Lecture Notes — Probability Theory

Manuel Cabral Morais

Department of Mathematics

Instituto Superior Técnico

Lisbon, Sep. 2009/10 — Jan. 2010/11 (Revised in Jul./Dec. 2014)

Contents

0. Warm up **1**
0.1 Historical note 1
0.2 (Symmetric) random walk 2

1 Probability spaces **12**
1.1 Random experiments 16
1.2 Events and classes of sets 19
1.3 Probabilities and probability functions 31
1.4 Distribution functions; discrete, absolutely continuous and mixed probabilities 39
1.5 Conditional probability 50

2 Random variables **56**
2.1 Fundamentals 56
2.2 Combining random variables 63
2.3 Distributions and distribution functions 66
2.4 Key r.v. and random vectors and distributions 70
2.4.1 Discrete r.v. and random vectors 70
2.4.2 Absolutely continuous r.v. and random vectors 75
2.5 Transformation theory 82
2.5.1 Transformations of r.v., general case 82
2.5.2 Transformations of discrete r.v. 84
2.5.3 Transformations of absolutely continuous r.v. 86
2.5.4 Transformations of random vectors, general case 92
2.5.5 Transformations of discrete random vectors 92
2.5.6 Transformations of absolutely continuous random vectors 98
2.5.7 Random variables with prescribed distributions 105

3 Independence **111**
3.1 Fundamentals . . . 111
3.2 Independent r.v. . . . 116
3.3 Functions of independent r.v. . . . 121
3.4 Order statistics . . . 126
3.5 Constructing independent r.v. . . . 130
3.6 Bernoulli process . . . 131
3.7 Poisson process . . . 136
3.8 Generalizations of the Poisson process . . . 143

4 Expectation **147**
4.1 Definition and fundamental properties . . . 148
4.1.1 Simple r.v. . . . 148
4.1.2 Non negative r.v. . . . 152
4.1.3 Integrable r.v. . . . 157
4.1.4 Complex r.v. . . . 159
4.2 Integrals with respect to distribution functions . . . 160
4.2.1 On integration . . . 160
4.2.2 Generalities . . . 163
4.2.3 Discrete distribution functions . . . 165
4.2.4 Absolutely continuous distribution functions . . . 165
4.2.5 Mixed distribution functions . . . 166
4.3 Computation of expectations . . . 167
4.3.1 Non negative r.v. . . . 167
4.3.2 Integrable r.v. . . . 168
4.3.3 Mixed r.v. . . . 169
4.3.4 Functions of r.v. . . . 171
4.3.5 Functions of random vectors . . . 172
4.3.6 Functions of independent r.v. . . . 173
4.3.7 Sum of independent r.v. . . . 174
4.4 L^p spaces . . . 176
4.5 Key inequalities . . . 177
4.5.1 Young's inequality . . . 178
4.5.2 Hölder's moment inequality . . . 179
4.5.3 Cauchy-Schwarz's moment inequality . . . 181
4.5.4 Lyapunov's moment inequality . . . 182

4.5.5 Minkowski's moment inequality . . . 183
4.5.6 Jensen's moment inequality . . . 184
4.5.7 Chebyshev's inequality . . . 186
4.6 Moments . . . 190
4.6.1 Moments of r.v. . . . 190
4.6.2 Variance and standard deviation . . . 192
4.6.3 Skewness and kurtosis . . . 194
4.6.4 Covariance . . . 196
4.6.5 Correlation . . . 198
4.6.6 Moments of random vectors . . . 202
4.6.7 Multivariate normal distributions . . . 203
4.6.8 Multinomial distributions . . . 216

5 Convergence concepts and classical limit theorems 224
5.1 Modes of convergence . . . 225
5.1.1 Convergence of r.v. as functions on Ω . . . 225
5.1.2 Convergence in distribution . . . 232
5.1.3 Alternative criteria . . . 238
5.2 Relationships among the modes of convergence . . . 242
5.2.1 Implications always valid . . . 242
5.2.2 Counterexamples . . . 244
5.2.3 Implications of restricted validity . . . 247
5.3 Convergence under transformations . . . 249
5.3.1 Continuous mappings . . . 249
5.3.2 Algebraic operations . . . 249
5.4 Convergence of random vectors . . . 256
5.5 Limit theorems for Bernoulli summands . . . 259
5.5.1 Laws of large numbers for Bernoulli summands . . . 259
5.5.2 Central limit theorems for Bernoulli summands . . . 262
5.5.3 The Poisson limit theorem . . . 264
5.6 Weak law of large numbers . . . 267
5.7 Strong law of large numbers . . . 273
5.8 Characteristic functions . . . 275
5.9 The Central Limit Theorem . . . 283
5.10 The law of the iterated logarithm . . . 287
5.11 Applications of the limit theorems . . . 289

Warm up

0.1 Historical note

Mathematical probability has its origins in games of chance [...]. Early calculations involving dice were included in a well-known and widely distributed poem entitled De Vetula.[1] *Dice and cards continued as the main vessels of gambling in the 15th. and 16th. centuries [...]. [...] (G. Cardano) went so far as to write a book,* On games of chance*, sometime shortly after 1550. This was not published however until 1663, by which time probability theory had already had its official inauguration elsewhere.*

It was around 1654 that B. Pascal and P. de Fermat generated a celebrated correspondence about their solutions of the problem of the points. These were soon widely known, and C. Huygens developed these ideas in a book published in 1657, in Latin. [...] the intuitive notions underlying this work were similar to those commonly in force nowdays.

These first simple ideas were soon extended by Jacob Bernoulli in Ars conjectandi *(1713) and by A. de Moivre in* Doctrine of chances *(1718, 1738, 1756). [...]. Methods, results, and ideas were all greatly refined and generalized by P. Laplace [...]. Many other eminent mathematicians of this period wrote on probability: Euler, Gauss, Lagrange, Legendre, Poisson, and so on.*

However, as ever harder problems were tackled by ever more powerful mathematical techniques during the 19th. century, the lack of a well-defined axiomatic structure was recognized as a serious handicap. [...] A. Kolmogorov provided the axioms which today underpin most mathematical probability.

Grimmett and Stirzaker (2001, p. 571)

[1] De vetula ("The Old Woman") is a long thirteenth-century poem written in Latin. (For more details see http://en.wikipedia.org/wiki/De_vetula.)

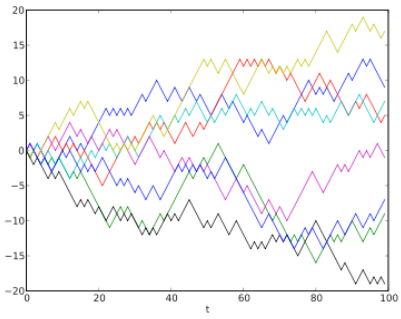
For more extensive and exciting accounts on the history of Statistics and Probability, we recommend:

- Hald, A. (1998). *A History of Mathematical Statistics from 1750 to 1930*. John Wiley & Sons. (QA273-280/2.HAL.50129);
- Stigler, S.M. (1986). *The History of Statistics: the Measurement of Uncertainty Before 1900*. Belknap Press of Harvard University Press. (QA273-280/2.STI.39095).

0.2 (Symmetric) random walk

This section is inspired by Karr (1993, pp. 1–14) and has the sole purpose of:

- illustrating concepts such as probability, random variables, independence, expectation and convergence of random sequences, and recall some limit theorems;
- drawing our attention to the fact that exploiting the special structure of a random process can provide answers for some of the questions raised.



It refers to the random walk, a mathematical formalization of path that consist of a succession of random steps (http://en.wikipedia.org/wiki/Random_walk), such as the ones portrayed above.

The term random walk was first introduced by Karl Pearson in 1905 (http://en.wikipedia.org/wiki/Random_walk).

Informal definition 0.1 — Symmetric random walk
The symmetric random walk (SRW) is a random experiment which can result from the observation of a particle moving randomly on $Z = \{\ldots, -1, 0, 1, \ldots\}$. Moreover, the particle starts at the origin at time 0, and then moves either one step up or one step down with equal likelihood. •

Remark 0.2 — Applications of random walk

The path followed by atom in a gas moving under the influence of collisions with other atoms can be described by a random walk (RW). Random walk has also been applied in other areas such as:

- economics (RW used to model shares prices and other factors);
- population genetics (RW describes the statistical properties of genetic drift);[2]
- mathematical ecology (RW used to describe individual animal movements, to empirically support processes of biodiffusion, and occasionally to model population dynamics);
- computer science (RW used to estimate the size of the Web);
- visual arts, such as Antony Gormley's *Quantum Cloud* sculpture in London which was designed by a computer using a random walk algorithm.[3]

•

The next proposition provides answers to the following questions:

- How can we model and analize the symmetric random walk?
- What random variables can arise from this random experiment and how can we describe them?

[2]Genetic drift is one of several evolutionary processes which lead to changes in allele frequencies over time.

[3]For more applications check http://en.wikipedia.org/wiki/Random_walk.

Proposition 0.3 — **Symmetric random walk** (Karr, 1993, pp. 1–4)

1. The model

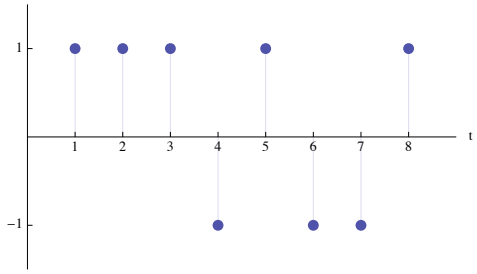
Let:

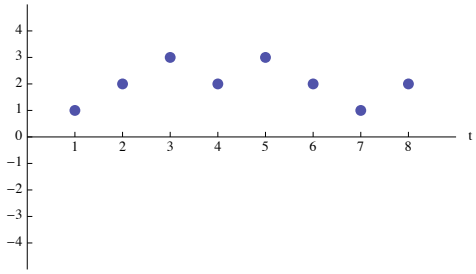
- ω_n be the step at time n ($\omega_n = \pm 1$);
- $\underline{\omega} = (\omega_1, \omega_2, \ldots)$ be a realization of the random walk;
- Ω be the sample space of the random experiment, i.e. the set of all possible realizations.

2. Random variables

Two random variables immediately arise:

- Y_n defined as $Y_n(\underline{\omega}) = \omega_n$, the size of the n^{th} step;[4]
- X_n which represents the position at time n and is defined as $X_n(\underline{\omega}) = \sum_{i=1}^{n} Y_i(\underline{\omega})$.
  A realization of $\{Y_n, n \in \mathbb{N}\}$ and the corresponding sample path of $\{X_n,\ n \in \mathbb{N}\}$ are shown below for $p = \frac{1}{2}$.





3. Probability and independence

The sets of outcomes of this random experiment are termed *events*. An event $A \subset \Omega$ occurs with probability $P(A)$.

Recall that a probability function is *countable additive*, i.e. for sequences of (pairwise) disjoint events $A_1, A_2, \ldots$ ($A_i \cap A_j = \emptyset$, $i \neq j$) we have

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

[4]Steps are functions defined on the sample space. Thus, steps are random variables.

Invoking the random and symmetric character of this walk, and assuming that the steps are independent and identically distributed, all 2^n possible values of $(Y_1, \ldots, Y_n)$ are equally likely, and, for every $(y_1, \ldots, y_n) \in \{-1, 1\}^n$,

$$P(Y_1 = y_1, \ldots, Y_n = y_n) = \prod_{i=1}^{n} P(Y_i = y_i) \tag{2}$$

$$= \left(\frac{1}{2}\right)^n. \tag{3}$$

The interpretation of (2) is *absence of probabilistic interaction* or independence.

4. First calculations

Let us assume from now on that $X_0 = 0$. Then:

- $|X_n| \leq n$, $\forall n \in I\!N$;
- X_n is even at even times ($n \, mod \, 2 = 0$) (e.g. X_2 cannot be equal to 1);
- X_n is odd at odd times ($n \, mod \, 2 = 1$) (e.g. X_1 cannot be equal to 0).

If $n \in I\!N$, $k \in \{-n, \ldots, 0, \ldots, n\}$, and $\frac{n+k}{2}$ is an integer ($n \, mod \, 2 = k \, mod \, 2$) then the event $\{X_n = k\}$ occurs if $\frac{n+k}{2}$ of the steps $Y_1, \ldots, Y_n$ are equal to 1 and the remainder are equal to -1.

In fact, $X_n = k$ if we observe a steps up and b steps down where

$$(a, b) \quad : \quad \begin{cases} a, b \in \{0, \ldots, n\} \\ a + b = n \\ a - b = k \end{cases} \tag{4}$$

that is, $a = \frac{n+k}{2}$ and a has to be an integer in $\{0, \ldots, n\}$.

As a consequence,

$$P(X_n = k) = \binom{n}{\frac{n+k}{2}} \times \left(\frac{1}{2}\right)^n, \tag{5}$$

for $n \in I\!N$, $k \in \{-n, \ldots, 0, \ldots, n\}$, $\frac{n+k}{2} \in \{0, 1, \ldots, n\}$. Recall that the *binomial coefficient* $\binom{n}{\frac{n+k}{2}}$ (it often reads as "n choose $\frac{n+k}{2}$") represents the number of **subsets of size** $\frac{n+k}{2}$ of a set of size n.

More generally,

$$\begin{aligned} P(X_n \in B) &= P\left(\{\underline{\omega} : X_n(\underline{\omega}) \in B\}\right) \\ &= \sum_{k \in B \cap \{-n,\ldots,0,\ldots,n\}} P(X_n = k), \end{aligned} \tag{6}$$

for any real set $B \subset \mathbb{R}$, by countable additivity of the probability function P.

(Rewrite (6) taking into account that n and k have to be both even or both odd.) •

Remark 0.4 — **Further properties of the symmetric random walk** (Karr, 1993, pp. 4–5)
Exploiting the special structure of the SRW lead us to conclude that:

- the SRW cannot move from one level to another without passing through all values between ("continuity");
- all 2^n length$-n$ paths are equally likely so two events containing the same number of paths have the same probability, $\frac{\text{no. paths}}{2^n}$, which allows the probability of one event to be determined by showing that the paths belonging to this event are in one-to-one correspondence with those of an event of known probability — in many cases this correspondence is established geometrically, **namely** via reasoning known as *reflection principle*.[5] •

Exercise 0.5 — **Symmetric random walk**
Prove that:

(a) $P(X_2 \neq 0) = P(X_2 = 0) = \frac{1}{2}$;

(b) $P(X_n = -k) = P(X_n = k)$, for each n and k. •

Proposition 0.6 — **Expectation and symmetric random walk** (Karr, 1993, p. 5)
The average value of any $i^{th}-$step of a SRW is equal to

$$\begin{aligned} E(Y_i) &= (-1) \times P(Y_i = -1) + (+1) \times P(Y_i = +1) \\ &= 0. \end{aligned} \tag{7}$$

Additivity of probability translates to linearity of expectation, thus the average position equals

[5] **For each n, there are as many paths of length $2n$ origination at $(0,0)$ that do not cross the $x-$axis before or at time $2n$ as there are paths from $(0,0)$ to $(2n,0)$ (Karr, 1993, pp. 7–8).**

$$\begin{aligned} E(X_n) &= E\left(\sum_{i=1}^{n} Y_i\right) \\ &= \sum_{i=1}^{n} E(Y_i) \\ &= 0. \end{aligned} \tag{8}$$

•

Proposition 0.7 — Conditioning and symmetric random walk (Karr, 1993, p. 6)
We can revise probability in light of the knowledge that some event has occurred. For example, we know that $P(X_{2n} = 0) = \binom{2n}{n} \times \left(\frac{1}{2}\right)^{2n}$. However, if we knew that $X_{2n-1} = 1$ then the event $\{X_{2n} = 0\}$ occurs with probability $\frac{1}{2}$. In fact,

$$\begin{aligned} P(X_{2n} = 0 | X_{2n-1} = 1) &= \frac{P(X_{2n-1} = 1, X_{2n} = 0)}{P(X_{2n-1} = 1)} \\ &= \frac{P(X_{2n-1} = 1, Y_{2n} = -1)}{P(X_{2n-1} = 1)} \\ &= \frac{P(X_{2n-1} = 1) \times P(Y_{2n} = -1)}{P(X_{2n-1} = 1)} \\ &= \frac{1}{2}. \end{aligned} \tag{9}$$

Note that, since the steps Y_i are independent random variables and $X_{2n-1} = \sum_{i=1}^{2n-1} Y_i$, we can state that Y_{2n} is independent of X_{2n-1}. •

Exercise 0.8 — Conditioning and asymmetric random walk[6]
Random walk models are often found in physics, from particle motion to a simple description of a polymer.

A physicist assumes that the position of a particle at time n, X_n, is governed by an **a**symmetric random walk — starting at 0 and with probability of an upward (resp. downward) step equal to p (resp. $1 - p$), where $p \in (0, 1) \backslash \{\frac{1}{2}\}$.

Derive $P(X_{2n} = 0 | X_{2n-2} = 0)$, for $n = 2, 3, \ldots$ •

Proposition 0.9 — Time of first return to the origin and symmetric random walk (Karr, 1993, pp. 7-9)
The time at which the SRW first returns to the origin,

[6] Exam 2010/01/19.

$$T^0 = \min\{n \in I\!N : X_n = 0\}, \tag{10}$$

is an important functional of the SRW (it maps the SRW into a scalar). It can represent the time to ruin.

Interestingly enough, for $n \in I\!N$, T^0 must be **a positive and even r.v.** (recall that $X_0 = 0$). And, for $n \in I\!N$:

$$P(T^0 > 2n) = P(X_1 \neq 0, \ldots, X_{2n} \neq 0) = \binom{2n}{n} \times \left(\frac{1}{2}\right)^{2n}; \tag{11}$$

$$P(T^0 = 2n) = \frac{1}{2n-1}\binom{2n}{n} \times \left(\frac{1}{2}\right)^{2n}. \tag{12}$$

Moreover, using the Stirling's approximation to $n!$, $n! \simeq \sqrt{2\pi}\, n^{n+\frac{1}{2}}\, e^{-n}$, we get

$$P(T^0 < +\infty) = 1. \tag{13}$$

If we note that $P(T^0 > 2n) \simeq \frac{1}{\sqrt{\pi n}}$ and recall that $\sum_{n=1}^{+\infty} \frac{1}{n^s}$ only converges for $s \geq 2$, we can conclude that T^0 assumes large values with probabilities large enough that

$$\sum_{n=1}^{+\infty} 2n\, P(T^0 = 2n) = +\infty \Rightarrow E(T^0) = +\infty. \tag{14}$$

•

Exercise 0.10 — Time of first return to the origin and symmetric random walk

(a) Prove result (12) using (11).

(b) Use the Stirling's approximation to $n!$, $n! \simeq \sqrt{2\pi}\, n^{n+\frac{1}{2}}\, e^{-n}$ to prove that

$$\lim_{n \to +\infty} P(T^0 > 2n) = \lim_{n \to +\infty} \frac{1}{\sqrt{\pi n}}.$$

(c) Use the previous result and the fact that

$$P(T^0 < +\infty) = 1 - \lim_{n \to +\infty} P(T^0 > 2n)$$

to derive (13).

(d) Verify that $\sum_{n=1}^{+\infty} 2n\, P(T^0 = 2n) = 1 + \sum_{n=1}^{+\infty} P(T^0 > 2n)$, even though we have $E(Z) = 2 \times \left[1 + \sum_{n=1}^{+\infty} P(Z > 2n)\right]$, for any positive and even random variable Z with finite expected value $E(Z) = \sum_{n=1}^{+\infty} 2n \times P(Z = 2n)$. •

Proposition 0.11 — **First passage times and symmetric random walk** (Karr, 1993, pp. 9–11)
Similarly, the first passage time

$$T^k = \min\{n \in \mathbb{N} : X_n = k\}, \tag{15}$$

has the following properties, for $n \in \mathbb{N}$, $k \in \{-n, \ldots, -1, 1, \ldots, n\}$ and $n \, mod \, 2 = k \, mod \, 2$:

$$\begin{aligned} P(T^k = n) &= \frac{|k|}{n} \times P(X_n = k); && (16)\\ P(T^k < \infty) &= 1; && (17)\\ E(T^k) &= +\infty. && (18) \end{aligned}$$

•

The following results pertain to the asymptotic behaviour of the position of a symmetric random walk and to the fraction of time spent positive.

Proposition 0.12 — **Law of large numbers** (Karr, 1993, p. 12)
Let Y_n and $X_n = \sum_{i=1}^n Y_i$ represent the size of the nth. step and the position at time n of a random walk, respectively. **Then**

$$P\left(\lim_{n \to +\infty} \frac{X_n}{n} = 0\right) = 1, \tag{19}$$

that is, the "empirical averages", $\frac{X_n}{n} = \frac{1}{n}\sum_{i=1}^n Y_i$, converge to the "theoretical average" $E(Y_1)$. •

Proposition 0.13 — **Central limit theorem** (Karr, 1993, pp. 12–13)

$$\begin{aligned} \lim_{n \to +\infty} P\left[\frac{\frac{X_n}{n} - E\left(\frac{X_n}{n}\right)}{\sqrt{V\left(\frac{X_n}{n}\right)}} \leq x\right] &= \lim_{n \to +\infty} P\left[\frac{\frac{X_n}{n} - E(Y_1)}{\sqrt{\frac{V(Y_1)}{n}}} \leq x\right] \\ &= \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} dy \\ &= \Phi(x), \, x \in \mathbb{R}. \end{aligned} \tag{20}$$

So, for large values of n, difficult-to-compute probabilities can be approximated. For instance, for $a < b$, we get:

$$
\begin{aligned}
P(a < X_n \leq b) &= \sum_{a<k\leq b} P(X_n = k) \\
&= P\left[\frac{\frac{a}{n} - 0}{\sqrt{\frac{1}{n}}} < \frac{\frac{X_n}{n} - 0}{\sqrt{\frac{1}{n}}} \leq \frac{\frac{b}{n} - 0}{\sqrt{\frac{1}{n}}}\right] \\
&\simeq \Phi(b/\sqrt{n}) - \Phi(a/\sqrt{n}). \qquad (21)
\end{aligned}
$$

•

Exercise 0.14 — Central limit theorem[7]
The words "symmetric random walk" refer to this situation.

> *The proverbial drunk* (PD) *is clinging to the lamppost. He decides to start walking. The road runs east and west. In his inebriated state he is as likely to take a step east (forward) as west (backward). In each new position he is again as likely to go forward as backward. Each of his steps are of the same length but of random direction — east or west.*
>
> http://www.physics.ucla.edu/∼chester/TECH/RandomWalk/3Pane.html

Admit that each step of PD has length equal to one meter and that he has already taken exactly 100 (a hundred) steps.

Find an approximate value for the probability that PD is within a five meters neighborhood of the lamppost. •

Proposition 0.15 — Arc sine law (Karr, 1993, pp. 13–14)
The fraction of time spent positive $\frac{W_n}{n} = \frac{1}{n}\sum_{i=1}^{n} I_{I\!N}(X_i + X_{i-1})$ has the following limiting law:[8]

$$\lim_{n\to+\infty} P\left(\frac{W_n}{n} \leq x\right) = \frac{2}{\pi}\arcsin\sqrt{x}. \qquad (22)$$

Moreover, the associated limiting density function, $\frac{1}{\pi\sqrt{x(1-x)}}$, is a $U-$shaped density. Thus, $\frac{W_n}{n}$ is more likely to be near 0 or 1 than near 1/2.

[7] Exam 2010/02/04.

[8] According to Karr (1993, p. 12), being *positive* at time i requires that either $X_i > 0$ or $X_{i-1} > 0$ (or both).

Please note that we can get the limiting distribution function by using the Stirling's approximation and the following result:

$$P(W_{2n} = 2k) = \binom{2k}{k} \times \binom{2n-2k}{n-k} \times \left(\frac{1}{2}\right)^{2n}. \tag{23}$$

•

Exercise 0.16 — **Arc sine law**

Prove result (22) **(Karr, 1993, p. 13)**. •

Exercise 0.17 — **Arc sine law**[9]

The *random walk hypothesis* is due to French economist Louis Bachelier (1870–1946) and asserts that the random nature of a commodity or stock prices cannot reveal trends and therefore current prices are no guide to future prices. Surprisingly, an investor assumes that his/her daily financial score is governed by a symmetric random walk starting at 0.

Obtain the corresponding approximate value for the probability that *the fraction of time the financial score is positive exceeds 50%*. •

Exercise 0.18 — **The cliff-hanger problem** (Mosteller, 1965, pp. 51–54)

From where he stands ($X_0 = 1$), one step toward the cliff would send the drunken man over the edge. He takes random steps, either toward or away from the cliff. At any step, his probability of taking a step away is p and of a step toward the cliff $1 - p$.

What is his chance of not escaping the cliff? (Write the results in terms of p.) •

References

- Grimmett, G.R. and Stirzaker, D.R. (2001). *Probability and Random Processes (3rd. edition)*. Oxford. (QA274.12-.76.GRI.40695 refers to the library code of the 1st. and 2nd. editions from 1982 and 1992, respectively.)
- Karr, A.F. (1993). *Probability*. Springer-Verlag.
- Konstantopoulos, T. (2009). *Introductory Lecture Notes on Markov Chains and Random Walks*. (www2.math.uu.se/∼takis/L/McRw/mcrw.pdf)
- Mosteller, F. (1965). *Fifty Challenging Problems in Probability with Solutions*. Dover Publications.

[9] Test 2009/11/07.

Chapter 1

Probability spaces

> *[...] have been taught that the universe evolves according to deterministic laws that specify exactly its future, and a probabilistic description is necessary only because of our ignorance. This deep-rooted skepticism in the validity of probabilistic results can be overcome only by proper interpretation of the meaning of probability.* Papoulis (1965, p. 3)

> **Probability is the mathematics of uncertainty**. *It has flourished under the stimulus of applications*, such as *insurance, demography, [...], clinical trials, signal processing, [...], spread of infectious diseases, [...], medical imaging*, etc. and *have furnished both mathematical questions and genuine interest in the answers.* Karr (1993, p. 15)

Much of our life is based on the belief that the future is largely unpredictable (Grimmett and Stirzaker, 2001, p. 1), nature is liable to change and chance governs life.

We express this belief in chance behaviour by the use of words such as *random*, *probable (probably)*, *probability*, *likelihood (likeliness)*, etc.

There are essentially four ways of defining probability (Papoulis, 1965, p. 7) and this is quite a controversial subject, proving that *not all of probability and statistics is cut-and-dried* (Righter, 200–):

- a priori definition as a ratio of favorable to total number of alternatives (classical definition; Laplace);[1]

[1] See the first principle of probability in http://en.wikipedia.org/wiki/Pierre-Simon_Laplace

- relative frequency (Von Mises);[2]
- probability as a measure of belief (inductive reasoning,[3] subjective probability; Bayesianism);[4]
- axiomatic (measure theory; Kolmogorov's axioms).[5]

Classical definition of probability

The classical definition of probability of an event A is found a priori without actual experimentation, by counting the total number $N = \#\Omega < +\infty$ of possible outcomes of the random experiment. If these outcomes are equally likely and $N_A = \#A$ of these outcomes the event A occurs, then

$$P(A) = \frac{N_A}{N} = \frac{\#A}{\#\Omega}. \tag{1.1}$$

Criticism of the classical definition of probability

It is only holds if $N = \#\Omega < +\infty$ and all the N outcomes are equally likely. Moreover,

- serious problems often arise in determining $N = \#\Omega$;
- it can be used only for a limited class of problems since the *equally likely condition* is often violated in practice;
- the classical definition, although presented as a priori logical necessity, makes implicit use of the relative-frequency interpretation of probability;
- in many problems the possible number of outcomes is infinite, so that to determine probabilities of various events one must introduce some measure of length or area.

[2]Kolmogorov said: "[...] mathematical theory of probability to real 'random phenomena' must depend on some form of the frequency concept of probability, [...] which has been established by von Mises [...]." (http://en.wikipedia.org/wiki/Richard_von_Mises)

[3]Inductive reasoning or inductive logic is a type of reasoning which involves moving from a set of specific facts to a general conclusion (http://en.wikipedia.org/wiki/Inductive_reasoning).

[4]Bayesianism uses probability theory as the framework for induction. Given new evidence, Bayes' theorem is used to evaluate how much the strength of a belief in a hypothesis should change with the data we collected.

[5]http://en.wikipedia.org/wiki/Kolmogorov_axioms

Relative frequency definition of probability

The relative frequency approach was developed by Von Mises in the beginning of the 20th. century; at that time the prevailing definition of probability was the classical one and his work was a healthy alternative (Papoulis, 1965, p. 9).

The relative frequency definition of probability used to be popular among engineers and physicists. A random experiment is repeated over and over again, N times; if the event A occurs N_A times out of N, then the probability of A is defined as the limit of the relative frequency of the occurrence of A:

$$P(A) = \lim_{N \to +\infty} \frac{N_A}{N}. \tag{1.2}$$

Criticism of the relative frequency definition of probability

This notion is meaningless in most important applications, e.g. finding the probability of the space shuttle blowing up, or of an earthquake (Righter, 200–), essentially because we cannot repeat the experiment.

It is also useless when dealing with hypothetical experiments (e.g. visiting Jupiter).

Subjective probability, personal probability, Bayesian approach; criticism

Each person determines for herself what the probability of an event is; this value is in $[0, 1]$ and expresses the personal belief on the occurrence of the event.

The Bayesian approach is the approach used by most engineers and many scientists and business people. It bothers some, because it is not "objective". For a Bayesian, anything that is unknown is random, and therefore has a probability, even events that have already occurred. (Someone flipped a fair coin in another room, the chance that it was heads or tails is .5 for a Bayesian. A non-Bayesian could not give a probability.)

