2\int_{u=0}^{1} I(0 < u < u + w < 1)\, du \\ &= 2\int_{u=0}^{1} I(0 < u < 1 - w)\, du \\ &= 2(1 - w), \quad 0 < w < 1. \end{aligned}$$

They can start to work when the second of them arrives, at time V, and this has expectation $\mathrm{E}(V) = \int_0^1 2v\, dv = 2/3$, i.e., 40 minutes after the agreed time.

6. Approximation and Convergence

Motivation

It is often difficult to calculate the exact probability p of an event of interest, and we have to **approximate**. Possible approaches:

- ☐ try to bound p;
- ☐ analytic approximation, often using the law of large numbers and the central limit theorem;
- ☐ numerical approximation, often using Monte Carlo methods.

The final approaches use the notion of **convergence** of sequences of random variables, which we will study in this chapter.
We have already seen examples of these ideas: normal approximation to the binomial distribution, law of small numbers, ...

6.1 Inequalities

Inequalities

Theorem 215. *If X is a random variable, $a > 0$ a constant, h a non-negative function and g a convex function, then*

$$\begin{aligned} \mathrm{P}\{h(X) \geq a\} &\leq \mathrm{E}\{h(X)\}/a, && \textit{(basic inequality)} \\ \mathrm{P}(|X| \geq a) &\leq \mathrm{E}(|X|)/a, && \textit{(Markov's inequality)} \\ \mathrm{P}(|X| \geq a) &\leq \mathrm{E}(X^2)/a^2, && \textit{(Chebyshov's inequality)} \\ \mathrm{E}\{g(X)\} &\geq g\{\mathrm{E}(X)\}. && \textit{(Jensen's inequality)} \end{aligned}$$

On replacing X by $X - \mathrm{E}(X)$, Chebyshov's inequality gives

$$\mathrm{P}\{|X - \mathrm{E}(X)| \geq a\} \leq \mathrm{var}(X)/a^2.$$

These inequalities are more useful for theoretical calculations than for practical use.

Example 216. *We are testing a classification method, in which the probability of a correct classification is p. Let $Y_1, \ldots, Y_n$ be the indicators of correct classifications in n test cases, and let $\overline{Y}$ be their average. For $\varepsilon = 0.2$ and $n = 100$, bound*

$$\mathrm{P}(|\overline{Y} - p| > \varepsilon).$$

Note to Theorem 215

(a) Let $Y = h(X)$. If $y \geq 0$, then for any $a > 0$, $y \geq yI(y \geq a) \geq aI(y \geq a)$. Therefore

$$\mathrm{E}\{h(X)\} = \mathrm{E}(Y) \geq \mathrm{E}\{YI(Y \geq a)\} \geq \mathrm{E}\{aI(Y \geq a)\} = a\mathrm{P}(Y \geq a) = a\mathrm{P}\{h(X) \geq a\},$$

and division by $a > 0$ gives the result.
(b) Note that $h(x) = |x|$ is a non-negative function on $\mathbb{R}$, and apply (a).
(c) Note that $h(x) = x^2$ is a non-negative function on $\mathbb{R}$, and that $\mathrm{P}(X^2 \geq a^2) = \mathrm{P}(|X| \geq a)$.
(d) A convex function has the property that, for all y, there exists a value $b(y)$ such that $g(x) \geq g(y) + b(y)(x - y)$ for all x. If $g(x)$ is differentiable, then we can take $b(y) = g'(y)$. (Draw a graph if need be.) To prove this result, we take $y = \mathrm{E}(X)$, and then have

$$g(X) \geq g\{\mathrm{E}(X)\} + b\{\mathrm{E}(X)\}\{X - \mathrm{E}(X)\},$$

and taking expectations of this gives $\mathrm{E}\{g(X)\} \geq g\{\mathrm{E}(X)\}$.

Note to Example 216

We note that $\sum_{j=1}^{n} Y_j \sim B(n, p)$, so has mean np and variance $np(1-p)$, write $X = \overline{Y} - p$, and note that $\mathrm{E}(X) = 0$ and $\mathrm{E}(X^2) = \mathrm{var}(X) = \mathrm{var}(\overline{Y}) = n^{-2} \times np(1-p)$. Now Chebyshov's inequality gives

$$\mathrm{P}(|\overline{Y} - p| > \varepsilon) = \mathrm{P}(|X| > \varepsilon) \leq \mathrm{P}(|X| \geq \varepsilon) \leq \mathrm{E}(X^2)/\varepsilon^2,$$

and since $p(1-p) \leq 1/4$ in the range $0 \leq p \leq 1$,

$$\mathrm{E}(X^2)/\varepsilon^2 = \mathrm{var}(\overline{Y})/\varepsilon^2 = p(1-p)/(n\varepsilon^2) \leq 1/4/(100 \times 0.2^2) = 1/16.$$

Hoeffding's inequality

Theorem 217. *(Hoeffding's inequality, no proof) Let $Z_1, \ldots, Z_n$ be independent random variables such that $\mathrm{E}(Z_i) = 0$ and $a_i \leq Z_i \leq b_i$ for constants $a_i < b_i$. If $\varepsilon > 0$, then for all $t > 0$,*

$$\mathrm{P}\left(\sum_{i=1}^{n} Z_i \geq \varepsilon\right) \leq e^{-t\varepsilon} \prod_{i=1}^{n} e^{t^2(b_i - a_i)^2/8}.$$

This inequality is much more useful than the others for finding powerful bounds in practical situations.

Example 218. *Show that if $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathrm{Bernoulli}(p)$ and $\varepsilon > 0$, then*

$$\mathrm{P}(|\overline{X} - p| > \varepsilon) \leq 2e^{-2n\varepsilon^2}.$$

For $\varepsilon = 0.2$ and $n = 100$, bound

$$\mathrm{P}(|\overline{X} - p| > \varepsilon).$$

Note to Example 218

For the theoretical part, take $Z_i = (X_i - p)/n$, and note that $-p/n \leq Z_i \leq (1-p)/n$, so $b_i - a_i = 1/n$. Then

$$\mathrm{P}(|\overline{X} - p| > \varepsilon) = \mathrm{P}(\sum Z_i > \varepsilon) + \mathrm{P}(-\sum Z_i > \varepsilon),$$

and each of these probabilities can be bounded by

$$e^{-t\varepsilon}\left\{e^{t^2(1/n)^2/8}\right\}^n = \exp\{t^2/(8n) - t\varepsilon\}.$$

To minimise this with respect to t, we take $t = 4n\varepsilon$, which leads to the result.
For the numerical part, just insert into the previous part and get 0.00067, which is much smaller than the bound obtained using the Chebyshov inequality (Example 216).

6.2 Convergence

Convergence

Definition 219 (Deterministic convergence). *If $x_1, x_2, \ldots, x$ are real numbers, then $x_n \to x$ iff for all $\varepsilon > 0$, there exists N_ε such that $|x_n - x| < \varepsilon$ for all $n > N_\varepsilon$.*

Probabilistic convergence is more complicated ... We could hope that (for example) $X_n \to X$ if either

$$\mathrm{P}(X_n \leq x) \to \mathrm{P}(X \leq x), \quad x \in \mathbb{R},$$

or

$$\mathrm{E}(X_n) \to \mathrm{E}(X)$$

when $n \to \infty$.

Example 220. *For $n = 1, 2, \ldots$ let X_n be the random variable such that*

$$\mathrm{P}(X_n = 0) = 1 - 1/n, \quad \mathrm{P}(X_n = n^2) = 1/n.$$

Then when $n \to \infty$,

$$\begin{aligned} \mathrm{P}(|X_n| > 0) &= \mathrm{P}(X_n = n^2) = 1/n \to 0, \\ \mathrm{E}(X_n) &= 0 \times (1 - 1/n) + n^2 \times 1/n = n \to \infty. \end{aligned}$$

Does $X_n \to 0$ or $X_n \to \infty$? What does 'converge' mean for random variables?

Modes of convergence of random variables

Definition 221. *Let $X, X_1, X_2, \ldots$ be random variables with cumulative distribution function $F, F_1, F_2, \ldots$. Then*
(a) X_n converges to X **almost surely**, *$X_n \xrightarrow{\text{a.s.}} X$, if*

$$\mathrm{P}\left(\lim_{n\to\infty} X_n = X\right) = 1;$$

(b) X_n converges to X **in mean square**, *$X_n \xrightarrow{2} X$, if*

$$\lim_{n\to\infty} \mathrm{E}\{(X_n - X)^2\} = 0, \quad \textit{where} \quad \mathrm{E}(X_n^2), \mathrm{E}(X^2) < \infty;$$

(c) X_n converges to X **in probability**, *$X_n \xrightarrow{P} X$, if for all $\varepsilon > 0$,*

$$\lim_{n\to\infty} \mathrm{P}(|X_n - X| > \varepsilon) = 0;$$

(d) X_n converges to X **in distribution**, *$X_n \xrightarrow{D} X$, if*

$$\lim_{n\to\infty} F_n(x) = F(x) \quad \textit{at each point } x \textit{ where } F(x) \textit{ is continuous.}$$

$X_n \xrightarrow{\text{a.s.}} X$

To understand this better:

- ☐ all the variables $\{X_n\}, X$ must be defined on the same probability space, $(\Omega, \mathcal{F}, \mathrm{P})$. It is not trivial to construct this space (we need 'Kolmogorov's extension theorem').
- ☐ Then to each $\omega \in \Omega$ corresponds a sequence

  $$X_1(\omega), X_2(\omega), \ldots, X_n(\omega), \ldots$$

  which will converge, or not, as a sequence of real numbers.
- ☐ If $X_n \xrightarrow{\text{a.s.}} X$, then
  $$\mathrm{P}\left(\left\{\omega : \lim_{n\to\infty} X_n(\omega) = X(\omega)\right\}\right) = 1 :$$
  the set of values of ω for which $X_n(\omega) \not\to X(\omega)$ has probability 0.

Example 222. *Let $U \sim U(0,1)$, where $\Omega = [0,1]$, $U(\omega) = \omega$, $X_n(\omega) = U(\omega)^n$, $n = 1, 2, \ldots$, and $X(\omega) = 0$. Show that $X_n \xrightarrow{\text{a.s.}} X$.*

Note to Example 222

Here we note that for any $0 \le \omega < 1$, $X_n(\omega) = U(\omega)^n = \omega^n \to 0$ as $n \to \infty$, so $X_n(\omega) \to X(\omega)$ for every $\omega \in [0, 1)$. The only ω for which $X_n(\omega) \not\to X(\omega)$ is $\omega = 1$, and this has zero probability of occurring, so

$$\mathrm{P}\left(\left\{\omega : \lim_{n\to\infty} X_n(\omega) = X(\omega)\right\}\right) = \mathrm{P}(U < 1) = 1,$$

as required.

Relations between modes of convergence

- If $X_n \xrightarrow{\text{a.s.}} X$, $X_n \xrightarrow{2} X$ or $X_n \xrightarrow{P} X$, then $X_1, X_2, \ldots, X$ must all be defined with respect to only one probability space. This is not the case for $X_n \xrightarrow{D} X$, which only concerns the probabilities. This last is thus weaker than the others.
- These modes of convergence are related to one another in the following way:

$$\begin{array}{lll} X_n \xrightarrow{\text{a.s.}} X & \Rightarrow & \\ & & X_n \xrightarrow{P} X \quad \Rightarrow \quad X_n \xrightarrow{D} X \\ X_n \xrightarrow{2} X & \Rightarrow & \end{array}$$

  All other implications are in general false.
- The most important modes of convergence in this course are $\xrightarrow{P}$ and $\xrightarrow{D}$, since we often wish to approximate probabilities, and $\xrightarrow{D}$ gives us a way to do so.

Example 223. *Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} (\mu, \sigma^2)$ with $0 < \sigma^2 < \infty$. Show that $\overline{X} = (X_1 + \cdots + X_n)/n \xrightarrow{2} \mu$.*

Example 224. *Let $X_n = (-1)^n Z$, where $Z \sim \mathcal{N}(0,1)$. Show that $X_n \xrightarrow{D} Z$, but that this is the only mode of convergence that applies here.*

Note to Example 223

Note that $\mathrm{E}(\overline{X}) = \mu$, so by definition of the variance as $\mathrm{var}(\overline{X}) = \mathrm{E}[\{\overline{X} - \mathrm{E}(\overline{X})\}^2]$, we have

$$\mathrm{E}\{(\overline{X} - \mu)^2\} = \mathrm{var}(\overline{X}) = \sigma^2/n \to 0, \quad n \to \infty,$$

which implies that $\overline{X} \xrightarrow{2} \mu$, as required.

Note to Example 224

For even n there is nothing to prove, since then $X_n = (-1)^n Z = Z$, and then $\mathrm{P}(X_n \le x) = \mathrm{P}(Z \le x)$.
For odd n, $X_n = (-1)^n Z = -Z$, so

$$\mathrm{P}(X_n \le x) = \mathrm{P}(-Z \le x) = \mathrm{P}(Z \ge -x) = 1 - \Phi(-x) = \Phi(x) = \mathrm{P}(Z \le x).$$

Hence $X_n \xrightarrow{D} Z$, though this is trivial because X_n and Z have the same distribution for every n.
Now for n odd,

$$\mathrm{P}(|X_n - Z| > \varepsilon) = \mathrm{P}(|-Z - Z| > \varepsilon) = \mathrm{P}(|Z| > \varepsilon/2) = 2\Phi(-\varepsilon/2) \not\to 0, \quad n \to \infty,$$

so X_n does not converge in probability to Z, and thus neither of the other modes of convergence can be true either.

Continuity theorem (reminder)

Theorem 225 (Continuity). *Let $\{X_n\}$, X be random variables with cumulative distribution functions $\{F_n\}$, F, whose MGFs $M_n(t)$, $M(t)$ exist for $0 \le |t| < b$. If there exists a $0 < a < b$ such that $M_n(t) \to M(t)$ for $|t| \le a$ when $n \to \infty$, then $X_n \xrightarrow{D} X$, that is to say, $F_n(x) \to F(x)$ at each $x \in \mathbb{R}$ where F is continuous.*

- ☐ We could replace $M_n(t)$ and $M(t)$ by the cumulant-generating functions $K_n(t) = \log M_n(t)$ and $K(t) = \log M(t)$.
- ☐ We established the law of small numbers (Theorem 104 and Example 186, Poisson approximation of the binomial distribution) by using this result.
- ☐ Here is another example:

Example 226. *Let X be a random variable which has a geometric distribution with a probability of success p. Calculate the limit distribution of pX when $p \to 0$.*

Note to Example 226

Recall that if $|a| < 1$, then $\sum_{r=0}^{\infty} a^r = 1/(1-a)$.
The MGF of pX is

$$\begin{aligned} \mathrm{E}(e^{tpX}) &= \sum_{x=1}^{\infty} e^{tpx} p(1-p)^{x-1} \\ &= pe^{tp} \sum_{x=0}^{\infty} \{e^{tp}(1-p)\}^x \\ &= \frac{pe^{tp}}{1-(1-p)e^{tp}} = \frac{1}{p^{-1}e^{-tp} - (1-p)/p} = \frac{1}{1+(e^{-tp}-1)/p} \to \frac{1}{1-t}, \quad p \to 0, \end{aligned}$$

which is the MGF of $Y \sim \exp(1)$. We need $t < 1$.

Combinations of convergent sequences

Theorem 227 (Combination of convergent sequences). *Let x_0, y_0 be constants, $X, Y, \{X_n\}, \{Y_n\}$ random variables, and h a function continuous at x_0. Then*

$$\begin{aligned} X_n \xrightarrow{D} x_0 &\Rightarrow X_n \xrightarrow{P} x_0, \\ X_n \xrightarrow{P} x_0 &\Rightarrow h(X_n) \xrightarrow{P} h(x_0), \\ X_n \xrightarrow{D} X \text{ and } Y_n \xrightarrow{P} y_0 &\Rightarrow X_n + Y_n \xrightarrow{D} X + y_0,\ X_n Y_n \xrightarrow{D} X y_0. \end{aligned}$$

The third line is known as **Slutsky's lemma**. *It is very useful in statistical applications.*

Example 228. *Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} (\mu_X, \sigma_X^2)$, $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} (\mu_Y, \sigma_Y^2)$, $\mu_X \neq 0$, $\sigma_X^2, \sigma_Y^2 < \infty$, and define*

$$R_n = \overline{Y}/\overline{X}, \quad \overline{Y} = n^{-1} \sum_{j=1}^{n} Y_j, \quad \overline{X} = n^{-1} \sum_{j=1}^{n} X_j.$$

Show that $R_n \xrightarrow{P} \mu_Y/\mu_X$ when $n \to \infty$.

Note to Example 228

Note that since $\sigma_X^2 < \infty$, by Example 223, $\overline{X} \xrightarrow{2} \mu_X$, and likewise $\overline{Y} \xrightarrow{2} \mu_Y$. Hence $\overline{X} \xrightarrow{P} \mu_X$, by the contents of slide 243, and since the function $h(x) = 1/x$ is continuous at $\mu_X \neq 0$, it must be true using line 2 of the theorem that $1/\overline{X} \xrightarrow{P} 1/\mu_X$, a constant. Therefore we have by line 3 that

$$R_n = \overline{Y} \times 1/\overline{X} \xrightarrow{D} \mu_Y \times 1/\mu_X,$$

and as this is a constant, line 1 implies that $R_n \xrightarrow{P} \mu_Y \times 1/\mu_X$, as required.