With a Bayesian approach it is possible to include nonstatistical information (such as expert opinions) to come up with a probability. The general Bayesian approach is to come up with a prior probability, collect data, and use the data to update the probability (using Bayes' Law, which we will study later). (Righter, 200–)

To understand the (axiomatic) definition of probability we shall need the following concepts:

- random experiment, whose outcome cannot be determined in advance;
- sample space Ω, the set of all (conceptually) possible outcomes;

- outcomes ω, elements of the sample space, also referred to as sample points or realizations;
- events A, a set of outcomes;
- $\sigma-$algebra on Ω, a family of subsets of Ω containing Ω and closed under complementation and countable union.

1.1 Random experiments

Definition 1.1 — Random experiment
A random experiment consists of both a procedure and observations,[6] and its outcome cannot be determined in advance. •

There is some uncertainty in what will be observed in the random experiment, otherwise performing the experiment would be unnecessary.

Example 1.2 — Random experiments

	Random experiment
E_1	Give a lecture. Observe the number of students seated in the 4th. row, which has 7 seats.
E_2	Choose a highway junction. Observe the number of car accidents in 12 hours.
E_3	Walk to a bus stop. Observe the time (in minutes) you wait for the arrival of a bus.
E_4	Give n lectures. Observe the number of students seated in the forth row in each of those n lectures.
E_5	Consider a particle in a gas modeled by a random walk. Observe the steps at times $1, 2, \ldots$
E_6	Consider a cremation chamber. Observe the temperature in the center of the chamber over the interval of time $[0, 1]$.

•

Exercise 1.3 — Random experiment
Identify at least one random experiment based on your daily schedule. •

Definition 1.4 — Sample space (Yates and Goodman, 1999, p. 8)
The sample space Ω of a random experiment is the finest-grain, mutually exclusive, collectively exhaustive set of all possible outcomes of the random experiment. •

[6] Yates and Goodman (1999, p. 7).

The finest-grain property simply means that all possible distinguishable outcomes are identified separately. Moreover, Ω is (usually) known before the random experiment takes place. The choice of Ω balances fidelity to reality with mathematical convenience (Karr, 1993, p. 12).

Remark 1.5 — Categories of sample spaces (Karr, 1993, pp. 16–17)
In practice, most sample spaces fall into one of the six categories:

- **Finite set**

  The simplest random experiment has two outcomes.

  A random experiment with n possible outcomes may be modeled with a sample space consisting of n integers.

- **Countable set**

  The sample space for an experiment with countably many possible outcomes is ordinarily the set $\mathbb{N} = \{1, 2, \ldots\}$ of positive integers or the set of $\{\ldots, -1, 0, +1, \ldots\}$ of all integers.

  Whether a finite or a countable sample space better describes a given phenomenon is a matter of judgement and compromise. (Comment!)

- **The real line $\mathbb{R}$ (and intervals in $\mathbb{R}$)**

  The most common sample space is the real line $\mathbb{R}$ (or the unit interval $[0, 1]$ the nonnegative half-line $\mathbb{R}_0^+$), which is used for most all numerical phenomena that are not inherently integer-valued.

- **Finitely many replications**

  Some random experiments result from the n ($n \in \mathbb{N}$) replications of a basic experiment with sample space Ω_0. In this case the sample space is the Cartesian product $\Omega = \Omega_0^n$.

- **Infinitely many replications**

  If a basic random experiment is repeated infinitely many times we deal with the sample space $\Omega = \Omega_0^{\mathbb{N}}$.

- **Function spaces**

  In some random experiments the outcome is a trajectory followed by a system over an interval of time. In this case the outcomes are functions. •

Example 1.6 — Sample spaces

The sample spaces defined below refer to the random experiments defined in Example 1.2:

Random experiment	Sample space (Ω)	Classification of Ω
E_1	$\{0, 1, 2, 3, 4, 5, 6, 7\}$	Finite set
E_2	$\mathbb{N}_0 = \{0, 1, 2, \ldots\}$	Countable set
E_3	$\mathbb{R}_0^+$	Interval in $\mathbb{R}$
E_4	$\{0, 1, 2, 3, 4, 5, 6, 7\}^n$	Finitely many replications
E_5	$\{-1, +1\}^{\mathbb{N}}$	Infinitely many replications
E_6	$\mathbf{C}([0, 1])$	Function space

Note that $\mathbf{C}([0, 1])$ represents the vector space of continuous, real-valued functions on $[0, 1]$. •

1.2 Events and classes of sets

Definition 1.7 — **Event** (Karr, 1993, p. 18)
Given a random experiment with sample space Ω, an event can be provisionally defined as a subset of Ω whose probability is defined. •

Remark 1.8 — An event A occurs if the outcome ω of the random experiment belongs to A, i.e. $\omega \in A$. •

Example 1.9 — **Events**
Some events associated to the six random experiments described in examples 1.2 and 1.6:

E.A.	Event
E_1	$A =$ "observe at least 3 students in the 4th. row" $= \{3, \ldots, 7\}$
E_2	$B =$ "observe more than 4 car accidents in 12 hours" $= \{5, 6, \ldots\}$
E_3	$C =$ "wait more than 8 minutes" $= (8, +\infty)$
E_4	$D =$ "observe at least 3 students in the 4th. row, in 5 consecutive days" $= \{3, \ldots, 7\}^5$
E_5	$E =$ "an ascending path" $= \{(1, 1, \ldots)\}$
E_6	$F =$ "temperature above 250^o over the interval $[0, 1]$" $= \{f \in \mathbf{C}([0, 1]) : f(x) > 250,\ x \in [0, 1]\}$

•

Definition 1.10 — **Set operations** (Resnick, 1999, p. 3)
As subsets of the sample space Ω, events can be manipulated using set operations. The set operations which you should know and will be commonly used are listed next:

- **Complementation**

  The complement of an event $A \subset \Omega$ is

  $$A^c = \{\omega \in \Omega : \omega \notin A\}. \tag{1.3}$$

- **Intersection**

  The intersection of events A and B ($A, B \subset \Omega$) is

  $$A \cap B = \{\omega \in \Omega : \omega \in A \text{ and } \omega \in B\}. \tag{1.4}$$

  The events A and B are disjoint (mutually exclusive) if $A \cap B = \emptyset$, i.e. they have no outcomes in common, therefore they never happen at the same time.

- **Union**

  The union of events A and B ($A, B \subset \Omega$) is

  $$A \cup B = \{\omega \in \Omega : \omega \in A \text{ or } \omega \in B\}. \tag{1.5}$$

  Karr (1993) uses $A + B$ to denote $A \cup B$ when A and B are disjoint.

- **Set difference**

  Given two events A and B ($A, B \subset \Omega$), the set difference between B and A consists of those outcomes in B but not in A:

  $$B \backslash A = B \cap A^c. \tag{1.6}$$

- **Symmetric difference**

  Let A and B be two events ($A, B \subset \Omega$). Then the outcomes that are in one but not in both sets consist on the symmetric difference:

  $$A \Delta B = (A \backslash B) \cup (B \backslash A). \tag{1.7}$$

•

Exercise 1.11 — Set operations
Represent the five set operations in Definition 1.10 pictorially by Venn diagrams. •

Proposition 1.12 — Properties of set operations (Resnick, 1999, pp. 4–5)
Set operations satisfy well known properties such as commutativity, associativity, De Morgan's laws, etc., providing now and then connections between set operations. These properties have been condensed in the following table:

Set operation	**Property**
Complementation	$(A^c)^c = A$ $\emptyset^c = \Omega$ $\Omega^c = \emptyset$
Intersection and union	**Commutativity** $A \cap B = B \cap A$, $A \cup B = B \cup A$ $A \cap \emptyset = \emptyset$, $A \cup \emptyset = A$ $A \cap A = A$, $A \cup A = A$ $A \cap \Omega = A$, $A \cup \Omega = \Omega$ $A \cap A^c = \emptyset$, $A \cup A^c = \Omega$ **Associativity** $(A \cap B) \cap C = A \cap (B \cap C)$ $(A \cup B) \cup C = A \cup (B \cup C)$ **De Morgan's laws** $(A \cap B)^c = A^c \cup B^c$ $(A \cup B)^c = A^c \cap B^c$ **Distributivity** $(A \cap B) \cup C = (A \cup C) \cap (B \cup C)$ $(A \cup B) \cap C = (A \cap C) \cup (B \cap C)$

•

Definition 1.13 — Relations between sets (Resnick, 1999, p. 4)
Now we list ways sets A and B can be compared:

- **Set containment or inclusion**

  A is a subset of B, written $A \subset B$ or $B \supset A$, iff $A \cap B = A$. This means that

  $$\omega \in A \Rightarrow \omega \in B. \tag{1.8}$$

  So if A occurs then B also occurs. However, the occurrence of B does not imply the occurrence of A.

- **Equality**

  Two events A and B are equal, written $A = B$, iff $A \subset B$ and $B \subset A$. This means

  $$\omega \in A \Leftrightarrow \omega \in B. \tag{1.9}$$

•

Proposition 1.14 — Properties of set containment (Resnick, 1999, p. 4)
These properties are straightforward but we stated them for the sake of completeness and their utility in the comparison of the probabilities of events:

- $A \subset A$
- $A \subset B$, $B \subset C \Rightarrow A \subset C$
- $A \subset C$, $B \subset C \Rightarrow (A \cup B) \subset C$
- $A \supset C$, $B \supset C \Rightarrow (A \cap B) \supset C$
- $A \subset B \Leftrightarrow B^c \subset A^c$. •

These properties will be essential to calculate or relate probabilities of (sophisticated) events.

Remark 1.15 — The jargon of set theory and probability theory
What follows results from minor changes of Table 1.1 from Grimmett and Stirkazer (2001, p. 3):

Typical notation	**Set jargon**	**Probability jargon**
Ω	Collection of objects	Sample space
ω	Member of Ω	Outcome
A	Subset of Ω	Event (that some outcome in A occurs)
A^c	Complement of A	Event (that no outcome in A occurs)
$A \cap B$	Intersection	A and B occur
$A \cup B$	Union	Either A or B or both A and B occur
$B \backslash A$	Difference	B occurs but not A
$A \Delta B$	Symmetric difference	Either A or B, but not both, occur
$A \subset B$	Inclusion	If A occurs then B occurs too
$\emptyset$	Empty set	Impossible event
Ω	Whole space	Certain event

•

Functions on the sample space (such as random variables defined in the next chapter) are even more important than events themselves.

An indicator function is the simplest way to associate a set with a (binary) function.

Definition 1.16 — Indicator function (Karr, 1993, p. 19)
The indicator function of the event $A \subset \Omega$ is the function on Ω given by

$$\mathbf{1}_A(w) = \begin{cases} 1 & \text{if } w \in A \\ 0 & \text{if } w \notin A \end{cases} \tag{1.10}$$

Therefore, $\mathbf{1}_A$ indicates whether A occurs. •

The indicator function of an event, which resulted from a set operation on events A and B, can often be written in terms of the indicator functions of these two events.

Proposition 1.17 — Properties of indicator functions (Karr, 1993, p. 19)
Simple algebraic operations on the indicator functions of the events A and B translate set operations on these two events:

$$\begin{aligned} \mathbf{1}_{A^c} &= 1 - \mathbf{1}_A && (1.11) \\ \mathbf{1}_{A\cap B} &= \min\{\mathbf{1}_A, \mathbf{1}_B\} \\ &= \mathbf{1}_A \times \mathbf{1}_B && (1.12) \\ \mathbf{1}_{A\cup B} &= \max\{\mathbf{1}_A, \mathbf{1}_B\}; && (1.13) \\ \mathbf{1}_{B\setminus A} &= \mathbf{1}_{B\cap A^c} \\ &= \mathbf{1}_B \times (1 - \mathbf{1}_A) && (1.14) \\ \mathbf{1}_{A\Delta B} &= |\mathbf{1}_A - \mathbf{1}_B|. && (1.15) \end{aligned}$$

•

Exercise 1.18 — Indicator functions
Solve exercises 1.1 and 1.7 of Karr (1993, p. 40). •

The definition of indicator function quickly yields the following result when we are able compare events A and B.

Proposition 1.19 — Another property of indicator functions (Resnick, 1999, p. 5)
Let A and B be two events of Ω. Then

$$A \subseteq B \Leftrightarrow \mathbf{1}_A \leq \mathbf{1}_B. \tag{1.16}$$

Note here that we use the convention that for two functions f, g with domain Ω and range $\mathbb{R}$, we have $f \leq g$ iff $f(\omega) \leq g(\omega)$ for all $\omega \in \Omega$. •

Motivation 1.20 — Limits of sets (Resnick, 1999, p. 6)
The definition of convergence concepts for random variables rests on manipulations of sequences of events which require the definition of limits of sets. •

Definition 1.21 — Operations on sequences of sets (Karr, 1993, p. 20)
Let $(A_n)_{n\in\mathbb{N}}$ be a sequence of events of Ω. Then the union and the intersection of $(A_n)_{n\in\mathbb{N}}$ are defined as follows

$$\bigcup_{n=1}^{+\infty} A_n = \{\omega : \omega \in A_n \text{ for some } n\} \tag{1.17}$$

$$\bigcap_{n=1}^{+\infty} A_n = \{\omega : \omega \in A_n \text{ for all } n\}. \tag{1.18}$$

The sequence $(A_n)_{n\in\mathbb{N}}$ is said to be pairwise disjoint if $A_i \cap A_j = \emptyset$ whenever $i \neq j$. •

Definition 1.22 — Lim sup, lim inf and limit set (Karr, 1993, p. 20)
Let $(A_n)_{n\in\mathbb{N}}$ be a sequence of events of Ω. Then we define the two following limit sets:

$$\begin{aligned} \limsup A_n &= \bigcap_{k=1}^{+\infty} \bigcup_{n=k}^{+\infty} A_n \qquad (1.19)\\ &= \{\omega \in \Omega : \omega \in A_n \text{ for infinitely many values of } n\} \\ &= \{A_n, \text{i.o.}\} \end{aligned}$$

$$\begin{aligned} \liminf A_n &= \bigcup_{k=1}^{+\infty} \bigcap_{n=k}^{+\infty} A_n \qquad (1.20)\\ &= \{\omega \in \Omega : \omega \in A_n \text{ for all but finitely many values of } n\} \\ &= \{A_n, \text{ult.}\}, \end{aligned}$$

where *i.o.* and *ult.* stand for *infinitely often* and *ultimately*, respectively.

Let A be an event of Ω. Then the sequence $(A_n)_{n\in\mathbb{N}}$ is said to converge to A, written $A_n \to A$ or $\lim_{n\to+\infty} A_n = A$, if

$$\liminf A_n = \limsup A_n = A. \tag{1.21}$$

•

Example 1.23 — Lim sup, lim inf and limit set
Let $(A_n)_{n\in\mathbb{N}}$ be a sequence of events of Ω such that

$$A_n = \begin{cases} A & \text{for } n \text{ even} \\ A^c & \text{for } n \text{ odd.} \end{cases} \tag{1.22}$$

Then

$$\begin{aligned} \limsup A_n &= \bigcap_{k=1}^{+\infty} \bigcup_{n=k}^{+\infty} A_n \\ &= \Omega \qquad (1.23) \\ &\neq \\ \liminf A_n &= \bigcup_{k=1}^{+\infty} \bigcap_{n=k}^{+\infty} A_n \\ &= \emptyset, \qquad (1.24) \end{aligned}$$

so there is no limit set $\lim_{n\to} A_n$. •

Exercise 1.24 — Lim sup, lim inf and limit set
Solve Exercise 1.3 of Karr (1993, p. 40). •

Proposition 1.25 — Properties of lim sup and lim inf (Resnick, 1999, pp. 7–8)
Let $(A_n)_{n\in\mathbb{N}}$ be a sequence of events of Ω. Then

$$\liminf A_n \subset \limsup A_n \tag{1.25}$$

$$(\liminf A_n)^c = \limsup(A_n^c). \tag{1.26}$$

•

Definition 1.26 — Monotone sequences of events (Resnick, 1999, p. 8)
Let $(A_n)_{n\in\mathbb{N}}$ be a sequence of events of Ω. It is said to be monotone non-decreasing, written $A_n \uparrow$, if

$$A_1 \subseteq A_2 \subseteq A_3 \subseteq \dots. \tag{1.27}$$

$(A_n)_{n\in\mathbb{N}}$ is monotone non-increasing, written $A_n \downarrow$, if

$$A_1 \supseteq A_2 \supseteq A_3 \supseteq \dots. \tag{1.28}$$

•

Proposition 1.27 — Properties of monotone sequences of events (Karr, 1993, pp. 20–21)
Suppose $(A_n)_{n\in\mathbb{N}}$ be a monotone sequence of events. Then

$$A_n \uparrow \;\Rightarrow\; \lim_{n\to+\infty} A_n = \bigcup_{n=1}^{+\infty} A_n \tag{1.29}$$

$$A_n \downarrow \;\Rightarrow\; \lim_{n\to+\infty} A_n = \bigcap_{n=1}^{+\infty} A_n. \tag{1.30}$$

•

Exercise 1.28 — Properties of monotone sequences of events
Prove Proposition 1.27. •

Example 1.29 — Monotone sequences of events
The Galton-Watson process is a branching stochastic process arising from Francis Galton's statistical investigation of the extinction of family names. Modern applications include the survival probabilities for a new mutant gene, [...], or the dynamics of disease outbreaks in their first generations of spread, or the chances of extinction of small populations of organisms. (http://en.wikipedia.org/wiki/Galton-Watson_process)

Let $(X_n)_{\mathbb{N}_0}$ be a stochastic process, where X_n represents the size of generation n. A $(X_n)_{\mathbb{N}_0}$ is Galton-Watson process if it evolves according to the recurrence formula:

- $X_0 = 1$ (we start with one individual); and
- $X_{n+1} = \sum_{i=1}^{X_n} Z_i^{(n)}$, where, for each n, $Z_i^{(n)}$ represents the number of descendants of the individual i from generation n and $\left(Z_i^{(n)}\right)_{i\in\mathbb{N}}$ is a sequence of i.i.d. non-negative random variables.

Let $A_n = \{X_n = 0\}$. Since $A_1 \Rightarrow A_2 \Rightarrow \ldots$, i.e. $(A_n)_{n \in \mathbb{N}}$ is a non-decreasing monotone sequence of events, written $A_n \uparrow$, we get $A_n \to A = \bigcup_{n=1}^{+\infty} A_n$. Moreover, the extinction probability is given by

$$\begin{aligned} P(\{X_n = 0 \text{ for some } n\}) &= P\left(\bigcup_{n=1}^{+\infty} \{X_n = 0\}\right) = P\left(\lim_{n \to +\infty} \{X_n = 0\}\right) \\ &= P\left(\bigcup_{n=1}^{+\infty} A_n\right) \\ &= P\left(\lim_{n \to +\infty} A_n\right). \end{aligned} \tag{1.31}$$

Later on, we shall conclude that we can conveniently interchange the limit sign and the probability function and add: $P(X_n = 0 \text{ for some } n) = P\left(\lim_{n \to +\infty} \{X_n = 0\}\right) = \lim_{n \to +\infty} P(\{X_n = 0\})$. •

Proposition 1.30 — Limits of indicator functions (Karr, 1993, p. 21)
In terms of indicator functions,

$$A_n \to A \Leftrightarrow \mathbf{1}_{A_n}(w) \to \mathbf{1}_A(w), \forall w \in \Omega. \tag{1.32}$$

Thus, the convergence of sets is the same as pointwise convergence of their indicator functions. •

Exercise 1.31 — Limits of indicator functions (Exercise 1.8, Karr, 1993, p. 40)
Prove Proposition 1.30. •

Motivation 1.32 — Closure under set operations (Resnick, 1999, p. 12)
We need the notion of closure because we want to combine and manipulate events to make more complex events via set operations and we require that certain set operations do not carry events outside the family of events. •

Definition 1.33 — Closure under set operations (Resnick, 1999, p. 12)
Let $\mathcal{C}$ be a collection of subsets of Ω. $\mathcal{C}$ is closed under a set operation[7] if the set obtained by performing the set operation on sets in $\mathcal{C}$ yields a set in $\mathcal{C}$. •

[7] Be it a countable union, finite union, countable intersection, finite intersection, complementation, monotone limits, etc.

Example 1.34 — **Closure under set operations** (Resnick, 1999, p. 12)

- $\mathcal{C}$ is closed under finite union if for any finite collection $A_1, \ldots, A_n$ of sets in $\mathcal{C}$, $\bigcup_{i=1}^{n} A_i \in \mathcal{C}$.
- Suppose $\Omega = \mathbb{R}$ and $\mathcal{C} = \{\text{finite real intervals}\} = \{(a, b] : -\infty < a < b < +\infty\}$. Then $\mathcal{C}$ is not closed under finite unions since $(1, 2] \cup (36, 37]$ is not a finite interval.

  However, $\mathcal{C}$ is closed under intersection since $(a, b] \cap (c, d] = (\max\{a, c\}, \min\{b, d\}] = (a \vee c, b \wedge d]$.
- Consider now $\Omega = \mathbb{R}$ and $\mathcal{C} = \{\text{open real subsets}\}$. $\mathcal{C}$ is not closed under complementation since the complement of an open set is not open. •

Definition 1.35 — **Algebra** (Resnick, 1999, p. 12)
$\mathcal{A}$ is an algebra (or field) on Ω if it is a non-empty class of subsets of Ω closed under finite union, finite intersection and complementation.

A minimal set of postulates for $\mathcal{A}$ to be an algebra on Ω is:

1. $\Omega \in \mathcal{A}$
2. $A \in \mathcal{A} \Rightarrow A^c \in \mathcal{A}$
3. $A, B \in \mathcal{A} \Rightarrow A \cup B \in \mathcal{A}$. •

Remark 1.36 — **Algebra**
Please note that, by the De Morgan's laws, $\mathcal{A}$ is closed under finite intersection ($(A \cup B)^c = A^c \cap B^c \in \mathcal{A}$), thus we do not need a postulate concerning finite intersection. •

Motivation 1.37 — **σ-algebra** (Karr, 1993, p. 21)
To define a probability function dealing with an algebra is not enough: we need to define a collection of sets which is closed under *countable* union, *countable* intersection, and complementation. •

Definition 1.38 — **$\sigma-$algebra** (Resnick, 1999, p. 12)
$\mathcal{F}$ is a $\sigma-$algebra on Ω if it is a non-empty class of subsets of Ω closed under countable union, countable intersection and complementation.

A minimal set of postulates for $\mathcal{F}$ to be an $\sigma-$algebra on Ω is:

1. $\Omega \in \mathcal{F}$
2. $A \in \mathcal{F} \Rightarrow A^c \in \mathcal{F}$
3. $A_1, A_2, \ldots \in \mathcal{F} \Rightarrow \bigcup_{i=1}^{+\infty} A_i \in \mathcal{F}$. •

Example 1.39 — **$\sigma-$algebra** (Karr, 1993, p. 21)

- **Trivial $\sigma-$algebra**

  $\mathcal{F} = \{\emptyset, \Omega\}$

- **Power set**

  $\mathcal{F} = I\!P(\Omega)$ = class of all subsets of Ω

In general, neither of these two $\sigma-$algebras is specially interesting or useful — we need something in between. •

Definition 1.40 — **Generated $\sigma-$algebra** (http://en.wikipedia.org/wiki/Sigma-algebra)
If $\mathcal{U}$ is an arbitrary family of subsets of Ω then we can form a special $\sigma-$algebra containing $\mathcal{U}$, called the $\sigma-$algebra generated by $\mathcal{U}$ and denoted by $\sigma(\mathcal{U})$, by intersecting all $\sigma-$algebras containing $\mathcal{U}$.

Defined in this way $\sigma(\mathcal{U})$ is the smallest/minimal $\sigma-$algebra on Ω that contains $\mathcal{U}$. •

Example 1.41 — **Generated $\sigma-$algebra** (http://en.wikipedia.org/wiki/Sigma-algebra; Karr, 1993, p. 22)

- **Trivial example**

  Let $\Omega = \{1, 2, 3\}$ and $\mathcal{U} = \{\{1\}\}$. Then $\sigma(\mathcal{U}) = \{\emptyset, \{1\}, \{2, 3\}, \Omega\}$ is a $\sigma-$algebra on Ω.

- **$\sigma-$algebra generated by a finite partition**

  If $\mathcal{U} = \{A_1, \ldots, A_n\}$ is a finite partition of Ω — that is, $A_1, \ldots, A_n$ are disjoint and $\bigcup_{i=1}^{n} A_i = \Omega$ — then $\sigma(\mathcal{U}) = \{\bigcup_{i \in I} A_i : I \subseteq \{1, \ldots, n\}\}$ which includes $\emptyset$.

- **$\sigma-$algebra generated by a countable partition**

  If $\mathcal{U} = \{A_1, A_2, \ldots\}$ is a countable partion of Ω — that is, $A_1, A_2, , \ldots$ are disjoint and $\bigcup_{i=1}^{+\infty} A_i = \Omega$ — then $\sigma(\mathcal{U}) = \{\bigcup_{i \in I} A_i : I \subseteq I\!N\}$ which also includes $\emptyset$. •

Since we tend to deal with real random variables we have to define a $\sigma-$algebra on $\Omega = \mathbb{R}$ and the power set on $\mathbb{R}$, $\mathcal{P}(\mathbb{R})$ is not an option. The most important $\sigma-$algebra on $\mathbb{R}$ is the one defined as follows.

Definition 1.42 — **Borel $\sigma-$algebra on $\mathbb{R}$** (Karr, 1993, p. 22)
The Borel $\sigma-$algebra on $\mathbb{R}$, denoted by $\mathcal{B}(\mathbb{R})$, is generated by the class of intervals

$$\mathcal{U} = \{(a, b] : -\infty < a < b < +\infty\}, \tag{1.33}$$

that is, $\sigma(\mathcal{U}) = \mathcal{B}(\mathbb{R})$. Its elements are called Borel sets.[8] •

Remark 1.43 — **Borel $\sigma-$algebra on $\mathbb{R}$** (Karr, 1993, p. 22)

- Every "reasonable" set of $\mathbb{R}$ — such as intervals, closed sets, open sets, finite sets, and countable sets — belong to $\mathcal{B}(\mathbb{R})$. For instance, $\{x\} = \bigcap_{n=1}^{+\infty}(x - 1/n, x]$.
- Moreover, the Borel $\sigma-$algebra on $\mathbb{R}$, $\mathcal{B}(\mathbb{R})$, can also be generated by the class of intervals $\{(-\infty, a] : -\infty < a < +\infty\}$ or $\{(b, +\infty) : -\infty < b < +\infty\}$.
- $\mathcal{B}(\mathbb{R}) \neq \mathcal{P}(\mathbb{R})$.
- An example of a subset of the reals which is not a Borel set is due to Lusin (1927, pp. 76–78) and is described in some detail in http://en.wikipedia.org/wiki/Borel_set#Non-Borel_sets. •

Definition 1.44 — **Borel $\sigma-$algebra on $\mathbb{R}^d$** (Karr, 1993, p. 22)
The Borel $\sigma-$algebra on $\mathbb{R}^d$, $d \in \mathbb{N}$, $\mathcal{B}(\mathbb{R}^d)$, is generated by the class of rectangles that are Cartesian products of real intervals

$$\mathcal{U} = \left\{ \prod_{i=1}^{d} (a_i, b_i] : -\infty < a_i < b_i < +\infty, i = 1, \ldots, d \right\}. \tag{1.34}$$

•

Exercise 1.45 — **Generated $\sigma-$algebra** (Exercise 1.9, Karr, 1993, p. 40)
Given sets A and B of Ω, identify all sets in $\sigma(\{A, B\})$. •

Exercise 1.46 — **Borel $\sigma-$algebra on $\mathbb{R}$** (Exercise 1.10, Karr, 1993, p. 40)
Prove that $\{x\}$ is a Borel set for every $x \in \mathbb{R}$. •

[8] Borel sets are named after Émile Borel. *Along with René-Louis Baire and Henri Lebesgue, he was among the pioneers of measure theory and its application to probability theory* (http://en.wikipedia.org/wiki/Émile_Borel).

1.3 Probabilities and probability functions

Motivation 1.47 — Probability function (Karr, 1993, p. 23)
A probability is a set function, defined for events; it should be countably additive (i.e. $\sigma-$additive), that is, the probability of a countable union of disjoint events is the sum of their individual probabilities. •

Definition 1.48 — Probability function (Karr, 1993, p. 24)
Let Ω be the sample space and $\mathcal{F}$ be the $\sigma-$algebra of events of Ω. A probability on $(\Omega, \mathcal{F})$ is a function $P : \Omega \to I\!R$ such that:

1. **Axiom 1** [9] — $P(A) \geq 0$, $\forall A \in \mathcal{F}$.
2. **Axiom 2** — $P(\Omega) = 1$.
3. **Axiom 3** (countable additivity or $\sigma-$additivity)

   Whenever $A_1, A_2, \ldots$ are (pairwise) disjoint events in $\mathcal{F}$,

$$P\left(\bigcup_{i=1}^{+\infty} A_i\right) = \sum_{i=1}^{+\infty} P(A_i). \tag{1.35}$$

•

Remark 1.49 — Probability function
The probability function P transforms events in real numbers in $[0, 1]$. •

Definition 1.50 — Probability space (Karr, 1993, p. 24)
The triple $(\Omega, \mathcal{F}, P)$ is a probability space. •

Example 1.51 — Probability function (Karr, 1993, p. 25)
Let

- $\{A_1, \ldots, A_n\}$ be a finite partition of Ω — that is, $A_1, \ldots, A_n$ are (nonempty and pairwise) disjoint events and $\bigcup_{i=1}^{n} A_i = \Omega$;
- $\mathcal{F}$ be the $\sigma-$algebra generated by the finite partition $\{A_1, A_2, \ldots, A_n\}$, i.e. $\mathcal{F} = \sigma(\{A_1, \ldots, A_n\})$;
- $p_1, \ldots, p_n$ positive numbers such that $\sum_{i=1}^{n} p_i = 1$.

[9] Righter (200—) called the first and second axioms *duh rules*.

Then the function defined as

$$P\left(\bigcup_{i\in I} A_i\right) = \sum_{i\in I} p_i,\ \forall I \subseteq \{1,\ldots,n\}, \tag{1.36}$$

where $p_i = P(A_i)$, is a probability function on $(\Omega, \mathcal{F})$. •

Exercise 1.52 — **Probability function** (Exercise 1.11, Karr, 1993, p. 40)
Let A, B and C be disjoint events such that: $A \cup B \cup C = \Omega$; $P(A) = 0.6$, $P(B) = 0.3$ and $P(C) = 0.1$. Calculate all probabilities of all events in $\sigma(\{A, B, C\})$. •

Motivation 1.53 — **Elementary properties of probability functions**
The axioms do not teach us how to calculate the probabilities of events. However, they establish rules for their calculation such as the following ones. •

Proposition 1.54 — **Elementary properties of probability functions** (Karr, 1993, p. 25)
Let $(\Omega, \mathcal{F}, P)$ be a probability space then:

1. **Probability of the empty set**

   $$P(\emptyset) = 0. \tag{1.37}$$

2. **Finite additivity**

   If $A_1, \ldots, A_n$ are (pairwise) disjoint events then

   $$P\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} P(A_i). \tag{1.38}$$

   Probability of the complement of A

   Consequently, for each A,

   $$P(A^c) = 1 - P(A). \tag{1.39}$$

3. **Monotonicity of the probability function**

   If $A \subseteq B$ then

   $$P(B \setminus A) = P(B) - P(A). \tag{1.40}$$

   Therefore if $A \subseteq B$ then

   $$P(A) \leq P(B). \tag{1.41}$$

4. **Addition rule**

   For all A and B (disjoint or not),

$$P(A \cup B) = P(A) + P(B) - P(A \cap B). \tag{1.42}$$

•

Remark 1.55 — **Elementary properties of probability functions**
According to Righter (200—), (1.41) is another *duh rule* but adds *one of Kahneman and Tversky's most famous examples, the Linda problem.*

Subjects were told the story (in the 70's):

- *Linda is 31 years old, single, outspoken, and very bright. She majored in philosophy. As a student she was deeply concerned with issues of discrimination and social justice and also participated in anti nuclear demonstrations.*

They are asked to rank the following statements by their probabilities:

- *Linda is a bank teller.*
- *Linda is a bank teller who is active in the feminist movement.*

Kahneman and Tversky found that about 85% *of the subjects ranked "Linda is a bank teller and is active in the feminist movement" as more probable than "Linda is a bank teller".* •

Exercise 1.56 — **Elementary properties of probability functions**
Prove properties 1. through 4. of Proposition 1.54 and that

$$P(B \setminus A) = P(B) - P(A \cap B) \tag{1.43}$$
$$P(A \Delta B) = P(A \cup B) - P(A \cap B) = P(A) + P(B) - 2 \times P(A \cap B). \tag{1.44}$$

Hints (Karr, 1993, p. 25):

- property 1. can be also proved by using the finite additivity;
- property 2. by considering $A_{n+1} = A_{n+2} = \ldots = \emptyset$;
- property 3. by rewriting B as $(B \setminus A) \cup (A \cap B) = (B \setminus A) \cup A$;
- property 4. by rewriting $A \cup B$ as $(A \setminus B) \cup (A \cap B) \cup (B \setminus A)$. •

We proceed with some advanced properties of probability functions.

Proposition 1.57 — Boole's inequality or $\sigma-$subadditivity (Karr, 1993, p. 26)
Let $A_1, A_2, \ldots$ be events in $\mathcal{F}$. Then

$$P\left(\bigcup_{n=1}^{+\infty} A_n\right) \leq \sum_{n=1}^{+\infty} P(A_n). \tag{1.45}$$

•

Exercise 1.58 — Boole's inequality or $\sigma-$subadditivity
Prove Boole's inequality by using the *disjointification* technique (Karr, 1993, p. 26),[10] the fact that $B_n = A_n \setminus \left(\bigcup_{i=1}^{n-1} A_i\right) \subseteq A_n$, and by applying the $\sigma-$additivity and monotonicity of probability functions. •

Proposition 1.59 — Finite subadditivity (Resnick, 1999, p. 31)
The probability function P is finite subadditive in the sense that

$$P\left(\bigcup_{i=1}^{n} A_i\right) \leq \sum_{i=1}^{n} P(A_i), \tag{1.46}$$

for all events $A_1, \ldots, A_n$. •

Remark 1.60 — Finite additivity
Finite additivity is a consequence of Boole's inequality (i.e. $\sigma-$subadditivity). However, finite additivity does not imply $\sigma-$subadditivity. •

Proposition 1.61 — Inclusion-exclusion formula (Resnick, 1999, p. 30)
If $A_1, \ldots, A_n$ are events, then the probability of their union can be written as follows:

$$\begin{aligned} P\left(\bigcup_{i=1}^{n} A_i\right) &= \sum_{i=1}^{n} P(A_i) - \sum_{1\leq i<j\leq n} P(A_i \cap A_j) + \sum_{1\leq i<j<k\leq n} P(A_i \cap A_j \cap A_k) \\ &\quad - \ldots - (-1)^n \times P(A_1 \cap \ldots \cap A_n). \end{aligned} \tag{1.47}$$

•

Remark 1.62 — Inclusion-exclusion formula

- The terms on the right side of (1.47) alternate in sign and give inequalities called *Bonferroni inequalities*[11] when we neglect the remainders. Two examples:

[10] Note that $\bigcup_{n=1}^{+\infty} A_n = \bigcup_{n=1}^{+\infty} B_n$, where $B_n = A_n \setminus \left(\bigcup_{i=1}^{n-1} A_i\right)$ are disjoint events.

[11] They are named after Italian mathematician Carlo Emilio Bonferroni.

$$P\left(\bigcup_{i=1}^{n} A_i\right) \leq \sum_{i=1}^{n} P(A_i) \tag{1.48}$$

$$P\left(\bigcup_{i=1}^{n} A_i\right) \geq \sum_{i=1}^{n} P(A_i) - \sum_{1\leq i<j\leq n} P(A_i \cap A_j) \tag{1.49}$$

(Resnick, 1999, p. 30).

- Let the event A_i represent the rejection of the (simple) null hypothesis $H_{0,i}$ $(i = 1, \ldots, n)$. Then if we test the (multiple or simultaneous) null hypothesis $H_0 : \cap_{i=1}^{n} H_{0,i}$, the probability of rejecting H_0 is equal to the probability of rejecting at least one of the (simple) null hypotheses. Moreover, this probability does not exceed the sum of the probabilities of individually rejecting each of the (simple) null hypotheses:

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

  Consequently, if the desired significance level for the test involving H_0 is set to be equal to α_0, then the Bonferroni correction leads to the conclusion that each individual null hypothesis should be tested at a significance level of α_0/n (http://en.wikipedia.org/wiki/Bonferroni_correction). •

Exercise 1.63 — **Inclusion-exclusion formula**
Prove the inclusion-exclusion formula by induction using the addition rule for $n = 2$ (Resnick, 1999, p. 30). •

Proposition 1.64 — **Monotone continuity** (Resnick, 1999, p. 31)
Probability functions are continuous for monotone sequences of events in the sense that:

1. If $A_n \uparrow A$, where $A_n \in \mathcal{F}$, then $P(A_n) \uparrow P(A)$.
2. If $A_n \downarrow A$, where $A_n \in \mathcal{F}$, then $P(A_n) \downarrow P(A)$. •

Exercise 1.65 — **Monotone continuity**
Prove Proposition 1.64 by using the *disjointification* technique, the monotone character of the sequence of events and $\sigma-$additivity (Resnick, 1999, p. 31).

For instance property 1. can be proved as follows.

- $A_1 \subset A_2 \subset A_3 \subset \ldots \subset A_n \subset \ldots$;

- $B_1 = A_1$, $B_2 = A_2 \setminus A_1$, $B_3 = A_3 \setminus (A_1 \cup A_2)$, $\ldots$, $B_n = A_n \setminus \left(\bigcup_{i=1}^{n-1} A_i\right)$ are disjoint events;
- since $A_1, A_2, \ldots$ is a non-decreasing sequence of events $A_n \uparrow A = \bigcup_{n=1}^{+\infty} A_n = \bigcup_{n=1}^{+\infty} B_n$, $B_n = A_n \setminus A_{n-1}$, and $\bigcup_{i=1}^{n} B_i = A_n$; if we add to this $\sigma-$additivity, we conclude that

$$\begin{aligned} P(A) &= P\left(\bigcup_{n=1}^{+\infty} A_n\right) = P\left(\bigcup_{n=1}^{+\infty} B_n\right) = \sum_{n=1}^{+\infty} P(B_n) \\ &= \lim_{n\to+\infty} \uparrow \sum_{i=1}^{n} P(B_i) = \lim_{n\to+\infty} \uparrow P\left(\bigcup_{i=1}^{n} B_i\right) = \lim_{n\to+\infty} \uparrow P(A_n). \end{aligned}$$

•

Motivation 1.66 — $\sigma-$additivity as a result of finite additivity and monotone continuity (Karr, 1993, p. 26)
The next theorem shows that $\sigma - additivity$ is equivalent to the confluence of finite additivity (which is reasonable) and monotone continuity (which is convenient and desirable mathematically). •

Theorem 1.67 — $\sigma-$additivity as a result of finite additivity and monotone continuity (Karr, 1993, p. 26)
Let P be a nonnegative, finitely additive set function on $\mathcal{F}$ with $P(\Omega) = 1$. Then, the following are equivalent:

1. P is $\sigma-$additive (thus a probability function).
2. Whenever $A_n \uparrow A$ in $\mathcal{F}$, $P(A_n) \uparrow P(A)$.
3. Whenever $A_n \downarrow A$ in $\mathcal{F}$, $P(A_n) \downarrow P(A)$.
4. Whenever $A_n \downarrow \emptyset$ in $\mathcal{F}$, $P(A_n) \downarrow 0$. •

Exercise 1.68 — $\sigma-$additivity as a result of finite additivity and monotone continuity
Prove Theorem 1.67.

Note that we need to prove 1. $\Rightarrow$ 2. $\Rightarrow$ 3. $\Rightarrow$ 4. $\Rightarrow$ 1. But since 2. $\Leftrightarrow$ 3. by complementation and 4. is a special case of 3. we just need to prove that 1. $\Rightarrow$ 2. and 4. $\Rightarrow$ 1. (Karr, 1993, pp. 26-27). •

Remark 1.69 — Inf, sup, lim inf and lim sup
Let $a_1, a_2, \ldots$ be a sequence of real numbers. Then

- **Infimum**

  The infimum of the set $\{a_1, a_2, \ldots\}$ — written $\inf a_n$ — corresponds to the greatest element (not necessarily in $\{a_1, a_2, \ldots\}$) that is less than or equal to all elements of $\{a_1, a_2, \ldots\}$.[12]

- **Supremum**

  The supremum of the set $\{a_1, a_2, \ldots\}$ — written $\sup a_n$ — corresponds to the smallest element (not necessarily in $\{a_1, a_2, \ldots\}$) that is greater than or equal to every element of $\{a_1, a_2, \ldots\}$.[13]

- **Limit inferior and limit superior of a sequence of real numbers** [14]

  $\liminf a_n = \sup_{k \geq 1} \inf_{n \geq k} a_n$

  $\limsup a_n = \inf_{k \geq 1} \sup_{n \geq k} a_n$.

Let $A_1, A_2, \ldots$ be a sequence of events. Then

- **Limit inferior and limit superior of a sequence of sets**

  $\liminf A_n = \bigcup_{k=1}^{+\infty} \bigcap_{n=k}^{+\infty} A_n$

  $\limsup A_n = \bigcap_{k=1}^{+\infty} \bigcup_{n=k}^{+\infty} A_n$. •

Motivation 1.70 — A special case of the Fatou's lemma
This result plays a vital role in the proof of continuity of probability functions. •

Proposition 1.71 — A special case of the Fatou's lemma (Resnick, 1999, p. 32)
Suppose $A_1, A_2, \ldots$ is a sequence of events in $\mathcal{F}$. Then

$$P(\liminf A_n) \leq \liminf P(A_n) \leq \limsup P(A_n) \leq P(\limsup A_n). \tag{1.50}$$

•

Exercise 1.72 — A special case of the Fatou's lemma
Prove Proposition 1.71 (Resnick, 1999, pp. 32-33; Karr, 1993, p. 27). •

[12] For more details check http://en.wikipedia.org/wiki/Infimum
[13] http://en.wikipedia.org/wiki/Supremum
[14] http://en.wikipedia.org/wiki/Limit_superior_and_limit_inferior

Theorem 1.73 — Continuity (Karr, 1993, p. 27)
If $A_n \to A$ then $P(A_n) \to P(A)$. •

Exercise 1.74 — Continuity
Prove Theorem 1.73 by using Proposition 1.71 (Karr, 1993, p. 27). •

Motivation 1.75 — (1st.) Borel-Cantelli Lemma (Resnick, 1999, p. 102)
This result is simple but still is a basic tool for proving almost sure convergence of sequences of random variables (see Chapter 5). •

Theorem 1.76 — (1st.) Borel-Cantelli Lemma (Resnick, 1999, p. 102; Karr, 1993, p. 27)
Let $A_1, A_2, \ldots$ be any events in $\mathcal{F}$. Then

$$\sum_{n=1}^{+\infty} P(A_n) < +\infty \quad \Rightarrow \quad P(\limsup A_n) = 0. \tag{1.51}$$

•

Exercise 1.77 — (1st.) Borel-Cantelli Lemma
Prove Theorem 1.76 (Resnick, 1999, p. 102; Karr, 1993, p. 27). •

1.4 Distribution functions; discrete, absolutely continuous and mixed probabilities

Motivation 1.78 — Distribution function (Karr, 1993, pp. 28-29)
A probability function P on the Borel $\sigma-$algebra $\mathcal{B}(\mathbb{R})$ is determined by its values $P((-\infty, x])$, for all intervals $(-\infty, x]$.

Probability functions on the real line play an important role as distribution functions of random variables. •

Definition 1.79 — Distribution function (Karr, 1993, p. 29)
Let P be a probability function defined on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. The distribution function associated to P is represented by F_P and defined by

$$F_P(x) = P((-\infty, x]),\ x \in \mathbb{R}. \tag{1.52}$$

•

Theorem 1.80 — Some properties of distribution functions (Karr, 1993, pp. 29-30)
Let F_P be the distribution function associated to P. Then

1. F_P is **non-decreasing**. Hence, the left-hand limit

$$F_P(x^-) = \lim_{s \uparrow x,\, s<x} F_P(s) \tag{1.53}$$

and the right-hand limit

$$F_P(x^+) = \lim_{s \downarrow x,\, s>x} F_P(s) \tag{1.54}$$

exist for each x, and

$$F_P(x^-) \leq F_P(x) \leq F_P(x^+). \tag{1.55}$$

2. F_P is **right-continuous**, i.e.

$$F_P(x^+) = F_P(x) \tag{1.56}$$

for each x.

3. F_P has the following **limits**:

$$\lim_{x \to -\infty} F_P(x) = 0 \tag{1.57}$$

$$\lim_{x \to +\infty} F_P(x) = 1. \tag{1.58}$$

•

Definition 1.81 — Distribution function (Resnick, 1999, p. 33)
A function $F_P : \mathbb{R} \to [0,1]$ satisfying properties 1., 2. and 3. from Theorem 1.80 is called a distribution function. •

Exercise 1.82 — Some properties of distribution functions
Prove Theorem 1.80 (Karr, 1993, p. 30). •

Definition 1.83 — Survival (or survivor) function (Karr, 1993, p. 31)
The survival (or survivor) function associated to P is

$$S_P(x) = 1 - F_P(x) = P((x, +\infty)),\ x \in \mathbb{R}. \tag{1.59}$$

$S_P(x)$ are also termed tail probabilities. •

Proposition 1.84 — Probabilities in terms of the distribution function (Karr, 1993, p. 30)
The following table condenses the probabilities of various intervals in terms of the distribution function

Interval I	Probability $P(I)$
$(-\infty, x]$	$F_P(x)$
$(x, +\infty)$	$1 - F_P(x)$
$(-\infty, x)$	$F_P(x^-)$
$[x, +\infty)$	$1 - F_P(x^-)$
$(a, b]$	$F_P(b) - F_P(a)$
$[a, b)$	$F_P(b^-) - F_P(a^-)$
$[a, b]$	$F_P(b) - F_P(a^-)$
(a, b)	$F_P(b^-) - F_P(a)$
$\{x\}$	$F_P(x) - F_P(x^-)$

where $x \in \mathbb{R}$ and $-\infty < a < b < +\infty$. •

Example 1.85 — Point mass (Karr, 1993, p. 31)
Let P be defined as

$$P(A) = \epsilon_s(A) = \begin{cases} 1, & \text{if } s \in A \\ 0, & \text{otherwise,} \end{cases} \tag{1.60}$$

for every event $A \in \mathcal{F}$, i.e. P is a point mass at s. Then

$$\begin{aligned} F_P(x) &= P((-\infty, x]) \\ &= \begin{cases} 0, & x < s \\ 1, & x \geq s. \end{cases} \end{aligned} \tag{1.61}$$

The property that $F_P(x)$ only takes values 0 or 1 characterizes point masses. •

Example 1.86 — Uniform distribution on $[0, 1]$ (Karr, 1993, p. 31)
Let P be defined as

$$P((a, b]) = \text{Length}((a, b] \cap [0, 1]), \tag{1.62}$$

for any real interval $(a, b]$ with $-\infty < a < b < +\infty$. Then

$$\begin{aligned} F_P(x) &= P((-\infty, x]) \\ &= \begin{cases} 0, & x < 0 \\ x, & 0 \leq x \leq 1 \\ 1, & x > 1. \end{cases} \end{aligned} \tag{1.63}$$

•

We are going to revisit the discrete and absolutely continuous probabilities and introduce mixed distributions.

Definition 1.87 — Discrete probabilities (Karr, 1993, p. 32)
A probability function P defined on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ is said to be discrete if there is a countable set C such that $P(C) = 1$. •

Remark 1.88 — Discrete probabilities (Karr, 1993, p. 32)
Discrete probabilities are finite or countable convex combinations of point masses. The associated distribution functions do not increase "smoothly" — they increase only by means of jumps. •

Proposition 1.89 — Discrete probabilities (Karr, 1993, p. 32)
Let P be a probability function on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. Then the following are equivalent:

1. P is a discrete probability.
2. There is a real sequence $x_1, x_2, \ldots$ and numbers $p_1, p_2, \ldots$ with $p_n > 0$, for all n, and $\sum_{n=1}^{+\infty} p_n = 1$ such that

$$P(A) = \sum_{n=1}^{+\infty} p_n \times \epsilon_{x_n}(A), \tag{1.64}$$

for all $A \in \mathcal{B}(\mathbb{R})$.

3. There is a real sequence $x_1, x_2, \ldots$ and numbers $p_1, p_2, \ldots$ with $p_n > 0$, for all n, and $\sum_{n=1}^{+\infty} p_n = 1$ such that

$$F_P(x) = \sum_{n=1}^{+\infty} p_n \times \mathbf{1}_{[x_n, +\infty)}(x), \tag{1.65}$$

for all $x \in \mathbb{R}$. •

Remark 1.90 — **Discrete probabilities** (Karr, 1993, p. 33)
The distribution function associated to a discrete probability increases only by jumps of size p_n at x_n. •

Exercise 1.91 — **Discrete probabilities**
Prove Proposition 1.89 (Karr, 1993, p. 32). •

Example 1.92 — **Discrete probabilities**
Let p_x represent from now on $P(\{x\})$.

- **Uniform distribution** on a finite set C

  $p_x = \frac{1}{\#C}$, $x \in C$

  $P(A) = \frac{\#A}{\#C}$, $A \subseteq C$.

  This distribution is also known as the Laplace distribution.

- **Bernoulli distribution** with parameter p ($p \in [0, 1]$)

  $C = \{0, 1\}$

  $p_x = p^x(1-p)^{1-x}$, $x \in C$.

  This probability function arises in what we call a Bernoulli trial — a yes/no random experiment which yields success with probability p.

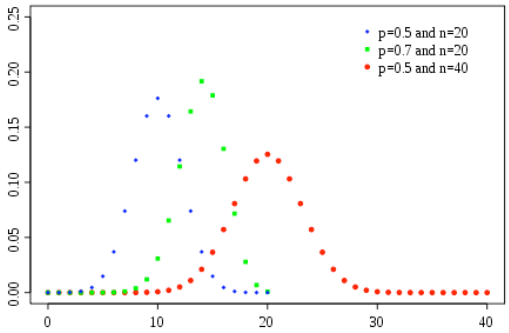
- **Binomial distribution** with parameters n and p ($n \in \mathbb{N}$, $p \in [0, 1]$)

  $C = \{0, 1, \ldots, n\}$

  $p_x = \binom{n}{x} p^x (1-p)^{n-x}$, $x \in C$.

  The binomial distribution is the discrete probability distribution of the number of successes in a sequence of n independent yes/no experiments, each of which yields success with probability p.

  Moreover, the result $\sum_{x=0}^{n} p_x = 1$ follows from the binomial theorem (http://en.wikipedia.org/wiki/Binomial_theorem).



- **Geometric distribution** with parameter p ($p \in [0,1]$)

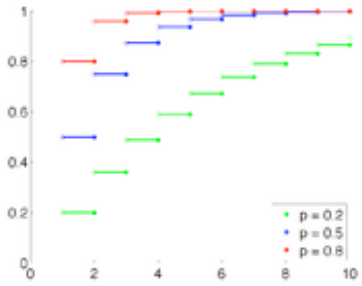
  $C = I\!N$

  $p_x = (1-p)^{x-1}p$, $x \in C$.

  This probability function is associated to the total number of (i.i.d.) Bernoulli trials needed to get one sucess — the first sucess (http://en.wikipedia.org/wiki/Geometric_distribution). The graph of

$$F_P(x) = \begin{cases} 0, & x < 1 \\ \sum_{i=1}^{[x]} (1-p)^{i-1}\, p = 1-(1-p)^{[x]}, & x \geq 1, \end{cases} \tag{1.66}$$

  where $[x]$ represents the integer part of the real number x, follows:



- **Negative binomial distribution** with parameters r and p ($r \in I\!N$, $p \in [0,1]$)

  $C = \{r, r+1, \ldots\}$

  $p_x = \binom{x-1}{r-1}(1-p)^{x-r}p^r$, $x \in C$.

  This probability function is associated to the total number of (i.i.d.) Bernoulli trials needed to get a pre-specified number r of sucesses (http://en.wikipedia.org/wiki/Negative_binomial_distribution). The geometric distribution is a particular case: $r = 1$.

- **Hypergeometric distribution** with parameters N, M, n ($N, M, n \in \mathbb{N}$ and $M, n \leq N$)

  $C = \{x \in \mathbb{N}_0 : \max\{0, n - N + M\} \leq x \leq \min\{n, M\}\}$

  $p_x = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}}$, $x \in C$.

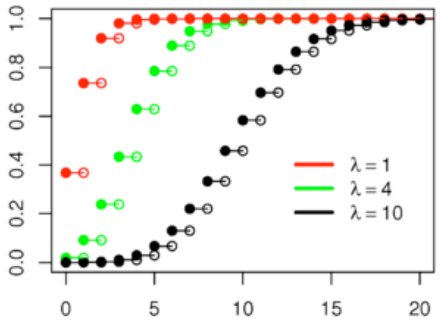
  It is a discrete probability that describes the number of successes in a sequence of n draws from a finite population with size N without replacement (http://en.wikipedia.org/wiki/Hypergeometric_distribution).

- **Poisson distribution** with parameter λ ($\lambda \in \mathbb{R}^+$)

  $C = \mathbb{N}_0$

  $p_x = e^{-\lambda}\frac{\lambda^x}{x!}$, $x \in C$.

  It is discrete probability that expresses the probability of a number of events occurring in a fixed period of time if these events occur with a known average rate and independently of the time since the last event. The Poisson distribution can



  also be used for the number of events in other specified intervals such as distance, area or volume (http://en.wikipedia.org/wiki/Poisson_distribution).

  The figure comprises the distribution function for three different values of λ. •

Motivation 1.93 — **Absolutely continuous probabilities** (Karr, 1993, p. 33)
Absolutely continuous probabilities are the antithesis of discrete probabilities in the sense that they have "smooth" distribution functions. •

Definition 1.94 — **Absolutely continuous probabilities** (Karr, 1993, p. 33)
A probability function P on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ is absolutely continuous if there is a non-negative function f_P on $\mathbb{R}$ such that

$$P((a, b]) = \int_a^b f_P(x)dx, \tag{1.67}$$

for every interval $(a, b] \in \mathcal{B}(\mathbb{R})$. •

Remark 1.95 — Absolutely continuous probabilities
If P is an absolutely continuous probability then $F_P(x)$ is an absolutely continuous real function. •

Remark 1.96 — Continuous, uniformly continuous and absolutely continuous functions

- **Continuous function**

  A real function f is continuous in x if for any sequence $\{x_1, x_2, \ldots\}$ such that $\lim_{n\to\infty} x_n = x$, it holds that $\lim_{n\to\infty} f(x_n) = f(x)$.

  One can say, briefly, that a function is continuous iff it preserves limits.

  For the Cauchy definition (epsilon-delta) of continuous function see http://en.wikipedia.org/wiki/Continuous_function

- **Uniformly continuous function**

  Given metric spaces (X, d_1) and (Y, d_2), a function $f : X \to Y$ is called uniformly continuous on X if for every real number $\epsilon > 0$ there exists $\delta > 0$ such that for every $x, y \in X$ with $d_1(x, y) < \delta$, we have that $d_2(f(x), f(y)) < \epsilon$.

  If X and Y are subsets of the real numbers, d_1 and d_2 can be the standard Euclidean norm, $|.|$, yielding the definition: for all $\epsilon > 0$ there exists a $\delta > 0$ such that for all $x, y \in X, |x - y| < \delta$ implies $|f(x) - f(y)| < \epsilon$.

  The difference between being uniformly continuous, and simply being continuous at every point, is that in uniform continuity the value of δ depends only on ϵ and not on the point in the domain (http://en.wikipedia.org/wiki/Uniform_continuity).

- **Absolutely continuous function**

  Let (X, d) be a metric space and let I be an interval in the real line $\mathbb{R}$. A function $f : I \to X$ is absolutely continuous on I if for every positive number ϵ, there is a positive number δ such that whenever a (finite or infinite) sequence of pairwise disjoint sub-intervals $[x_k, y_k]$ of I satisfies $\sum_k |y_k - x_k| < \delta$ then $\sum_k d\left(f(y_k), f(x_k)\right) < \epsilon$.

  Absolute continuity is a smoothness property which is stricter than continuity and uniform continuity.

  The two following functions are continuous everywhere but not absolutely continuous:

  1. $f(x) = x^2$ on an unbounded interval;

2. $f(x) = \begin{cases} 0, \text{ if } x = 0 \\ x \sin(1/x), \text{ if } x \neq 0, \end{cases}$
on a finite interval containing the origin.

(http://en.wikipedia.org/wiki/Absolute_continuity) •

Proposition 1.97 — Absolutely continuous probabilities (Karr, 1993, p. 34)
A probability function P is absolutely continuous iff there is a non-negative function f on $\mathbb{R}$ such that

$$\int_{-\infty}^{+\infty} f(s)ds = 1 \tag{1.68}$$

$$F_P(x) = \int_{-\infty}^{x} f(s)ds, \; x \in \mathbb{R}. \tag{1.69}$$

•

Exercise 1.98 — Absolutely continuous probabilities
Prove Proposition 1.97 (Karr, 1993, p. 34). •

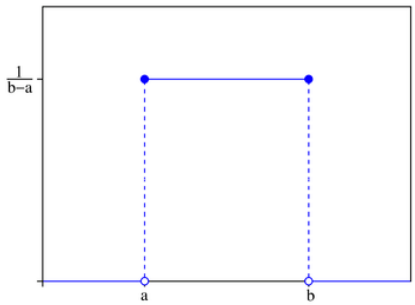
Example 1.99 — Absolutely continuous probabilities

- **Uniform distribution** on $[a, b]$ $(a, b \in \mathbb{R}, \; a < b)$

$$f_P(x) = \begin{cases} \frac{1}{b-a}, & a \leq x \leq b \\ 0, & \text{otherwise.} \end{cases}$$

$$F_P(x) = \begin{cases} 0, & x < a \\ \frac{x-a}{b-a}, & a \leq x \leq b \\ 1, & x > b. \end{cases}$$

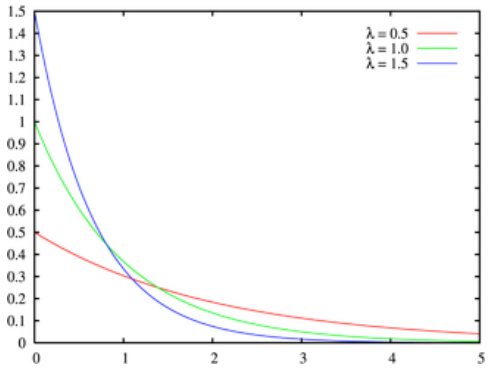
This absolutely continuous probability is such that all intervals of the same length on the distribution's support are equally probable. The support is defined



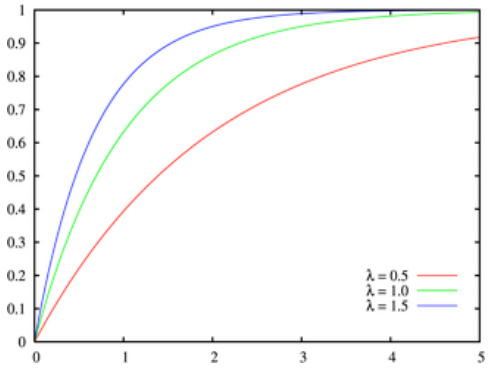
by the two parameters, a and b, which are its minimum and maximum values (http://en.wikipedia.org/wiki/Uniform_distribution_(continuous)).