Convergence in distribution: Limits for maxima

- In applications, we often have to take into account the greatest or the smallest random variables considered.
- A system of n composants can break down when any composant of the system becomes faulty. What is the distribution of the failure time?
- Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} F$, and $M_n = \max\{X_1, \ldots, X_n\}$. Then

$$\mathrm{P}(M_n \leq x) = \mathrm{P}(X_1 \leq x, \ldots, X_n \leq x) = F(x)^n \to \begin{cases} 0, & F(x) < 1, \\ 1, & F(x) = 1. \end{cases}$$

- Hence M_n must be renormalised to get a non-degenerate limit distribution. Let $\{a_n\} > 0$ and $\{b_n\}$ be sequences of constants, and consider the convergence in distribution of

$$Y_n = (M_n - b_n)/a_n,$$

where a_n, b_n are chosen so that a non-degenerate limit distribution for Y_n exists.

Examples

Example 229. *Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \exp(\lambda)$, and let M_n be their maximum. Find a_n, b_n such that $Y_n = (M_n - b_n)/a_n \xrightarrow{D} Y$, where Y has a non-degenerate distribution.*

Example 230. *Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} U(0,1)$, and let M_n be their maximum. Find a_n, b_n such that $Y_n = (M_n - b_n)/a_n \xrightarrow{D} Y$, where Y has a non-degenerate distribution.*

Note to Example 229

We have

$$\mathrm{P}(Y_n \leq y) = F(b_n + a_n y)^n = \{1 - \exp(-b_n\lambda - a_n\lambda y)\}^n,$$

and on setting $a_n = 1/\lambda$, $b_n = \log n/\lambda$, we have

$$\mathrm{P}(Y_n \leq y) = \{1 - \exp(-y)/n\}^n \to \exp\{-\exp(-y)\},$$

which is the Gumbel distribution function.

Note to Example 230

We have

$$\mathrm{P}(Y_n \leq y) = F(b_n + a_n y)^n = (b_n + a_n y)^n,$$

and on setting $a_n = 1/n$, $b_n = 1$, we have (since $M_n < 1$) that

$$\mathrm{P}(Y_n \leq y) = \mathrm{P}\{n(M_n - 1) \leq y\} = (1 + y/n)^n \to \exp(y), \quad y < 0$$

which is the distribution function of $-Z$, where $Z \sim \exp(1)$.

Fisher–Tippett theorem

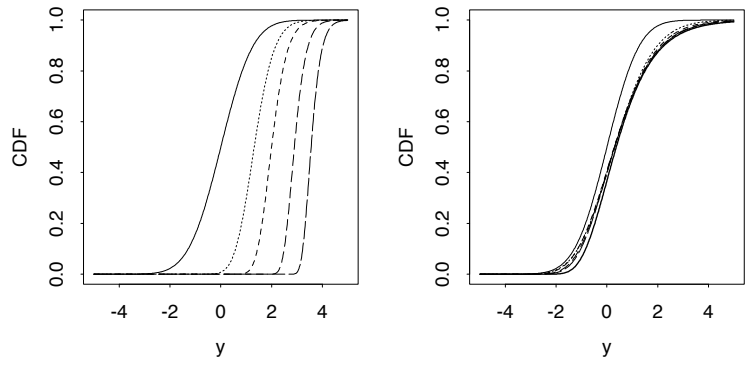
Theorem 231. *Suppose that $X_1, \ldots, X_n \overset{\text{iid}}{\sim} F$, where F is a continuous cumulative distribution function. Let $M_n = \max\{X_1, \ldots, X_n\}$, and suppose that the sequences of constants $\{a_n\} > 0$ and $\{b_n\}$ can be chosen so that $Y_n = (M_n - b_n)/a_n \xrightarrow{D} Y$, where Y has a non-degenerate limit distribution $H(y)$ when $n \to \infty$. Then H must be the* **generalised extreme-value (GEV) distribution**,

$$H(y) = \begin{cases} \exp\left[-\{1 + \xi(y-\eta)/\tau\}_+^{-1/\xi}\right], & \xi \neq 0, \\ \exp\left[-\exp\{-(y-\eta)/\tau\}\right], & \xi = 0, \end{cases}$$

where $u_+ = \max(u, 0)$, and $\eta, \xi \in \mathbb{R}$, $\tau > 0$.

Example

The graph below shows the distributions of M_n and of Y_n for $n = 1, 7, 30, 365, 3650$, from left to right, for $X_1, \ldots, X_n \overset{\text{iid}}{\sim} N(0, 1)$. The panel on the right also shows the limit distribution (bold), $H(y) = \exp\{-\exp(-y)\}$.



Law of large numbers

The first part of our limit results concern the behaviour of averages of independent random variables.

Theorem 232. *(Weak law of large numbers) Let $X_1, X_2, \ldots$ be a sequence of independent identically distributed random variables with finite expectation μ, and write their average as*

$$\overline{X} = n^{-1}(X_1 + \cdots + X_n).$$

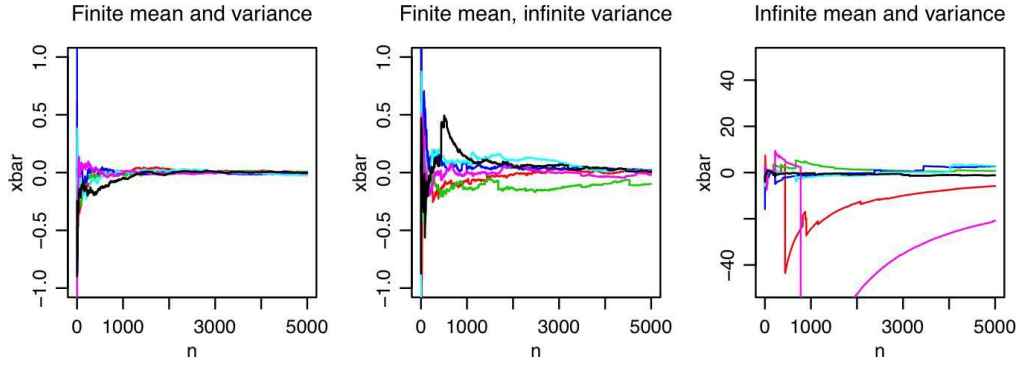
Then $\overline{X} \xrightarrow{P} \mu$; i.e., for all $\varepsilon > 0$,

$$\mathrm{P}(|\overline{X} - \mu| > \varepsilon) \to 0, \quad n \to \infty.$$

- ☐ Thus, under mild conditions, the averages of samples of important size converge towards the expectation of the distribution from which the sample is taken.
- ☐ If the X_i are independent Bernoulli trials, we return to our primitive notion of probability as a limit of relative frequencies. The circle is complete.

Weak law of large numbers

- ☐ The graphs below show the behaviour of $\overline{X}$ when X_i has two finite moments (on the left), only $\mathrm{E}(|X_i|) < \infty$ (centre), $\mathrm{E}(X_i)$ doesn't exist (and so $\mathrm{var}(X)$ does not exist either) (on the right).
- ☐ When $\mathrm{E}(X_i)$ does not exist, the possibility of huge values of X_i implies that $\overline{X}$ cannot converge.



Remarks

- The weak law is easy to prove under the supplementary hypothesis that $\mathrm{var}(X_j) = \sigma^2 < \infty$. We calculate $\mathrm{E}(\overline{X})$ and $\mathrm{var}(\overline{X})$, then we apply Chebyshov's inequality. For any $\varepsilon > 0$,

$$\mathrm{P}(|\overline{X} - \mu| > \varepsilon) \leq \mathrm{var}(\overline{X})/\varepsilon^2 = \frac{\sigma^2}{n\varepsilon^2} \to 0, \quad n \to \infty.$$

- The same result applies to smooth functions of averages, empirical quantiles, and other statistics.
- Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} F$, where F is a continuous cumulative distribution function, and let $x_p = F^{-1}(p)$ be the p quantile of F. By noting that

$$X_{(\lceil np \rceil)} \leq x_p \quad \Leftrightarrow \quad \sum_{j=1}^{n} I(X_j \leq x_p) \geq \lceil np \rceil$$

  and applying the weak law to the sum on the right, we have $X_{(\lceil np \rceil)} \xrightarrow{P} x_p$.

Strong law of large numbers

In fact, a stronger result is true:

Theorem 233. *(Strong law of large numbers) Under the conditions of the last theorem,* $\overline{X} \xrightarrow{\text{a.s.}} \mu$*:*

$$\mathrm{P}\left(\lim_{n\to\infty} \overline{X} = \mu\right) = 1.$$

- This is stronger in the sense that for all $\varepsilon > 0$, the weak law allows the event $|\overline{X} - \mu| > \varepsilon$ to occur an infinite number of times, though with smaller and smaller probabilities. The strong law excludes this possibility: it implies that the event $|\overline{X} - \mu| > \varepsilon$ can only occur a finite number of times.
- The weak and strong laws remain valid under certain types of dependence amongst the X_j.

6.4 Central Limit Theorem

Standardisation of an average

The law of large numbers shows us that the average $\overline{X}$ approaches μ when $n \to \infty$. If $\mathrm{var}(X_j) < \infty$, then Lemma 166 tells us that

$$\mathrm{E}(\overline{X}) = \mu, \quad \mathrm{var}(\overline{X}) = \sigma^2/n,$$

so, for all n, the difference between $\overline{X}$ and its expectation relative to its standard deviation,

$$Z_n = \frac{\overline{X} - \mathrm{E}(\overline{X})}{\mathrm{var}(\overline{X})^{1/2}} = \frac{\overline{X} - \mu}{\sqrt{\sigma^2/n}} = \frac{n^{1/2}(\overline{X} - \mu)}{\sigma}$$

has expected value zero and unit variance.

What is the limiting behaviour of Z_n?

Central limit theorem

Theorem 234 (Central limit theorem (CLT)). *Let $X_1, X_2, \ldots$ be independent random variables with expectation μ and variance $0 < \sigma^2 < \infty$. Then*

$$Z_n = \frac{n^{1/2}(\overline{X} - \mu)}{\sigma} \xrightarrow{D} Z, \quad n \to \infty,$$

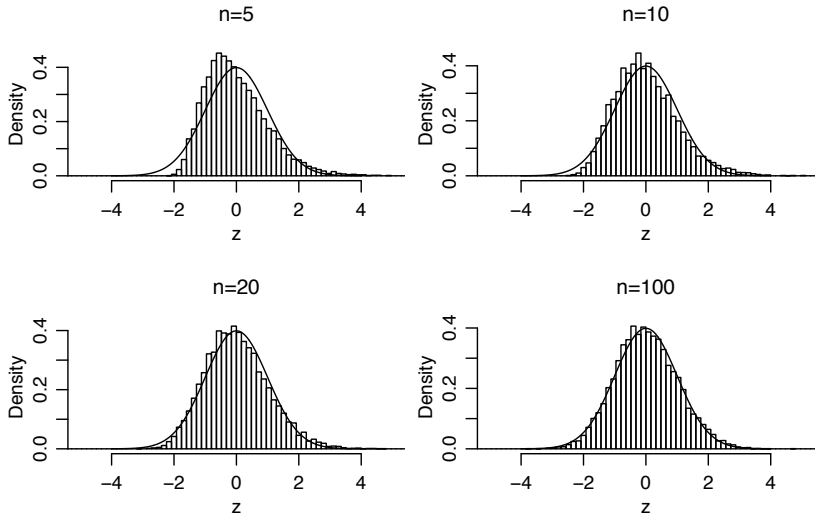
where $Z \sim N(0,1)$.

Thus

$$\mathrm{P}\left\{\frac{n^{1/2}(\overline{X} - \mu)}{\sigma} \leq z\right\} \doteq \mathrm{P}(Z \leq z) = \Phi(z)$$

for large n.

The following page shows this effect for $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \exp(1)$; the histograms show how the empirical densities of Z_n approach the density of Z.



Note to Theorem 234

The cumulant-generating function of

$$Z_n = \left(\overline{X} - \mu\right) / (\sigma^2/n)^{1/2} = \sum_{j=1}^{n} (n^{-1/2}/\sigma) X_j - n^{1/2} \frac{\mu}{\sigma}$$

is

$$K_{Z_n}(t) = \sum_{j=1}^{n} K_{X_j}(tn^{-1/2}/\sigma) - n^{1/2} \frac{\mu}{\sigma} t,$$

where

$$K_{X_j}(t) = t\mu + \tfrac{1}{2} t^2 \sigma^2 + o(t^2), \quad t \to 0.$$

Thus

$$K_{Z_n}(t) = n \left[tn^{-1/2}\mu/\sigma + \tfrac{1}{2}(tn^{-1/2}/\sigma)^2 \sigma^2 + o\{t^2/(n\sigma^2)\} \right] - n^{1/2} t \frac{\mu}{\sigma} \to t^2/2, \quad n \to \infty,$$

is the CGF of $Z \sim \mathcal{N}(0,1)$. Thus the result follows by the continuity theorem, Theorem 185.

Use of the CLT

The CLT is used to approximate probabilities involving the sums of independent random variables. Under the previous conditions, we have

$$\mathrm{E}\left(\sum_{j=1}^{n} X_j \right) = n\mu, \quad \mathrm{var}\left(\sum_{j=1}^{n} X_j \right) = n\sigma^2,$$

so

$$\frac{\sum_{j=1}^{n} X_j - n\mu}{\sqrt{n\sigma^2}} = \frac{n(\overline{X} - \mu)}{\sqrt{n\sigma^2}} = \frac{n^{1/2}(\overline{X} - \mu)}{\sigma} = Z_n$$

can be approximated using a normal variable:

$$\mathrm{P}\left(\sum_{j=1}^{n} X_j \le x \right) = \mathrm{P}\left\{ \frac{\sum_{j=1}^{n} X_j - n\mu}{\sqrt{n\sigma^2}} \le \frac{x - n\mu}{(n\sigma^2)^{1/2}} \right\} \doteq \Phi\left\{ \frac{x - n\mu}{(n\sigma^2)^{1/2}} \right\}.$$

The accuracy of the approximation depends on the underlying variables: it is (of course) exact for normal X_j, works better if the X_j are symmetrically distributed (e.g., uniform), and typically is adequate if $n > 25$ or so.

Example

Example 235. *A book of 640 pages has a number of random errors on each page. If the number of errors on each page follows a Poisson distribution with expectation $\lambda = 0.1$, what is the probability that the book contains less than 50 errors?*

When $\sum_{j=1}^{n} X_j$ takes whole values, we can obtain a better approximation using a continuity correction:

$$\mathrm{P}\left(\sum_{j=1}^{n} X_j \leq x\right) \doteq \Phi\left\{\frac{x + \frac{1}{2} - n\mu}{(n\sigma^2)^{1/2}}\right\};$$

this can be important when the distribution of $\sum_{j=1}^{n} X_j$ is quite discrete.

Note to Example 235

We take $\mu = \sigma^2 = 0.1$ and $n = 640$. The expected number of errors is $n\mu = 640\lambda = 64$, and the variance is $n\sigma^2 = 64$, as the variable is Poisson. Thus we seek

$$\mathrm{P}\left(\sum_{j=1}^{n} X_j \leq 49\right) = \mathrm{P}\left(\frac{\sum_{j=1}^{n} X_j - 64}{\sqrt{64}} \leq \frac{49 - 64}{\sqrt{64}}\right) \doteq \Phi(-15/8) = 0.03.$$

The true number is 0.031. With continuity correction we take $\Phi\{(-15 + 0.5)/8\} = 0.035$.

6.5 Delta Method

slide 261

Delta method

We often need the approximate distribution of a smooth function of an average.

Theorem 236. *Let $X_1, X_2, \ldots$ be independent random variables with expectation μ and variance $0 < \sigma^2 < \infty$, and let $g'(\mu) \neq 0$, where g' is the derivative of g. Then*

$$\frac{g(\overline{X}) - g(\mu)}{\{g'(\mu)^2\sigma^2/n\}^{1/2}} \xrightarrow{D} N(0,1), \quad n \to \infty.$$

This implies that for large n, we have $g(\overline{X}) \dot{\sim} N\left\{g(\mu), g'(\mu)^2\sigma^2/n\right\}$. Combined with Slutsky's lemma, we have

$$g(\overline{X}) \dot{\sim} N\left\{g(\mu), g'(\overline{X})^2 S^2/n\right\}, \quad S^2 = \frac{1}{n-1}\sum_{j=1}^{n}(X_j - \overline{X})^2.$$

Example 237. *If $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \exp(\lambda)$, find the approximate distribution of $\log \overline{X}$.*

8 Statistical Inference

slide 318

8.1 Introduction

slide 319

Introduction

The study of mathematics is based on **deduction**:

$$\text{axioms} \quad \Rightarrow \quad \text{consequences.}$$

In the case of probability, we have

$$(\Omega, \mathcal{F}, \mathrm{P}) \quad \Rightarrow \quad \mathrm{P}(A), \mathrm{P}(A \mid B), \mathrm{P}(X \leq x), \mathrm{E}(X^r), \ldots$$

Inferential statistics concern **induction**—having observed an event A, we want to say something about a probability space $(\Omega, \mathcal{F}, \mathrm{P})$ we suppose to be underlying the data:

$$A \quad \stackrel{?}{\Rightarrow} \quad (\Omega, \mathcal{F}, P).$$

In the past the term **inverse probability** was given to this process.