- **Exponential distribution** with parameter λ ($\lambda \in \mathbb{R}^+$)

$$f_P(x) = \begin{cases} \lambda e^{-\lambda x}, & x \geq 0 \\ 0, & \text{otherwise.} \end{cases}$$



$$F_P(x) = \begin{cases} 0, & x < 0 \\ 1 - e^{-\lambda x}, & x \geq 0. \end{cases}$$



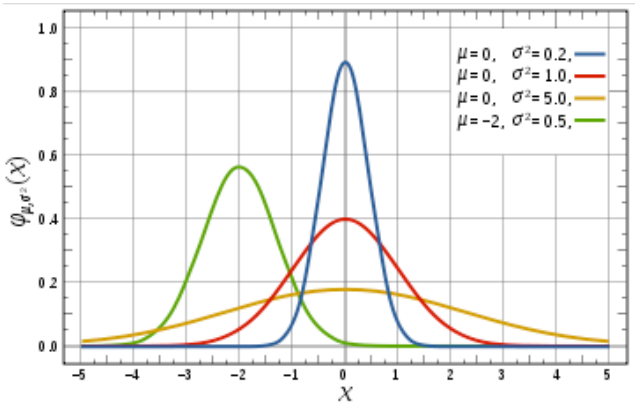
The exponential distribution is used to describe the times between consecutive events in a Poisson process.[15] (http://en.wikipedia.org/wiki/Exponential_distribution).

Let P^* be the (discrete) Poisson probability with parameter $\lambda\, x$. Then $P^*(\{0\}) = e^{-\lambda\, x} = P((x, +\infty)) = 1 - F_P(x)$.

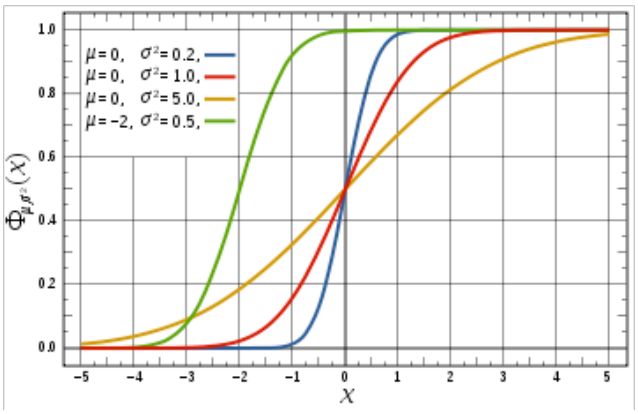
- **Normal distribution** with parameters μ ($\mu \in \mathbb{R}$) and σ^2 ($\sigma^2 \in \mathbb{R}^+$)

$f_P(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$, $x \in \mathbb{R}$

[15] I.e. a process in which events occur continuously and independently at a constant average rate.



$F_P(x) = \int_{-\infty}^{x} f_P(s)ds,\ x \in \mathbb{R}$



The normal distribution or Gaussian distribution is used describes data that cluster around a mean or average. The graph of the associated probability density function is bell-shaped, with a peak at the mean, and is known as the Gaussian function or bell curve http://en.wikipedia.org/wiki/Normal_distribution). •

Motivation 1.100 — **Mixed distributions** (Karr, 1993, p. 34)
A probability function need not to be discrete or absolutely continuous... •

Definition 1.101 — **Mixed distributions** (Karr, 1993, p. 34)
A probability function P is mixed if there is a discrete probability P_d, an absolutely continuous probability P_c and $\alpha \in (0, 1)$ such that P is a convex combination of P_d and P_c, that is,

$$P = \alpha P_d + (1 - \alpha)P_c. \tag{1.70}$$

•

Example 1.102 — Mixed distributions

Let $M(\lambda)/M(\mu)/1$ represent a queueing system with Poisson arrivals (rate $\lambda > 0$) and exponential service times (rate $\mu > 0$) and one server.

In the equilibrium state, the probability function associated to the waiting time of an arriving customer is

$$P(A) = (1-\rho)\epsilon_{\{0\}}(A) + \rho P_{Exp(\mu(1-\rho))}(A),\ A \in \mathcal{B}(I\!R), \tag{1.71}$$

where $0 < \rho = \frac{\lambda}{\mu} < 1$ and

$$P_{Exp(\mu(1-\rho))}(A) = \int_A \mu(1-\rho)\, e^{-\mu(1-\rho)\, s} ds. \tag{1.72}$$

The associated distribution function is given by

$$F_P(x) = \begin{cases} 0, & x < 0 \\ (1-\rho) + \rho\left[1 - e^{-\mu(1-\rho)\, x}\right], & x \geq 0. \end{cases} \tag{1.73}$$

•

1.5 Conditional probability

Motivation 1.103 — Conditional probability (Karr, 1993, p. 35)
We shall revise probabilities to account for the knowledge that an event has occurred, using a concept known as conditional probability. •

Definition 1.104 — Conditional probability (Karr, 1993, p. 35)
Let A and B be events. If $P(A) > 0$ the *conditionally probability of B given A* is equal to

$$P(B|A) = \frac{P(B \cap A)}{P(A)}. \tag{1.74}$$

In case $P(A) = 0$, we make the convention that $P(B|A) = P(B)$. •

Remark 1.105 — Conditional probability (Karr, 1993, p. 35)
$P(B|A)$ can be interpreted as the relative likelihood that B occurs given that A is known to have occured. •

Exercise 1.106 — Conditional probability
Solve exercises 1.23 and 1.24 of Karr (1993, p. 40). •

Example 1.107 — Conditional probability (Grimmett and Stirzaker, 2001, p. 9)
A family has two children.

- What is the probability that both are boys, given that at least one is a boy?

The older and younger child may each be male or female, so there are four possible combination of sexes, which we assume to be equally likely. Therefore

- $\Omega = \{GG, GB, BG, BB\}$

where $G =$ girl, $B =$ boy, and $P(GG) = P(GB) = P(BG) = P(BB) = \frac{1}{4}$.

From the definition of conditional probability

$$\begin{aligned} P(BB|\text{one boy at least}) &= P[BB|(GB \cup BG \cup BB)] \\ &= \frac{P[BB \cap (GB \cup BG \cup BB)]}{P(GB \cup BG \cup BB)} \\ &= \frac{P(BB)}{P(GB) + P(BG) + P(BB)} \\ &= \frac{\frac{1}{4}}{\frac{1}{4} + \frac{1}{4} + \frac{1}{4}} \\ &= \frac{1}{3}. \end{aligned} \tag{1.75}$$

A popular but incorrect answer to this question is $\frac{1}{2}$.

This is the correct answer to another question:

- For a family with two children, what is the probability that both are boys given that the younger is a boy?

In this case

$$\begin{aligned} P(BB|\text{younger child is a boy}) &= P[BB|(GB \cup BB)] \\ &= \dots \\ &= \frac{P(BB)}{P(GB) + P(BB)} \\ &= \frac{\frac{1}{4}}{\frac{1}{4} + \frac{1}{4}} \\ &= \frac{1}{2}. \end{aligned} \tag{1.76}$$

•

Exercise 1.108 — Conditional probability (Grimmett and Stirzaker, 2001, p. 9) The prosecutor's fallacy[16] — Let G be the event that an accused is guilty, and T the event that some testimony is true.

Some lawyers have argued on the assumption that $P(G|T) = P(T|G)$. Show that this holds iff $P(G) = P(T)$. •

Motivation 1.109 — Multiplication rule (Montgomery and Runger, 2003, p. 42) The definition of conditional probability can be rewritten to provide a general expression for the probability of the intersection of (two) events. This formula is referred to as a multiplication rule for probabilities. •

Proposition 1.110 — Multiplication rule (Montgomery and Runger, 2003, p. 43) Let A and B be two events. Then

$$P(A \cap B) = P(B|A) \times P(A) = P(A|B) \times P(B). \tag{1.77}$$

More generally: let $A_1, \dots, A_n$ be events then

$$\begin{aligned} P(A_1 \cap A_2 \cap \dots \cap A_{n-1} \cap A_n) &= P(A_1) \times P(A_2|A_1) \times P[A_3|(A_1 \cap A_2)] \\ &\quad \dots \times P[A_n|(A_1 \cap A_2 \cap \dots A_{n-1})]. \end{aligned} \tag{1.78}$$

•

[16] The prosecution made this error in the famous Dreyfus affair (http://en.wikipedia.org/wiki/Alfred_Dreyfus) in 1894.

Example 1.111 — Multiplication rule (Montgomery and Runger, 2003, p. 43)
The probability that an automobile battery, subject to high engine compartment temperature, suffers low charging current is 0.7. The probability that a battery is subject to high engine compartment temperature 0.05.

What is the probability that a battery suffers low charging current and is subject to high engine compartment temperature?

- **Table of events and probabilities**

Event	Probability
C = battery suffers low charging current	$P(C) = ?$
T = battery subject to high engine compartment temperature	$P(T) = 0.05$
$C\|T$ = battery suffers low charging current given that it is subject to high engine compartment temperature	$P(C\|T) = 0.7$

- **Probability**

$$\begin{aligned} P(C \cap T) &\overset{mult.\ rule}{=} P(C|T) \times P(T) \\ &= 0.7 \times 0.05 \\ &= 0.035. \end{aligned} \tag{1.79}$$

•

Motivation 1.112 — Law of total probability (Karr, 1993, p. 35)
The next law expresses the probability of an event in terms of its conditional probabilities given elements of a partition of Ω. •

Proposition 1.113 — Law of total probability (Karr, 1993, p. 35)
Let $\{A_1, A_2, \ldots\}$ a countable partition of Ω. Then, for each event B,

$$P(B) = \sum_{i=1}^{+\infty} P(B|A_i) \times P(A_i). \tag{1.80}$$

•

Exercise 1.114 — Law of total probability
Prove Proposition 1.113 by using $\sigma-$additivity of a probability function and the fact that $B = \bigcup_{i=1}^{+\infty}(B \cap A_i)$ (Karr, 1993, p. 36). •

Corollary 1.115 — Law of total probability (Montgomery and Runger, 2003, p. 44)
For any events A and B,

$$P(B) = P(B|A) \times P(A) + P(B|A^c) \times P(A^c). \tag{1.81}$$

•

Example 1.116 — Law of total probability (Grimmett and Stirzaker, 2001, p. 11)
Only two factories manufacture zoggles. 20% of the zoggles from factory I and 5% from factory II are defective. Factory I produces twice as many zoggles as factory II each week.

What is the probability that a zoggle, randomly chosen from a week's production, is not defective?

- **Table of events and probabilities**

Event	Probability
D = defective zoggle	$P(D) = ?$
A = zoggle made in factory I	$P(A) = 2 \times [1 - P(A)] = \frac{2}{3}$
A^c = zoggle made in factory II	$P(A^c) = 1 - P(A) = \frac{1}{3}$
$D\|A$ = defective zoggle given that it is made in factory I	$P(D\|A) = 0.20$
$D\|A^c$ = defective zoggle given that it is made in factory II	$P(D\|A^c) = 0.05$

- **Probability**

$$\begin{aligned} P(D^c) &= 1 - P(D) \\ &\overset{lawtotalprob}{=} 1 - [P(D|A) \times P(A) + P(D|A^c) \times P(A^c)] \\ &= 1 - \left(0.20 \times \frac{2}{3} + 0.05 \times \frac{1}{3}\right) \\ &= \frac{51}{60}. \end{aligned}$$

•

Motivation 1.117 — Bayes' theorem (Karr, 1993, p. 36)
Traditionally (and probably incorrectly) attributed to the English cleric Thomas Bayes (http://en.wikipedia.org/wiki/Thomas_Bayes), the theorem that bears his name is used to compute conditional probabilities "the other way around". •

Proposition 1.118 — Bayes' theorem (Karr, 1993, p. 36)
Let $\{A_1, A_2, \ldots\}$ be a countable partition of Ω. Then, for each event B with $P(B) > 0$ and each n,

$$\begin{aligned} P(A_n|B) &= \frac{P(B|A_n)P(A_n)}{P(B)} \\ &= \frac{P(B|A_n)P(A_n)}{\sum_{i=1}^{+\infty} P(B|A_i)P(A_i)} \end{aligned} \tag{1.82}$$

•

Exercise 1.119 — Bayes' theorem (Karr, 1993, p. 36)
Prove Bayes' theorem by using the definition of conditional probability and the law of total probability (Karr, 1993, p. 36). •

Example 1.120 — Bayes' theorem (Grimmett and Stirzaker, 2003, p. 11)
Resume Example 1.116...

If the chosen zoggle is defective, what is the probability that it came from factory II.

- **Probability**

$$\begin{aligned} P(A|D) &= \frac{P(D|A) \times P(A)}{P(D)} \\ &= \frac{0.20 \times \frac{2}{3}}{1 - \frac{51}{60}} \\ &= \frac{8}{9}. \end{aligned} \tag{1.83}$$

•

References

- Grimmett, G.R. and Stirzaker, D.R. (2001). *Probability and Random Processes (3rd. edition)*. Oxford. (QA274.12-.76.GRI.30385 and QA274.12-.76.GRI.40695 refer to the library code of the 1st. and 2nd. editions from 1982 and 1992, respectively.)

- Karr, A.F. (1993). *Probability*. Springer-Verlag.

- Lusin, N. (1927). Sur les ensembles analytiques. *Fundamenta Mathematicae* **10**, 1–95.

- Montgomery, D.C. and Runger, G.C. (2003). *Applied statistics and probability for engineers*. John Wiley & Sons, New York. (QA273-280/3.MON.64193)

- Papoulis, A. (1965). *Probability, Random Variables and Stochastic Processes*. McGraw-Hill Kogakusha, Ltd. (QA274.12-.76.PAP.28598)

- Resnick, S.I. (1999). *A Probability Path*. Birkhäuser. (QA273.4-.67.RES.49925)

- Righter, R. (200–). Lectures notes for the course *Probability and Risk Analysis for Engineers*. Department of Industrial Engineering and Operations Research, University of California at Berkeley.

- Yates, R.D. and Goodman, D.J. (1999). Probability and Stochastic Processes: A friendly Introduction for Electrical and Computer Engineers. John Wiley & Sons, Inc. (QA273-280/4.YAT.49920)

Chapter 2

Random variables

2.1 Fundamentals

Motivation 2.1 — **Inverse image of sets** (Karr, 1993, p. 43)
Before we introduce the concept of random variable (r.v.) we have to talk rather extensively on inverse images of sets and inverse image mapping. •

Definition 2.2 — **Inverse image** (Karr, 1993, p. 43)
Let:

- X be a function with domain Ω and range Ω', i.e. $X : \Omega \to \Omega'$;
- $\mathcal{F}$ and $\mathcal{F}'$ be the $\sigma -$ algebras on Ω and Ω', respectively.

(Frequently $\Omega' = I\!R$ and $\mathcal{F}' = \mathcal{B}(I\!R)$.) Then the inverse image under X of the set $B \in \mathcal{F}'$ is the subset of Ω given by

$$X^{-1}(B) = \{\omega : X(\omega) \in B\}, \tag{2.1}$$

written from now on $\{X \in B\}$ (graph!).

•

Remark 2.3 — **Inverse image mapping** (Karr, 1993, p. 43)
The inverse image mapping X^{-1} maps subsets of Ω' to subsets of Ω. X^{-1} preserves all set operations, as well as disjointness. •

Proposition 2.4 — Properties of inverse image mapping (Karr, 1993, p. 43; Resnick, 1999, p. 72)
Let:

- $X : \Omega \to \Omega'$;
- $\mathcal{F}$ and $\mathcal{F}'$ be the $\sigma-$algebras on Ω and Ω', respectively;
- B, B' and $\{B_i : i \in I\}$ be sets in $\mathcal{F}'$.

Then:

1. $X^{-1}(\emptyset) = \emptyset$
2. $X^{-1}(\Omega') = \Omega$
3. $B \subseteq B' \Rightarrow X^{-1}(B) \subseteq X^{-1}(B')$
4. $X^{-1}(\bigcup_{i\in I} B_i) = \bigcup_{i\in I} X^{-1}(B_i)$
5. $X^{-1}(\bigcap_{i\in I} B_i) = \bigcap_{i\in I} X^{-1}(B_i)$
6. $B \cap B' = \emptyset \Rightarrow X^{-1}(B) \cap X^{-1}(B') = \emptyset$
7. $X^{-1}(B^c) = [X^{-1}(B)]^c$. •

Exercise 2.5 — Properties of inverse image mapping
Prove Proposition 2.4 (Karr, 1993, p. 43). •

Proposition 2.6 — $\sigma-$algebras and inverse image mapping (Resnick, 1999, pp. 72–73)
Let $X : \Omega \to \Omega'$ be a mapping with inverse image. If $\mathcal{F}'$ is a $\sigma-$algebra on Ω' then

$$X^{-1}(\mathcal{F}') = \{X^{-1}(B) : B \in \mathcal{F}'\} \tag{2.2}$$

is a $\sigma-$algebra on Ω.

Exercise 2.7 — $\sigma-$algebras and inverse image mapping
Prove Proposition 2.6 by verifying the 3 postulates for a $\sigma-$algebra (Resnick, 1999, p. 73). •

Proposition 2.8 — Inverse images of $\sigma-$ algebras generated by classes of subsets (Resnick, 1999, p. 73)
Let $\mathcal{C}'$ be a class of subsets of Ω'. Then

$$X^{-1}(\sigma(\mathcal{C}')) = \sigma(\{X^{-1}(\mathcal{C}')\}), \tag{2.3}$$

i.e., the inverse image of the $\sigma-$ algebra generated by $\mathcal{C}'$ is the same as the $\sigma-$ algebra on Ω generated by the inverse images. •

Exercise 2.9 — Inverse images of $\sigma-$ algebras generated by classes of subsets
Prove Proposition 2.8. This proof comprises the verification of the 3 postulates for a $\sigma-$ algebra (Resnick, 1999, pp. 73–74) and much more. •

Definition 2.10 — Measurable space (Resnick, 1999, p. 74)
The pair $(\Omega, \mathcal{F})$ consisting of a set Ω and a $\sigma-$ algebra on Ω is called a measurable space. •

Definition 2.11 — Measurable map (Resnick, 1999, p. 74)
Let $(\Omega, \mathcal{F})$ and $(\Omega', \mathcal{F}')$ be two measurable spaces. Then a map $X : \Omega \to \Omega'$ is called a measurable map if

$$X^{-1}(\mathcal{F}') \subseteq \mathcal{F}. \tag{2.4}$$

•

Remark 2.12 — Measurable maps/ Random variables (Karr, 1993, p. 44)
A special case occurs when $(\Omega', \mathcal{F}') = (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ — in this case X is called a random variable. That is, random variables are functions on the sample space Ω for which inverse images of Borel sets are events of Ω. •

Definition 2.13 — Random variable (Karr, 1993, p. 44)
Let $(\Omega, \mathcal{F})$ and $(\Omega', \mathcal{F}') = (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ be two measurable spaces. A random variable (r.v.) is a function $X : \Omega \to \mathbb{R}$ such that

$$X^{-1}(B) \in \mathcal{F}, \forall B \in \mathcal{B}(\mathbb{R}). \tag{2.5}$$

•

Remark 2.14 — Random variables (Karr, 1993, p. 44)
A r.v. is a function on the sample space: it transforms events into real sets.

The technical requirement that sets $\{X \in B\} = X^{-1}(B)$ be events of Ω is needed in order that the probability

$$P(\{X \in B\}) = P(X^{-1}(B)) \tag{2.6}$$

be defined. •

Motivation 2.15 — Checking if X is a r.v. (Karr, 1993, p. 47)
To verify that X is a r.v. it is not necessary to check that $\{X \in B\} = X^{-1}(B) \in \mathcal{F}$ for all Borel sets B. •

Proposition 2.16 — Checking if X is a r.v. (Resnick, 1999, p. 77; Karr, 1993, p. 47)
The real function $X : \Omega \to I\!R$ is a r.v. iff

$$\{X \leq x\} = X^{-1}((-\infty, x]) \in \mathcal{F}, \forall x \in I\!R. \tag{2.7}$$

Similarly if we replace $\{X \leq x\}$ by $\{X > x\}$, $\{X < x\}$ or $\{X \geq x\}$. •

Example 2.17 — Random variable

- **Random experiment**

  Throw a traditional fair die and observe the number of points.

- **Sample space**

  $\Omega = \{1, 2, 3, 4, 5, 6\}$

- **$\sigma-$algebra on Ω**

  Let us consider a non trivial one:

  $\mathcal{F} = \{\emptyset, \{1, 3, 5\}, \{2, 4, 6\}, \Omega\}$

- **Random variable**

  $X : \Omega \to I\!R$ such that: $X(1) = X(3) = X(5) = 0$ and $X(2) = X(4) = X(6) = 1$

- **Inverse image mapping**

  Let $B \in \mathcal{B}(\mathbb{R})$. Then

$$\begin{aligned} X^{-1}(B) &= \begin{cases} \emptyset, & \text{if } 0 \notin B, 1 \notin B \\ \{1,3,5\}, & \text{if } 0 \in B, 1 \notin B \\ \{2,4,6\}, & \text{if } 0 \notin B, 1 \in B \\ \Omega, & \text{if } 0 \in B, 1 \in B \end{cases} \\ &\in \mathcal{F}, \forall B \in \mathcal{B}(\mathbb{R}). \end{aligned} \tag{2.8}$$

  Therefore X is a r.v. defined in $\mathcal{F}$.

- **A function which is not a r.v.**

  $Y : \Omega \to \mathbb{R}$ such that: $Y(1) = Y(2) = Y(3) = 1$ and $Y(4) = Y(5) = Y(6) = 0$.

  Y is not a r.v. defined in $\mathcal{F}$ because $Y^{-1}(\{1\}) = \{1,2,3\} \notin \mathcal{F}$. •

There are generalizations of r.v.

Definition 2.18 — **Random vector** (Karr, 1993, p. 45)
A $d-$dimensional random vector is a function $\underline{X} = (X_1, \ldots, X_d) : \Omega \to \mathbb{R}^d$ such that each component X_i, $i = 1, \ldots, d$, is a random variable. •

Remark 2.19 — **Random vector** (Karr, 1993, p. 45)
Random vectors will sometimes be treated as finite sequences of random variables. •

Definition 2.20 — **Stochastic process** (Karr, 1993, p. 45)
A stochastic process with index set (or parameter space) T is a collection $\{X_t : t \in T\}$ of r.v. (indexed by T). •

Remark 2.21 — **Stochastic process** (Karr, 1993, p. 45)
Typically:

- $T = \mathbb{N}_0$ and $\{X_n : n \in \mathbb{N}_0\}$ is called a discrete time stochastic process;
- $T = \mathbb{R}_0^+$ and $\{X_t : t \in \mathbb{R}_0^+\}$ is called a continuous time stochastic process. •

Proposition 2.22 — **$\sigma-$algebra generated by a r.v.** (Karr, 1993, p. 46)
The family of events that are inverse images of Borel sets under a r.v is a $\sigma-$algebra on Ω. In fact, given a r.v. X, the family

$$\sigma(X) = \{X^{-1}(B) : B \in \mathcal{B}(\mathbb{R})\} \tag{2.9}$$

is a $\sigma-$algebra on Ω, known as the $\sigma-$algebra generated by X. •

Remark 2.23 — **$\sigma-$algebra generated by a r.v.**

- Proposition 2.22 is a particular case of Proposition 2.6 when $\mathcal{F}' = \mathcal{B}(\mathbb{R})$.
- Moreover, $\sigma(X)$ is a $\sigma-$algebra for every function $X : \Omega \to \mathbb{R}$; and X is a r.v. iff $\sigma(X) \subseteq \mathcal{F}$, i.e., iff X is a measurable map (Karr, 1993, p. 46). •

Example 2.24 — **$\sigma-$algebra generated by a constant r.v.**
Let $X : \Omega \to \mathbb{R}$ such that $X(\omega) = c$, $\forall\omega \in \Omega$. Then

$$X^{-1}(B) = \begin{cases} \emptyset, & \text{if } c \notin B \\ \Omega, & \text{if } c \in B, \end{cases} \tag{2.10}$$

for any $B \in \mathcal{B}(\mathbb{R})$, and $\sigma(X) = \{\emptyset, \Omega\}$ (trivial $\sigma-$algebra). •

Example 2.25 — **$\sigma-$algebra generated by an indicator r.v.** (Karr, 1993, p. 46)
Let:

- A be a subset of the sample space Ω;
- $X : \Omega \to \mathbb{R}$ such that

$$X(\omega) = \mathbf{1}_A(w) = \begin{cases} 1, & \omega \in A \\ 0, & \omega \notin A. \end{cases} \tag{2.11}$$

Then X is the indicator r.v. of an event A. In addition,

$$\sigma(X) = \sigma(\mathbf{1}_A) = \{\emptyset, A, A^c, \Omega\} \tag{2.12}$$

since

$$X^{-1}(B) = \begin{cases} \emptyset, & \text{if } 0 \notin B, 1 \notin B \\ A^c, & \text{if } 0 \in B, 1 \notin B \\ A, & \text{if } 0 \notin B, 1 \in B \\ \Omega, & \text{if } 0 \in B, 1 \in B, \end{cases} \tag{2.13}$$

for any $B \in \mathcal{B}(\mathbb{R})$. •

Example 2.26 — $\sigma-$**algebra generated by a simple r.v.** (Karr, 1993, pp. 45-46) A simple r.v. takes only finitely many values and has the form

$$X = \sum_{i=1}^{n} a_i \times \mathbf{1}_{A_i}, \tag{2.14}$$

where a_i, $i = 1, \ldots, n$, are (not necessarily distinct) real numbers and A_i, $i = 1, \ldots, n$, are events that constitute a partition of Ω. X is a r.v. since

$$\{X \in B\} = \bigcup_{i=1}^{n} \{A_i : a_i \in B\}, \tag{2.15}$$

for any $B \in \mathcal{B}(\mathbb{R})$.

For this simple r.v. we get

$$\begin{aligned} \sigma(X) &= \sigma(\{A_1, \ldots, A_n\}) \\ &= \{\bigcup_{i \in I} A_i : I \subseteq \{1, \ldots, n\}\}, \end{aligned} \tag{2.16}$$

regardless of the values of $a_1, \ldots, a_n$. •

Definition 2.27 — $\sigma-$**algebra generated by a random vector** (Karr, 1993, p. 46) The $\sigma-$algebra generated by the $d-$dimensional random vector $(X_1, \ldots, X_d) : \Omega \to \mathbb{R}^d$ is given by

$$\sigma((X_1, \ldots, X_d)) = \{(X_1, \ldots, X_d)^{-1}(B) : B \in \mathcal{B}(\mathbb{R}^d)\}. \tag{2.17}$$

•

2.2 Combining random variables

To work with r.v., we need assurance that algebraic, limiting and transformation operations applied to them yield other r.v.

In the next proposition we state that the set of r.v. is closed under:

- addition and scalar multiplication;[1]
- maximum and minimum;
- multiplication;
- division.

Proposition 2.28 — Closure under algebraic operations (Karr, 1993, p. 47)
Let X and Y be r.v. Then:

1. $aX + bY$ is a r.v., for all $a, b \in \mathbb{R}$;
2. $\max\{X, Y\}$ and $\min\{X, Y\}$ are r.v.;
3. XY is a r.v.;
4. $\frac{X}{Y}$ is a r.v. provided that $Y(\omega) \neq 0$, $\forall \omega \in \Omega$. •

Exercise 2.29 — Closure under algebraic operations
Prove Proposition 2.28 (Karr, 1993, pp. 47–48). •

Corollary 2.30 — Closure under algebraic operations (Karr, 1993, pp. 48–49)
Let $X : \Omega \to \mathbb{R}$ be a r.v. Then

$$X^+ = \max\{X, 0\} \tag{2.18}$$
$$X^- = -\min\{X, 0\}, \tag{2.19}$$

the positive and negative parts of X (respectively), are non-negative r.v., and so is

$$|X| = X^+ + X^-. \tag{2.20}$$

•

Remark 2.31 — Canonical representation of a r.v. (Karr, 1993, p. 49)
A r.v. can be written as a difference of its positive and negative parts:

$$X = X^+ - X^-. \tag{2.21}$$

•

[1] I.e. the set of r.v. is a vector space.

Theorem 2.32 — Closure under limiting operations (Karr, 1993, p. 49)
Let $X_1, X_2, \ldots$ be r.v. Then $\sup X_n$, $\inf X_n$, $\limsup X_n$ and $\liminf X_n$ are r.v. Consequently if

$$X(\omega) = \lim_{n \to +\infty} X_n(\omega) \tag{2.22}$$

exists for every $\omega \in \Omega$, then X is also a r.v. •

Exercise 2.33 — Closure under limiting operations
Prove Theorem 2.32 by noting that

$$\begin{aligned} \{\sup X_n \leq x\} &= (\sup X_n)^{-1}((-\infty, x]) \\ &= \bigcap_{n=1}^{+\infty} \{X_n \leq x\} \\ &= \bigcap_{n=1}^{+\infty} (X_n)^{-1}((-\infty, x]) \end{aligned} \tag{2.23}$$

$$\begin{aligned} \{\inf X_n \geq x\} &= (\inf X_n)^{-1}([x, +\infty)) \\ &= \bigcap_{n=1}^{+\infty} \{X_n \geq x\} \\ &= \bigcap_{n=1}^{+\infty} (X_n)^{-1}([x, +\infty)) \end{aligned} \tag{2.24}$$

$$\limsup X_n = \inf_k \sup_{n \geq k} X_n \tag{2.25}$$

$$\liminf X_n = \sup_k \inf_{n \geq k} X_n \tag{2.26}$$

and that when $X = \lim_{n \to +\infty} X_n$ exists, $X = \limsup X_n = \liminf X_n$ (Karr, 1993, p. 49). •

Corollary 2.34 — Series of r.v. (Karr, 1993, p. 49)
If $X_1, X_2, \ldots$ are r.v., then provided that $X(\omega) = \sum_{n=1}^{+\infty} X_n(\omega)$ converges for each ω, X is a r.v. •

Motivation 2.35 — Transformations of r.v. and random vectors (Karr, 1993, p. 50)
Another way of constructing r.v. is as functions of other r.v. •

Definition 2.36 — Borel measurable function (Karr, 1993, p. 66)
A function $g : \mathbb{R}^n \to \mathbb{R}^m$ (for fixed $n, m \in \mathbb{N}$) is Borel measurable if

$$g^{-1}(B) \in \mathcal{B}(\mathbb{R}^n), \forall B \in \mathcal{B}(\mathbb{R}^m). \tag{2.27}$$

•

Remark 2.37 — Borel measurable function (Karr, 1993, p. 66)

- In order that $g : \mathbb{R}^n \to \mathbb{R}$ be Borel measurable it suffices that

$$g^{-1}((-\infty, x]) \in \mathcal{B}(\mathbb{R}^n), \forall x \in \mathbb{R}. \tag{2.28}$$

- A function $g : \mathbb{R}^n \to \mathbb{R}^m$ Borel measurable iff each of its components is Borel measurable as a function from $\mathbb{R}^n$ to $\mathbb{R}$.
- Indicator functions, monotone functions and continuous functions are Borel measurable.
- Moreover, the class of Borel measurable function has the closure properties under algebraic and limiting operations as the family of r.v. on a probability space $(\Omega, \mathcal{F}, P)$. •

Theorem 2.38 — Transformations of random vectors (Karr, 1993, p. 50)
Let:

- $X_1, \ldots, X_d$ be r.v.;
- $g : \mathbb{R}^d \to \mathbb{R}$ be a Borel measurable function.

Then $Y = g(X_1, \ldots, X_d)$ is a r.v. •

Exercise 2.39 — Transformations of r.v.
Prove Theorem 2.38 (Karr, 1993, p. 50). •

Corollary 2.40 — Transformations of r.v. (Karr, 1993, p. 50)
Let:

- X be r.v.;
- $g : \mathbb{R} \to \mathbb{R}$ be a Borel measurable function.

Then $Y = g(X)$ is a r.v. •

2.3 Distributions and distribution functions

The main importance of probability functions on $\mathbb{R}$ is that they are distributions of r.v.

Proposition 2.41 — R.v. and probabilities on $\mathbb{R}$ (Karr, 1993, p. 52)
Let X be a r.v. and P a p.f. defined of $(\Omega, \mathcal{F})$. Then the set function

$$P_X(B) = P(X^{-1}(B)) = P(\{X \in B\}) \tag{2.29}$$

is a probability function on $\mathbb{R}$. •

Exercise 2.42 — R.v. and probabilities on $\mathbb{R}$
Prove Proposition 2.41 by checking if the three axioms in the definition of probability function hold (Karr, 1993, p. 52). •

Definition 2.43 — Distribution, distribution and survival function of a r.v. (Karr, 1993, p. 52)
Let X be a r.v. Then

1. the probability function on $\mathbb{R}$

   $P_X(B) = P(X^{-1}(B)) = P(\{X \in B\})$, $B \in \mathcal{B}(\mathbb{R})$, is the distribution of X;

2. $F_X(x) = P_X((-\infty, x]) = P(X^{-1}((-\infty, x]) = P(\{X \leq x\})$, $x \in \mathbb{R}$, is the distribution function of X;

3. $S_X(x) = 1 - F_X(x) = P_X((x, +\infty)) = P(X^{-1}((x, +\infty)) = P(\{X > x\})$, $x \in \mathbb{R}$, is the survival (or survivor) function of X. •

Definition 2.44 — Discrete/absolutely continuous/mixed r.v. (Karr, 1993, p. 52)
X is said to be a discrete/absolutely continuous/mixed r.v. if P_X is a discrete/absolutely continuous/mixed p.f. •

Motivation 2.45 — Confronting r.v.
How can we confront two r.v. X and Y ? •

Definition 2.46 — **Identically distributed r.v.** (Karr, 1993, p. 52)
Let X and Y be two r.v. Then X and Y are said to be identically distributed — written $X \stackrel{d}{=} Y$ — if

$$\begin{aligned} P_X(B) &= P(\{X \in B\}) \\ &= P(\{Y \in B\}) = P_Y(B),\ B \in \mathcal{B}(\mathbb{R}), \end{aligned} \tag{2.30}$$

i.e. if $F_X(x) = P(\{X \leq x) = P(\{Y \leq x\}) = F_Y(x),\ x \in \mathbb{R}$. •

Definition 2.47 — **Equal r.v. almost surely** (Karr, 1993, p. 52; Resnick, 1999, p. 167)
Let X and Y be two r.v. Then X is equal to Y almost surely — written $X \stackrel{a.s.}{=} Y$ — if

$$\begin{aligned} P(\{X = Y\}) &= P(\{\omega \in \Omega : X(\omega) = Y(\omega)\}) \\ &= 1. \end{aligned} \tag{2.31}$$

•

Remark 2.48 — **Identically distributed r.v. vs. equal r.v. almost surely** (Karr, 1993, p. 52)
Equality in distribution of X and Y has no bearing on their equality as functions on Ω, i.e.

$$X \stackrel{d}{=} Y \not\Rightarrow X \stackrel{a.s.}{=} Y, \tag{2.32}$$

even though

$$X \stackrel{a.s.}{=} Y \Rightarrow X \stackrel{d}{=} Y. \tag{2.33}$$

•

Example 2.49 — **Identically distributed r.v. vs. equal r.v. almost surely**

- $X \sim$ Bernoulli(0.5)

  $P(\{X = 0\}) = P(\{X = 1\}) = 0.5$

- $Y = 1 - X \sim$ Bernoulli(0.5) since

  $P(\{Y = 0\}) = P(\{1 - X = 0\}) = P(\{X = 1\}) = 0.5$

  $P(\{Y = 1\}) = P(\{1 - X = 1\}) = P(\{X = 0\}) = 0.5$

- $X \stackrel{d}{=} Y$ but $X \stackrel{a.s.}{\neq} Y$. •

Exercise 2.50 — **Identically distributed r.v. vs. equal r.v. almost surely**
Prove that $X \overset{a.s.}{=} Y \Rightarrow X \overset{d}{=} Y$. •

Definition 2.51 — **Distribution and distribution function of a random vector** (Karr, 1993, p. 53)
Let $\underline{X} = (X_1, \ldots, X_d)$ be a $d-$dimensional random vector. Then

1. the probability function on $\mathbb{R}^d$

   $P_{\underline{X}}(B) = P(\underline{X}^{-1}(B)) = P(\{\underline{X} \in B\})$, $B \in \mathcal{B}(\mathbb{R}^d)$, is the distribution of $\underline{X}$;

2. the distribution function of $\underline{X} = (X_1, \ldots, X_d)$, also known as the joint distribution function of $X_1, \ldots, X_d$ is the function $F_{\underline{X}} : \mathbb{R}^d \to [0, 1]$ given by

$$\begin{aligned} F_{\underline{X}}(\underline{x}) &= F_{(X_1,\ldots,X_d)}(x_1, \ldots, x_d) \\ &= P(\{X_1 \leq x_1, \ldots, X_d \leq x_d\}), \end{aligned} \tag{2.34}$$

   for any $\underline{x} = (x_1, \ldots, x_d) \in \mathbb{R}^d$. •

Remark 2.52 — **Distribution function of a random vector** (Karr, 1993, p. 53)
The distribution $P_{\underline{X}}$ is determined uniquely by $F_{\underline{X}}$. •

Motivation 2.53 — **Marginal distribution function** (Karr, 1993, p. 53)
Can we obtain the distribution of X_i from the joint distribution function? •

Proposition 2.54 — **Marginal distribution function** (Karr, 1993, p. 53)
Let $\underline{X} = (X_1, \ldots, X_d)$ be a $d-$dimensional random vector. Then, for each i ($i = 1, \ldots, d$) and x ($x \in \mathbb{R}$),

$$F_{X_i}(x) = \lim_{x_j \to +\infty, j \neq i} F_{(X_1,\ldots,X_{i-1},X_i,X_{i+1},\ldots,X_d)}(x_1, \ldots, x_{i-1}, x, x_{i+1}, \ldots, x_d). \tag{2.35}$$

•

Exercise 2.55 — **Marginal distribution function**
Prove Proposition 2.54 by noting that $\{X_1 \leq x_1, \ldots, X_{i-1} \leq x_{i-1}, X_i \leq x, X_{i+1} \leq x_{i+1}, \ldots, X_d \leq x_d\} \uparrow \{X_i \leq x\}$ when $x_j \to +\infty$, $j \neq i$, and by considering the monotone continuity of probability functions (Karr, 1993, p. 53). •

Definition 2.56 — **Discrete random vector** (Karr, 1993, pp. 53–54)
The random vector $\underline{X} = (X_1, \ldots, X_d)$ is said to be discrete if $X_1, \ldots, X_d$ are discrete r.v. i.e. if there is a countable set $\mathcal{C} \subset \mathbb{R}^d$ such that $P(\{\underline{X} \in \mathcal{C}\}) = 1$. •

Definition 2.57 — **Absolutely continuous random vector** (Karr, 1993, pp. 53–54)
The random vector $\underline{X} = (X_1, \ldots, X_d)$ is absolutely continuous if there is a non-negative function $f_{\underline{X}} : \mathbb{R}^d \to \mathbb{R}_0^+$ such that

$$F_{\underline{X}}(\underline{x}) = \int_{-\infty}^{x_1} \ldots \int_{-\infty}^{x_d} f_{\underline{X}}(s_1, \ldots, s_d)\, ds_d \ldots ds_1, \tag{2.36}$$

for every $\underline{x} = (x_1, \ldots, x_d) \in \mathbb{R}^d$. $f_{\underline{X}}$ is called the joint density function of $(X_1, \ldots, X_d)$. •

Proposition 2.58 — **Absolutely continuous random vector; marginal density function** (Karr, 1993, p. 54)
If $\underline{X} = (X_1, \ldots, X_d)$ is absolutely continuous then, for each i $(i = 1, \ldots, d)$, X_i is absolutely continuous and

$$f_{X_i}(x) = \int_{-\infty}^{+\infty} \ldots \int_{-\infty}^{+\infty} f_{\underline{X}}(s_1, \ldots, s_{i-1}, x, s_{i+1}, \ldots, s_d)\, ds_d \ldots ds_{i-1} ds_{i+1} \ldots ds_1. \tag{2.37}$$

f_{X_i} is termed the marginal density function of X_i. •

Remark 2.59 — **Absolutely continuous random vector** (Karr, 1993, p. 54)
If the random vector is absolutely continuous then any "sub-vector" is absolutely continuous. Moreover, the converse of Proposition 2.58 is not true, that is, the fact that $X_1, \ldots, X_d$ are absolutely continuous does not imply that $(X_1, \ldots, X_d)$ is an absolutely continuous random vector. •

2.4 Key r.v. and random vectors and distributions

2.4.1 Discrete r.v. and random vectors

Integer-valued r.v. like the Bernoulli, binomial, geometric, negative binomial, hypergeometric and Poisson, and integer-valued random vectors like the multinomial are discrete r.v. and random vectors of great interest.

- **Uniform distribution** on a finite set

Notation	$X \sim \text{Uniform}(\{x_1, x_2, \ldots, x_n\})$
Parameter	$\{x_1, x_2, \ldots, x_n\}$ $(x_i \in I\!R,\ i = 1, \ldots, n)$
Range	$\{x_1, x_2, \ldots, x_n\}$
P.f.	$P(\{X = x\}) = \frac{1}{n},\ x = x_1, x_2, \ldots, x_n$

  This simple r.v. has the form $X = \sum_{i=1}^{n} x_i \times \mathbf{1}_{\{x_i\}}$.

- **Bernoulli distribution**

Notation	$X \sim \text{Bernoulli}(p)$
Parameter	$p = P(\text{sucess})$ $(p \in [0, 1])$
Range	$\{0, 1\}$
P.f.	$P(\{X = x\}) = p^x(1-p)^{1-x},\ x = 0, 1$

  A Bernoulli distributed r.v. X is the indicator function of the event $\{X = 1\}$.

- **Binomial distribution**

Notation	$X \sim \text{Binomial}(n, p)$
Parameters	$n =$ number of Bernoulli trials $(n \in I\!N)$
	$p = P(\text{sucess})$ $(p \in [0, 1])$
Range	$\{0, 1, \ldots, n\}$
P.f.	$P(\{X = x\}) = \binom{n}{x} p^x(1-p)^{n-x},\ x = 0, 1, \ldots, n$

  The binomial r.v. results from the sum of n i.i.d. Bernoulli distributed r.v.

- **Geometric distribution**

Notation	$X \sim \text{Geometric}(p)$
Parameter	$p = P(\text{sucess})$ $(p \in [0,1])$
Range	$\mathbb{N} = \{1, 2, 3, \ldots\}$
P.f.	$P(\{X = x\}) = (1-p)^{x-1}\, p$, $x = 1, 2, 3, \ldots$

This r.v. satisfies the *lack of memory property*:

$$P(\{X > k + x\} | \{X > k\}) = P(\{X > x\}),\ \forall k, x \in \mathbb{N}. \tag{2.38}$$

- **Negative binomial distribution**

Notation	$X \sim \text{NegativeBinomial}(r, p)$
Parameters	$r =$ pre-specified number of sucesses $(r \in \mathbb{N})$ $p = P(\text{sucess})$ $(p \in [0,1])$
Range	$\{r, r+1, \ldots\}$
P.f.	$P(\{X = x\}) = \binom{x-1}{r-1}(1-p)^{x-r}p^r$, $x = r, r+1, \ldots$

The negative binomial r.v. results from the sum of r i.i.d. geometrically distributed r.v.

- **Hypergeometric distribution**

Notation	$X \sim \text{Hypergeometric}(N, M, n)$
Parameters	$N =$ population size $(N \in \mathbb{N})$ $M =$ sub-population size $(M \in \mathbb{N}, M \leq N)$ $n =$ sample size $(n \in \mathbb{N}, n \leq N)$
Range	$\{\max\{0, n-N+M\}, \ldots, \min\{n, M\ \}\}$
P.f.	$P(\{X = x\}) = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}}$, $x = \max\{0, n-N+M\}, \ldots, \min\{n, M\ \}$

Note that the sample is collected without replacement. Otherwise $X \sim \text{Binomial}(n, \frac{M}{N})$.

- **Poisson distribution**

Notation	$X \sim \text{Poisson}(\lambda)$
Parameter	λ ($\lambda \in \mathbb{R}^+$)
Range	$\mathbb{N}_0 = \{0, 1, 2, 3, \ldots\}$
P.f.	$P(\{X = x\}) = e^{-\lambda}\frac{\lambda^x}{x!}$, $x = 0, 1, 2, 3, \ldots$

The distribution was proposed by Siméon-Denis Poisson (1781–1840) and published, together with his probability theory, in 1838 in his work *Recherches sur la probabilité des jugements en matiéres criminelles et matiére civile* (Research on the probability of judgments in criminal and civil matters). The Poisson distribution can be derived as a limiting case of the binomial distribution.[2]

In 1898 Ladislaus Josephovich Bortkiewicz (1868–1931) published a book titled *The Law of Small Numbers*. In this book he first noted that events with low frequency in a large population follow a Poisson distribution even when the probabilities of the events varied. It was that book that made the Prussian horse-kick data famous. Some historians of mathematics have even argued that the Poisson distribution should have been named the *Bortkiewicz distribution*.[3]

- **Multinomial distribution**

In probability theory, the multinomial distribution is a generalization of the binomial distribution when we are dealing not only with two types of events — a success with probability p and a failure with probability $1-p$ — but with d types of events with probabilities $p_1, \ldots, p_d$ such that $p_1, \ldots, p_d \geq 0, \sum_{i=1}^{d} p_i = 1$.[4]

Notation	$\underline{X} = (X_1, \ldots, X_d) \sim \text{Multinomial}_{d-1}(n, (p_1, \ldots, p_d))$
Parameters	n = number of Bernoulli trials ($n \in \mathbb{N}$) $(p_1, \ldots, p_d)$ where $p_i = P(\text{event of type } i)$ $(p_1, \ldots, p_d \geq 0, \sum_{i=1}^{d} p_i = 1)$
Range	$\{(n_1, \ldots, n_d) \in \mathbb{N}_0^d : \sum_{i=1}^{d} n_i = n\}$
P.f.	$P(\{X_1 = n_1, \ldots, X_d = n_d\}) = \frac{n!}{\prod_{i=1}^{d} n_i!} \prod_{i=1}^{d} p_i^{n_i}$, $(n_1, \ldots, n_d) \in \mathbb{N}_0^d : \sum_{i=1}^{d} n_i = n$

[2] http://en.wikipedia.org/wiki/Poisson_distribution
[3] http://en.wikipedia.org/wiki/Ladislaus_Bortkiewicz
[4] http://en.wikipedia.org/wiki/Multinomial_distribution

Exercise 2.60 — Binomial r.v. (Grimmett and Stirzaker, 2001, p. 25)
DNA fingerprinting — In a certain style of detective fiction, the sleuth is required to declare *the criminal has the unusual characteristics...; find this person you have your man.* Assume that any given individual has these unusual characteristics with probability 10^{-7} (independently of all other individuals), and the city in question has 10^7 inhabitants.

Given that the police inspector finds such person, what is the probability that there is at least one other? •

Exercise 2.61 — Binomial r.v. (Righter, 200–)
A student (Fred) is getting ready to take an important oral exam and is concerned about the possibility of having an *on* day or an *off* day. He figures that if he has an *on* day, then each of his examiners will pass him independently of each other, with probability 0.8, whereas, if he has an *off* day, this probability will be reduced to 0.4.

Suppose the student will pass if a majority of examiners pass him. If the student feels that he is twice as likely to have an *off* day as he is to have an *on* day, should he request an examination with 3 examiners or with 5 examiners? •

Exercise 2.62 — Geometric r.v.
Prove that the distribution function of $X \sim \text{Geometric}(p)$ is given by

$$F_X(x) = P(X \leq x) = \begin{cases} 0, & x < 1 \\ \sum_{i=1}^{[x]} (1-p)^{i-1} p = 1 - (1-p)^{[x]}, & x \geq 1, \end{cases} \tag{2.39}$$

where $[x]$ represents the integer part of x. •

Exercise 2.63 — Hypergeometric r.v. (Righter, 200–)
From a mix of 50 widgets from supplier 1 and 100 from supplier 2, 10 widgets are randomly selected and shipped to a customer.

What is the probability that all 10 came from supplier 1? •

Exercise 2.64 — Poisson r.v. (Grimmett and Stirzaker, 2001, p. 19)
In your pocket is a random number N of coins, where $N \sim \text{Poisson}(\lambda)$. You toss each coin once, with heads showing with probability p each time.

Show that the total number of heads has a Poisson distribution with parameter λp. •

Exercise 2.65 — Negative hypergeometric r.v. (Grimmett and Stirzaker, 2001, p. 19)

Capture-recapture — A population of N animals has had a number M of its members captured, marked, and released. Let X be the number of animals it is necessary to recapture (without re-release) in order to obtain r marked animals.

Show that

$$P(\{X = x\}) = \frac{\frac{M}{N}\binom{M-1}{r-1}\binom{N-M}{x-r}}{\binom{N-1}{x-1}}. \tag{2.40}$$

•

Exercise 2.66 — Discrete random vectors

Prove that if

- $Y \sim \text{Poisson}(\lambda)$
- $(X_1, \ldots, X_d)|\{Y = n\} \sim \text{Multinomial}_{d-1}(n, (p_1, \ldots, p_d))$

then $X_i \sim \text{Poisson}(\lambda p_i)$, $i = 1, \ldots, d$. •

Exercise 2.67 — Relating the p.f. of the negative binomial and binomial r.v.

Let $X \sim \text{NegativeBinomial}(r, p)$ and $Y \sim \text{Binomial}(x - 1, p)$. Prove that, for $x = r, r + 1, r + 2, \ldots$ and $r = 1, 2, 3, \ldots$, we get

$$\begin{aligned} P(X = x) &= p \times P(Y = r - 1) \\ &= p \times \left[F_{Binomial(x-1,p)}(r - 1) - F_{Binomial(x-1,p)}(r - 2)\right]. \end{aligned} \tag{2.41}$$

•

Exercise 2.68 — Relating the d.f. of the negative binomial and binomial r.v.

Let $X \sim \text{NegativeBinomial}(r, p)$, $Y \sim \text{Binomial}(x, p)$ e $Z = x - Y \sim \text{Binomial}(x, 1 - p)$. Prove that, for $x = r, r + 1, r + 2, \ldots$ and $r = 1, 2, 3, \ldots$, we have

$$\begin{aligned} F_{NegativeBinomial(r,p)}(x) &= P(X \leq x) \\ &= P(Y \geq r) \\ &= 1 - F_{Binomial(x,p)}(r - 1) \\ &= P(Z \leq x - r) \\ &= F_{Binomial(x,1-p)}(x - r). \end{aligned} \tag{2.42}$$

•

2.4.2 Absolutely continuous r.v. and random vectors

- **Uniform distribution** on the interval $[a, b]$

Notation	$X \sim \text{Uniform}(a, b)$
Parameters	$a =$ minimum value ($a \in \mathbb{R}$) $b =$ maximum value ($b \in \mathbb{R}$, $a < b$)
Range	$[a, b]$
P.d.f.	$f_X(x) = \frac{1}{b-a}$, $a \leq x \leq b$

Let X be an absolutely continuous r.v. with d.f. $F_X(x)$. Then $Y = F_X(X) \sim \text{Uniform}(0, 1)$.

- **Beta distribution**

In probability theory and statistics, the beta distribution is a family of continuous probability distributions defined on the interval [0, 1] parameterized by two positive shape parameters, typically denoted by α and β. In Bayesian statistics, it can be seen as the posterior distribution of the parameter p of a binomial distribution, if the prior distribution of p was uniform. It is also used in information theory, particularly for the information theoretic performance analysis for a communication system.

Notation	$X \sim \text{Beta}(\alpha, \beta)$
Parameters	α ($\alpha \in \mathbb{R}^+$) β ($\beta \in \mathbb{R}^+$)
Range	$[0, 1]$
P.d.f.	$f_X(x) = \frac{1}{B(\alpha,\beta)} x^{\alpha-1} (1-x)^{\beta-1}$, $0 \leq x \leq 1$

where

$$B(\alpha, \beta) = \int_0^1 x^{\alpha-1} (1-x)^{\beta-1} dx \tag{2.43}$$

represents the beta function. Note that

$$B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}, \tag{2.44}$$

where

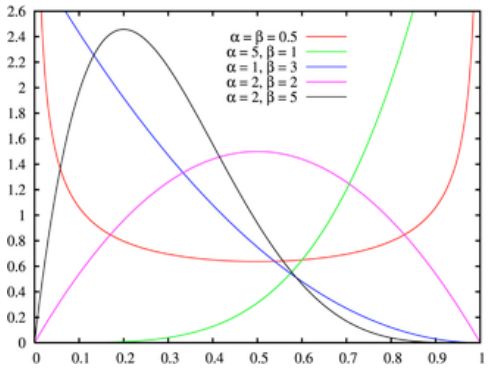
$$\Gamma(\alpha) = \int_0^{+\infty} y^{\alpha-1}\, e^{-y} dy \tag{2.45}$$

is the Euler's gamma function.

The uniform distribution on $[0, 1]$ is a particular case of the beta distribution — $\alpha = \beta = 1$. Moreover, the beta distribution can be generalized to the interval $[a, b]$:

$$f_Y(y) = \frac{1}{B(\alpha, \beta)} \frac{(y-a)^{\alpha-1}\,(b-y)^{\beta-1}}{(b-a)^{\alpha+\beta-1}},\ a \leq y \leq b. \tag{2.46}$$

The p.d.f. of this distribution can take various forms on account of the "shape" parameters a and b, as illustrated by the following graph and table:



Parameters	Shape of the beta p.d.f.
$\alpha,\ \beta > 1$	Unique mode at $x = \frac{\alpha-1}{\alpha+\beta-2}$
$\alpha < 1,\ \beta > 1$	Unique anti-mode at $x = \frac{\alpha-1}{\alpha+\beta-2}$ $(U-\text{shape})$
$(\alpha-1)(\beta-1) \leq 0$	$J-\text{shape}$
$\alpha = \beta$	Symmetric around 1/2 (e.g. constant ou parabolic)
$\alpha < \beta$	Positively assymmetric
$\alpha > \beta$	Negatively assymmetric

Exercise 2.69 — Relating the Beta and Binomial distributions

(a) Prove that the d.f. of the r.v. $X \sim \text{Beta}(\alpha, \beta)$ can be written in terms of the d.f. of Binomial r.v. when α and β are integer-valued:

$$F_{Beta(\alpha,\beta)}(x) = 1 - F_{Binomial(\alpha+\beta-1,x)}(\alpha - 1). \tag{2.47}$$

(b) Prove that the p.d.f. of the r.v. $X \sim \text{Beta}(\alpha, \beta)$ can be rewritten in terms of the p.f. of the r.v. $Y \sim \text{Binomial}(\alpha + \beta - 2, x)$, when α and β are integer-valued:

$$\begin{aligned} f_{Beta(\alpha,\beta)}(x) &= (\alpha+\beta-1) \times P(Y = \alpha - 1) \\ &= (\alpha+\beta-1) \times \left[F_{Binomial(\alpha+\beta-2,x)}(\alpha-1) \right. \\ &\qquad \left. - F_{Binomial(\alpha+\beta-2,x)}(\alpha-2)\right]. \end{aligned} \tag{2.48}$$

•

- **Normal distribution**

The normal distribution or Gaussian distribution is a continuous probability distribution that describes data that cluster around a mean or average. The graph of the associated probability density function is bell-shaped, with a peak at the mean, and is known as the Gaussian function or bell curve. The Gaussian distribution is one of many things named after Carl Friedrich Gauss, who used it to analyze astronomical data, and determined the formula for its probability density function. However, Gauss was not the first to study this distribution or the formula for its density function that had been done earlier by Abraham de Moivre.

Notation	$X \sim \text{Normal}(\mu, \sigma^2)$
Parameters	μ ($\mu \in I\!R$) σ^2 ($\sigma^2 \in I\!R^+$)
Range	$I\!R$
P.d.f.	$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$, $-\infty < x < +\infty$

The normal distribution can be used to describe, at least approximately, any variable that tends to cluster around the mean. For example, the heights of adult males in the United States are roughly normally distributed, with a mean of about 1.8 m. Most men have a height close to the mean, though a small number of outliers have a height significantly above or below the mean. A histogram of male heights will appear similar to a bell curve, with the correspondence becoming closer if more data are used. (http://en.wikipedia.org/wiki/Normal_distribution).

Standard normal distribution — Let $X \sim \text{Normal}(\mu, \sigma^2)$. Then the r.v. $Z = \frac{X-E(X)}{\sqrt{V(X)}} = \frac{X-\mu}{\sigma}$ is said to have a standard normal distribution, i.e. $Z \sim \text{Normal}(0, 1)$. Moreover, Z has d.f. given by

$$F_Z(z) = P(Z \leq z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt = \Phi(z), \tag{2.49}$$

and

$$\begin{aligned} F_X(x) &= P(X \leq x) \\ &= P\left(Z = \frac{X-\mu}{\sigma} \leq \frac{x-\mu}{\sigma}\right) \\ &= \Phi\left(\frac{x-\mu}{\sigma}\right). \end{aligned} \tag{2.50}$$

- **Exponential distribution**

  The exponential distributions are a class of continuous probability distributions. They tend to be used to describe the times between events in a Poisson process, i.e. a process in which events occur continuously and independently at a constant average rate (http://en.wikipedia.org/wiki/Exponential_distribution).

Notation	$X \sim \text{Exponential}(\lambda)$
Parameter	$\lambda =$ inverse of the scale parameter ($\lambda \in I\!R^+$)
Range	$I\!R_0^+ = [0, +\infty)$
P.d.f.	$f_X(x) = \lambda\, e^{-\lambda x}, x \geq 0$

  Consider $X \sim$ Exponencial(λ). Then

$$P(X > t + x | X > t) = P(X > x),\ \forall t, x \in I\!R_0^+. \tag{2.51}$$

  Equivalently,

$$(X - t | X > t) \sim \text{Exponencial}(\lambda),\ \forall t \in I\!R_0^+. \tag{2.52}$$

  This property is referred as to *lack of memory*: no matter how old your equipment is, its remaining life has same distribution as a new one.

  The exponential (resp. geometric) distribution is the only absolutely continuous (resp. discrete) r.v. satisfying this property.