Statistical model

- ☐ We assume that the observed data, or data to be observed, can be considered as realisations of a random process, and we aim to say something about this process based on the data.
- ☐ Since the data are finite, and the process is unknown, there will be many uncertainties in our analysis, and we must try to quantify them as well as possible.
- ☐ Several problems must be addressed:
  - **specification** of a model (or of models) for the data;
  - **estimation** of the unknowns of the model (parameters, ...);
  - **tests** of hypotheses concerning a model;
  - **planning** of the data collection and analysis, to answer the key questions as effectively as possible (i.e., minimise uncertainty for a given cost);
  - **decision** when faced with uncertainties;
  - **prediction** of future unknowns;
  - behind the other problems lies the **relevance** of the data to the question we want to answer.

Definitions

Notation: we will use y and Y to represent the data $y_1, \ldots, y_n$ and $Y_1, \ldots, Y_n$.

Definition 244. *A* **statistical model** *is a probability distribution $f(y)$ chosen or constructed to learn from observed data y or from potential data Y.*

- ☐ *If $f(y) = f(y; \theta)$ is determined by a parameter θ of finite dimension, it is a* **parametric model**, *and otherwise it is a* **nonparametric model**.
- ☐ *A perfectly known model is called* **simple**, *otherwise it is* **composite**.

Statistical models are (almost) always composite in practice, but simple models are useful when developing theory.

Definition 245. *A* **statistic** *$T = t(Y)$ is a known function of the data Y.*

Definition 246. *The* **sampling distribution** *of a statistic $T = t(Y)$ is its distribution when $Y \sim f(y)$.*

Definition 247. *A* **random sample** *is a set of independent and identically distributed random variables $Y_1, \ldots, Y_n$, or their realisations $y_1, \ldots, y_n$.*

Examples

Example 248. *Assume that $y_1, \ldots, y_n$ is a random sample from a Bernoulli distribution with unknown parameter $p \in (0, 1)$. Then the statistic*

$$t = \sum_{j=1}^{n} y_j$$

is considered to be a realisation of the random variable

$$T = \sum_{j=1}^{n} Y_j,$$

whose sampling distribution is $B(n, p)$.

Example 249. *Assume that $y_1, \ldots, y_n$ is a random sample from the $\mathcal{N}(\mu, \sigma^2)$ distribution, with μ, σ^2 unknown. Then $\overline{y} = n^{-1}(y_1 + \cdots + y_n)$ and $s^2 = (n-1)^{-1} \sum_{j=1}^{n} (y_j - \overline{y})^2$ are statistics, realisations of the random variables*

$$\overline{Y} = n^{-1}(Y_1 + \cdots + Y_n), \quad S^2 = \frac{1}{n-1} \sum_{j=1}^{n} (Y_j - \overline{Y})^2.$$

Find the sampling distribution of $\overline{Y}$.

Note to Example 249

If μ and σ^2 are finite, then elementary computations (see Lemma 166) give

$$\mathrm{E}(\overline{Y}) = \mathrm{E}\left(n^{-1}\sum_{j=1}^{n} Y_j\right) = n^{-1}n\mathrm{E}(Y_j) = \mu, \quad \mathrm{var}(\overline{Y}) = \sum_{j=1}^{n} n^{-2}\mathrm{var}(Y_j) = \sigma^2/n,$$

since the Y_j are independent and all have variance σ^2. These results do not rely on normality of the Y_j, but the variance computation does need independence. We see that the larger n is, the smaller is the variance of $\overline{Y}$. This backs up our intuition that a larger sample is more informative about the underlying phenomenon—but the data must be sampled independently, and the variance must be finite! If in addition the Y_j are normal, then $\overline{Y}$ is a linear combination of normal variables, and so has a normal distribution,

$$\overline{Y} \sim \mathcal{N}(\mu, \sigma^2/n),$$

so we have a very precise idea of how $\overline{Y}$ will behave (or, rather, we would have, if we knew μ and σ^2).

8.2 Point Estimation

Statistical models

We would like to study a set of individuals or elements called a **population** based on a subset of this set called a **sample**:

- **statistical model**: the unknown distribution F or density f of Y;
- **parametric statistical model**: the distribution of Y is known except for the values of parameters θ, so we can write $F(y) = F(y;\theta)$, but with θ unknown;
- **sample** (must be representative of the population): "data" $y_1, \ldots, y_n$, often supposed to be a **random sample**, i.e., $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} F$;
- **statistic**: any function $T = t(Y_1, \ldots, Y_n)$ of the random variables $Y_1, \ldots, Y_n$;
- **estimator**: a statistic $\widehat{\theta}$ used to estimate a parameter θ of f.

Example

Example 250. *If we assume that* $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$ *but with* μ, σ^2 *unknown, then*

- *this is a parametric statistical model;*
- $\widehat{\mu} = \overline{Y}$ *is an estimator of* μ*, whose observed value is* $\overline{y}$*;*
- $\widehat{\sigma}^2 = n^{-1}\sum_{j=1}^{n}(Y_j - \overline{Y})^2$ *is an estimator of* σ^2*, whose observed value is* $n^{-1}\sum_{j=1}^{n}(y_j - \overline{y})^2$.

Note that:

- a statistic T is a function of the random variables $Y_1, \ldots, Y_n$, so T is itself a random variable;
- the **sampling distribution of** T depends on the distribution of the Y_j;
- if we cannot deduce the exact distribution of T from that of the Y_j, we must sometimes make do with knowing $\mathrm{E}(T)$ and $\mathrm{var}(T)$, which give partial information on the distribution of T, and thus may allow us to approximate the distribution of T (often using the central limit theorem).

Estimation methods

There are many methods for estimating the parameters of models. The choice among them depends on various criteria, such as:

- ease of calculation;
- efficiency (getting estimators that are as precise as possible);
- robustness (getting estimators that don't fail calamitously when the model is wrong, e.g., when outliers appear).

The trade-off between these criteria depends on what assumptions we are willing to make in a given context.
Examples of common methods are:

- **method of moments** (simple, can be inefficient);
- **maximum likelihood estimation** (general, optimal in many parametric models);
- **M-estimation** (even more general, can be robust, but loses efficiency compared to maximum likelihood).

Method of moments

- The **method of moments estimate** of a parameter θ is the value $\tilde{\theta}$ that matches the theoretical and empirical moments.
- For a model with p unknown parameters, we set the theoretical moments of the population equal to the empirical moments of the sample $y_1, \ldots, y_n$, and solve the resulting equations, i.e.,

$$\mathrm{E}(Y^r) = \int y^r f(y;\theta)\, dy = \frac{1}{n}\sum_{j=1}^{n} y_j^r, \quad r = 1, \ldots, p.$$

- We thus need as many (finite!) moments of the underlying model as there are unknown parameters.
- We may have more than one choice of moments to use, so in principle the estimate is not unique, but in practice we usually use the first r moments, because they give the most stable estimates.

Example 251. *If $y_1, \ldots, y_n$ is a random sample from the $U(0,\theta)$ distribution, estimate θ.*

Example 252. *If $y_1, \ldots, y_n$ is a random sample from the $\mathcal{N}(\mu, \sigma^2)$ distribution, estimate μ and σ^2.*

Example 251

- Standard computations show that if $Y \sim U(0,\theta)$, then $\mathrm{E}(Y) = \theta/2$. To find the moments estimate of θ, we therefore solve the equation

$$\mathrm{E}(Y) = \overline{y}, \quad \text{i.e.,} \quad \theta/2 = \overline{y},$$

  to get the estimate $\tilde{\theta} = 2\overline{y}$.
- Simulations show that with $n \geq 12$ the distribution of the random variable $\tilde{\theta}$ is very close to normality, as we would expect, because the central limit theorem gives a good approximation to the distribution of $\tilde{\theta}$ for small n, owing to the symmetry of the uniform distribution.

Example 252

The theoretical values of the first two moments are

$$\mathrm{E}(Y) = \mu, \quad \mathrm{E}(Y^2) = \mathrm{var}(Y) + \mathrm{E}(Y)^2 = \sigma^2 + \mu^2,$$

and the corresponding sample versions are

$$\overline{y} = \tilde{\mu}, \quad n^{-1}\sum_{j=1}^{n} y_j^2 = \tilde{\sigma}^2 + \tilde{\mu}^2.$$

Solving these gives

$$\tilde{\mu} = \overline{y}, \quad \tilde{\sigma}^2 = n^{-1}\left(\sum_{j=1}^{n} y_j^2 - n\,\overline{y}^2\right) = n^{-1}\sum_{j=1}^{n}(y_j - \overline{y})^2,$$

as can be seen by expanding out the right-hand expression:

$$\sum_{j=1}^{n}(y_j - \overline{y})^2 = \sum_{j=1}^{n} y_j^2 - \sum_{j=1}^{n} 2\overline{y}y_j + n\overline{y}^2 = \sum_{j=1}^{n} y_j^2 - 2n\overline{y}^2 + n\overline{y}^2 = \sum_{j=1}^{n} y_j^2 - n\overline{y}^2.$$

Maximum likelihood estimation

This is a much more general and powerful method of estimation, but in practice it usually requires numerical methods of optimisation.

Definition 253. *If $y_1, \ldots, y_n$ is a random sample from the density $f(y;\theta)$, then the* **likelihood** *for θ is*

$$L(\theta) = f(y_1, \ldots, y_n; \theta) = f(y_1;\theta) \times f(y_2;\theta) \times \cdots \times f(y_n;\theta).$$

The data are treated as fixed, and the likelihood $L(\theta)$ is regarded as a function of θ.

Definition 254. *The* **maximum likelihood estimate (MLE)** *$\widehat{\theta}$ of a parameter θ is the value that gives the observed data the highest likelihood. Thus*

$$L(\widehat{\theta}) \geq L(\theta) \text{ for each } \theta.$$

Calculation of the MLE $\widehat{\theta}$

We simplify the calculations by maximising $\ell(\theta) = \log L(\theta)$ rather than $L(\theta)$.
The approach is:

- ☐ calculate the log-likelihood $\ell(\theta)$ (and plot it if possible);
- ☐ find the value $\widehat{\theta}$ maximising $\ell(\theta)$, which often satisfies $d\ell(\widehat{\theta})/d\theta = 0$;
- ☐ check that $\widehat{\theta}$ gives a maximum, often by checking that $d^2\ell(\widehat{\theta})/d\theta^2 < 0$.

Example 255. *Suppose that $y_1, \ldots, y_n$ is a random sample from an exponential density with unknown λ. Find $\widehat{\lambda}$.*

Example 256. *Suppose that $y_1, \ldots, y_n$ is a random sample from a uniform density, $U(0, \theta)$, with unknown θ. Find $\widehat{\theta}$.*

Note to Example 255

The likelihood is

$$L(\lambda) = \lambda\, e^{-\lambda y_1} \times \cdots \times \lambda\, e^{-\lambda y_2} = \lambda^n e^{-\lambda(y_1 + \cdots + y_n)}, \quad \lambda > 0,$$

so the log likelihood is

$$\ell(\lambda) = \log L(\lambda) = n\, \log \lambda - n\lambda \overline{y}.$$

Thus the maximum likelihood estimate $\widehat{\lambda}$ is the solution to

$$\frac{d\ell(\lambda)}{d\lambda} = \frac{n}{\lambda} - n\overline{y} = 0,$$

and so $\widehat{\lambda} = 1/\overline{y}$.
To check that $\widehat{\lambda}$ gives a maximum, we note that the second derivative of $\ell(\lambda)$ is

$$\frac{d^2\ell(\lambda)}{d\lambda^2} = -\frac{n}{\lambda^2} < 0, \quad \lambda > 0,$$

so the log likelihood is concave, and therefore $\widehat{\lambda}$ gives the unique maximum.

Note to Example 256

The density is $f(y; \theta) = \theta^{-1} I(0 < y < \theta)$, so since the observations are independent, the likelihood is

$$L(\theta) = \prod_{j=1}^{n} \theta^{-1} I(0 < y_j < \theta) = \theta^{-n} I(0 < y_1, \ldots, y_n < \theta) = \theta^{-n} I(\theta > m), \quad \theta > 0,$$

where $m = \max(y_1, \ldots, y_n)$; note that $\prod_j I(0 < y_j < \theta) = I(m < \theta)$. Viewed as a function of θ this is maximised at $\widehat{\theta} = m$, which is therefore the MLE.
In this case the maximum is NOT found by differentiation of the likelihood, which is not differentiable at $\widehat{\theta}$.

M-estimation

- This generalises maximum likelihood estimation. We maximise a function of the form

$$\rho(\theta; Y) = \sum_{j=1}^{n} \rho(\theta; Y_j),$$

  where $\rho(\theta; y)$ is (if possible) concave as a function of θ for all y. Equivalently we minimise $-\rho(\theta; Y)$.
- We choose the function ρ to give estimators with suitable properties, such as small variance or robustness to outliers.
- Taking $\rho(\theta; y) = \log f(y; \theta)$ gives the maximum likelihood estimator.

Example 257. *Let* $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} f$ *with* $\mathrm{E}(Y_j) = \theta$*, and take* $\rho(y; \theta) = -(y - \theta)^2$*. Find the* **least squares estimator** *of* θ*.*

Example 258. *Let* $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} f$ *such that* $\mathrm{E}(Y_j) = \theta$*, and take* $\rho(y; \theta) = -|y - \theta|$*. Find the corresponding estimator of* θ*.*

Note to Example 257

We want to maximise

$$\rho(\theta; y) = -\sum_{j=1}^{n} (y_j - \theta)^2,$$

and this is equivalent to minimising the sum of squares

$$-\rho(\theta; y) = \sum_{j=1}^{n} (y_j - \theta)^2$$

with respect to θ. Differentiation gives

$$-\frac{d\rho(\theta; y)}{d\theta} = -\sum_{j=1}^{n} 2(y_j - \theta),$$

and setting this equal to zero gives $\widehat{\theta} = \overline{y}$. The second derivative is

$$-\frac{d^2\rho(\theta; y)}{d^2\theta} = 2n > 0,$$

so the minimum is unique.

Note to Example 258

We want to maximise

$$\rho(\theta; y) = -\sum_{j=1}^{n} |y_j - \theta|,$$

and we note that if $\theta > y$ then $-|y - \theta| = y - \theta$ and if $\theta < y$ then $-|y - \theta| = \theta - y$, so the respective derivatives with respect to θ are -1 and $+1$. This implies that

$$-\frac{d\rho(\theta; y)}{d\theta} = P(\theta) - N(\theta),$$

where $P(\theta)$ is the number of y_j for which $\theta < y_j$ and $N(\theta) = n - P(\theta)$ is the number of y_j for which $\theta > y_j$. Hence when regarded as a function of θ,

$$-\frac{d\rho(\theta; y)}{d\theta} = 2P(\theta) - n$$

is a step function that has initial value n for $\theta = -\infty$, drops by 2 at each y_j, and takes value $-n$ when $\theta = +\infty$. If n is odd, then $2P(\theta) - n$ equals zero when θ is the median of the sample, and if n is even, then $2P(\theta) - n$ equals zero on the interval $y_{(n/2)} \leq \theta \leq y_{(n/2+1)}$. In this latter case we can take the median to be $(y_{(n/2)} + y_{(n/2+1)})/2$ for uniqueness.
Thus this choice of function ρ yields the sample median as an estimator.

Bias

How should we compare estimators?

Definition 259. *The bias of the estimator $\widehat{\theta}$ of θ is*

$$b(\theta) = \mathrm{E}(\widehat{\theta}) - \theta.$$

- Interpretation of the bias:
  - if $b(\theta) < 0$ for all θ, then on average $\widehat{\theta}$ underestimates θ;
  - if $b(\theta) > 0$ for all θ, then on average $\widehat{\theta}$ overestimates θ;
  - if $b(\theta) = 0$ for all θ, then $\widehat{\theta}$ is said to be **unbiased**.
- If $b(\theta) \approx 0$, then $\widehat{\theta}$ is 'in the right place' on average.

Example 260. *Let $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$. Find the bias and variance of $\widehat{\mu} = \overline{Y}$ and the bias of $\widehat{\sigma}^2 = n^{-1} \sum_j (Y_j - \overline{Y})^2$.*

Note to Example 260

☐ In Example 249 we saw that

$$\mathrm{E}(\overline{Y}) = \mu, \quad \mathrm{var}(\overline{Y}) = \sigma^2/n,$$

so the bias of $\widehat{\mu} = \overline{Y}$ as an estimator of μ is $\mathrm{E}(\overline{Y}) - \mu = 0$.