- **Gamma distribution**

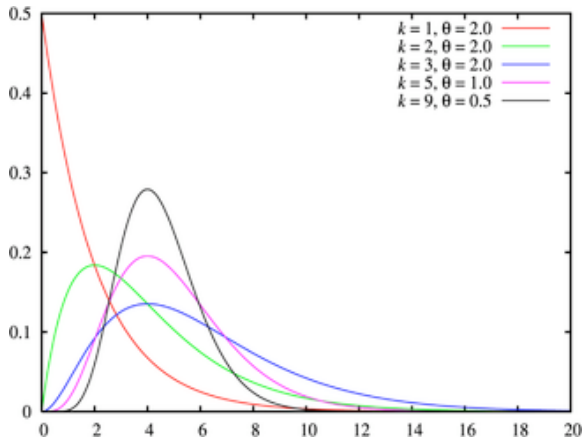
  The gamma distribution is frequently a probability model for waiting times; for instance, in life testing, the waiting time until death is a random variable that is frequently modeled with a gamma distribution (http://en.wikipedia.org/wiki/Gamma_distribution).

Notation	$X \sim \text{Gamma}(\alpha, \beta)$
Parameters	$\alpha =$ shape parameter ($\alpha \in \mathbb{R}^+$)
	$\beta =$ inverse of the scale parameter ($\beta \in \mathbb{R}^+$)
Range	$\mathbb{R}_0^+ = [0, +\infty)$
P.d.f.	$f_X(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}, x \geq 0$

Special cases

- Exponential — $\alpha = 1$ which has the *lack of memory property* as the geometric distribution in the discrete case;
- Erlang — $\alpha \in \mathbb{N}$;[5]
- Chi-square with n degrees of freedom — $\alpha = n/2$, $\beta = 1/2$.

This distribution has a shape parameter α, therefore it comes as no surprise the sheer variety of forms of the gamma p.d.f. in the following graph.



Parameters	Shape of the gamma p.d.f.
$\alpha < 1$	Unique supremum at $x = 0$
$\alpha = 1$	Unique mode at $x = 0$
$\alpha > 1$	Unique mode at $x = \frac{\alpha-1}{\beta}$ and positively assymmetric

The gamma distribution stand in the same relation to exponential as negative binomial to geometric: sums of i.i.d exponential r.v. have gamma distribution. χ^2 distributions result from sums of squares of independent standard normal r.v.

[5]The Erlang distribution was developed by Agner Krarup Erlang (1878–1929) to examine the number of telephone calls which might be made at the same time to the operators of the switching stations. This work on telephone traffic engineering has been expanded to consider waiting times in queueing systems in general. The distribution is now used in the fields of stochastic processes and of biomathematics (http://en.wikipedia.org/wiki/Erlang_distribution)

It is possible to relate the d.f. of $X \sim \text{Erlang}(n, \beta)$ with the d.f. of a Poisson r.v.:

$$\begin{aligned} F_{Erlang(n,\beta)}(x) &= \sum_{i=n}^{\infty} e^{-\beta x}(\beta x)^i / i! \\ &= 1 - F_{Poisson(\beta x)}(n-1),\ x > 0,\ n \in I\!N. \end{aligned} \tag{2.53}$$

- **$d-$dimensional uniform distribution**

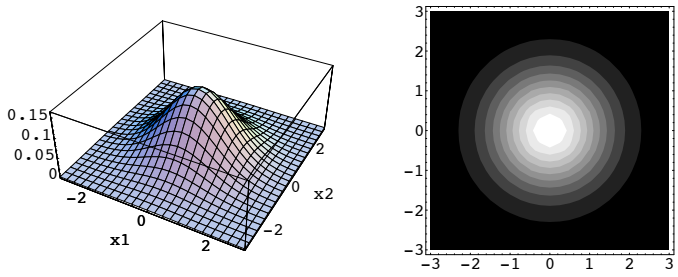
Notation	$\underline{X} \sim \text{Uniform}([0,1]^d)$
Range	$[0,1]^d$
P.d.f.	$f_{\underline{X}}(\underline{x}) = 1,\ \underline{x} \in [0,1]^d$

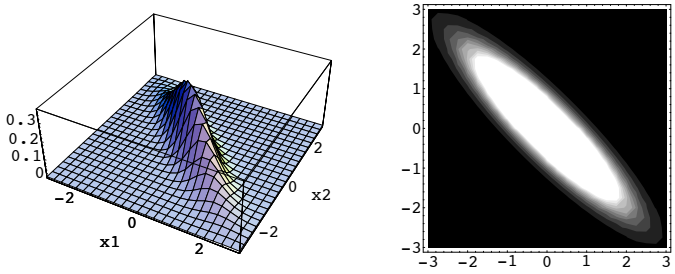
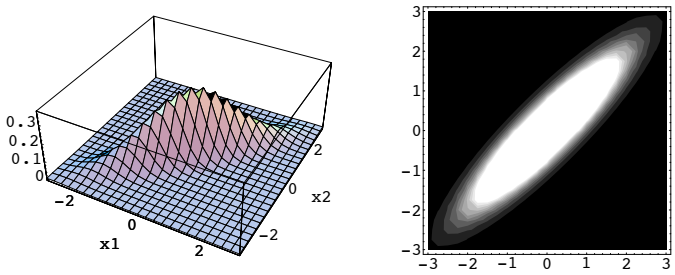
- **Bivariate STANDARD normal distribution**

Notation	$\underline{X} \sim \text{Normal}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}\right)$
Parameter	$\rho =$ correlation between X_1 and X_2 $(-1<\rho<1)$
Range	$I\!R^2$
P.d.f.	$f_{\underline{X}}(\underline{x}) = f_{(X_1,X_2)}(x_1,x_2) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2}\frac{x_1^2 - 2\rho x_1 x_2 + x_2^2}{1-\rho^2}\right),\ \underline{x} \in I\!R^2$

The graphical representation of the joint density of a random vector with a bivariate standard normal distribution follows — it depends on the parameter ρ.

Case	**Graph and contour plot of the joint p.d.f. of a bivariate STANDARD normal**
$\rho = 0$	Circumferences centered in $(0,0)$



Case	**Graph and contour plot of the joint p.d.f. of a bivariate STANDARD normal** (cont.)
$\rho < 0$	Ellipses centered in $(0, 0)$ and asymmetric in relation to the axes, suggesting that X_2 decreases when X_1 increases 
$\rho > 0$	Ellipses centered in $(0, 0)$ and asymmetric in relation to the axes, suggesting that X_2 increases when X_1 increases 

Both components of $\underline{X} = (X_1, X_2)$ have standard normal marginal densities and X_1 and X_2 are independent iff $\rho = 0$.

2.5 Transformation theory

2.5.1 Transformations of r.v., general case

Motivation 2.70 — Transformations of r.v., general case (Karr, 1993, p. 60)
Let:

- X be a r.v. with d.f. F_X;
- $Y = g(X)$ be a transformation of X under g, where $g : I\!R \to I\!R$ is a Borel measurable function.

Then we know that $Y = g(X)$ is also a r.v. But this is manifestly not enough: we wish to know

- how the d.f. of Y relates to that of X?

This question admits an obvious answer when g is invertible and in a few other cases described below. •

Proposition 2.71 — D.f. of a transformation of a r.v., general case (Rohatgi, 1976, p. 68; Murteira, 1979, p. 121)
Let:

- X be a r.v. with d.f. F_X;
- $Y = g(X)$ be a transformation of X under g, where $g : I\!R \to I\!R$ is a Borel measurable function;
- $g^{-1}((-\infty, y]) = \{x \in I\!R : g(x) \leq y\}$ be the inverse image of the Borel set $(-\infty, y]$ under g.

Then

$$\begin{aligned} F_Y(y) &= P(\{Y \leq y\}) \\ &= P(\{X \in g^{-1}((-\infty, y])\}). \end{aligned} \tag{2.54}$$

•

Exercise 2.72 — D.f. of a transformation of a r.v., general case
Prove Proposition 2.71 (Rohatgi, 1976, p. 68).

Note that if g is a Borel measurable function then

$$g^{-1}(B) \in \mathcal{B}(I\!R),\ \forall B = (-\infty, y] \in \mathcal{B}(I\!R). \tag{2.55}$$

Thus, we are able to write

$$P(\{Y \in B\}) = P(\{g(X) \in B\}) = P(\{X \in g^{-1}(B)\}). \tag{2.56}$$

•

Remark 2.73 — D.f. of a transformation of a r.v., general case
Proposition 2.71 relates the d.f. of Y to that of X.

The inverse image $g^{-1}((-\infty, y])$ is a Borel set and tends to be a "reasonable" set — a real interval or a union of real intervals. •

Exercise 2.74 — D.f. of a transformation of a r.v., general case (Karr, 1993, p. 70, Exercise 2.20(a))
Let X be a r.v. and $Y = X^2$. Prove that

$$F_Y(y) = F_X(\sqrt{y}) - F_X[-(\sqrt{y})^-], \tag{2.57}$$

for $y \geq 0$. •

Exercise 2.75 — D.f. of a transformation of a r.v., general case (Rohatgi, 1976, p. 68)
Let X be a r.v. with d.f. F_X. Derive the d.f. of the following r.v.:

(a) $|X|$

(b) $aX + b$

(c) e^X. •

Exercise 2.76 — D.f. of a transformation of a r.v., absolutely continuous case
The electrical resistance[6] (X) of an object and its electrical conductance[7] (Y) are related as follows: $Y = X^{-1}$.

Assuming that $X \sim \text{Uniform}(900\,ohm, 1100\,ohm)$:

(a) Identify the range of values of the r.v. Y.

(b) Derive the survival function of Y, $P(Y > y)$, and calculate $P(Y > 10^{-3}\,mho)$. •

[6]The electrical resistance of an object is a measure of its opposition to the passage of a steady electric current. The SI unit of electrical resistance is the *ohm* (http://en.wikipedia.org/wiki/Electrical_resistance).

[7]Electrical conductance is a measure of how easily electricity flows along a certain path through an electrical element. The SI derived unit of conductance is the *siemens* (also called the *mho*, because it is the reciprocal of electrical resistance, measured in ohms). Oliver Heaviside coined the term in September 1885 (http://en.wikipedia.org/wiki/Electrical_conductance).

Exercise 2.77 — D.f. of a transformation of a r.v., absolutely continuous case Let $X \sim \text{Uniform}(0, 2\pi)$ and $Y = \sin X$. Prove that

$$F_Y(y) = \begin{cases} 0, & y < -1 \\ \frac{1}{2} + \frac{\arcsin y}{\pi}, & -1 \leq y \leq 1 \\ 1, & y > 1. \end{cases} \tag{2.58}$$

•

2.5.2 Transformations of discrete r.v.

Proposition 2.78 — P.f. of a one-to-one transformation of a discrete r.v. (Rohatgi, 1976, p. 69)
Let:

- X be a discrete r.v. with p.f. $P(\{X = x\})$;
- $\mathcal{R}_X$ be a countable set such that $P(\{X \in \mathcal{R}_X\}) = 1$ and $P(\{X = x\}) > 0$, $\forall x \in \mathcal{R}_X$;
- $Y = g(X)$ be a transformation of X under g, where $g : \mathbb{R} \to \mathbb{R}$ is a one-to-one Borel measurable function that transforms $\mathcal{R}_X$ onto some set $\mathcal{R}_Y = g(\mathcal{R}_X)$.

Then the inverse map, g^{-1}, is a single-valued function of y and

$$P(\{Y = y\}) = \begin{cases} P(\{X = g^{-1}(y)\}), & y \in \mathcal{R}_Y \\ 0, & \text{otherwise.} \end{cases} \tag{2.59}$$

•

Exercise 2.79 — P.f. of a one-to-one transformation of a discrete r.v. (Rohatgi, 1976, p. 69)
Let $X \sim \text{Poisson}(\lambda)$. Obtain the p.f. of $Y = X^2 + 3$. •

Exercise 2.80 — P.f. of a one-to-one transformation of a discrete r.v.
Let $X \sim \text{Binomial}(n, p)$ and $Y = n - X$. Prove that:

- $Y \sim \text{Binomial}(n, 1 - p)$;
- $F_Y(y) = 1 - F_X(n - y - 1)$, $y = 0, 1, \ldots, n$. •

Remark 2.81 — P.f. of a transformation of a discrete r.v. (Rohatgi, 1976, p. 69)
Actually the restriction of a single-valued inverse on g is not necessary. If g has a finite (or even a countable) number of inverses for each y, from the countable additivity property of probability functions we can obtain the p.f. of the r.v. $Y = g(X)$. •

Proposition 2.82 — P.f. of a transformation of a discrete r.v. (Murteira, 1979, p. 122)
Let:

- X be a discrete r.v. with p.f. $P(\{X = x\})$;
- $\mathcal{R}_X$ be a countable set such that $P(\{X \in \mathcal{R}_X\}) = 1$ and $P(\{X = x\}) > 0$, $\forall x \in \mathcal{R}_X$;
- $Y = g(X)$ be a transformation of X under g, where $g : I\!R \to I\!R$ is a Borel measurable function that transforms $\mathcal{R}_X$ onto some set $\mathcal{R}_Y = g(\mathcal{R}_X)$;
- $\mathcal{A}_y = \{x \in \mathcal{R}_X : g(x) = y\}$ be a non empty set, for $y \in \mathcal{R}_Y$.

Then

$$\begin{aligned} P(\{Y = y\}) &= P(\{X \in \mathcal{A}_y\}) \\ &= \sum_{x \in \mathcal{A}_y} P(\{X = x\}), \end{aligned} \tag{2.60}$$

for $y \in \mathcal{R}_Y$. •

Exercise 2.83 — P.f. of a transformation of a discrete r.v. (Rohatgi, 1976, pp. 69–70)
Let X be a discrete r.v. with p.f.

$$P(\{X = x\}) = \begin{cases} \frac{1}{5}, & x = -2 \\ \frac{1}{6}, & x = -1 \\ \frac{1}{5}, & x = 0 \\ \frac{1}{15}, & x = 1 \\ \frac{11}{30}, & x = 2 \\ 0, & \text{otherwise} \end{cases} \tag{2.61}$$

Derive the p.f. of $Y = X^2$. •

2.5.3 Transformations of absolutely continuous r.v.

Proposition 2.84 — D.f. of a strictly monotonic transformation of an absolutely continuous r.v. (Karr, 1993, pp. 60 and 68)
Let:

- X be an absolutely continuous r.v. with d.f. F_X and p.d.f. f_X;
- $\mathcal{R}_X$ be the range of the r.v. X, i.e. $\mathcal{R}_X = \{x \in I\!R : f_X(x) > 0\}$;
- $Y = g(X)$ be a transformation of X under g, where $g : I\!R \to I\!R$ is a continuous, **strictly increasing**, Borel measurable function that transforms $\mathcal{R}_X$ onto some set $\mathcal{R}_Y = g(\mathcal{R}_X)$;
- g^{-1} be the pointwise inverse of g.

Then

$$F_Y(y) = F_X[g^{-1}(y)], \tag{2.62}$$

for $y \in \mathcal{R}_Y$. Similarly, if

- g is a continuous, **strictly decreasing**, Borel measurable function

then

$$F_Y(y) = 1 - F_X[g^{-1}(y)], \tag{2.63}$$

for $y \in \mathcal{R}_Y$. •

Exercise 2.85 — D.f. of a strictly monotonic transformation of an absolutely continuous r.v.
Prove Proposition 2.84 (Karr, 1993, p. 60). •

Exercise 2.86 — D.f. of a strictly monotonic transformation of an absolutely continuous r.v.
Let $X \sim \text{Normal}(0, 1)$. Derive the d.f. of

(a) $Y = e^X$

(b) $Y = \mu + \sigma X$, where $\mu \in I\!R$ and $\sigma \in I\!R^+$

(Karr, 1993, p. 60). •

Remark 2.87 — Transformations of absolutely continuous and discrete r.v. (Karr, 1993, p. 61)
in general, $Y = g(X)$ need not be absolutely continuous even when X is, as shown in the next exercise, while if X is a discrete r.v. then so is $Y = g(X)$ regardless of the Borel measurable function g. •

Exercise 2.88 — A mixed r.v. as a transformation of an absolutely continuous r.v.
Let $X \sim \text{Uniform}(-1, 1)$. Prove that $Y = X^+ = \max\{0, X\}$ is a mixed r.v. whose d.f. is given by

$$F_Y(y) = \begin{cases} 0, & y < 0 \\ \frac{1}{2}, & y = 0 \\ \frac{1}{2} + \frac{y}{2}, & 0 < y \leq 1 \\ 1, & y > 1 \end{cases} \tag{2.64}$$

(Rohatgi, 1976, p. 70). •

Exercise 2.88 shows that we need some conditions on g to ensure that $Y = g(X)$ is also an absolutely continuous r.v. This will be the case when g is a continuous monotonic function.

Theorem 2.89 — P.d.f. of a strictly monotonic transformation of an absolutely continuous r.v. (Rohatgi, 1976, p. 70; Karr, 1993, p. 61)
Suppose that:

- X is an absolutely continuous r.v. with p.d.f. f_X;
- there is an open subset $\mathcal{R}_X \subset \mathbb{R}$ such that $P(\{X \in \mathcal{R}_X\}) = 1$;
- $Y = g(X)$ is a transformation of X under g, where $g : \mathbb{R} \to \mathbb{R}$ is a continuously differentiable, Borel measurable function such that either $\frac{dg(x)}{dx} > 0$, $\forall x \in \mathcal{R}_X$, or $\frac{dg(x)}{dx} < 0$, $\forall x \in \mathcal{R}_X$;[8]
- g transforms $\mathcal{R}_X$ onto some set $\mathcal{R}_Y = g(\mathcal{R}_X)$;
- g^{-1} represents the pointwise inverse of g.

[8]This implies that $\frac{dg(x)}{dx} \neq 0$, $\forall x \in \mathcal{R}_X$.

Then $Y = g(X)$ is an absolutely continuous r.v. with p.d.f. given by

$$f_Y(y) = f_X[g^{-1}(y)] \times \left| \frac{dg^{-1}(y)}{dy} \right|, \tag{2.65}$$

for $y \in \mathcal{R}_Y$. •

Exercise 2.90 — P.d.f. of a strictly monotonic transformation of an absolutely continuous r.v.
Prove Theorem 2.89 by considering the case $\frac{dg(x)}{dx} > 0, \forall x \in \mathcal{R}_X$, applying Proposition 2.84 to derive the d.f. of $Y = g(X)$, and differentiating it to obtain the p.d.f. of Y (Rohatgi, 1976, p. 70). •

Remark 2.91 — P.d.f. of a strictly monotonic transformation of an absolutely continuous r.v. (Rohatgi, 1976, p. 71)
The key to computation of the induced d.f. of $Y = g(X)$ from the d.f. of X is $P(\{Y \leq y\}) = P(\{X \in g^{-1}((-\infty, y])\})$. If the conditions of Theorem 2.89 are satisfied, we are able to identify the set $\{X \in g^{-1}((-\infty, y])\}$ as $\{X \leq g^{-1}(y)\}$ or $\{X \geq g^{-1}(y)\}$, according to whether g in strictly increasing or strictly decreasing. •

Exercise 2.92 — P.d.f. of a strictly monotonic transformation of an absolutely continuous r.v.
Let $X \sim \text{Normal}(0, 1)$. Identify the p.d.f. and the distribution of

(a) $Y = e^X$

(b) $Y = \mu + \sigma X$, where $\mu \in \mathbb{R}$ and $\sigma \in \mathbb{R}^+$

(Karr, 1993, p. 61). •

Corollary 2.93 — P.d.f. of a strictly monotonic transformation of an absolutely continuous r.v. (Rohatgi, 1976, p. 71)
Under the conditions of Theorem 2.89, and by noting that

$$\frac{dg^{-1}(y)}{dy} = \left. \frac{1}{\frac{dg(x)}{dx}} \right|_{x=g^{-1}(y)}, \tag{2.66}$$

we conclude that the p.d.f. of $Y = g(X)$ can be rewritten as follows:

$$f_Y(y) = \left. \frac{f_X(x)}{\left| \frac{dg(x)}{dx} \right|} \right|_{x=g^{-1}(y)}, \tag{2.67}$$

$\forall y \in \mathcal{R}_Y$. •

Remark 2.94 — P.d.f. of a non monotonic transformation of an absolutely continuous r.v. (Rohatgi, 1976, p. 71)
In practice Theorem 2.89 is quite useful, but whenever its conditions are violated we should return to $P(\{Y \leq y\}) = P(\{X \in g^{-1}((-\infty, y])\})$ to obtain the $F_Y(y)$ and then differentiate this d.f. to derive the p.d.f. of the transformation Y. This is the case in the next two exercises. •

Exercise 2.95 — P.d.f. of a non monotonic transformation of an absolutely continuous r.v.
Let $X \sim \text{Normal}(0, 1)$ and $Y = g(X) = X^2$. Prove that $Y \sim \chi^2_{(1)}$ by noting that

$$F_Y(y) = F_X(\sqrt{y}) - F_X(-\sqrt{y}),\ y > 0 \tag{2.68}$$

$$f_Y(y) = \frac{dF_Y(y)}{dy} = \begin{cases} \frac{1}{2\sqrt{y}} \times \left[f_X(\sqrt{y}) + f_X(-\sqrt{y})\right], & y \geq 0 \\ 0, & y < 0 \end{cases} \tag{2.69}$$

(Rohatgi, 1976, p. 72). •

Exercise 2.96 — P.d.f. of a non monotonic transformation of an absolutely continuous r.v.
Let X be an absolutely continuous r.v. with p.d.f.

$$f_X(x) = \begin{cases} \frac{2x}{\pi^2}, & 0 < x < \pi \\ 0, & \text{otherwise} \end{cases} \tag{2.70}$$

Prove that $Y = \sin X$ has p.d.f. given by

$$f_Y(y) = \begin{cases} \frac{2}{\pi\sqrt{1-y^2}}, & 0 < y < 1 \\ 0, & \text{otherwise} \end{cases} \tag{2.71}$$

(Rohatgi, 1976, p. 73). •

Motivation 2.97 — P.d.f. of a sum of monotonic restrictions of a function g of an absolutely continuous r.v. (Rohatgi, 1976, pp. 73–74)
in the two last exercises the function $y = g(x)$ can be written as the sum of two monotonic restrictions of g in two disjoint intervals. Therefore we can apply Theorem 2.89 to each of these monotonic summands.

In fact, these two exercises are special cases of the following theorem. •

Theorem 2.98 — P.d.f. of a finite sum of monotonic restrictions of a function g of an absolutely continuous r.v. (Rohatgi, 1976, pp. 73–74)
Let:

- X be an absolutely continuous r.v. with p.d.f. f_X;
- $Y = g(X)$ be a transformation of X under g, where $g : \mathbb{R} \to \mathbb{R}$ is a Borel measurable function that transforms $\mathcal{R}_X$ onto some set $\mathcal{R}_Y = g(\mathcal{R}_X)$.

Moreover, suppose that:

- $g(x)$ is differentiable for all $x \in \mathcal{R}_X$;
- $\frac{dg(x)}{dx}$ is continuous and nonzero at all points of $\mathcal{R}_X$ but a finite number of x.

Then, for every real number $y \in \mathcal{R}_Y$,

(a) there is a positive integer $n = n(y)$ and real numbers (inverses) $g_1^{-1}(y), \ldots, g_n^{-1}(y)$ such that

$$g(x)\big|_{x=g_k^{-1}(y)} = y \quad \text{and} \quad \left.\frac{dg(x)}{dx}\right|_{x=g_k^{-1}(y)} \neq 0, \;\; k = 1, \ldots, n(y), \tag{2.72}$$

or

(b) there is not an x such that $g(x) = y$ and $\frac{dg(x)}{dx} \neq 0$, in which case we write $n = n(y) = 0$.

In addition, $Y = g(X)$ is an absolutely continuous r.v. with p.d.f. given by

$$f_Y(y) = \begin{cases} \sum_{k=1}^{n(y)} f_X[g_k^{-1}(y)] \times \left|\frac{dg_k^{-1}(y)}{dy}\right|, & n = n(y) > 0 \\ 0, & n = n(y) = 0, \end{cases} \tag{2.73}$$

for $y \in \mathcal{R}_Y$. •

Exercise 2.99 — P.d.f. of a finite sum of monotonic restrictions of a function g of an absolutely continuous r.v.
Let $X \sim \text{Uniform}(-1, 1)$. Use Theorem 2.98 to prove that $Y = |X| \sim \text{Uniform}(0, 1)$ (Rohatgi, 1976, p. 74). •

Exercise 2.100 — P.d.f. of a finite sum of monotonic restrictions of a function g of an absolutely continuous r.v.
Let $X \sim \text{Uniform}(0, 2\pi)$ and $Y = \sin X$. Use Theorem 2.98 to prove that

$$f_Y(y) = \begin{cases} \frac{1}{\pi\sqrt{1-y^2}}, & -1 < y < 1 \\ 0, & \text{otherwise.} \end{cases} \tag{2.74}$$

•

Motivation 2.101 — P.d.f. of a countable sum of monotonic restrictions of a function g of an absolutely continuous r.v.

The formula $P(\{Y \leq y\}) = P(\{X \in g^{-1}((-\infty, y])\})$ and the countable additivity of probability functions allows us to compute the p.d.f. of $Y = g(X)$ in some instance even if g has a countable number of inverses. •

Theorem 2.102 — P.d.f. of a countable sum of monotonic restrictions of a function g of an absolutely continuous r.v. (Rohatgi, 1976, pp. 74–75)

Let g be a Borel measurable function that maps $\mathcal{R}_X$ onto some set $\mathcal{R}_Y = g(\mathcal{R}_X)$. Suppose that $\mathcal{R}_X$ can be represented as a countable union of disjoint sets A_k, $k = 1, 2, \ldots$ Then $Y = g(X)$ is an absolutely continuous r.v. with d.f. given by

$$\begin{aligned} F_Y(y) &= P(\{Y \leq y\}) \\ &= P(\{X \in g^{-1}((-\infty, y])\}) \\ &= P\left(\left\{X \in \bigcup_{k=1}^{+\infty} \left[g^{-1}((-\infty, y]) \cap A_k\right]\right\}\right) \\ &= \sum_{k=1}^{+\infty} P\left(\left\{X \in \left[g^{-1}((-\infty, y]) \cap A_k\right]\right\}\right), \; y \in \mathcal{R}_Y. \end{aligned} \tag{2.75}$$

If the conditions of Theorem 2.89 are satisfied by the restriction of g to each A_k, g_k, we may obtain the p.d.f. of $Y = g(X)$ on differentiating the d.f. of Y.[9] In this case

$$f_Y(y) = \sum_{k=1}^{+\infty} f_X[g_k^{-1}(y)] \times \left|\frac{dg_k^{-1}(y)}{dy}\right|, \; y \in \mathcal{R}_Y. \tag{2.76}$$

•

Exercise 2.103 — P.d.f. of a countable sum of monotonic restrictions of a function g of an absolutely continuous r.v.

Let $X \sim \text{Exponential}(\lambda)$ and $Y = \sin X$. Prove that

$$F_Y(y) = 1 + \frac{e^{-\lambda\pi + \lambda \arcsin y} - e^{-\lambda \arcsin y}}{1 - e^{-2\pi\lambda}}, \; 0 < y < 1 \tag{2.77}$$

$$f_Y(y) = \begin{cases} \frac{\lambda e^{-\lambda\pi}}{(1-e^{-2\lambda\pi}) \times \sqrt{1-y^2}} \times \left(e^{\lambda \arcsin y} + e^{-\lambda\pi - \lambda \arcsin y}\right), & -1 \leq y < 0 \\ \frac{\lambda}{(1-e^{-2\lambda\pi}) \times \sqrt{1-y^2}} \times \left(e^{-\lambda \arcsin y} + e^{-\lambda\pi + \lambda \arcsin y}\right), & 0 \leq y < 1 \\ 0, & \text{otherwise} \end{cases} \tag{2.78}$$

(Rohatgi, 1976, p. 75). •

[9] We remind the reader that term-by-term differentiation is permissible if the differentiated series is uniformly convergent.

2.5.4 Transformations of random vectors, general case

What follows is the analogue of Proposition 2.71 in a multidimensional setting.

Proposition 2.104 — D.f. of a transformation of a random vector, general case
Let:

- $\underline{X} = (X_1, \ldots, X_d)$ be a random vector with joint d.f. $F_{\underline{X}}$;
- $\underline{Y} = (Y_1, \ldots, Y_m) = \underline{g}(\underline{X}) = (g_1(X_1, \ldots, X_d), \ldots, g_m(X_1, \ldots, X_d))$ be a transformation of $\underline{X}$ under $\underline{g}$, where $\underline{g} : \mathbb{R}^d \to \mathbb{R}^m$ is a Borel measurable function;
- $\underline{g}^{-1}(\prod_{i=1}^m (-\infty, y_i]) = \{\underline{x} = (x_1, \ldots, x_d) \in \mathbb{R}^d : g_1(x_1, \ldots, x_d) \leq y_1, \ldots, g_m(x_1, \ldots, x_d) \leq y_m\}$ be the inverse image of the Borel set $\prod_{i=1}^m (-\infty, y_i]$ under $\underline{g}$.[10]

Then

$$\begin{aligned} F_{\underline{Y}}(\underline{y}) &= P(\{Y_1 \leq y_1, \ldots, Y_m \leq y_m\}) \\ &= P\left(\left\{\underline{X} \in \underline{g}^{-1}\left(\prod_{i=1}^m (-\infty, y_i]\right)\right\}\right). \end{aligned} \tag{2.79}$$

•

Exercise 2.105 — D.f. of a transformation of a random vector, general case
Let $\underline{X} = (X_1, \ldots, X_d)$ be an absolutely continuous random vector such that $X_i \overset{indep}{\sim}$ Exponential(λ_i), $i = 1, \ldots, d$.