☐ To find the expectation of $\widehat{\sigma}^2 = n^{-1}\sum_j (Y_j - \overline{Y})^2$, note that

$$\begin{aligned}
\sum_{j=1}^{n}(Y_j - \overline{Y})^2 &= \sum_{j=1}^{n}\left\{Y_j - \mu - (\overline{Y} - \mu)\right\}^2 \\
&= \sum_{j=1}^{n}(Y_j - \mu)^2 - 2\sum_{j=1}^{n}(Y_j - \mu)(\overline{Y} - \mu) + \sum_{j=1}^{n}(\overline{Y} - \mu)^2 \\
&= \sum_{j=1}^{n}(Y_j - \mu)^2 - 2n(\overline{Y} - \mu)^2 + n(\overline{Y} - \mu)^2 \\
&= \sum_{j=1}^{n}(Y_j - \mu)^2 - n(\overline{Y} - \mu)^2,
\end{aligned}$$

which implies that

$$\begin{aligned}
\mathrm{E}\left\{\sum_{j=1}^{n}(Y_j - \overline{Y})^2\right\} &= \mathrm{E}\left\{\sum_{j=1}^{n}(Y_j - \mu)^2\right\} - n\mathrm{E}\left\{(\overline{Y} - \mu)^2\right\} \\
&= n\mathrm{var}(Y_j) - n\mathrm{var}(\overline{Y}) \\
&= n\sigma^2 - n\sigma^2/n \\
&= (n-1)\sigma^2.
\end{aligned}$$

Therefore

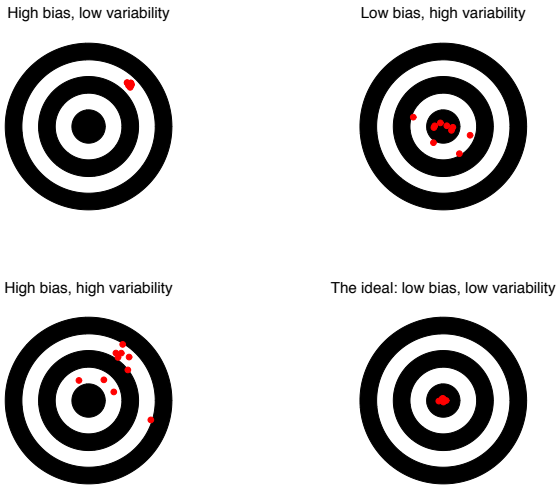
$$\mathrm{E}\left(\widehat{\sigma}^2\right) = n^{-1}\mathrm{E}\left\{\sum_{j=1}^{n}(Y_j - \overline{Y})^2\right\} = \frac{n-1}{n}\sigma^2,$$

and the bias of $\widehat{\sigma}^2$ is

$$\mathrm{E}(\widehat{\sigma}^2) - \sigma^2 = \frac{(n-1)\sigma^2}{n} - \sigma^2 = -\frac{\sigma^2}{n}.$$

Therefore on average $\widehat{\sigma}^2$ underestimates σ^2, by an amount that should be small for large n.

Bias and variance



- θ = bullseye, supposed to be the real value
- $\widehat{\theta}$ = red dart thrown at the bullseye, value estimated using the data

Mean square error

Definition 261. *The* **mean square error (MSE)** *of the estimator* $\widehat{\theta}$ *of* θ *is*

$$\mathrm{MSE}(\widehat{\theta}) = \mathrm{E}\{(\widehat{\theta} - \theta)^2\} = \cdots = \mathrm{var}(\widehat{\theta}) + b(\theta)^2.$$

This is the average squared distance between $\widehat{\theta}$ and its target value θ.

Definition 262. *Let* $\widehat{\theta}_1$ *and* $\widehat{\theta}_2$ *be two unbiased estimators of the same parameter* θ*. Then*

$$\begin{aligned} \mathrm{MSE}(\widehat{\theta}_1) &= \mathrm{var}(\widehat{\theta}_1) + b_1(\theta)^2 = \mathrm{var}(\widehat{\theta}_1) \\ \mathrm{MSE}(\widehat{\theta}_2) &= \mathrm{var}(\widehat{\theta}_2) + b_2(\theta)^2 = \mathrm{var}(\widehat{\theta}_2), \end{aligned}$$

and we say that $\widehat{\theta}_1$ *is* **more efficient** *than* $\widehat{\theta}_2$ *if*

$$\mathrm{var}(\widehat{\theta}_1) \leq \mathrm{var}(\widehat{\theta}_2).$$

If so, then we prefer $\widehat{\theta}_1$ *to* $\widehat{\theta}_2$*.*

Example 263. *Let* $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$*, with large* n*. Find the bias and variance of the median* M *and the average* $\overline{Y}$*. Which is preferable? What if outliers might appear?*

Note to Example 263

- We've already seen in Lemma 166 that

$$\mathrm{E}(\overline{Y}) = \mu, \quad \mathrm{var}(\overline{Y}) = \sigma^2/n,$$

  so the bias of $\overline{Y}$ as an estimator of μ is $\mathrm{E}(\overline{Y}) - \mu = 0$.
- Results from Example 240 give that for large n,

$$\mathrm{E}(M) \doteq \mu, \quad \mathrm{var}(M) = \frac{\pi\sigma^2}{2n},$$

  so both estimators are (approximately) unbiased (in fact exactly unbiased), but

$$\frac{\mathrm{var}(M)}{\mathrm{var}(\overline{Y})} = \frac{\pi}{2} > 1,$$

  so M is less efficient than $\overline{Y}$, because the latter has a smaller variance.
  However if there are outliers, we have seen that the median M is little changed, whereas the average $\overline{Y}$ can be badly affected. Our choice between these estimators will depend on how much we fear that our data will be contaminated by bad values.

Delta method

In practice, we often consider functions of estimators, and so we appeal to another version of the delta method (Theorem 236).

Theorem 264 (Delta method). *Let $\widehat{\theta}$ be an estimator based on a sample of size n, such that*

$$\widehat{\theta} \dot{\sim} \mathcal{N}(\theta, v/n),$$

for large n, and let g be a smooth function such that $g'(\theta) \neq 0$. Then

$$g(\widehat{\theta}) \dot{\sim} \mathcal{N}\left\{g(\theta) + vg''(\theta)/(2n), vg'(\theta)^2/n\right\}.$$

This implies that the mean square error of $g(\widehat{\theta})$ as an estimator of $g(\theta)$ is

$$\mathrm{MSE}\left\{g(\widehat{\theta})\right\} \approx \left\{\frac{vg''(\theta)}{2n}\right\}^2 + \frac{vg'(\theta)^2}{n}.$$

Thus for large n we can disregard the bias contribution.

Example 265. *Let $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} \mathrm{Poiss}(\theta)$. Find two estimators of $\mathrm{P}(Y = 0)$, and compare their biases and variances.*

Note to Example 265

- Let $\psi = g(\theta) = \exp(-\theta) = \mathrm{P}(Y = 0)$.
- The two estimators are $T_1 = n^{-1} \sum I(Y_i = 0)$ and $T_2 = \exp(-\overline{Y})$.
- Simple computations (e.g., noting that $nT_1 \sim B(n, \psi)$) give

$$\mathrm{E}(T_1) = \psi, \quad \mathrm{var}(T_1) = \psi(1 - \psi)/n.$$

  Thus T_1 is unbiased and has MSE $\psi(1 - \psi)/n$.
- For T_2 we note that $\widehat{\theta} = \overline{Y}$ has mean and variance θ and θ/n, and hence

$$\mathrm{E}(T_2) \doteq \exp(-\theta) + \theta \exp(-\theta)/(2n), \quad \mathrm{var}(T_2) \doteq \theta \exp(-2\theta)/n.$$

  Therefore T_2 has positive bias $\theta \exp(-\theta)/(2n)$ but

$$\frac{\mathrm{var}(T_2)}{\mathrm{var}(T_1)} = \frac{\theta \exp(-2\theta)}{\exp(-\theta)\{1 - \exp(-\theta)\}} = \frac{\theta}{e^\theta - 1} < 1$$

  for all $\theta > 0$.
  Therefore T_2 is preferable to T_1 in terms of variance (especially if θ is large).

Efficiency and robustness

- Under certain conditions, notably that $y_1, \ldots, y_n$ are really from the assumed model $f(y; \theta)$, and if f is 'nice', the maximum likelihood estimator $\widehat{\theta}$ has good properties: for large n, $\mathrm{E}(\widehat{\theta}) \doteq \theta$, and $\mathrm{var}(\widehat{\theta})$ is minimal, so no estimator is better than $\widehat{\theta}$.
- In reality we are never certain of the model, and often we sacrifice some efficiency (small variance under an ideal model) for robustness (good estimation even if there are outliers, or if the assumed model is incorrect).
- If θ is a $p \times 1$ vector, the same ideas apply. For example, for M-estimation we maximise

$$\sum_{j=1}^{n} \rho(\theta; y_j)$$

  with respect to the vector $\theta_{p\times 1}$, giving an estimator $\widehat{\theta}_{p\times 1}$, which often has an approximate $\mathcal{N}_p(\theta, V)$ distribution.

8.3 Interval Estimation

Pivots

A key element of statistical thinking is to assess uncertainty of results and conclusions.
Let $t = 1$ be an estimate of an unknown parameter θ based on a sample of size n:

- if $n = 10^5$ we are much more sure that $\theta \approx t$ than if $n = 10$;
- as well as t we would thus like to give an interval which will be wider when $n = 10$ than when $n = 10^5$, to make the uncertainty of t explicit.

We suppose that we have

- **data** $y_1, \ldots, y_n$, which are regarded as a realisation of a
- **random sample** $Y_1, \ldots, Y_n$ drawn from a
- **statistical model** $f(y; \theta)$ whose unknown
- **parameter** θ is estimated by the
- **estimate** $t = t(y_1, \ldots, y_n)$, which is regarded as a realisation of the
- **estimator** $T = t(Y_1, \ldots, Y_n)$.

We therefore need to link θ and $Y_1, \ldots, Y_n$.

Definition 266. *Let $Y = (Y_1, \ldots, Y_n)$ be sampled from a distribution F with parameter θ. Then a* **pivot** *is a function $Q = q(Y, \theta)$ of the data and the parameter θ, where the distribution of Q is known and does not depend on θ. We say that Q is* **pivotal**.

Example

Example 267. *Let $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} U(0, \theta)$ with θ unknown,*

$$M = \max(Y_1, \ldots, Y_n), \quad \overline{Y} = n^{-1} \sum Y_j.$$

- *Show that $Q_1 = M/\theta$ is a pivot.*
- *Use the central limit theorem to find an approximate pivot Q_2 for large n, based on $\overline{Y}$.*

Note to Example 267

- We first note that Q_1 is a function of the data and the parameter, and that

$$\mathrm{P}(M \le x) = F_Y(x)^n = (x/\theta)^n, \quad 0 < x < \theta,$$

so

$$\mathrm{P}(Q_1 \le q) = \mathrm{P}(M/\theta \le q) = \mathrm{P}(M \le \theta q) = (\theta q/\theta)^n = q^n, \quad 0 < q < 1.$$

which is known and does not depend on θ. Hence Q_1 is a pivot.

- In Example 119(a) we saw that if $Y \sim U(0, \theta)$, then $\mathrm{E}(Y) = \theta/2$ and $\mathrm{var}(Y) = \theta^2/12$. Hence Lemma 166(c) gives that $\overline{Y}$ has mean $\theta/2$ and variance $\theta^2/(12n)$, and for large n, $\overline{Y} \dot{\sim} \mathcal{N}\{\theta/2, \theta^2/(12n)\}$ using the central limit theorem. Therefore

$$Q_2 = \frac{\overline{Y} - \theta/2}{\sqrt{\theta^2/(12n)}} = (3n)^{1/2}(2\overline{Y}/\theta - 1) \dot{\sim} \mathcal{N}(0, 1).$$

Thus Q_2 depends on both data and θ, and has an (approximately) known distribution: hence Q_2 is an (approximate) pivot. (In fact it is exact, if we could know the distribution of $\overline{Y}$ exactly.)

Confidence intervals

Definition 268. *Let $Y = (Y_1, \ldots, Y_n)$ be data from a parametric statistical model with scalar parameter θ. A* **confidence interval (CI)** (L, U) **for** θ *with lower confidence bound L and upper confidence bound U is a random interval that contains θ with a specified probability, called the* **(confidence) level** *of the interval.*

- $L = l(Y)$ and $U = u(Y)$ are statistics that can be computed from the data $Y_1, \ldots, Y_n$. They do not depend on θ.
- In a continuous setting (so $<$ gives the same probabilities as $\le$), and if we write the probabilities that θ lies below and above the interval as

$$\mathrm{P}(\theta < L) = \alpha_L, \quad \mathrm{P}(U < \theta) = \alpha_U,$$

then (L, U) has confidence level

$$\mathrm{P}(L \le \theta \le U) = 1 - \mathrm{P}(\theta < L) - \mathrm{P}(U < \theta) = 1 - \alpha_L - \alpha_U.$$

- Often we seek an interval with equal probabilities of not containing θ at each end, with $\alpha_L = \alpha_U = \alpha/2$, giving an **equi-tailed** $(1 - \alpha) \times 100\%$ **confidence interval**.
- We usually take standard values of α, such that $1 - \alpha = 0.9, 0.95, 0.99, \ldots$

Construction of a CI

- We use pivots to construct CIs:
  - we find a pivot $Q = q(Y, \theta)$ involving θ;
  - we obtain the quantiles q_{α_U}, $q_{1-\alpha_L}$ of Q;
  - then we transform the equation

$$\mathrm{P}\{q_{\alpha_U} \le q(Y,\theta) \le q_{1-\alpha_L}\} = (1-\alpha_L) - \alpha_U$$

    into the form

$$\mathrm{P}(L \le \theta \le U) = 1 - \alpha_L - \alpha_U,$$

    where the bounds L, U depend on Y, $q_{1-\alpha_L}$ and q_{α_U}, but not on θ.
- In many cases, the bounds are of a standard form (see below).

Example 269. *In Example 267, find CIs based on Q_1 and on Q_2.*

Example 270. *A sample of $n = 16$ Vaudois number plates has maximum 523308 and average 320869. Give two-sided 95% CIs for the number of cars in canton Vaud.*

Note to Example 269

- The p quantile of $Q_1 = M/\theta$ is given by $p = \mathrm{P}(Q_1 \le q_p) = q_p^n$, so $q_p = p^{1/n}$. Thus

$$\mathrm{P}\{\alpha_U^{1/n} \le M/\theta \le (1-\alpha_L)^{1/n}\} = 1 - \alpha_L - \alpha_U,$$

  and a little algebra gives that

$$\mathrm{P}\{M/(1-\alpha_L)^{1/n} \le \theta \le M/\alpha_U^{1/n}\} = 1 - \alpha_L - \alpha_U,$$

  so

$$L = M/(1-\alpha_L)^{1/n}, \quad U = M/\alpha_U^{1/n}.$$

- For $Q_2 = (3n)^{1/2}(2\overline{Y}/\theta - 1) \dot{\sim} \mathcal{N}(0,1)$, the quantiles are $z_{1-\alpha_L}$ and z_{α_U}, so

$$\mathrm{P}\{z_{\alpha_U} \le (3n)^{1/2}(2\overline{Y}/\theta - 1) \le z_{1-\alpha_L}\} = 1 - \alpha_L - \alpha_U,$$

  and hence we obtain

$$L = \frac{2\overline{Y}}{1 + z_{1-\alpha_L}/(3n)^{1/2}}, \quad U = \frac{2\overline{Y}}{1 + z_{\alpha_U}/(3n)^{1/2}};$$

  note that for large n these are $L \approx 2\overline{Y}\{1 - z_{1-\alpha_L}/(3n)^{1/2}\}$ and $U \approx 2\overline{Y}\{1 - z_{\alpha_U}/(3n)^{1/2}\}$.

Note to Example 270

- We set $\alpha_U = \alpha_L = 0.025$, with M and $\overline{Y}$ observed to be $m = 523308$ and $\overline{y} = 320869$.
- For Q_1 with $n = 16$ we have $\alpha_U^{1/n} = 0.025^{1/16} = 0.794$, $(1-\alpha_L)^{1/n} = 0.975^{1/16} = 0.998$, so

$$L = m/(1-\alpha_L)^{1/n} = 524135, \quad U = m/\alpha_U^{1/n} = 659001.$$

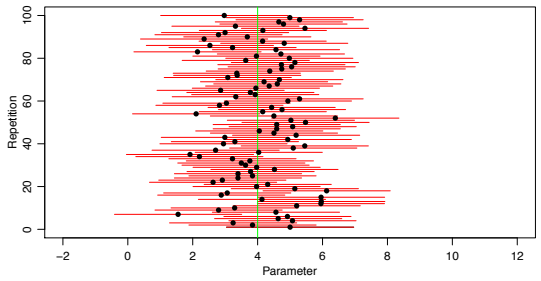
  Note that this CI does not contain m (and this makes sense).
- For $Q_2 = (3n)^{1/2}(2\overline{Y}/\theta - 1) \dot{\sim} \mathcal{N}(0,1)$, the quantiles are $z_{\alpha_U} = -z_{1-\alpha_L} = -1.96$, so we obtain

$$L = \frac{2\overline{y}}{1 + 1.96/(3n)^{1/2}} = 500226, \quad U = \frac{2\overline{y}}{1 - 1.96/(3n)^{1/2}} = 894903.$$

  This is much wider than the other CI, and includes impossible values, as we already know that $\theta \geq m$.
- Clearly we prefer the interval based on Q_1.