Prove that $Y = \min_{i=1,\ldots,d} X_i \sim$ Exponential$(\sum_{i=1}^d \lambda_i)$. •

2.5.5 Transformations of discrete random vectors

Theorem 2.106 — Joint p.f. of a one-to-one transformation of a discrete random vector (Rohatgi, 1976, p. 131)
Let:

- $\underline{X} = (X_1, \ldots, X_d)$ be a discrete random vector with joint p.f. $P(\{\underline{X} = \underline{x}\})$;
- $\mathcal{R}_{\underline{X}}$ be a countable set of points such that $P(\underline{X} \in \mathcal{R}_{\underline{X}}) = 1$ and $P(\{\underline{X} = \underline{x}) > 0$, $\forall \underline{x} \in \mathbb{R}_{\underline{X}}$;

[10] Let us remind the reader that since $\underline{g}$ is a Borel measurable function we have $\underline{g}^{-1}(B) \in \mathcal{B}(\mathbb{R}^d)$, $\forall B \in \mathcal{B}(\mathbb{R}^m)$.

- $\underline{Y} = (Y_1, \ldots, Y_d) = \underline{g}(\underline{X}) = (g_1(X_1, \ldots, X_d), \ldots, g_d(X_1, \ldots, X_d))$ be a transformation of $\underline{X}$ under $\underline{g}$, where $\underline{g} : \mathbb{R}^d \to \mathbb{R}^d$ is a one-to-one Borel measurable function that maps $\mathcal{R}_{\underline{X}}$ onto some set $\mathcal{R}_{\underline{Y}} \subset \mathbb{R}^d$;
- $\underline{g}^{-1}$ be the inverse mapping such that $\underline{g}^{-1}(\underline{y}) = (g_1^{-1}(\underline{y}), \ldots, g_d^{-1}(\underline{y}))$.

Then the joint p.f. of $\underline{Y} = (Y_1, \ldots, Y_d)$ is given by

$$\begin{aligned} P(\{\underline{Y} = \underline{y}\}) &= P(\{Y_1 = y_1, \ldots, Y_d = y_d\}) \\ &= P(\{X_1 = g_1^{-1}(\underline{y}), \ldots, X_d = g_d^{-1}(\underline{y})\}), \end{aligned} \tag{2.80}$$

for $\underline{y} = (y_1, \ldots, y_d) \in \mathcal{R}_{\underline{Y}}$. •

Remark 2.107 — Joint p.f. of a one-to-one transformation of a discrete random vector (Rohatgi, 1976, pp. 131–132)
The marginal p.f. of any Y_j (resp. the joint p.f. of any subcollection of $Y_1, \ldots, Y_d$, say $(Y_j)_{j \in I \subset \{1,\ldots,d\}}$) is easily computed by summing on the remaining y_i, $i \neq j$ (resp. $(Y_i)_{i \notin I}$). •

Theorem 2.108 — Joint p.f. of a transformation of a discrete random vector
Let:

- $\underline{X} = (X_1, \ldots, X_d)$ be a discrete random vector with range $\mathcal{R}_{\underline{X}} \subset \mathbb{R}^d$;
- $\underline{Y} = (Y_1, \ldots, Y_m) = \underline{g}(\underline{X}) = (g_1(X_1, \ldots, X_d), \ldots, g_m(X_1, \ldots, X_d))$ be a transformation of $\underline{X}$ under $\underline{g}$, where $\underline{g} : \mathbb{R}^d \to \mathbb{R}^m$ is a Borel measurable function that maps $\mathcal{R}_{\underline{X}}$ onto some set $\mathcal{R}_{\underline{Y}} \subset \mathbb{R}^m$;
- $\mathcal{A}_{y_1,\ldots,y_m} = \{\underline{x} = (x_1, \ldots, x_d) \in \mathcal{R}_{\underline{X}} : g_1(x_1, \ldots, x_d) = y_1, \ldots, g_m(x_1, \ldots, x_d) = y_m\}$.

Then the joint p.f. of $\underline{Y} = (Y_1, \ldots, Y_m)$ is given by

$$\begin{aligned} P(\{\underline{Y} = \underline{y}\}) &= P(\{Y_1 = y_1, \ldots, Y_m = y_m\}) \\ &= \sum_{\underline{x}=(x_1,\ldots,x_d) \in \mathcal{A}_{y_1,\ldots,y_m}} P(\{X_1 = x_1, \ldots, X_d = x_d\}), \end{aligned} \tag{2.81}$$

for $\underline{y} = (y_1, \ldots, y_d) \in \mathcal{R}_Y$. •

Exercise 2.109 — Joint p.f. of a transformation of a discrete random vector
Let $\underline{X} = (X_1, X_2)$ be a discrete random vector with joint p.f. $P(X = x, Y = y)$ given in the following table:

X_1	X_2		
	-2	0	2
-1	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{12}$
0	$\frac{1}{12}$	$\frac{1}{12}$	0
1	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{12}$

Derive the joint p.f. of $Y_1 = |X_1|$ and $Y_2 = X_2^2$. •

Theorem 2.110 — P.f. of the sum, difference, product and division of two discrete r.v.
Let:

- (X, Y) be a discrete bidimensional random vector with joint p.f. $P(X = x, Y = y)$;
- $Z = X + Y$
- $U = X - Y$
- $V = X\,Y$
- $W = X/Y$, provided that $P(\{Y = 0\}) = 0$.

Then

$$\begin{aligned} P(Z = z) &= P(X + Y = z) \\ &= \sum_x P(X = x, X + Y = z) \\ &= \sum_x P(X = x, Y = z - x) \\ &= \sum_y P(X + Y = z, Y = y) \\ &= \sum_y P(X = z - y, Y = y) \end{aligned} \tag{2.82}$$

$$\begin{aligned} P(U = u) &= P(X - Y = u) \\ &= \sum_x P(X = x, X - Y = u) \\ &= \sum_x P(X = x, Y = x - u) \end{aligned}$$

$$
\begin{aligned}
&= \sum_y P(X - Y = u, Y = y) \\
&= \sum_y P(X = u + y, Y = y) \qquad (2.83)
\end{aligned}
$$

$$
\begin{aligned}
P(V = v) &= P(XY = v) \\
&= \sum_x P(X = x, XY = v) \\
&= \sum_x P(X = x, Y = v/x) \\
&= \sum_y P(XY = v, Y = y) \\
&= \sum_y P(X = v/y, Y = y) \qquad (2.84)
\end{aligned}
$$

$$
\begin{aligned}
P(W = w) &= P(X/Y = w) \\
&= \sum_x P(X = x, X/Y = w) \\
&= \sum_x P(X = x, Y = x/w) \\
&= \sum_y P(X/Y = w, Y = y) \\
&= \sum_y P(X = wy, Y = y). \qquad (2.85)
\end{aligned}
$$

•

Exercise 2.111 — P.f. of the difference of two discrete r.v.
Let (X, Y) be a discrete random vector with joint p.f. $P(X = x, Y = y)$ given in the following table:

X	Y		
	1	2	3
1	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{2}{12}$
2	$\frac{2}{12}$	0	0
3	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{4}{12}$

(a) Prove that X and Y are identically distributed but are not independent.

(b) Obtain the p.f. of $U = X - Y$

(c) Prove that $U = X - Y$ is not a symmetric r.v., that is U and $-U$ are not identically distributed. •

Corollary 2.112 — P.f. of the sum, difference, product and division of two INDEPENDENT discrete r.v.
Let:

- X and Y be two INDEPENDENT discrete r.v. with joint p.f. $P(X = x, Y = y) = P(X = x) \times P(Y = y)$, $\forall x, y$
- $Z = X + Y$
- $U = X - Y$
- $V = X\,Y$
- $W = X/Y$, provided that $P(\{Y = 0\}) = 0$.

Then

$$\begin{aligned} P(Z = z) &= P(X + Y = z) \\ &= \sum_x P(X = x) \times P(Y = z - x) \\ &= \sum_y P(X = z - y) \times P(Y = y) \end{aligned} \tag{2.86}$$

$$\begin{aligned} P(U = u) &= P(X - Y = u) \\ &= \sum_x P(X = x) \times P(Y = x - u) \\ &= \sum_y P(X = u + y) \times P(Y = y) \end{aligned} \tag{2.87}$$

$$\begin{aligned} P(V = v) &= P(X\,Y = v) \\ &= \sum_x P(X = x) \times P(Y = v/x) \\ &= \sum_y P(X = v/y) \times P(Y = y) \end{aligned} \tag{2.88}$$

$$\begin{aligned} P(W = w) &= P(X/Y = w) \\ &= \sum_x P(X = x) \times P(Y = x/w) \\ &= \sum_y P(X = wy) \times P(Y = y). \end{aligned} \tag{2.89}$$

•

Exercise 2.113 — P.f. of the sum of two INDEPENDENT r.v. with three well known discrete distributions

Let X and Y be two independent discrete r.v. Prove that

(a) $X \sim \text{Binomial}(n_X, p) \perp\!\!\!\perp Y \sim \text{Binomial}(n_Y, p) \Rightarrow (X+Y) \sim \text{Binomial}(n_X + n_Y, p)$

(b) $X \sim \text{NegativeBinomial}(n_X, p) \perp\!\!\!\perp Y \sim \text{NegativeBinomial}(n_Y, p) \Rightarrow (X+Y) \sim \text{NegativeBinomial}(n_X + n_Y, p)$

(c) $X \sim \text{Poisson}(\lambda_X) \perp\!\!\!\perp Y \sim \text{Poisson}(\lambda_Y) \Rightarrow (X+Y) \sim \text{Poisson}(\lambda_X + \lambda_Y)$,

i.e. the families of Poisson, Binomial and Negative Binomial distributions are closed under summation of independent members.[11] •

Exercise 2.114 — P.f. of the difference of two INDEPENDENT Poisson r.v.

Let $X \sim \text{Poisson}(\lambda_X) \perp\!\!\!\perp Y \sim \text{Poisson}(\lambda_Y)$. Then $(X-Y)$ has p.f. given by

$$\begin{aligned} P(X-Y=u) &= \sum_{y=0}^{+\infty} P(X=u+y) \times P(Y=y) \\ &= e^{-(\lambda_X+\lambda_Y)} \sum_{y=\max\{0,-u\}}^{+\infty} \frac{\lambda_X^{u+y}\,\lambda_Y^{y}}{(u+y)!\,y!},\ u=\ldots,-1,0,1,\ldots \end{aligned} \tag{2.90}$$

•

Remark 2.115 — Skellam distribution (http://en.wikipedia.org/wiki/Skellam_distribution)

The Skellam distribution is the discrete probability distribution of the difference of independent r.v. X and Y having Poisson distributions with parameters λ_X and λ_Y. It is useful in describing the statistics of the difference of two images with simple photon noise, as well as describing the point spread distribution in certain sports where all scored points are equal, such as baseball, hockey and soccer.

When $\lambda_X = \lambda_Y = \lambda$ and u is also large, and of order of the square root of 2λ,

$$P(X-Y=u) \simeq \frac{e^{-\frac{u^2}{2\times 2\lambda}}}{\sqrt{2\pi \times 2\lambda}}, \tag{2.91}$$

the p.d.f. of a Normal distribution with parameters $\mu = 0$ and $\sigma^2 = 2\lambda$.

Please note that the expression of the p.f. of the Skellam distribution that can be found in http://en.wikipedia.org/wiki/Skellam_distribution is not correct. •

[11] Use the Vandermonde's identity to prove result (a). In combinatorial mathematics, Vandermonde's identity, named after Alexandre-Théophile Vandermonde (1772), states that the equality $\binom{m+n}{r} = \sum_{k=0}^{r} \binom{m}{k}\binom{n}{r-k}$, $m,n,r \in \mathbb{N}_0$, for binomial coefficients holds; this identity was given already in 1303 by the Chinese mathematician Zhu Shijie (Chu Shi-Chieh) (http://en.wikipedia.org/wiki/Vandermonde's_identity).

2.5.6 Transformations of absolutely continuous random vectors

Motivation 2.116 — P.d.f. of a transformation of an absolutely continuous random vector (Karr, 1993, p. 62)
Recall that a random vector $\underline{X} = (X_1, \ldots, X_d)$ is absolutely continuous if there is a function $f_{\underline{X}}$ on $\mathbb{R}^d$ satisfying

$$\begin{aligned} F_{\underline{X}}(\underline{x}) &= F_{X_1,\ldots,X_d}(x_1, \ldots, x_d) \\ &= \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_d} f_{X_1,\ldots,X_d}(s_1, \ldots, s_d)\, ds_d \ldots ds_1. \end{aligned} \tag{2.92}$$

Computing the density of $\underline{Y} = \underline{g}(\underline{X})$ requires that $\underline{g}$ be invertible, except for the special case that $X_1, \ldots, X_d$ are independent (and then only for particular choices of $\underline{g}$). •

Theorem 2.117 — P.d.f. of a one-to-one transformation of an absolutely continuous random vector (Rohatgi, 1976, p. 135; Karr, 1993, p. 62)
Let:

- $\underline{X} = (X_1, \ldots, X_d)$ be an absolutely continuous random vector with joint p.d.f. $f_{\underline{X}}(\underline{x})$;
- $\mathcal{R}_{\underline{X}}$ be an open set of $\mathbb{R}^d$ such that $P(\underline{X} \in \mathcal{R}_{\underline{X}}) = 1$;
- $\underline{Y} = (Y_1, \ldots, Y_d) = \underline{g}(\underline{X}) = (g_1(X_1, \ldots, X_d), \ldots, g_d(X_1, \ldots, X_d))$ be a transformation of $\underline{X}$ under $\underline{g}$, where $\underline{g} : \mathbb{R}^d \to \mathbb{R}^d$ is a one-to-one Borel measurable function that maps $\mathcal{R}_{\underline{X}}$ onto some set $\mathcal{R}_{\underline{Y}} \subset R^d$;
- $\underline{g}^{-1}(\underline{y}) = (g_1^{-1}(\underline{y}), \ldots, g_d^{-1}(\underline{y}))$ be the inverse mapping defined over the range $\mathcal{R}_{\underline{Y}}$ of the transformation.

Assume that:

- both $\underline{g}$ and its inverse $\underline{g}^{-1}$ are continuous;
- the partial derivatives, $\frac{\partial g_i^{-1}(\underline{y})}{\partial y_j}$, $1 \leq i, j \leq d$, exist and are continuous;
- the Jacobian of the inverse transformation $\underline{g}^{-1}$ (i.e. the determinant of the matrix of partial derivatives $\frac{\partial g_i^{-1}(\underline{y})}{\partial y_j}$) is such that

$$J(\underline{y}) = \begin{vmatrix} \frac{\partial g_1^{-1}(\underline{y})}{\partial y_1} & \cdots & \frac{\partial g_1^{-1}(\underline{y})}{\partial y_d} \\ \vdots & \cdots & \vdots \\ \frac{\partial g_d^{-1}(\underline{y})}{\partial y_1} & \cdots & \frac{\partial g_d^{-1}(\underline{y})}{\partial y_d} \end{vmatrix} \neq 0, \tag{2.93}$$

for $\underline{y} = (y_1, \ldots, y_d) \in \mathcal{R}_{\underline{Y}}$.

Then the random vector $\underline{Y} = (Y_1, \ldots, Y_d)$ is absolutely continuous and its joint p.d.f. is given by

$$f_{\underline{Y}}(\underline{y}) = f_{\underline{X}}\left[\underline{g}^{-1}(\underline{y})\right] \times |J(\underline{y})|, \tag{2.94}$$

for $\underline{y} = (y_1, \ldots, y_d) \in \mathcal{R}_{\underline{Y}}$. •

Exercise 2.118 — P.d.f. of a one-to-one transformation of an absolutely continuous random vector

Prove Theorem 2.117 (Rohatgi, 1976, pp. 135–136). •

Exercise 2.119 — P.d.f. of a one-to-one transformation of an absolutely continuous random vector

Let

- $\underline{X} = (X_1, \ldots, X_d)$ be an absolutely continuous random vector with joint p.d.f. $f_{\underline{X}}(\underline{x})$;
- $\underline{Y} = (Y_1, \ldots, Y_d) = \underline{g}(\underline{X}) = \mathbf{A}\underline{X} + \underline{b}$ be an invertible affine mapping of $\mathbb{R}^d$ into itself, where $\mathbf{A}$ is a nonsingular $d \times d$ matrix and $\underline{b} \in \mathbb{R}^d$.

Derive the inverse mapping $\underline{g}^{-1}$ and the joint p.d.f. of $\underline{Y}$ (Karr, 1993, p. 62). •

Exercise 2.120 — P.d.f. of a one-to-one transformation of an absolutely continuous random vector

Let

- $\underline{X} = (X_1, X_2, X_3)$ such that $X_i \overset{i.i.d.}{\sim}$ Exponential(1);
- $\underline{Y} = (Y_1, Y_2, Y_3) = \left(X_1 + X_2 + X_3, \frac{X_1+X_2}{X_1+X_2+X_3}, \frac{X_1}{X_1+X_2}\right)$.

Derive the joint p.d.f. of $\underline{Y}$ and conclude that Y_1, Y_2, and Y_3 are also independent (Rohatgi, 1976, p. 137). •

Remark 2.121 — P.d.f. of a one-to-one transformation of an absolutely continuous random vector (Rohatgi, 1976, p. 136)

In actual applications, we tend to know just k functions, $Y_1 = g_1(\underline{X}), \ldots, Y_k = g_k(\underline{X})$. In this case, we introduce arbitrarily $(d-k)$ (convenient) r.v., $Y_{k+1} = g_{k+1}(\underline{X}), \ldots, Y_d = g_d(\underline{X})$), such that the conditions of Theorem 2.117 are satisfied.

To find the joint density of the k r.v. we simply integrate the joint p.d.f. $f_{\underline{Y}}$ over all the $(d-k)$ r.v. that were arbitrarily introduced. •

We can state a similar result to Theorem 2.117 when $\underline{g}$ is not a one-to-one transformation.

Theorem 2.122 — P.d.f. of a transformation, with a finite number of inverses, of an absolutely continuous random vector (Rohatgi, 1976, pp. 136–137)
Assume the conditions of Theorem 2.117 and suppose that:

- for each $\underline{y} \in \mathcal{R}_{\underline{Y}} \subset I\!R^d$, the transformation $\underline{g}$ has a finite number $k = k(\underline{y})$ of inverses;
- $\mathcal{R}_{\underline{X}} \subset I\!R^d$ can be partitioned into k disjoint sets, $A_1, \ldots, A_k$, such that the transformation $\underline{g}$ from A_i $(i = 1, \ldots, k)$ into $I\!R^d$, say $\underline{g}_i$, is one-to-one with inverse transformation $\underline{g}_i^{-1} = (g_{1\,i}^{-1}(\underline{y}), \ldots, g_{d\,i}^{-1}(\underline{y}))$, $i = 1, \ldots, k$;
- the first partial derivatives of $\underline{g}_i^{-1}$ exist, are continuous and that each Jacobian

$$J_i(\underline{y}) = \begin{vmatrix} \frac{\partial g_{1\,i}^{-1}(\underline{y})}{\partial y_1} & \cdots & \frac{\partial g_{1\,i}^{-1}(\underline{y})}{\partial y_d} \\ \vdots & \cdots & \vdots \\ \frac{\partial g_{d\,i}^{-1}(\underline{y})}{\partial y_1} & \cdots & \frac{\partial g_{d\,i}^{-1}(\underline{y})}{\partial y_d} \end{vmatrix} \neq 0, \tag{2.95}$$

  for $\underline{y} = (y_1, \ldots, y_d)$ in the range of the transformation $\underline{g}_i$.

Then the random vector $\underline{Y} = (Y_1, \ldots, Y_d)$ is absolutely continuous and its joint p.d.f. is given by

$$f_{\underline{Y}}(\underline{y}) = \sum_{i=1}^{k} f_{\underline{X}}\left[\underline{g}_i^{-1}(\underline{y})\right] \times |J_i(\underline{y})|, \tag{2.96}$$

for $\underline{y} = (y_1, \ldots, y_d) \in \mathcal{R}_{\underline{Y}}$. •

Theorem 2.123 — P.d.f. of the sum, difference, product and division of two absolutely continuous r.v. (Rohatgi, 1976, p. 141)
Let:

- (X, Y) be an absolutely continuous bidimensional random vector with joint p.d.f. $f_{X,Y}(x, y)$;
- $Z = X + Y$, $U = X - Y$, $V = X\,Y$ and $W = X/Y$.

Then

$$\begin{aligned} f_Z(z) &= f_{X+Y}(z) \\ &= \int_{-\infty}^{+\infty} f_{X,Y}(x, z - x)\, dx \\ &= \int_{-\infty}^{+\infty} f_{X,Y}(z - y, y)\, dy \end{aligned} \tag{2.97}$$

$$\begin{aligned} f_U(u) &= f_{X-Y}(u) \\ &= \int_{-\infty}^{+\infty} f_{X,Y}(x, x-u)\,dx \\ &= \int_{-\infty}^{+\infty} f_{X,Y}(u+y, y)\,dy \end{aligned} \tag{2.98}$$

$$\begin{aligned} f_V(v) &= f_{XY}(v) \\ &= \int_{-\infty}^{+\infty} f_{X,Y}(x, v/x) \times \frac{1}{|x|}\,dx \\ &= \int_{-\infty}^{+\infty} f_{X,Y}(v/y, y) \times \frac{1}{|y|}\,dy \end{aligned} \tag{2.99}$$

$$\begin{aligned} f_W(w) &= f_{X/Y}(w) \\ &= \int_{-\infty}^{+\infty} f_{X,Y}(x, x/w) \times \frac{|x|}{w^2}\,dx \\ &= \int_{-\infty}^{+\infty} f_{X,Y}(wy, y) \times |y|\,dy. \end{aligned} \tag{2.100}$$

•

Remark 2.124 — P.d.f. of the sum and product of two absolutely continuous r.v.

It is interesting to note that:

$$\begin{aligned} f_Z(z) &= \frac{d\,F_Z(z)}{dz} \\ &= \frac{d\,P(Z = X+Y \leq z)}{dz} \\ &= \frac{d}{dz}\left[\int\int_{\{(x,y):\, x+y\leq z\}} f_{X,Y}(x,y)\,dy\,dx\right] \\ &= \frac{d}{dz}\left[\int_{-\infty}^{+\infty}\int_{-\infty}^{z-x} f_{X,Y}(x,y)\,dy\,dx\right] \\ &= \int_{-\infty}^{+\infty}\frac{d}{dz}\left[\int_{-\infty}^{z-x} f_{X,Y}(x,y)\,dy\right]dx \\ &= \int_{-\infty}^{+\infty} f_{X,Y}(x, z-x)\,dx; \end{aligned} \tag{2.101}$$

$$\begin{aligned} f_V(v) &= \frac{d\,F_V(v)}{dv} \\ &= \frac{d\,P(V = XY \leq v)}{dv} \end{aligned}$$

$$
\begin{aligned}
&= \frac{d}{dv}\left[\int\int_{\{(x,y):\, xy\le v\}} f_{X,Y}(x,y)\,dy\,dx\right] \\
&= \begin{cases} \int_{-\infty}^{+\infty} \frac{d}{dv}\left[\int_{-\infty}^{v/x} f_{X,Y}(x,y)\,dy\right] dx, & x>0 \\ \int_{-\infty}^{+\infty} \frac{d}{dv}\left[\int_{v/x}^{+\infty} f_{X,Y}(x,y)\,dy\right] dx, & x<0 \end{cases} \\
&= \int_{-\infty}^{+\infty} \frac{1}{|x|}\, f_{X,Y}(x,v/x)\,dx. \qquad (2.102)
\end{aligned}
$$

•

Corollary 2.125 — P.d.f. of the sum, difference, product and division of two INDEPENDENT **absolutely continuous r.v.** (Rohatgi, 1976, p. 141)
Let:

- X and Y be two INDEPENDENT absolutely continuous r.v. with joint p.d.f. $f_{X,Y}(x,y) = f_X(x) \times f_Y(y)$, $\forall x, y$;
- $Z = X + Y$, $U = X - Y$, $V = X\,Y$ and $W = X/Y$.

Then

$$
\begin{aligned}
f_Z(z) &= f_{X+Y}(z) \\
&= \int_{-\infty}^{+\infty} f_X(x) \times f_Y(z-x)\,dx \\
&= \int_{-\infty}^{+\infty} f_X(z-y) \times f_Y(y)\,dy \qquad (2.103)
\end{aligned}
$$

$$
\begin{aligned}
f_U(u) &= f_{X-Y}(u) \\
&= \int_{-\infty}^{+\infty} f_X(x) \times f_Y(x-u)\,dx \\
&= \int_{-\infty}^{+\infty} f_X(u+y) f_Y(y)\,dy \qquad (2.104)
\end{aligned}
$$

$$
\begin{aligned}
f_V(v) &= f_{XY}(v) \\
&= \int_{-\infty}^{+\infty} f_X(x) \times f_Y(v/x) \times \frac{1}{|x|}\,dx \\
&= \int_{-\infty}^{+\infty} f_X(v/y) \times f_Y(y) \times \frac{1}{|y|}\,dy \qquad (2.105)
\end{aligned}
$$

$$
\begin{aligned}
f_W(w) &= f_{X/Y}(w) \\
&= \int_{-\infty}^{+\infty} f_X(x) \times f_Y(x/w) \times \frac{|x|}{w^2}\,dx
\end{aligned}
$$

$$= \int_{-\infty}^{+\infty} f_X(wy) \times f_Y(y) \times |y| \, dy. \tag{2.106}$$

•

Exercise 2.126 — P.d.f. of the sum and difference of two INDEPENDENT absolutely continuous r.v.
Let X and Y be two r.v. which are independent and uniformly distributed in $(0, 1)$. Derive the p.d.f. of:

(a) $(X + Y, X - Y)$ (Rohatgi, 1976, pp. 137–138);

(b) $X + Y$;

(c) $X - Y$. •

Exercise 2.127 — P.d.f. of the mean of two INDEPENDENT absolutely continuous r.v.
Let X and Y be two independent r.v. with standard normal distribution. Prove that their mean $\frac{X+Y}{2} \sim \text{Normal}(0, 2^{-1})$. •

Remark 2.128 — D.f. and p.d.f. of the sum, difference, product and division of two absolutely continuous r.v.
In several cases it is simpler to obtain the d.f. of those four algebraic functions of X and Y than to derive the corresponding p.d.f. It suffices to apply Proposition 2.104 and then differentiate the d.f. to get the p.d.f., as seen in the next exercises. •

Exercise 2.129 — D.f. and p.d.f. of the difference of two absolutely continuous r.v.
Choosing adequate underkeel clearance (UKC) is one of the most crucial and most difficult problems in the navigation of large ships, especially very large crude oil carriers.

Let X be the water depth in a passing shallow waterway, say a harbour or a channel, and Y be the maximum ship draft. Then the probability of a safe passing a shallow waterway can be expressed as $P(\text{UKC} = X - Y > 0)$.

Assume that X and Y are independent r.v. such that $X \sim \text{Gamma}(n, \beta)$ and $Y \sim \text{Gamma}(m, \beta)$, where $n, m \in \mathbb{N}$ and $m < n$. Derive an expression for $P(\text{UKC} = X - Y > 0)$ taking into account that $F_{Gamma(k,\beta)}(x) = \sum_{i=k}^{\infty} e^{-\beta x}(\beta x)^i / i!$, $k \in \mathbb{N}$. •

Exercise 2.130 — D.f. and p.d.f. of the sum of two absolutely continuous r.v. Let X and Y be the durations of two independent system components set in what is called a *stand by connection*.[12] In this case the system duration is given by $X + Y$.

Prove that the p.d.f. of $X + Y$ equals

$$f_{X+Y}(z) = \frac{\alpha\beta\left(e^{-\beta z} - e^{-\alpha z}\right)}{\alpha - \beta},\ z > 0,$$

if $X \sim$ Exponencial(α) and $Y \sim$ Exponencial(β), where $\alpha, \beta > 0$ and $\alpha \neq \beta$. •

Exercise 2.131 — D.f. of the division of two absolutely continuous r.v. Let X and Y be the intensity of a transmitted signal and its damping until its reception, respectively. Moreover, $W = X/Y$ represents the intensity of the received signal.

Assume that the joint p.d.f. of (X, Y) equals $f_{X,Y}(x, y) = \lambda\mu e^{-(\lambda x + \mu y)} \times I_{(0,+\infty)\times(0,+\infty)}(x, y)$. Prove that the d.f. of $W = X/Y$ is given by:

$$F_W(w) = \left(1 - \frac{\mu}{\mu + \lambda w}\right) \times I_{(0,+\infty)}(w). \tag{2.107}$$

•

[12] At time 0, only the component with duration X is on. The component with duration Y replaces the other one as soon as it fails.

2.5.7 Random variables with prescribed distributions

Motivation 2.132 — Construction of a r.v. with a prescribed distribution (Karr, 1993, p. 63)
Can we construct (or simulate) explicitly individual r.v., random vectors or sequences of r.v. with prescribed distributions? •

Proposition 2.133 — Construction of a r.v. with a prescribed d.f. (Karr, 1993, p. 63)
Let F be a d.f. on $\mathbb{R}$. Then there is a probability space $(\Omega, \mathcal{F}, P)$ and a r.v. X defined on it such that $F_X = F$. •

Exercise 2.134 — Construction of a r.v. with a prescribed d.f.
Prove Proposition 2.133 (Karr, 1993, p. 63). •

The construction of a r.v. with a prescribed d.f. depends on the following definition.

Definition 2.135 — Quantile function (Karr, 1993, p. 63)
The inverse function of F, F^{-1}, or quantile function associated with F, is defined by

$$F^{-1}(p) = \inf\{x : F(x) \geq p\},\ p \in (0, 1). \tag{2.108}$$

This function is often referred to as the generalized inverse of the d.f. •

Exercise 2.136 — Quantile functions of an absolutely continuous and a discrete r.v.
Obtain and draw the graphs of the d.f. and quantile function of:

(a) $X \sim$ Exponential(λ);

(b) $X \sim$ Bernoulli(θ).