Interpretation of a CI

- (L, U) is a random interval that contains θ with probability $1 - \alpha$.
- We imagine an infinite sequence of repetitions of the experiment that gave (L, U).
- In that case, the CI that we calculated is one of an infinity of possible CIs, and we can consider that our CI was chosen at random from among them.
- Although we do not know whether our particular CI contains θ, the event $\theta \in (L, U)$ has probability $1 - \alpha$, matching the confidence level of the CI.
- In the figure below, the parameter θ (green line) is contained (or not) in realisations of the 95% CI (red). The black points show the corresponding estimates.



One- and two-sided intervals

- A **two-sided confidence interval** (L, U) is generally used, but **one-sided confidence intervals**, of the form $(-\infty, U)$ or (L, ∞), are also sometimes required.
- For one-sided CIs, we take $\alpha_U = 0$ or $\alpha_L = 0$, giving respective intervals (L, ∞) or $(-\infty, U)$.
- To get a one-sided $(1-\alpha) \times 100\%$ interval, we can compute a two-sided interval with $\alpha_L = \alpha_U = \alpha$, and then replace the unwanted limit by $\pm\infty$ (or another value if required in the context).

Example 271. *A sample of $n = 16$ Vaudois number plates has maximum 523308. Use the pivot Q_1 to give one-sided 95% CIs for the number of cars in canton Vaud.*

Note to Example 271

- We set $\alpha_U = \alpha_L = 0.05$, with M observed to be $m = 523308$.
- For Q_1 with $n = 16$ we have $\alpha_U^{1/n} = 0.05^{1/16} = 0.829$, $(1-\alpha_L)^{1/n} = 0.95^{1/16} = 0.997$, so

$$L = m/(1-\alpha_L)^{1/n} = 524988.3, \quad U = m/\alpha_U^{1/n} = 631061.6.$$

- For the interval of form (L, ∞), we have have $(524988.3, \infty)$, with the interpretation that we are 95% sure that the number of cars in the canton is at least 524988.3 (which we would interpret as 524988, for practical purposes).
- For the interval of form $(-\infty, U)$, we have have $(-\infty, 631061.6)$, but since we have observed $m = 523308$, we replace the lower bound, giving $(523308, 631061.6)$. We are 95% sure that the number of cars in the canton is lower than 631062 but it must be at least 523308.

Standard errors

In most cases we use approximate pivots, based on estimators whose variances we must estimate.

Definition 272. *Let $T = t(Y_1, \ldots, Y_n)$ be an estimator of θ, let $\tau_n^2 = \mathrm{var}(T)$ be its variance, and let $V = v(Y_1, \ldots, Y_n)$ be an estimator of τ_n^2. Then we call $V^{1/2}$, or its realisation $v^{1/2}$, a* **standard error** *for T.*

Theorem 273. *Let T be an estimator of θ based on a sample of size n, with*

$$\frac{T-\theta}{\tau_n} \xrightarrow{D} Z, \quad \frac{V}{\tau_n^2} \xrightarrow{P} 1, \quad n \to \infty,$$

where $Z \sim \mathcal{N}(0,1)$. Then by Theorem 227 we have

$$\frac{T-\theta}{V^{1/2}} = \frac{T-\theta}{\tau_n} \times \frac{\tau_n}{V^{1/2}} \xrightarrow{D} Z, \quad n \to \infty.$$

Hence, when basing a CI on the Central Limit Theorem, we can replace τ_n by $V^{1/2}$.

Approximate normal confidence intervals

- We can often construct approximate CIs using the CLT, since many statistics that are based on averages of $Y = (Y_1, \ldots, Y_n)$ have approximate normal distributions for large n. If $T = t(Y)$ is an estimator of θ with standard error $\sqrt{V}$, and if Theorem 273 applies, then

$$T \;\dot\sim\; N(\theta, V),$$

and so $(T-\theta)/\sqrt{V} \dot\sim N(0,1)$. Thus

$$\mathrm{P}\left\{ z_{\alpha_U} < (T-\theta)/\sqrt{V} \le z_{1-\alpha_L} \right\} \doteq \Phi(z_{1-\alpha_L}) - \Phi(z_{\alpha_U}) = 1 - \alpha_L - \alpha_U,$$

implying that an approximate $(1-\alpha_L-\alpha_U)\times 100\%$ CI for θ is

$$(L, U) = (T - \sqrt{V} z_{1-\alpha_L}, T - \sqrt{V} z_{\alpha_U}).$$

Recall that if $\alpha_L, \alpha_U < 1/2$, then $z_{1-\alpha_L} > 0$ and $z_{\alpha_U} < 0$, so $L < U$.

- Example 269 is an example of this, with $T = 2\overline{Y}$ and $V = T^2/(3n)$, since for large n,

$$L \approx T - T z_{1-\alpha_L}/(3n)^{1/2}, \quad U \approx T - T z_{\alpha_U}/(3n)^{1/2}.$$

- Often we take $\alpha_L = \alpha_U = 0.025$, and then $z_{1-\alpha_L} = -z_{\alpha_U} = 1.96$, giving the 'rule of thumb' $(L, U) \approx T \pm 2\sqrt{V}$ for a two-sided 95% confidence interval.

Normal random sample

An important case where exact CIs are available is the normal random sample.

Theorem 274. *If* $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$, *then*

$$\left.\begin{array}{c} \overline{Y} \sim \mathcal{N}(\mu, \sigma^2/n) \\ (n-1)S^2 = \sum_{j=1}^n (Y_j - \overline{Y})^2 \sim \sigma^2 \chi^2_{n-1} \end{array}\right\} \textit{ independent}$$

where χ^2_ν *represents the* **chi-square distribution with ν degrees of freedom**.

The first result here implies that if σ^2 is known, then

$$Z = \frac{\overline{Y} - \mu}{\sqrt{\sigma^2/n}} \sim \mathcal{N}(0,1).$$

is a pivot that provides an exact $(1-\alpha_L-\alpha_U)$ confidence interval for μ, of the form

$$(L, U) = \left(\overline{Y} - \frac{\sigma}{\sqrt{n}} z_{1-\alpha_L}, \overline{Y} - \frac{\sigma}{\sqrt{n}} z_{\alpha_U} \right), \tag{4}$$

where z_p denotes the p quantile of the standard normal distribution.

Unknown variance

☐ In applications σ^2 is usually unknown. If so, Theorem 274 implies that

$$\frac{\overline{Y}-\mu}{\sqrt{S^2/n}} \sim t_{n-1}, \quad \frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$$

are pivots that provide confidence intervals for μ and σ^2, respectively, i.e.,

$$(L,U) = \left(\overline{Y} - \frac{S}{\sqrt{n}} t_{n-1}(1-\alpha_L), \overline{Y} - \frac{S}{\sqrt{n}} t_{n-1}(\alpha_U)\right), \tag{5}$$

$$(L,U) = \left(\frac{(n-1)S^2}{\chi^2_{n-1}(1-\alpha_L)}, \frac{(n-1)S^2}{\chi^2_{n-1}(\alpha_U)}\right), \tag{6}$$

where:

- $t_\nu(p)$ is the p quantile of the **Student t distribution with ν degrees of freedom**;
- $\chi^2_\nu(p)$ is the p quantile of the **chi-square distribution with ν degrees of freedom**.

☐ For symmetric densities such as the normal and the Student t, the quantiles satisfy

$$z_p = -z_{1-p}, \quad t_\nu(p) = -t_\nu(1-p),$$

so equi-tailed $(1-\alpha)\times 100\%$ CIs have the forms

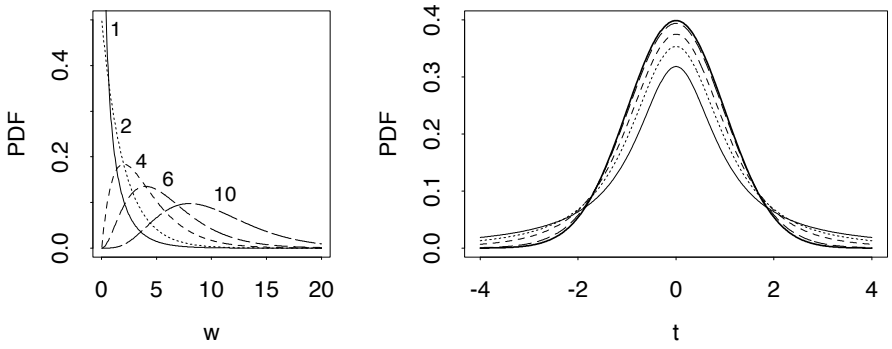
$$\overline{Y} \pm n^{-1/2}\,\sigma\, z_{1-\alpha/2}, \quad \overline{Y} \pm n^{-1/2}\, S\, t_{n-1}(1-\alpha/2).$$

Two giants of 20th century statistics

Left: William Sealy Gosset ('Student') (1876–1937)
Right: Ronald Aylmer Fisher (1890–1962)

(Source: Wikipedia)

Chi-square and Student probability densities



Left: χ^2_ν densities with $\nu = 1, 2, 4, 6, 10$. Right: t_ν densities with $\nu = 1$ (bottom centre), 2, 4, 20, ∞ (top centre).

Example

Example 275. *Suppose that the resistance X of a certain type of electrical equipment has an approximate $\mathcal{N}(\mu, \sigma^2)$ distribution. A random sample of size $n = 9$ has average $\overline{x} = 5.34$ ohm and variance $s^2 = 0.12^2$ ohm^2.*

- ☐ *Find an equi-tailed two-sided 95% CI for μ.*
- ☐ *Find an equi-tailed two-sided 95% CI for σ^2.*
- ☐ *How does the interval for μ change if we are later told that $\sigma^2 = 0.12^2$?*
- ☐ *How does the calculation change if we want a 95% confidence interval for μ of form (L, ∞)?*

Note to Example 275

- ☐ We want $1 - \alpha = 0.95$, so $\alpha = 0.05$ and we take $\alpha_U = \alpha_L = 0.025$. The formula (5) gives $(5.25, 5.43)$ ohms.
- ☐ Formula (6) gives $(0.0066, 0.0529)$ ohms^2 as the interval for σ^2, giving $(\sqrt{0.0066}, \sqrt{0.0529}) = (0.081, 0.230)$ ohms as the interval for σ (which must be positive).
- ☐ In this case σ^2 is known, so we should use (4). We replace $t_9(0.975) = 2.306$ with $z_{0.975} = 1.96$, giving $(5.26, 5.42)$ ohm. This interval is a factor $2.306/1.96 = 1.18$ shorter, because there is no uncertainty about the value of σ.
- ☐ Now we want $U = \infty$, so we take $\alpha_U = 0$ and $\alpha_L = 0.05$, and replace the first interval above by

$$\left(\overline{Y} + \frac{S}{\sqrt{n}} t_{n-1}(\alpha_L), \infty\right) \quad \text{ohms},$$

which gives $(5.27, \infty)$ ohms.

Comments

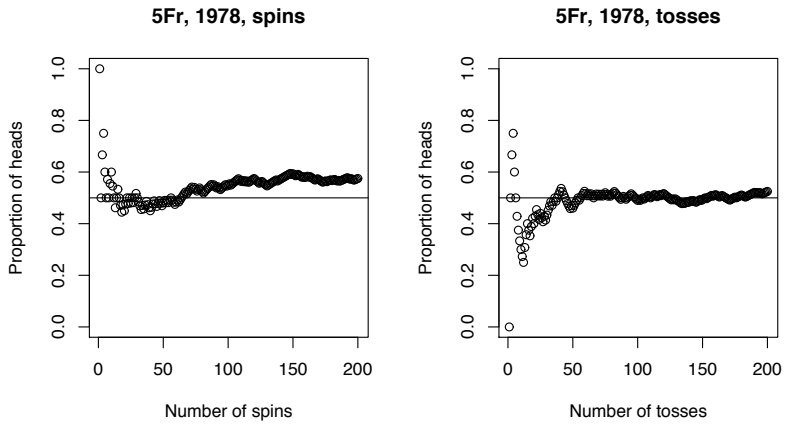
- ☐ The construction of confidence intervals is based on pivots, often using the central limit theorem to approximate the distribution of an estimator, and thus giving approximate intervals.
- ☐ A confidence interval (L, U) not only suggests where an unknown parameter is situated, but its width $U - L$ gives an idea of the precision of the estimate.
- ☐ In most cases
$$U - L \propto \sqrt{V} \propto n^{-1/2},$$
so multiplying the sample size by 100 increases precision only by a factor of 10.
- ☐ Having to estimate the variance using V decreases precision, and thus increases the width.
- ☐ To get a one-sided $(1 - \alpha) \times 100\%$ interval, we can compute a two-sided interval with $\alpha_L = \alpha_U = \alpha$, and then replace the unwanted limit by $\pm\infty$ (or another suitable limit).
- ☐ In some cases, especially normal models, exact CIs are available.

8.4 Hypothesis Tests

Statistical tests

Example 276. *I observe 115 heads when spinning it a 5Fr coin 200 times, and 105 heads when tossing it.*

- ☐ *Give a statistical model for this problem.*
- ☐ *Is the coin fair?*



Note to Example 276

- On the assumption that the spins are independent, and that heads occurs with probability θ, the total number of heads $R \sim B(n = 200, \theta)$, and if the coin is fair, $\theta = 1/2$.
- One way to see if the coin is fair is to compute a 95% CI for the unknown θ, and see if the value $\theta = 1/2$ lies in the interval.
- An unbiased estimator for θ is $\widehat{\theta} = R/n$ (and in fact this is the MLE, and the moments estimator), and its variance is $\theta(1-\theta)/n$, which we can estimate by $V = \widehat{\theta}(1-\widehat{\theta})/n$, so our discussion of confidence intervals tells us that an approximate 95% confidence for θ is

$$\widehat{\theta} \pm z_{1-\alpha/2}\sqrt{V} = \widehat{\theta} \pm 1.96\sqrt{V},$$

  which gives

$$\text{tosses: } (0.456, 0.594) \qquad \text{spins: } (0.506, 0.644),$$

  suggesting that since the 95% confidence interval for spins does not contain $1/2$, the coin is not fair for spins, but that it is fair for tosses.
- Note that if we had had $R = 85$ for tosses, then we would get interval $(0.356, 0.494)$, and would also have concluded that the coin is not fair for tosses.
- Similar computations for the CI with $\alpha = 99\%$ give

$$\text{tosses: } (0.434, 0.616) \qquad \text{spins: } (0.485, 0.665),$$

  so if we take a wider confidence interval, we conclude that the coin is fair for spins also.

Confidence intervals and tests

- We can use confidence intervals (CIs) to assess the plausibility of a value θ^0 of θ:
  - If θ^0 lies inside a $(1-\alpha) \times 100\%$ CI, then we **cannot reject** the hypothesis that $\theta = \theta^0$, at **significance level** α.
  - If θ^0 lies outside a $(1-\alpha) \times 100\%$ CI, then we **reject** the hypothesis that $\theta = \theta^0$, at **significance level** α.
- The discussion of the scientific method at the start of §7 (slide 267) tells us that data cannot prove correctness of a theory (hypothesis), because we can always imagine that future data or a new experiment might undermine it, but data can falsify theory. Hence we can **reject** or **not reject (provisionally accept)** a hypothesis, but we cannot **prove** it.
- The decision to reject or not depends on the chosen significance level α: we will reject less often if α is small, since then the CI will be wider.
- If α is small and we do reject, this gives stronger evidence against θ^0.
- Use of a two-sided CI (L, U) implies that seeing either $\theta^0 < L$ or $\theta^0 > U$ would be evidence against the theory. This is true for Example 276, but in general we should consider whether to use $(-\infty, U)$ or (L, ∞) instead.

Null and alternative hypotheses

In a general testing problem we aim to use the data to decide between two hypotheses.

- The **null hypothesis** H_0, which represents the theory/model we want to test.
  - For the coin tosses, H_0 is that the coin is fair, i.e., $\mathrm{P}(\text{heads}) = \theta = \theta^0 = 1/2$.
- The **alternative hypothesis** H_1, which represents what happens if H_0 is false.
  - For the coin tosses, H_1 is that the coin is not fair, i.e., $\mathrm{P}(\text{heads}) \neq \theta^0$.
- When we decide between the hypotheses, we can make two sorts of error:

  Type I error (false positive): H_0 is true, but we wrongly reject it (and choose H_1);

  Type II error (false negative): H_1 is true, but we wrongly accept H_0.

		Decision	
		Accept H_0	Reject H_0
State of Nature	H_0 true	Correct choice (True negative)	Type I Error (False positive)
	H_1 true	Type II Error (False negative)	Correct choice (True positive)

Taxonomy of hypotheses

Definition 277. *A* **simple hypothesis** *entirely fixes the distribution of the data Y, whereas a* **composite hypothesis** *does not fix the distribution of Y.*

Example 278. *If*

$$H_0 : Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} \mathcal{N}(0,1), \quad H_1 : Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} \mathcal{N}(0,3),$$

then both hypotheses are simple.

Example 279. *If θ_0 is fixed (e.g., $\theta_0 = 1/2$) and*

$$H_0 : R \sim B(n, \theta_0), \quad H_1 : R \sim B(n, \theta), \quad \theta \in (0, \theta_0) \cup (\theta_0, 1),$$

then H_0 ('the coin is fair') is simple but H_1 ('the coin is not fair') is composite.

Example 280. *If μ, σ^2 are unknown and F is a unknown (but non-normal) distribution, and*

$$H_0 : Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2), \quad H_1 : Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} F,$$

then both H_0 ('the data are normally distributed') and H_1 ('the data are not normally distributed') are composite.

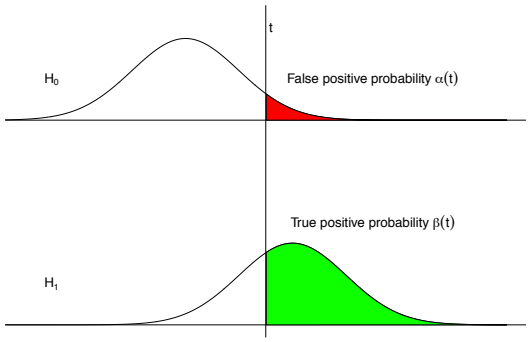
True and false positives: Example

- $H_0 : T \sim \mathcal{N}(0,1)$ and $H_1 : T \sim \mathcal{N}(\mu,1)$, with $\mu > 0$.
- Reject H_0 if $T > t$, where t is some cut-off, so we
  - reject H_0 incorrectly (false positive) with probability

$$\alpha(t) = \mathrm{P}_0(T > t) = 1 - \Phi(t) = \Phi(-t)$$

  - reject H_0 correctly (true positive) with probability

$$\beta(t) = \mathrm{P}_1(T > t) = \mathrm{P}(T - \mu > t - \mu) = 1 - \Phi(t - \mu) = \Phi(\mu - t).$$



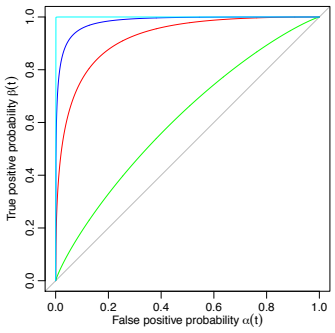
ROC curve

Definition 281. *The* **receiver operating characteristic (ROC) curve** *of a test plots* $\beta(t)$ *against* $\alpha(t)$ *as the cut-off* t *varies, i.e., it shows* $(\mathrm{P}_0(T \geq t), \mathrm{P}_1(T > t))$*, when* $t \in \mathbb{R}$.

- In the example above, we have $\alpha = \Phi(-t)$, so $t = -\Phi^{-1}(\alpha) = -z_\alpha$, so equivalently we graph

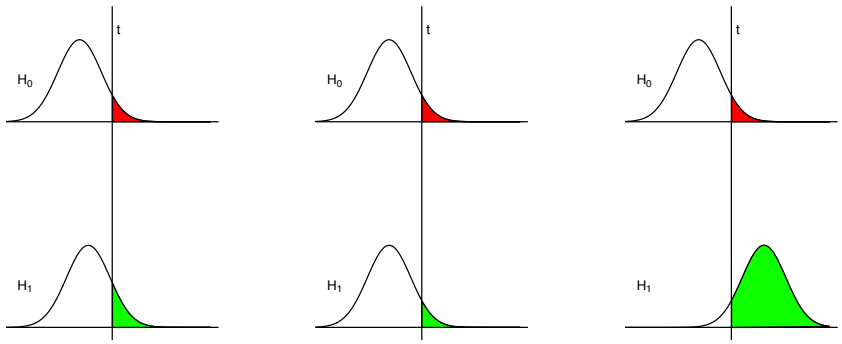
$$\beta(t) = \Phi(\mu + z_\alpha) \equiv \beta(\alpha) \text{ against } \alpha \in [0,1].$$

- Here is the ROC curve for the example above, which has $\mu = 2$ (in red). Also shown are the ROC curves for $\mu = 0, 0.4, 3, 6$. Which is which?



Example, II

- In case you need help, here are the densities for three of the cases:



Size and power

- As μ increases, it becomes easier to detect when H_0 is false, because the densities under H_0 and H_1 become more separated, and the ROC curve moves 'further north-west'.
- When H_0 and H_1 are the same, i.e., $\mu = 0$, then the curve lies on the diagonal. Then the hypotheses cannot be distinguished.
- In applications, μ is usually unknown, so we fix α (often at some conventional value,e.g., 0.05, 0.01) and then accept the resulting $\beta(\alpha)$.
- We also call (particularly in statistics books and papers)
  - the **false positive probability** the **size** α of the test, and
  - the **true positive probability** the **power** β of the test.

Definition 282. *Let $\mathrm{P}_0(\cdot)$ and $\mathrm{P}_1(\cdot)$ denote probabilities computed under null and alternative hypotheses H_0 and H_1 respectively. Then the* **size** *and* **power** *of a statistical test of H_0 against H_1 are*

$$\textit{size } \alpha = \mathrm{P}_0(\textit{reject } H_0), \quad \textit{power } \beta = \mathrm{P}_1(\textit{reject } H_0).$$

Power and confidence intervals

- If the test is based on a $(1-\alpha)\times 100\%$ CI, the size is the probability that the true value of the parameter lies outside the CI, so it is α.
- Taking a smaller value of α gives a wider interval, so it must decrease the power.
- Usually the width of the interval (L, U) satisfies

$$U - L \propto n^{-1/2},$$

  so increasing n gives a narrower interval and will increase the power of the test. This makes sense, because having more data should allow us to be more certain in our conclusions.
- Unfortunately, not all tests correspond to confidence intervals, so we need a more general approach.
- For example, checking the fit of a model is not usually possible using a confidence interval ...

Testing goodness of fit

We may want to assess whether a statistical model fits data appropriately.

Example 283. *In a legal dispute, it was claimed that the numbers below were faked:*

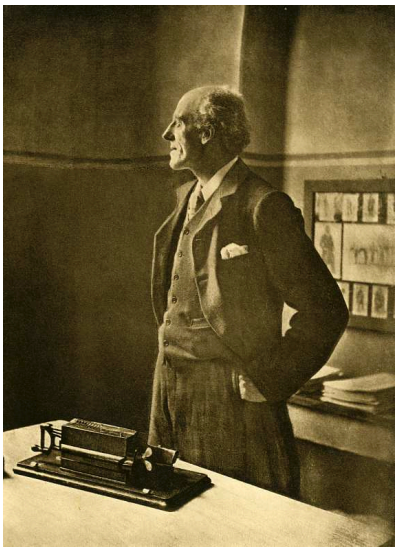
```
261 289 291 265 281 291 285 283 280 261 263 281 291 289 280
292 291 282 280 281 291 282 280 286 291 283 282 291 293 291
300 302 285 281 289 281 282 261 282 291 291 282 280 261 283
291 281 246 249 252 253 241 281 282 280 261 265 281 283 280
242 260 281 261 281 282 280 241 249 251 281 273 281 261 281
282 260 281 282 241 245 253 260 261 281 280 261 265 281 241
260 241
```

Real data could be expected to have final digits uniformly distributed on $\{0, 1, \ldots, 9\}$*, but here we have*

```
 0  1  2  3  4  5  6  7  8  9
14 42 14  9  0  6  2  0  0  5
```

How strong is the evidence that the final digits are not uniform?

Karl Pearson (1857–1936)



(Source: University College London)

Pearson statistic

Definition 284. *Let $O_1, \ldots, O_k$ be the number of observations of a random sample of size $n = n_1 + \cdots + n_k$ falling into the categories $1, \ldots, k$, whose expected numbers are $E_1, \ldots, E_k$, where $E_i > 0$. Then the* **Pearson statistic (or chi-square statistic)** *is*

$$T = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}.$$

Definition 285. *Let $Z_1, \ldots, Z_\nu \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$, then $W = Z_1^2 + \cdots + Z_\nu^2$ follows the* **chi-square distribution with ν degrees of freedom**, *whose density function is*

$$f_W(w) = \frac{1}{2^{\nu/2}\Gamma(\nu/2)} w^{\nu/2-1} e^{-w/2}, \quad w > 0,\ \nu = 1, 2, \ldots,$$

where $\Gamma(a) = \int_0^\infty u^{a-1} e^{-u}\, du$, $a > 0$, is the gamma function.

- ☐ If the joint distribution of $O_1, \ldots, O_k$ is multinomial with denominator n and probabilities $p_1 = E_1/n, \ldots, p_k = E_k/n$, then $T \dot{\sim} \chi^2_{k-1}$, the approximation being good if $k^{-1} \sum E_i \geq 5$.
- ☐ We can use T to check the agreement between the data $O_1, \ldots, O_k$ and the theoretical probabilities $p_1, \ldots, p_k$.

Pearson statistic: Rationale

- ☐ If $O_i \approx E_i$ for all i, then T will be small, otherwise it will tend to be bigger.
- ☐ If the joint distribution of $O_1, \ldots, O_k$ is multinomial with denominator n and probabilities $p_i = E_i/n$, then each $O_i \sim B(n, p_i)$, and thus

$$\mathrm{E}(O_i) = np_i = E_i, \quad \mathrm{var}(O_i) = np_i(1-p_i) = E_i(1 - E_i/n) \approx E_i,$$

thus $Z_i = (O_i - E_i)/\sqrt{E_i} \dot{\sim} \mathcal{N}(0,1)$, for large n, and we would imagine that

$$T = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} = \sum_{i=1}^{k} Z_i^2 \dot{\sim} \chi^2_k,$$

but the constraint $\sum_i O_i = n$ means that only $k-1$ of the Z_i can vary independently, thus reducing the degrees of freedom to $k-1$.

Null and alternative hypotheses for Example 283

- **Null hypothesis, H_0**: the final digits are independent and distributed according to the uniform distribution on $0, \ldots, 9$. This simple null hypothesis implies that $O_0, \ldots, O_9$ have the multinomial distribution with probabilities $p_0 = \cdots = p_9 = 0.1$, and since $\sum E_j/10 > 5$, we have

$$\mathrm{P}_0(T \leq t) \doteq \mathrm{P}(\chi^2_9 \leq t), \quad t > 0.$$

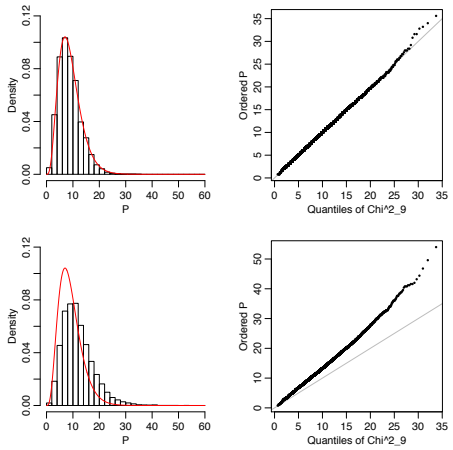
- **Alternative hypothesis, H_1**: the final digits are independent but not uniform, so $O_0, \ldots, O_9$ follow a multinomial distribution with unequal probabilities, $p_0, \ldots, p_9$. This hypothesis is composite, and the parameter $\theta \equiv (p_1, \ldots, p_9)$ is of dimension 9, as $p_0 = 1 - p_1 - \cdots - p_9$. Under this model,

$$\mathrm{P}_1(T > t) \geq \mathrm{P}(\chi^2_9 > t), \quad t > 0.$$

- Since values of T tend to be smaller under H_0 than under H_1, we should large values of T to be evidence against H_0 in favour of H_1.
- We verify this on the following slides.

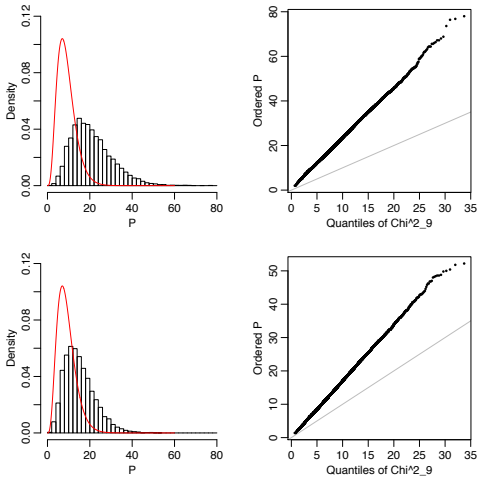
Monte Carlo simulations of T, $n = 50$

Pearson's statistics for 10,000 sets of data when testing $H_0 : p_0 = \cdots = p_9 = 0.1$, when: (a) (top) the data are generated under H_0; (b) (bottom) the data are generated with a multinomial distribution having $p_0 = p_1 = 0.15$, $p_2 = \cdots = p_9 = 0.0875$. The values of T tend to be bigger under (b). The red line shows the χ^2_9 density.



Monte Carlo simulations of T, $n = 100, 50$

Pearson's statistics for 10,000 sets of data when testing $H_0 : p_0 = \cdots = p_9 = 0.1$, when: (a) (top) the data are generated with $p_0 = p_1 = 0.15$, $p_2 = \cdots = p_9 = 0.0875$, and $n = 100$; (b) (bottom) the data are generated with $p_0 = p_1 = 0.2$, $p_2 = \cdots = p_9 = 0.075$ and $n = 50$. The red line shows the χ^2_9 density.



Example

The simulations in the previous figures show that

- ☐ under H_0, we indeed have $T \dot{\sim} \chi^2_9$, even with $n = 50$;
- ☐ under H_1, the distribution of T is shifted to the right;
- ☐ the size of the shift under H_1 will determine the power of the test, which depends on the sample size n and on the non-uniformity of $(p_0, \ldots, p_9)$.

Example 286 (Example 283, continued). *Our data*

0	1	2	3	4	5	6	7	8	9
14	42	14	9	0	6	2	0	0	5

give observed value of T equal to $t_{\text{obs}} \doteq 158$.

- ☐ *For a test of H_0 at significance level $\alpha = 0.05$, note that the $(1 - \alpha)$ quantile of the χ^2_9 distribution is 16.92. Since $t_{\text{obs}} > 16.92$, we can reject H_0 at significance level 0.05.*
- ☐ *In fact,*

$$\mathrm{P}_0(T \geq t_{\text{obs}}) \doteq \mathrm{P}(\chi^2_9 \geq 158) < 2.2 \times 10^{-16},$$

so seeing data like this would be essentially impossible under H_0. It is almost certain that the observed final digits did not come from a uniform distribution.

Evidence and P-values

A statistical hypothesis test has the following elements:

- a **null hypothesis** H_0, to be tested against an **alternative hypothesis** H_1;
- **data**, from which we compute a **test statistic** T, chosen such that large values of T provide evidence against H_0;
- the observed value of T is t_{obs}, which we compare with the **null distribution** of T, i.e., the sampling distribution of T under H_0;
- we measure the evidence against H_0 using the **P-value**

$$p_{\text{obs}} = \mathrm{P}_0(T \geq t_{\text{obs}}),$$

  where small values of p_{obs} suggest that either
  - H_0 is true but something unlikely has occurred, or
  - H_0 is false.
- If $p_{\text{obs}} < \alpha$, then we say that the test is **significant at level α** or **significant at the $\alpha \times 100\%$ level**.
- If we must make a decision, then we **reject** H_0 if $p_{\text{obs}} < \alpha$, where α is the significance level of the test, and we (provisionally) **accept** H_0 if $p_{\text{obs}} \geq \alpha$.

Examples

Example 287. *Recast Example 276 in terms of P-values.*

Example 288. *Ten new electricity meters are measured for quality control purposes, resulting in the data*

$$983 \quad 1002 \quad 998 \quad 996 \quad 1002 \quad 983 \quad 994 \quad 991 \quad 1005 \quad 986$$

Is there a systematic divergence from the standard value of 1000?

Note to Example 287

- Under H_0 we have $R \sim B(n, \theta^0)$, and therefore $R \dot{\sim} \mathcal{N}\{n\theta^0, n\theta^0(1-\theta^0)\}$ by the central limit theorem. Since values of R far from $n\theta^0$ in either direction would be evidence against H_0, this suggests taking

$$T = \{R - \mathrm{E}(R)\}^2/\mathrm{var}(R) = (R - n\theta^0)^2/\{n\theta^0(1-\theta^0\} = (R-100)^2/50,$$

  since here $n = 200$ and $\theta^0 = 1/2$ yield $\mathrm{E}(R) = 100$ and $\mathrm{var}(R) = 50$.
- Since $T = Z^2$, where $Z \sim \mathcal{N}(0,1)$ we have that $T \dot{\sim} \chi^2_1$ under H_0.
- This gives $t_{\text{obs}} = 0.5$ for the tosses, and $t_{\text{obs}} = 4.5$ for the spins, with corresponding P-values

$$\mathrm{P}_0(T \geq t_{\text{obs}}) \doteq \mathrm{P}(\chi^2_1 \geq 0.5) \doteq 0.480, \quad \mathrm{P}_0(T \geq t_{\text{obs}}) \doteq \mathrm{P}(\chi^2_1 \geq 4.5) \doteq 0.034.$$

- With $\alpha = 0.05$ we would accept H_0 for the tosses but reject it for the spins.
- With $\alpha = 0.01$ we would accept H_0 for both tosses and spins.

Note to Example 288

☐ We assume that $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$, with σ^2 unknown. We take

$$H_0 : \mu = \mu_0 = 1000, \quad H_1 : \mu \neq 1000.$$

☐ We know from Theorem 274 that under H_0,

$$Z = \frac{\overline{Y} - \mu_0}{\sqrt{S^2/n}} \sim t_{n-1}.$$

Here the alternative hypothesis H_1 is two-sided, i.e., we will reject if either $\overline{Y}$ is much larger or much smaller than μ_0, so we should take

$$T = \left| \frac{\overline{Y} - \mu_0}{\sqrt{S^2/n}} \right| = |Z|,$$

and for a test at significance level $\alpha = 0.05$ we therefore need to choose t_α such that

$$\alpha = \mathrm{P}_0(T > t_\alpha) = 1 - \mathrm{P}_0\left(-t_\alpha \leq Z \leq t_\alpha\right).$$

But $Z \sim t_{n-1}$ is a pivot under H_0, so $1 - \mathrm{P}_0\left(-t_\alpha \leq Z \leq t_\alpha\right) = 2\mathrm{P}_0(Z \leq -t_\alpha)$, and this implies that $t_\alpha = -t_{n-1}(\alpha/2)$. With $\alpha = 0.025$ and $n = 10$, we have $t_9(0.025) = -2.262$ from the tables, or R, as `qt(0.025, df=9)`.

☐ For the data above, $\overline{y} = 994$ and

$$s^2 = \frac{1}{9}\sum_{i=1}^{n}(y_i - \overline{y})^2 = 64.88.$$

☐ Now $t_{\text{obs}} = |(994 - 1000)/\sqrt{64.88/10}| = |-2.35| = 2.35 > t_\alpha = 2.262$, so we reject H_0 at level $\alpha = 5\%$.

☐ Alternatively we can compute the 95% confidence interval based on Z, which is $(988.238, 999.762)$. Since this does not contain μ_0, H_0 is rejected at the 5% level.

☐ If instead the alternative hypothesis is $H_1 : \mu > 1000$, then we take Z as the test statistic, since we are likely to have positive Z under H_1. In this case we need to choose t_α such that

$$\alpha = \mathrm{P}_0(Z > t_\alpha) = \mathrm{P}_0\left\{\frac{\overline{Y} - \mu_0}{\sqrt{S^2/n}} > t_\alpha\right\}.$$

Since $Z \sim t_{n-1}$, we have that $t_\alpha = t_9(0.95) = 1.833$, and since $z_{\text{obs}} = -2.35 < 1.833$, we cannot reject the null hypothesis at the 5% level. Indeed, having $\overline{y} = 994$ suggests that it is not true that $\mu > \mu_0$.

☐ If the alternative hypothesis is $H_1 : \mu < 1000$, then we take $T = -Z$ as the test statistic, since we are likely to have negative Z under H_1. In this case we need to choose t_α such that

$$\alpha = \mathrm{P}_0(-Z > t_\alpha) = \mathrm{P}_0\left\{\frac{\overline{Y} - \mu_0}{\sqrt{S^2/n}} < -t_\alpha\right\} = \mathrm{P}_0\left\{\frac{\overline{Y} - \mu_0}{\sqrt{S^2/n}} < t_{n-1}(\alpha)\right\},$$

implying that $t_\alpha = -t_{n-1}(\alpha) = t_{n-1}(1-\alpha)$. With $\alpha = 0.05$, we therefore have $t_\alpha = 1.833$, and since $-z_{\text{obs}} = 2.35 > t_\alpha = 1.833$, we reject the null hypothesis at the 5% level. Having $\overline{y} = 994 < \mu_0$ suggests that maybe $\mu < \mu_0$.

Decision procedures and measures of evidence

We can use a test of H_0 in two related ways:

- as a **decision procedure**, where we
  - choose a level α at which we want to test H_0, and then
  - reject H_0 (i.e., choose H_1) if the P-value is less than α, or
  - do not reject H_0 if the P-value is greater than α.
- as a **measure of evidence** against H_0, with
  - small values of p_{obs} suggesting stronger evidence against H_0, but
  - H_1 need not be explicit, though the type of departure from H_0 that we seek is implicit in the choice of T.
- Knowing the exact value of p_{obs} is more useful than knowing that H_0 has been rejected, so the measure of evidence is more informative.
- The strength of the evidence contained in a P-value can be summarised as follows:

α	Evidence against H_0
0.05	Weak
0.01	Positive
0.001	Strong
0.0001	Very strong

Choice of α

- As with CIs, conventional values are often used, such as $\alpha = 0.05, 0.01, 0.001$.
- The most common value is $\alpha = 0.05$, which corresponds to a Type I error probability of 5%, i.e., H_0 will be rejected once in every 20 tests, even when it is true.
- When many tests are performed, using large α can give many **false positives**, i.e., significant tests for which in fact H_0 is true.
- Consider a microarray experiment, where we test 1000 genes at significance level α, to see which genes influence some disease. If only 100 genes have effects, we can write

$$\mathrm{P}(H_0) = 900/1000, \quad \mathrm{P}(H_1) = 100/1000, \quad \mathrm{P}(S \mid H_0) = \alpha, \quad \mathrm{P}(S \mid H_1) = \beta,$$

where α is the size of the test, $\beta > \alpha$ is its power, and S denotes the event that the test is significant at level α. Bayes' theorem gives

$$\mathrm{P}(H_0 \mid S) = \frac{\mathrm{P}(H_0)\mathrm{P}(S \mid H_0)}{\mathrm{P}(H_0)\mathrm{P}(S \mid H_0) + \mathrm{P}(H_1)\mathrm{P}(S \mid H_1)} = \frac{0.9\alpha}{0.9\alpha + 0.1\beta}.$$

Hence with $\alpha = 0.05$, $\beta = 0.8$, say, $\mathrm{P}(H_0 \mid S) \doteq 0.36$, so over one-third of significant tests will not be interesting. If instead we set $\alpha = 0.005$, we have $\mathrm{P}(H_0 \mid S) \doteq 0.05$, which is more reasonable.

8.5 Comparison of Tests

Types of test

There are many different tests for different hypotheses. Two important classes of tests are:

- ☐ **parametric tests**, which are based on a parametric statistical model, such as $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$, and $H_0 : \mu = 0$;
- ☐ **nonparametric tests**, which are based on a more general statistical model, such as $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} f$, et $H_0 : \mathrm{P}(Y > 0) = \mathrm{P}(Y < 0) = 1/2$, i.e., the median of f is at $y = 0$

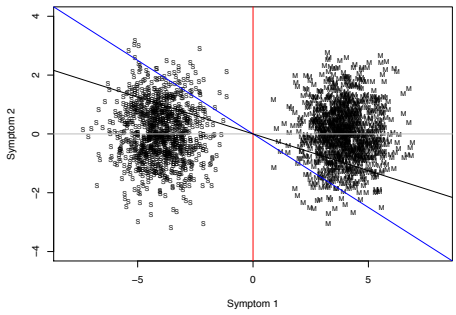
The main advantage of a parametric test is the possibility of finding a (nearly-)optimal test, if the underlying assumptions are correct, though such a test could perform badly in the presence of outliers. A nonparametric test is often more robust, but it will suffer a loss of power compared to a parametric test, used appropriately.

Medical analogy

We diagnose an illness based on symptoms presented by a patient:

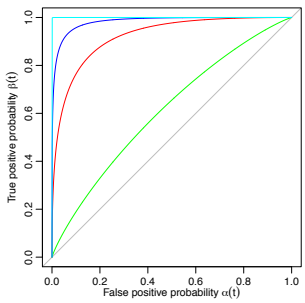
		Decision	
		Healthy	Diseased
Patient	Healthy	True negative	False positive
	Diseased	False negative	True positive

In the graphic below, Symptom 1 gives perfect diagnoses, but Symptom 2 is useless. Think how the probability of a correct diagnosis varies as the different lines move parallel to their slopes.



ROC curve, II

- We previously met the ROC curve as a summary of the properties of a test.
- A good test will have an ROC curve lying as close to the upper left corner as possible.
- A useless test has an ROC curve lying on (or close to) the diagonal.
- This suggests that if we have a choice of tests, we should choose one whose ROC curve is as close to the north-west as possible, i.e., we should choose the test that maximises the power for a given size.
- This leads us to the **Neyman–Pearson lemma**, which says how to do this (in ideal circumstances).



Most powerful tests

- We aim to choose our test statistic T to maximise the power of the test for a given size.
- A decision procedure corresponds to partitioning the sample space Ω containing the data Y into a **rejection region**, $\mathcal{Y}$, and its complement, $\overline{\mathcal{Y}}$, with

$$Y \in \mathcal{Y} \Rightarrow \text{Reject } H_0, \qquad Y \in \overline{\mathcal{Y}} \Rightarrow \text{Accept } H_0.$$

- In Example 287, $\mathcal{Y} = \{(y_1, \ldots, y_n) : |\sum y_j - 100|/50 > 1.96\}$.
- We aim to choose $\mathcal{Y}$ such that $\mathrm{P}_1(Y \in \mathcal{Y})$ is the largest possible such that $\mathrm{P}_0(Y \in \mathcal{Y}) = \alpha$.

Lemma 289 (Neyman–Pearson). *Let $f_0(y)$, $f_1(y)$ be the densities of Y under simple null and alternative hypotheses. Then if it exists, the set*

$$\mathcal{Y}_\alpha = \{y \in \Omega : f_1(y)/f_0(y) > t\}$$

such that $\mathrm{P}_0(Y \in \mathcal{Y}_\alpha) = \alpha$ maximises $\mathrm{P}_1(Y \in \mathcal{Y}_\alpha)$, amongst all the $\mathcal{Y}'$ such that $\mathrm{P}_0(Y \in \mathcal{Y}') \leq \alpha$. Thus to maximise the power of a given threshold, we must base the decision on $\mathcal{Y}_\alpha$.

Note to Lemma 289

Suppose that a region $\mathcal{Y}_\alpha$ such that $\mathrm{P}_0(Y \in \mathcal{Y}_\alpha) = \alpha$ does exist and let $\mathcal{Y}'$ be any other critical region of size α or less. Then for any density f,

$$\int_{\mathcal{Y}_\alpha} f(y)\,dy - \int_{\mathcal{Y}'} f(y)\,dy, \tag{7}$$

equals

$$\int_{\mathcal{Y}_\alpha \cap \mathcal{Y}'} f(y)\,dy + \int_{\mathcal{Y}_\alpha \cap \overline{\mathcal{Y}'}} f(y)\,dy - \int_{\mathcal{Y}' \cap \mathcal{Y}_\alpha} f(y)\,dy - \int_{\mathcal{Y}' \cap \overline{\mathcal{Y}_\alpha}} f(y)\,dy,$$

where $\overline{\mathcal{Y}_\alpha}$ is the complement of $\mathcal{Y}_\alpha$ in the sample space, and this is

$$\int_{\mathcal{Y}_\alpha \cap \overline{\mathcal{Y}'}} f(y)\,dy - \int_{\mathcal{Y}' \cap \overline{\mathcal{Y}_\alpha}} f(y)\,dy. \tag{8}$$

If $f = f_0$, (7) and hence (8) are non-negative, because $\mathcal{Y}'$ has size at most that of $\mathcal{Y}_\alpha$. Suppose that $f = f_1$. If $y \in \overline{\mathcal{Y}_\alpha}$, then $t_\alpha f_0(y) > f_1(y)$, while $f_1(y) \geq t_\alpha f_0(y)$ if $y \in \mathcal{Y}_\alpha$. Hence when $f = f_1$, (8) is no smaller than

$$t_\alpha \left\{ \int_{\mathcal{Y}_\alpha \cap \overline{\mathcal{Y}'}} f_0(y)\,dy - \int_{\mathcal{Y}' \cap \overline{\mathcal{Y}_\alpha}} f_0(y)\,dy \right\} \geq 0.$$

Thus the power of $\mathcal{Y}_\alpha$ is at least that of $\mathcal{Y}'$, and the result is established.

Example

Example 290. *(a) Construct an optimal test for the hypothesis $H_0 : \theta = 1/2$ in Example 276, with $\alpha = 0.05$.*
(b) Do you think that $\theta = 1/2$ for spins?

Note to Example 290

- The joint density of n independent Bernoulli variables can be written as

$$f(y) = \theta^r(1-\theta)^{n-r}, \quad 0 < \theta < 1, \quad r = \sum y_j,$$

and H_0 imposes $\theta = 1/2$. Thus for any fixed θ we have

$$\frac{f_1(y)}{f_0(y)} = \frac{\theta^r(1-\theta)^{n-r}}{(1/2)^r(1-1/2)^{n-r}} = \{2(1-\theta)\}^n\{\theta/(1-\theta)\}^r,$$

which is increasing in r if $\theta > 1/2$ and is decreasing in r if $\theta < 1/2$. Hence if $\theta > 1/2$ we must take

$$\mathcal{Y}_1 = \{y_1, \ldots, y_n : \sum y_j \geq r_1\}$$

for some r_1, and if $\theta < 1/2$ we must take

$$\mathcal{Y}_2 = \{y_1, \ldots, y_n : \sum y_j \leq r_2\}$$

for some r_2. So if we want to test H_0 against (say) $H_1 : \theta = 0.6$, we take $\mathcal{Y}_1$, and if we want to test H_0 against (say) $H_1 : \theta = 0.4$, we take $\mathcal{Y}_2$.

- Suppose that we take $H_1 : \theta = 0.6$. Then we need to choose r_1 such that

$$\alpha = \mathrm{P}_0(Y \in \mathcal{Y}_1) = \mathrm{P}_0(R \geq r_1) = \mathrm{P}_0\left\{\frac{R - n/2}{\sqrt{n/4}} \geq \frac{r_1 - n/2}{\sqrt{n/4}}\right\} \doteq 1 - \Phi\left(\frac{r_1 - n/2}{\sqrt{n/4}}\right)$$

and this implies that $r_1 \doteq n/2 + \sqrt{n}z_{1-\alpha}/2$. With $n = 200$ and $\alpha = 0.05$ this is $r_1 \doteq 111.6$. Since we observed $R = 115 > r_1$, we reject H_0 at the 5% significance level, and conclude that the coin is biased upwards (but not downwards).

- Since the result does not depend on the value of θ chosen, provided $\theta > 0.5$, we would also reject against any other H_1 setting $\theta > 1/2$.
- A similar computation gives $r_2 = 88.37$.
- If we are not sure of the value of θ, then we take a region of the form $\mathcal{Y}_1 \cup \mathcal{Y}_2$. But in order for it to have overall size α, we take $\alpha/2$ for each of the regions, giving $r_1 = 113.86$ and $r_2 = 86.14$. Since $Y \in \mathcal{Y}_1 \cup \mathcal{Y}_2$, we still reject H_0 at the 5% significance level, and conclude that the coin is biased, without being sure in which direction it is biased.

Power and distance

- A canonical example is where $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$, and

$$H_0 : \mu = \mu_0, \qquad H_1 : \mu = \mu_1.$$

- If σ^2 is known, then the Neyman–Pearson lemma can be applied, and we find that the most powerful test is based on $\overline{Y}$ and its power is $\Phi(z_\alpha + \delta)$, where $\Phi(z_\alpha) = \alpha$, and

$$\delta = n^{1/2} \frac{|\mu_1 - \mu_0|}{\sigma}$$

  is the **standardized distance** between the models.
- We see that
  - the power increases if n increases, or if $|\mu_1 - \mu_0|$ increases, since in either case the difference between the hypotheses is easier to detect,
  - the power decreases if σ increases, since then the data become noisier,
  - if $\delta = 0$, then the power equals the size, because the two hypotheses are the same, and therefore $\mathrm{P}_0(\cdot) = \mathrm{P}_1(\cdot)$.
- Many other situations are analogous to this, with power depending on generalised versions of δ.

Summary

- We have considered the situation where we have to make a binary choice between
  - the null hypothesis, H_0, against which we want to test
  - the alternative hypothesis, H_1,

  using a test statistic T whose observed value is t_{obs}, computing the P-value,

$$p_{\text{obs}} = \mathrm{P}_0(T \geq t_{\text{obs}}),$$

  which is computed assuming that H_0 is true.
- We can consider p_{obs} as a measure of the evidence in the data against H_0.
- For a test with significance level α, we reject H_0 and choose H_1 if $p_{\text{obs}} < \alpha$.
- We must accept that we can make mistakes:

		Decision	
		Accept H_0	Reject H_0
State of Nature	H_0 true	Good choice	Type I Error
	H_1 true	Type II Error	Good choice

- If we try to minimise the probability of Type II error (i.e., maximise power) for a given probability of Type I error (fixed size), we can construct an optimal test, but this is only possible in simple cases. Otherwise we usually have to compare tests numerically.

9 Likelihood

slide 387

9.1 Motivation

slide 388

Motivation

Likelihood is one of the basic ideas of statistical inference and modelling. It gives a general and powerful framework for dealing with all kinds of applications, in particular for

- ☐ finding estimators with the smallest variances in large samples; and
- ☐ constructing powerful tests.

Illustration

- ☐ When we toss a coin, small asymmetries influence the probability of obtaining heads, which is not necessarily $1/2$. If $Y_1, \ldots, Y_n$ denote the results of independent Bernoulli trials, then we can write

$$\mathrm{P}(Y_j = 1) = \theta, \quad \mathrm{P}(Y_j = 0) = 1 - \theta, \quad 0 \leq \theta \leq 1, \quad j = 1, \ldots, n.$$

- ☐ Below is such a sequence for a 5Fr coin with $n = 10$:

$$1 \quad 1 \quad 1 \quad 1 \quad 1 \quad 0 \quad 1 \quad 1 \quad 1 \quad 1$$

Which values of θ seem to you the most and least credible:

$$\theta = 0, \quad \theta = 0.3, \quad \theta = 0.9, \quad \theta = 0.99?$$

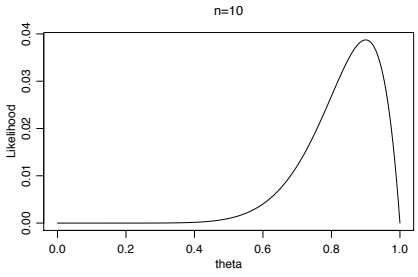
- ☐ How can we find the most plausible θ value(s)?

Basic Idea

For a value of θ which is not very credible, the density of the data will be smaller: the higher the density, the more credible the corresponding θ. Since the $y_1, \ldots, y_{10}$ result from independent trials, we have

$$\begin{aligned} f(y_1, \ldots, y_{10}; \theta) &= \prod_{j=1}^{10} f(y_j; \theta) = f(y_1; \theta) \times \cdots \times f(y_{10}; \theta) = \theta^5 \times (1 - \theta) \times \theta^4 \\ &= \theta^9 (1 - \theta), \end{aligned}$$

which we will consider as a function of θ for $0 \leq \theta \leq 1$, called the **likelihood** $L(\theta)$.



Relative likelihood

- To compare values of θ, we only need to consider the ratio of the corresponding values of $L(\theta)$:

$$\frac{L(\theta_1)}{L(\theta_2)} = \frac{f(y_1, \ldots, y_{10}; \theta_1)}{f(y_1, \ldots, y_{10}; \theta_2)} = \frac{\theta_1^9(1-\theta_1)}{\theta_2^9(1-\theta_2)} = c$$

  implies that θ_1 is c times more plausible than θ_2.
- The most plausible value is $\widehat{\theta}$, which satisfies

$$L(\widehat{\theta}) \geq L(\theta), \quad 0 \leq \theta \leq 1;$$

  $\widehat{\theta}$ is called the **maximum likelihood estimate**.
- To find $\widehat{\theta}$, we can equivalently maximise the **log likelihood**

$$\ell(\theta) = \log L(\theta).$$

- The **relative likelihood** $RL(\theta) = L(\theta)/L(\widehat{\theta})$ gives the plausibility of θ with respect to $\widehat{\theta}$.

Example

Example 291. *Find $\widehat{\theta}$ and $RL(\theta)$ for a sequence of independent Bernoulli trials.*

The following graph represents $RL(\theta)$, for $n = 10, 20, 100$ and the sequence

```
1 1 1 1 1 0 1 1 1 1 0 1 1 1 1 0 1 0 1 1
1 1 1 1 1 1 0 1 0 1 0 0 1 1 0 1 1 1 0 1
1 1 1 0 0 1 0 1 1 1 1 1 0 0 1 1 1 1 1 1
1 0 1 0 1 1 0 1 1 1 0 0 1 1 1 0 1 1 1 1
1 0 0 0 0 1 0 1 0 0 1 0 0 1 1 1 1 1 1 0
```

- As n increases, $RL(\theta)$ gets closer to $\widehat{\theta}$: values of θ which are far away from $\widehat{\theta}$ become less credible with respect to $\widehat{\theta}$.
- This suggests that we could construct a CI by taking the set

$$\{\theta : RL(\theta) \geq c\},$$

  for some c. Later we will see how to choose c.

Note to Example 291

The likelihood is

$$L(\theta) = f(y;\theta) = \prod_{j=1}^{n} f(y_j;\theta) = \prod_{j=1}^{n} \theta^{y_j}(1-\theta)^{1-y_j} = \theta^s(1-\theta)^{n-s}, \quad 0 \le \theta \le 1,$$

where $s = \sum y_j$ and we have used the fact that the observations are independent. Therefore

$$\ell(\theta) = s\log\theta + (n-s)\log(1-\theta), \quad 0 \le \theta \le 1.$$

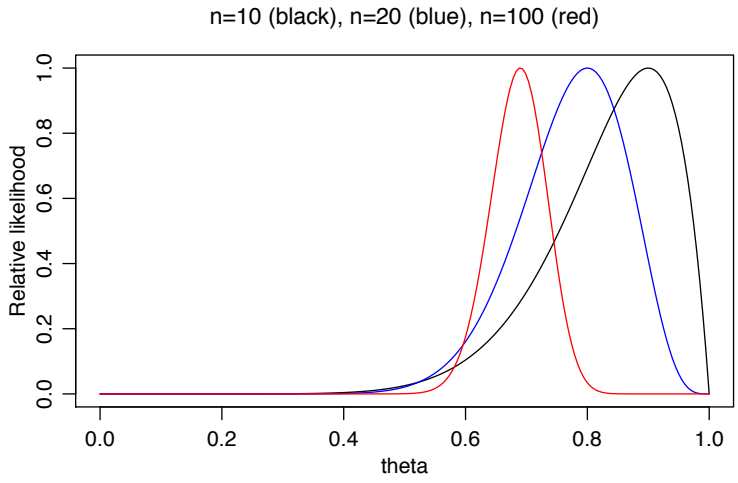
Differentiation of this yields

$$\frac{d\ell(\theta)}{d\theta} = \frac{s}{\theta} - \frac{n-s}{1-\theta}, \quad \frac{d^2\ell(\theta)}{d\theta^2} = -\frac{s}{\theta^2} - \frac{n-s}{(1-\theta)^2}.$$

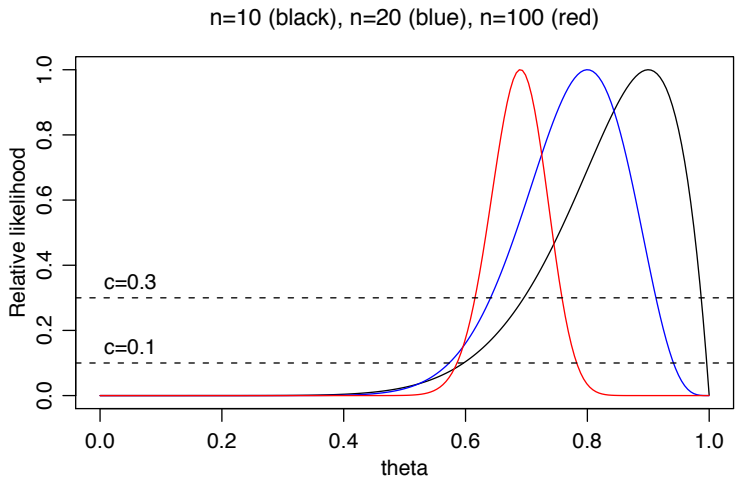
Setting $d\ell(\theta)/d\theta = 0$ gives just one solution, $\widehat{\theta} = s/n = \overline{y}$, and since the second derivative is always negative, this is clearly the maximum. Therefore

$$RL(\theta) = \frac{L(\theta)}{L(\widehat{\theta})} = \left(\frac{\theta}{\widehat{\theta}}\right)^s \left(\frac{1-\theta}{1-\widehat{\theta}}\right)^{n-s}, \qquad 0 \le \theta \le 1.$$

Bernoulli sequence



Bernoulli sequence



9.2 Scalar Parameter slide 396

Likelihood

Definition 292. *Let y be a set of data, whose joint probability density $f(y;\theta)$ depends on a parameter θ, then the* **likelihood** *and the* **log likelihood** *are*

$$L(\theta) = f(y;\theta), \quad \ell(\theta) = \log L(\theta),$$

considered a function of θ.

If $y = (y_1, \ldots, y_n)$ is a realisation of the independent random variables of $Y_1, \ldots, Y_n$, then

$$L(\theta) = f(y;\theta) = \prod_{j=1}^{n} f(y_j;\theta), \quad \ell(\theta) = \sum_{j=1}^{n} \log f(y_j;\theta),$$

where $f(y_j;\theta)$ represents the density of one of the y_j.

Maximum likelihood estimation

Definition 293. *The* **maximum likelihood estimate** $\widehat{\theta}$ *satisfies*

$$L(\widehat{\theta}) \geq L(\theta) \quad \text{for all } \theta,$$

which is equivalent to $\ell(\widehat{\theta}) \geq \ell(\theta)$*, since* $L(\theta)$ *and* $\ell(\theta)$ *have their maxima at the same value of* θ*. The corresponding random variable is called the* **maximum likelihood estimator (MLE)**.

- Often $\widehat{\theta}$ satisfies
$$\frac{d\ell(\widehat{\theta})}{d\theta} = 0, \quad \frac{d^2\ell(\widehat{\theta})}{d\theta^2} < 0.$$
In this course we will suppose that the first of these equations has only one solution (not always the case in reality).
- In realistic cases we use numerical algorithms to obtain $\widehat{\theta}$ and $d^2\ell(\widehat{\theta})/d\theta^2$.

Information

Definition 294. *The* **observed information** $J(\theta)$ *and the* **expected information (or Fisher information)** $I(\theta)$ *are*

$$J(\theta) = -\frac{d^2\ell(\theta)}{d\theta^2}, \quad I(\theta) = \mathrm{E}\{J(\theta)\} = \mathrm{E}\left\{-\frac{d^2\ell(\theta)}{d\theta^2}\right\}.$$

They measure the curvature of $-\ell(\theta)$: the larger $J(\theta)$ and $I(\theta)$, the more concentrated $\ell(\theta)$ and $L(\theta)$ are.

Example 295. *If* $y_1, \ldots, y_n \overset{\text{iid}}{\sim} \text{Bernoulli}(\theta)$*, calculate* $L(\theta)$*,* $\ell(\theta)$*,* $\widehat{\theta}$*,* $\text{var}(\widehat{\theta})$*,* $J(\theta)$ *and* $I(\theta)$*.*

Note to Example 295

We saw in Example 291 that

$$L(\theta) = \theta^s(1-\theta)^{n-s}, \quad \ell(\theta) = s\log\theta + (n-s)\log(1-\theta), \quad 0 \leq \theta \leq 1,$$

that the MLE is $\widehat{\theta} = s/n = \overline{y}$, and clearly

$$J(\theta) = -\frac{d^2\ell(\theta)}{d\theta^2} = \frac{s}{\theta^2} + \frac{n-s}{(1-\theta)^2}.$$

Now treating $\widehat{\theta}$ as a random variable, $\widehat{\theta} = S/n$, where $S \sim B(n, \theta)$, we see that since $\mathrm{E}(S) = n\theta$ and $\text{var}(S) = n\theta(1-\theta)$, we have after a little algebra that

$$\text{var}(\widehat{\theta}) = \frac{\theta(1-\theta)}{n}, \quad I(\theta) = \mathrm{E}\{J(\theta)\} = \frac{n}{\theta(1-\theta)}, \quad 0 < \theta < 1.$$

Note that $\text{var}(\widehat{\theta}) = 1/I(\theta)$.

Limit distribution of the MLE

Theorem 296. *Let $Y_1, \ldots, Y_n$ be a random sample from a parametric density $f(y;\theta)$, and let $\widehat{\theta}$ be the MLE of θ. If f satisfies* **regularity conditions** *(see below), then*

$$J(\widehat{\theta})^{1/2}(\widehat{\theta} - \theta) \xrightarrow{D} \mathcal{N}(0,1), \quad n \to \infty.$$

Thus for large n,

$$\widehat{\theta} \quad \dot{\sim} \quad \mathcal{N}\left\{\theta, J(\widehat{\theta})^{-1}\right\},$$

and a two-sided equi-tailed CI for θ with approximate level $(1-\alpha)$ is

$$\mathcal{I}^{\widehat{\theta}}_{1-\alpha} = (L, U) = (\widehat{\theta} - J(\widehat{\theta})^{-1/2} z_{1-\alpha/2}, \widehat{\theta} + J(\widehat{\theta})^{-1/2} z_{1-\alpha/2}).$$

We can show that for large n (and a regular model) no estimator has a smaller variance than $\widehat{\theta}$, which implies that the CIs $\mathcal{I}^{\widehat{\theta}}_{1-\alpha}$ are as narrow as possible.

Example 297. *Find the 95% CI for the coin data with $n = 10$, 20, 100.*

n	Tails	$\widehat{\theta}$	$J(\widehat{\theta})$	$\mathcal{I}^{\widehat{\theta}}_{0.95}$	$\mathcal{I}^{W}_{0.95}$
10	9	0.9	111.1	$(0.72, 1.08)$	$(0.63, 0.99)$
20	16	0.8	125.0	$(0.62, 0.98)$	$(0.59, 0.94)$
100	69	0.69	467.5	$(0.60, 0.78)$	$(0.60, 0.78)$

Likelihood ratio statistic

Sometimes a CI based on the normal limit distribution of $\widehat{\theta}$ is unreasonable. It is then better to use $\ell(\theta)$ itself.

Definition 298. *Let $\ell(\theta)$ be the log likelihood for a scalar parameter θ, whose MLE is $\widehat{\theta}$. Then the* **likelihood ratio statistic** *is*

$$W(\theta) = 2\left\{\ell(\widehat{\theta}) - \ell(\theta)\right\}.$$

Theorem 299. *If θ^0 is the value of θ that generated the data, then under the regularity conditions giving $\widehat{\theta}$ a normal limit distribution,*

$$W(\theta^0) \xrightarrow{D} \chi^2_1, \quad n \to \infty.$$

Hence $W(\theta^0) \dot{\sim} \chi^2_1$ for large n.

Example 300. *Find $W(\theta)$ when $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} \text{Bernoulli}(\theta^0)$.*

Note to Example 300

Since

$$\ell(\theta) = s \log \theta + (n - s) \log(1 - \theta), \quad 0 \leq \theta \leq 1,$$

and $\widehat{\theta} = s/n = \overline{y}$, we have

$$W(\theta) = 2 \left[n\widehat{\theta} \log(\widehat{\theta}/\theta) + n(1 - \widehat{\theta}) \log\{(1 - \widehat{\theta})/(1 - \theta)\} \right],$$

and if we write $\widehat{\theta} = \theta + n^{-1/2} a(\theta) Z$, where $a^2(\theta) = \theta(1 - \theta)$ and $Z \xrightarrow{D} \mathcal{N}(0, 1)$, we end up after a Taylor series or two with

$$W(\theta) \doteq Z^2 \xrightarrow{D} \chi^2_1.$$

Implications of Theorem 299

- Suppose we want to test the hypothesis $H_0 : \theta = \theta^0$, where θ^0 is fixed. If H_0 is true, the theorem implies that $W(\theta^0) \stackrel{\cdot}{\sim} \chi^2_1$. The larger $W(\theta^0)$ is, the more we doubt H_0. Thus we can take $W(\theta^0)$ as a test statistic, whose observed value is w_{obs}, and with

  $$p_{\text{obs}} = \mathrm{P}\left\{W(\theta^0) \geq w_{\text{obs}}\right\} \doteq \mathrm{P}\left\{\chi^2_1 \geq w_{\text{obs}}\right\}$$

  as significance level. The smaller p_{obs} is, the more we doubt H_0.
- Let $\chi^2_\nu(1 - \alpha)$ be the $(1 - \alpha)$ quantile of the χ^2_ν distribution. Theorem 299 implies that a CI for θ^0 at the $(1 - \alpha)$ level is the set

  $$\begin{aligned} \mathcal{I}^W_{1-\alpha} = \left\{\theta : W(\theta) \leq \chi^2_1(1 - \alpha)\right\} &= \left\{\theta : 2\left\{\ell(\widehat{\theta}) - \ell(\theta)\right\} \leq \chi^2_1(1 - \alpha)\right\} \\ &= \left\{\theta : \ell(\theta) \geq \ell(\widehat{\theta}) - \tfrac{1}{2}\chi^2_1(1 - \alpha)\right\}. \end{aligned}$$

- With $1 - \alpha = 0.95$ we have $\chi^2_1(0.95) = 3.84$. Thus the 95% CI for a scalar θ contains all θ such that $\ell(\theta) \geq \ell(\widehat{\theta}) - 1.92$. In this case we have

  $$RL(\theta) = L(\theta)/L(\widehat{\theta}) = \exp\{\ell(\theta) - \ell(\widehat{\theta})\} \geq \exp(-1.92) \approx 0.15;$$

  compare with slide 395.