•

Remark 2.137 — Existence of a quantile function (Karr, 1993, p. 63)
Even though F need be neither continuous nor strictly increasing, F^{-1} always exists.

As the figure of the quantile function (associated with the d.f.) of $X \sim \text{Bernoulli}(\theta)$, F^{-1} jumps where F is flat, and is flat where F jumps.

Although not necessarily a pointwise inverse of F, F^{-1} serves that role for many purposes and has a few interesting properties. •

Proposition 2.138 — Basic properties of the quantile function (Karr, 1993, p. 63)
Let F^{-1} be the (generalized) inverse of F or quantile function associated with F. Then

1. For each p and x,

$$F^{-1}(p) \leq x \text{ iff } p \leq F(x); \tag{2.109}$$

2. F^{-1} is non decreasing and left-continuous;

3. If F is absolutely continuous, then

$$F[F^{-1}(p)] = p, \forall p \in (0,1). \tag{2.110}$$

•

Motivation 2.139 — Quantile transformation (Karr, 1993, p. 63)
A r.v. with d.f. F can be constructed by applying F^{-1} to a r.v. with distribution on $(0,1)$. This is usually known as quantile transformation and is a very popular transformation in random numbers generation/simulation on computer. •

Proposition 2.140 — Quantile transformation (Karr, 1993, p. 64)
Let F be a d.f. on $\mathbb{R}$ and suppose $U \sim \text{Uniform}(0,1)$. Then

$$X = F^{-1}(U) \text{ has distribution function } F. \tag{2.111}$$

•

Exercise 2.141 — Quantile transformation
Prove Proposition 2.140 (Karr, 1993, p. 64). •

Example 2.142 — Quantile transformation
If $U \sim \text{Uniform}(0,1)$ then both $-\frac{1}{\lambda}\ln(1-U)$ and $-\frac{1}{\lambda}\ln(U)$ have exponential distribution with parameter λ ($\lambda > 0$). •

Remark 2.143 — **Quantile transformation** (Karr, 1993, p. 64)
R.v. with d.f. F can be simulated by applying F^{-1} to the (uniformly distributed) values produced by the random number generator.

Feasibility of this technique depends on either having F^{-1} available in closed form or being able to approximate it numerically. •

Proposition 2.144 — **The quantile transformation and the simulation of discrete and absolutely continuous distributions**
To generate (pseudo-)random numbers from a r.v. X with d.f. F, it suffices to:

1. Generate a (pseudo-)random number u from the Uniform$(0, 1)$ distribution.
2. Assign

$$x = F^{-1}(u) = \inf\{m \in I\!R : F(m) \geq u\}, \tag{2.112}$$

the quantile of order u of X, where F^{-1} represents the generalized inverse of F. •

For a detailed discussion on (pseudo-)random number generation/generators and their properties please refer to Gentle (1998, pp. 6–22). For a brief discussion — in Portuguese — on (pseudo-)random number generation and Monte Carlo simulation method we refer the reader to Morais (2003, Chapter 2).

Exercise 2.145 — **The quantile transformation and the generation of the Logistic distribution**
X is said to have a Logistic(μ, σ) if its p.d.f. is given by

$$f(x) = \frac{e^{-\frac{x-\mu}{\sigma}}}{\sigma\left(1 + e^{-\frac{x-\mu}{\sigma}}\right)^2}, -\infty < x < +\infty. \tag{2.113}$$

Define the quantile transformation to produce (pseudo-)random numbers with such a distribution. •

Exercise 2.146 — **The quantile transformation and the simulation of the Erlang distribution**
Describe a method to generate (pseudo-)random numbers from the Erlang(n, λ).[13] •

[13] Let us remind the reader that the sum of n independent exponential distributions with parameter λ has an Erlang(n, λ).

Exercise 2.147 — The quantile transformation and the generation of the Beta distribution

Let Y and Z be two independent r.v. with distributions Gamma(α, λ) and Gamma(β, λ), respectively $(\alpha, \beta, \lambda > 0)$.

(a) Prove that $X = Y/(Y + Z) \sim \text{Beta}(\alpha, \beta)$.

(b) Use this result to describe a random number generation method for the Beta(α, β), where $\alpha, \beta \in \mathbb{N}$.

(c) Use any software you are familiar with to generate and plot the histogram of 1000 observations from the Beta$(4, 5)$ distribution. •

Example 2.148 — The quantile transformation and the generation of the Bernoulli distribution (Gentle, 1993, p. 47)

To generate (pseudo-)random numbers from the Bernoulli(p) distribution, we should proceed as follows:

1. Generate a (pseudo-)random number u from the Uniform$(0, 1)$ distribution.
2. Assign

$$x = \begin{cases} 0, \text{ if } u \leq 1 - p \\ 1, \text{ if } u > 1 - p \end{cases} \tag{2.114}$$

or, equivalently,

$$x = \begin{cases} 0, \text{ if } u \geq p \\ 1, \text{ if } u < p. \end{cases} \tag{2.115}$$

(Is there any advantage of (2.115) over (2.114)?) •

Exercise 2.149 — The quantile transformation and the simulation of the Binomial distribution

Describe a method to generate (pseudo-)random numbers from a Binomial(n, p) distribution. •

Proposition 2.150 — The converse of the quantile transformation (Karr, 1993, p. 64)

A converse of the quantile transformation (Propositon 2.140) holds as well, under certain conditions. In fact, if F_X is continuous (not necessarily absolutely continuous) then

$$F_X(X) \sim \text{Uniform}(0, 1). \tag{2.116}$$

•

Exercise 2.151 — The converse of the quantile transformation
Prove Propositon 2.150 (Karr, 1993, p. 64). •

Motivation 2.152 — Construction of random vectors with a prescribed distribution (Karr, 1993, p. 65)
The construction of a random vector with an arbitrary d.f. is more complicated. We shall address this issue in the next chapter for a special case: when the random vector has independent components. However, we can state the following result. •

Proposition 2.153 — Construction of a random vector with a prescribed d.f. (Karr, 1993, p. 65)
Let $F : \mathbb{R}^d \to [0,1]$ be a $d-$dimensional d.f. Then there is a probability space $(\Omega, \mathcal{F}, P)$ and a random vector $\underline{X} = (X_1, \ldots, X_d)$ defined on it such that $F_{\underline{X}} = F$. •

Motivation 2.154 — Construction of a sequence of r.v. with a prescribed joint d.f. (Karr, 1993, p. 65)
How to construct a sequence $\{X_k\}_{k\in\mathbb{N}}$ of r.v. with a prescribed joint d.f. F_n where F_n is the joint d.f. of $\underline{X}_n = (X_1, \ldots, X_n)$, for each $n \in \mathbb{N}$. The d.f. F_n must satisfy certain consistency conditions since if such r.v. exists then

$$\begin{aligned} F_n(\underline{x}_n) &= P(X_1 \leq x_1, \ldots, X_n \leq x_n) \\ &= \lim_{x\to+\infty} P(X_1 \leq x_1, \ldots, X_n \leq x_n, X_{n+1} \leq x), \end{aligned} \tag{2.117}$$

for all $x_1, \ldots, x_n$. •

Theorem 2.155 — Kolmogorov existence Theorem (Karr, 1993, p. 65)
Let F_n be a d.f. on $\mathbb{R}^n$, and suppose that

$$\lim_{x\to+\infty} F_{n+1}(x_1, \ldots, x_n, x) = F_n(x_1, \ldots, x_n), \tag{2.118}$$

for each $n \in \mathbb{N}$ and $x_1, \ldots, x_n$. Then there is a probability space say $(\Omega, \mathcal{F}, P)$ and a sequence of $\{X_k\}_{k\in\mathbb{N}}$ of r.v. defined on it such that F_n is the d.f. of $(X_1, \ldots, X_n)$, for each $n \in \mathbb{N}$. •

Remark 2.156 — Kolmogorov existence Theorem (http://en.wikipedia.org/wiki/Kolmogorov_extension_theorem)
Theorem 2.155 guarantees that a suitably "consistent" collection of finite-dimensional distributions will define a stochastic process. This theorem is credited to soviet mathematician Andrey Nikolaevich Kolmogorov (1903–1987, http://en.wikipedia.org/wiki/Andrey_Kolmogorov). •

References

- Gentle, J.E. (1998). *Random Number Generation and Monte Carlo Methods*. Springer-Verlag, New York, Inc. (QA298.GEN.50103)

- Grimmett, G.R. and Stirzaker, D.R. (2001). *One Thousand Exercises in Probability*. Oxford University Press.

- Karr, A.F. (1993). *Probability*. Springer-Verlag.

- Morais, M.C. (2003). Estatística Computacional — Módulo 1: Notas de Apoio (Caps. 1 e 2), 141 pags.

  (http://www.math.ist.utl.pt/∼mjmorais/materialECMCM.html)

- Murteira, B.J.F. (1979). *Probabilidades e Estatística (volume I)*. Editora McGraw-Hill de Portugal, Lda. (QA273-280/3.MUR.5922, QA273-280/3.MUR.34472, QA273-280/3.MUR.34474, QA273-280/3.MUR.34476)

- Resnick, S.I. (1999). *A Probability Path*. Birkhäuser. (QA273.4-.67.RES.49925)

- Righter, R. (200–). Lectures notes for the course *Probability and Risk Analysis for Engineers*. Department of Industrial Engineering and Operations Research, University of California at Berkeley.

- Rohatgi, V.K. (1976). *An Introduction to Probability Theory and Mathematical Statistics*. John Wiley & Sons. (QA273-280/4.ROH.34909)

Chapter 3

Independence

Independence is a basic property of events and r.v. in a probability model.

3.1 Fundamentals

Motivation 3.1 — **Independence** (Resnick, 1999, p. 91; Karr, 1993, p. 71)
The intuitive appeal of independence stems from the easily envisioned property that the ocurrence of an event has no effect on the probability that an independent event will occur. Despite the intuitive appeal, it is important to recognize that independence is a technical concept/definition which must be checked with respect to a specific model.

Independence — or the absence of probabilistic interaction — sets probability apart as a distinct mathematical theory. •

A series of definitions of independence of increasingly sophistication will follow.

Definition 3.2 — **Independence for two events** (Resnick, 1999, p. 91)
Suppose $(\Omega, \mathcal{F}, P)$ is a fixed probability space. Events $A, B \in \mathcal{F}$ are independent if

$$P(A \cap B) = P(A) \times P(B). \tag{3.1}$$

•

Exercise 3.3 — **Independence**
Let A and B be two independent events. Show that:

(a) A^c and B are independent, and so are A and B^c, and A^c and B^c;

(b) A and B are independent iff $P(B|A) = P(B|A^c)$, where $P(A) \in (0, 1)$. •

Exercise 3.4 — **(In)dependence and disjoint events**
Let A and B two DISJOINT EVENTS with probabilities $P(A), P(B) > 0$. Show that these two events are NOT INDEPENDENT. •

Exercise 3.5 — **Independence** (Exercise 3.2, Karr, 1993, p. 95)
Show that:

(a) an event whose probability is either zero or one is independent of every event;

(b) an event that is independent of itself has probability zero or one. •

Definition 3.6 — **Independence for a finite/infinite number of events**
The events $A_1, \ldots, A_n \in \mathcal{F}$ are independent if

$$P\left(\bigcap_{i \in I} A_i\right) = \prod_{i \in I} P(A_i), \tag{3.2}$$

for all finite $I \subseteq \{1, \ldots, n\}$ (Resnick, 1999, p. 91).

The events $A_1, A_2, \ldots \in \mathcal{F}$ are said to be independent if the events A_i, $i \in I$, are independent for all finite $I \subset \mathbb{N}$. •

Remark 3.7 — **Independence for a finite number of events** (Resnick, 1999, p. 92)
Note that (3.2) represents $\sum_{k=2}^{n} \binom{n}{k} = 2^n - n - 1$ equations and can be rephrased as follows:

- the events $A_1, \ldots, A_n$ are independent if

$$P\left(\bigcap_{i=1}^{n} B_i\right) = \prod_{i=1}^{n} P(B_i), \tag{3.3}$$

where, for each $i = 1, \ldots, n$, B_i equals A_i or Ω. •

Corollary 3.8 — **Independence for a finite number of events** (Karr, 1993, p. 81)
Events $A_1, \ldots, A_n$ are independent iff $A_1^c, \ldots, A_n^c$ are independent. •

Exercise 3.9 — **Independence for a finite number of events** (Exercise 3.1, Karr, 1993, p. 95)
Let $A_1, \ldots, A_n$ be independent events.

(a) Prove that $P\left(\bigcup_{i=1}^{n} A_i\right) = 1 - \prod_{i=1}^{n}[1 - P(A_i)]$.

(b) Consider a parallel system with n components and assume that $P(A_i)$ is the reliability of the component i $(i = 1, \ldots, n)$. What is the system reliability? •

Motivation 3.10 — **(2nd.) Borel-Cantelli lemma** (Karr, 1993, p. 81)
For independent events, Theorem 1.76, the (1st.) Borel-Cantelli lemma, has a converse. It states that if the events $A_1, A_2, \ldots$ are independent and the sum of the probabilities of the A_n diverges to infinity, then the probability that infinitely many of them occur is 1. •

Theorem 3.11 — **(2nd.) Borel-Cantelli lemma** (Karr, 1993, p. 81)
Let $A_1, A_2, \ldots$ be independent events. Then

$$\sum_{n=1}^{+\infty} P(A_n) = +\infty \quad \Rightarrow \quad P(\limsup A_n) = 1. \tag{3.4}$$

(Moreover, $P(\limsup A_n) = 1 \Rightarrow \sum_{n=1}^{+\infty} P(A_n) = +\infty$ follows from the 1st. Borel-Cantelli lemma.) •

Exercise 3.12 — **(2nd.) Borel-Cantelli lemma**
Prove Theorem 3.11 (Karr, 1993, p. 82). •

Definition 3.13 — **Independent classes of events** (Resnick, 1999, p. 92)
Let $\mathcal{C}_i \subseteq \mathcal{F}$, $i = 1, \ldots, n$, be a class of events. Then the classes $\mathcal{C}_1, \ldots, \mathcal{C}_n$ are said to be independent if for any choice $A_1, \ldots, A_n$, with $A_i \in \mathcal{C}_i$, $i = 1, \ldots, n$, the events $A_1, \ldots, A_n$ are independent events according to Definition 3.6. •

Definition 3.14 — **Independent sub $\sigma-$algebras** (Karr, 1993, p. 94)
Sub $\sigma -$ algebras $\mathcal{G}_1, \ldots, \mathcal{G}_n$ of $\sigma -$ algebra $\mathcal{F}$ are independent if

$$P\left(\bigcap_{i=1}^{n} A_i\right) = \prod_{i=1}^{n} P(A_i), \tag{3.5}$$

for all $A_i \in \mathcal{G}_i$, $i = 1, \ldots, n$. •

Motivation 3.15 — **Independence of $\sigma -$ algebras** (Resnick, 1999, p. 92)
To provide a basic criterion for proving independence of $\sigma-$algebras, we need to introduce the notions of $\pi -$ system and $d -$ system. •

Definition 3.16 — $\pi-$**system** (Resnick, 1999, p. 32; Karr, 1993, p. 21)
Let $\mathcal{P}$ family of subsets of the sample space Ω. $\mathcal{P}$ is said to be a $\pi-$system if it is closed under finite intersection: $A, B \in \mathcal{P} \Rightarrow A \cap B \in \mathcal{P}$. •

Remark 3.17 — $\pi-$**system** (Karr, 1993, p. 21)
A $\sigma-$algebra is a $\pi-$system. •

Definition 3.18 — $d-$**system** (Karr, 1993, p. 21)
Let $\mathcal{D}$ family of subsets of the sample space Ω. $\mathcal{D}$ is said to be a $d-$system [1] if it

1. contains the sample space Ω,
2. is closed under proper difference,[2]
3. and is closed under countable increasing union.[3] •

Proposition 3.19 — **Relating** $\pi-$ **and** $d-$ **systems and** $\sigma-$**algebras** (Resnick, 1999, p. 38)
If a class $\mathcal{C}$ is both a $\pi-$system and $d-$system then it is a $\sigma-$algebra. •

Theorem 3.20 — **Basic independence criterion** (Resnick, 1999, p. 92)
If, for each $i = 1, \ldots, n$, $\mathcal{C}_i$ is a non-empty class of events satisfying

1. $\mathcal{C}_i$ is a $\pi-$system
2. $\mathcal{C}_i$, $i = 1, \ldots, n$, are independent

then the $\sigma-$algebras generated by these n classes of events, $\sigma(\mathcal{C}_i), \ldots, \sigma(\mathcal{C}_n)$, are independent. •

Exercise 3.21 — **Basic independence criterion**
Prove the basic independence criterion in Theorem 3.20 (Resnick, 1999, pp. 92–93). •

[1] Synonyms (Resnick, 1999, p. 36): $\lambda-$system, $\sigma-$additive, Dynkin class.
[2] If $A, B \in \mathcal{D}$ and $A \subseteq B$ then $B \backslash A \in \mathcal{D}$.
[3] If $A_1 \subseteq A_2 \subseteq \ldots$ and $A_i \in \mathcal{D}$ then $\bigcup_{i=1}^{+\infty} A_i \in \mathcal{D}$.

Definition 3.22 — Arbitrary number of independent classes (Resnick, 1999, p. 93; Karr, 1993, p. 94)
Let T be an arbitrary index set. The classes $\{\mathcal{C}_t,\, t \in T\}$ are independent if, for each finite I such that $I \subset T$, $\{\mathcal{C}_t,\, t \in I\}$ are independent.

An infinite collection of $\sigma -$ algebras is independent if every finite subcollection is independent. •

Corollary 3.23 — Arbitrary number of independent classes (Resnick, 1999, p. 93)
If $\{\mathcal{C}_t,\, t \in T\}$ are non-empty $\pi -$ systems that are independent then $\{\sigma(\mathcal{C}_t),\, t \in T\}$ are independent. •

Exercise 3.24 — Arbitrary number of independent classes
Prove Corollary 3.23 by using the basic independence criterion. •

3.2 Independent r.v.

The notion of independence for r.v. can be stated in terms of Borel sets. Moreover, basic independence criteria can be develloped based solely on intervals such as $(-\infty, x]$.

Definition 3.25 — Independence of r.v. (Karr, 1993, p. 71)
R.v. $X_1, \ldots, X_n$ are independent if

$$P\left(\{X_1 \in B_1, \ldots, X_n \in B_n\}\right) = \prod_{i=1}^{n} P\left(\{X_i \in B_i\}\right), \tag{3.6}$$

for all Borel sets $B_1, \ldots, B_n$. •

Independence for r.v. can also be defined in terms of the independence of $\sigma-$algebras.

Definition 3.26 — Independence of r.v. (Resnick, 1999, p. 93)
Let T be an arbitrary index set. Then $\{X_t,\, t \in T\}$ is a family of independent r.v. if $\{\sigma(X_t),\, t \in T\}$ is a family of independent $\sigma-$algebras as stated in Definition 3.22. •

Remark 3.27 — Independence of r.v. (Resnick, 1999, p. 93)
The r.v. are independent if their induced/generated $\sigma-$algebras are independent. The information provided by any individual r.v. should not affect the probabilistic behaviour of other r.v. in the family.

Since $\sigma(\mathbf{1}_A) = \{\emptyset, A, A^c, \Omega\}$ we have $\mathbf{1}_{A_1}, \ldots, \mathbf{1}_{A_n}$ independent iff $A_1, \ldots, A_n$ are independent. •

Definition 3.28 — Independence of an infinite set of r.v. (Karr, 1993, p. 71)
An infinite set of r.v. is independent if every finite subset of r.v. is independent. •

Motivation 3.29 — Independence criterion for a finite number of r.v. (Karr, 1993, pp. 71–72)
R.v. are independent iff their joint d.f. is the product of their marginal/individual d.f.

This result affirms the general principle that definitions stated in terms of all Borel sets need only be checked for intervals $(-\infty, x]$. •

Theorem 3.30 — Independence criterion for a finite number of r.v. (Karr, 1993, p. 72)
R.v. $X_1, \ldots, X_n$ are independent iff

$$F_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = \prod_{i=1}^{n} F_{X_i}(x_i), \tag{3.7}$$

for all $x_1, \ldots, x_n \in I\!R$. •

Remark 3.31 — Independence criterion for a finite number of r.v. (Resnick, 1999, p. 94)
Theorem 3.30 is usually referred to as factorization criterion. •

Exercise 3.32 — Independence criterion for a finite number of r.v.
Prove Theorem 3.30 (Karr, 1993, p. 72; Resnick, 1999, p. 94, provides a more straightforward proof of this result). •

Theorem 3.33 — Independence criterion for an infinite number of r.v. (Resnick, 1994, p. 94)
Let T be as arbitrary index set. A family of r.v. $\{X_t,\, t \in T\}$ is independent iff

$$F_I(x_t, t \in I) = \prod_{t \in I} F_{X_t}(x_t), \tag{3.8}$$

for all finite $I \subset T$ and $x_t \in I\!R$. •

Exercise 3.34 — Independence criterion for an infinite number of r.v.
Prove Theorem 3.33 (Resnick, 1994, p. 94). •

Specialized criteria for discrete and absolutely continuous r.v. follow from Theorem 3.30.

Theorem 3.35 — Independence criterion for discrete r.v. (Karr, 1993, p. 73; Resnick, 1999, p. 94)
The discrete r.v. $X_1, \ldots, X_n$, with countable ranges $\mathcal{R}_1, \ldots, \mathcal{R}_n$, are independent iff

$$P(\{X_1 = x_1, \ldots, X_n = x_n\}) = \prod_{i=1}^{n} P(\{X_i = x_i\}), \tag{3.9}$$

for all $x_i \in \mathcal{R}_i$, $i = 1, \ldots, n$. •

Exercise 3.36 — Independence criterion for discrete r.v.
Prove Theorem 3.35 (Karr, 1993, p. 73; Resnick, 1999, pp. 94–95). •

Exercise 3.37 — Independence criterion for discrete r.v.
The number of laptops (X) and PCs (Y) sold daily in a store have a joint p.f. partially described in the following table:

X	Y		
	0	1	2
0	0.1	0.1	0.3
1	0.2	0.1	0.1
2	0	0.1	a

Complete the table and prove that X and Y are not independent r.v. •

Theorem 3.38 — Independence criterion for absolutely continuous r.v. (Karr, 1993, p. 74)
Let $\underline{X} = (X_1, \ldots, X_n)$ be an absolutely continuous random vector. Then $X_1, \ldots, X_n$ are independent iff

$$f_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = \prod_{i=1}^{n} f_{X_i}(x_i), \tag{3.10}$$

for all $x_1, \ldots, x_n \in I\!R$. •

Exercise 3.39 — Independence criterion for absolutely continuous r.v.
Prove Theorem 3.38 (Karr, 1993, p. 74). •

Exercise 3.40 — Independence criterion for absolutely continuous r.v.
The r.v. X and Y represent the lifetimes (in 10^3 hours) of two components of a control system and have joint p.d.f. given by

$$f_{X,Y}(x, y) = \begin{cases} 1, & 0 < x < 1,\ 0 < y < 1 \\ 0, & \text{otherwise.} \end{cases} \tag{3.11}$$

Prove that X and Y are independent r.v. •

Exercise 3.41 — Independence criterion for absolutely continuous r.v.
Let X and Y be two r.v. that represent, respectively, the width (in dm) and the length (in dm) of a rectangular piece. Admit the joint p.d.f. of (X, Y) is given by

$$f_{X,Y}(x, y) = \begin{cases} 2, & 0 < x < y < 1 \\ 0, & \text{otherwise.} \end{cases} \tag{3.12}$$

Prove that X and Y are not independent r.v. •

Example 3.42 — Independent r.v. (Karr, 1993, pp. 75–76)
Independent r.v. are inherent to certain probability structures.

- **Binary expansions**[4]

  Let P be the uniform distribution on $\Omega = [0, 1]$. Each point $\omega \in \Omega$ has a binary expansion

  $$\omega \to \quad 0.\, X_1(\omega)\, X_2(\omega) \ldots, \tag{3.13}$$

  where the X_i are functions from Ω to $\{0, 1\}$.

  This expansion is "unique" and it can be shown that $X_1, X_2, \ldots$ are independent and with a Bernoulli($p = \frac{1}{2}$) distribution.[5] Moreover, $\sum_{n=1}^{+\infty} 2^{-n} X_n \sim \text{Uniform}(0, 1)$.

  According to Resnick (1999, pp. 98-99), the binary expansion of 1 is **0.111...** since

  $$\sum_{n=1}^{+\infty} 2^{-n} \times 1 = 1. \tag{3.14}$$

  In addition, if a number such a $\frac{1}{2}$ has two possible binary expansions, we agree to use the non terminating one. Thus, even though $\frac{1}{2}$ has two expansions **0.0111...** and **0.1000...** because

  $$2^{-1} \times 0 + \sum_{n=2}^{+\infty} 2^{-n} \times 1 \;=\; \frac{1}{2} \tag{3.15}$$

  $$2^{-1} \times 1 + \sum_{n=2}^{+\infty} 2^{-n} \times 0 \;=\; \frac{1}{2}, \tag{3.16}$$

  by convention, we use the first binary expansion.

- **Multidimensional uniform distribution**

  Suppose that P is the uniform distribution on $[0, 1]^n$. Then the coordinate r.v. $U_i((\omega_1, \ldots, \omega_n)) = \omega_i$, $i = 1, \ldots, n$, are independent, and each of them is uniformly distributed on $[0, 1]$. In fact, for intervals $I_1, \ldots, I_n$,

  $$P\left(\{U_1 \in I_1, \ldots, U_n \in I_n\}\right) = \prod_{i=1}^{n} P\left(\{U_i \in I_i\}\right). \tag{3.17}$$

[4] Or *dyadic expansions of uniform random numbers* (Resnick, 1999, pp. 98-99).
[5] The proof of this result can also be found in Resnick (1999, pp. 99-100).

In other cases, whether r.v. are independent depends on the value of a parameter.

- **Standard bivariate normal distribution**

  Let (X, Y) be a random vector with a standard bivariate normal distribution with p.d.f.

$$f_{X,Y}(x,y) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left[-\frac{x^2 - 2\rho xy + y^2}{2(1-\rho^2)}\right], (x,y) \in I\!R^2 \tag{3.18}$$

  X and Y have both marginal standard normal distributions then, by the factorization criterion, X and Y are independent iff $\rho = 0$. •

Exercise 3.43 — **Bivariate normal distributed r.v.** (Karr, 1993, p. 96, Exercise 3.8)
Let (X, Y) have the bivariate normal p.d.f.

$$f_{X,Y}(x,y) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left[-\frac{x^2 - 2\rho xy + y^2}{2(1-\rho^2)}\right], (x,y) \in I\!R^2. \tag{3.19}$$

(a) Prove that $X \sim Y \sim \text{Normal}(0, 1)$.

(b) Prove that X and Y are independent iff $\rho = 0$.

(c) Prove that $P(\{X \geq 0, Y \geq 0\}) = \frac{1}{4} + \frac{1}{2\pi}\arcsin(\rho)$ (Grimmett and Stirzaker, 2001, pp. 196–197). •

Exercise 3.44 — **I.i.d. r.v. with absolutely continuous distributions** (Karr, 1993, p. 96, Exercise 3.9)
Let (X, Y) be an absolutely continuous random vector where X and Y are i.i.d. r.v. with absolutely continuous d.f. F. Prove that:

(a) $P(\{X = Y\}) = 0$;

(b) $P(\{X < Y\}) = \frac{1}{2}$. •

3.3 Functions of independent r.v.

Motivation 3.45 — Disjoint blocks theorem (Karr, 1993, p. 76)
R.v. that are functions of disjoint subsets of a family of independent r.v. are also independent. •

Theorem 3.46 — Disjoint blocks theorem (Karr, 1993, p. 76)
Let:

- $X_1, \ldots, X_n$ be independent r.v.;
- $J_1, \ldots, J_k$ be disjoint subsets of $\{1, \ldots, n\}$;
- $Y_l = g_l\left(X^{(l)}\right)$, where g_l is a Borel measurable function and $X^{(l)} = \{X_i,\, i \in J_l\}$ is a subset of the family of the independent r.v., for each $l = 1, \ldots, k$.

Then

$$Y_1 = g_1\left(X^{(1)}\right), \ldots, Y_k = g_k\left(X^{(k)}\right) \tag{3.20}$$

are independent r.v. •

Remark 3.47 — Disjoint blocks theorem (Karr, 1993, p. 77)
According to Definitions 3.25 and 3.28, the disjoint blocks theorem can be extended to (countably) infinite families and blocks. •

Exercise 3.48 — Disjoint blocks theorem
Prove Theorem 3.46 (Karr, 1993, pp. 76–77). •

Example 3.49 — Disjoint blocks theorem
Let $X_1, \ldots, X_5$ be five independent r.v., and $J_1 = \{1, 2\}$ and $J_2 = \{3, 4\}$ two disjoint subsets of $\{1, \ldots, 5\}$. Then

- $Y_1 = X_1 + X_2 = g_1(X_i,\, i \in J_1 = \{1, 2\})$ and
- $Y_2 = X_3 - X_4 = g_2(X_i,\, i \in J_2 = \{3, 4\})$

are independent r.v. •

Corollary 3.50 — Disjoint blocks theorem (Karr, 1993, p. 77)
Let:

- $X_1, \ldots, X_n$ be independent r.v.;
- $Y_i = g_i(X_i)$, $i = 1, \ldots, n$, where $g_1, \ldots, g_n$ are (Borel measurable) functions from $\mathbb{R}$ to $\mathbb{R}$.

Then $Y_1, \ldots, Y_n$ are independent r.v. •

We have already addressed the p.d.f. (or p.f.) of a sum, difference, product or division of two INDEPENDENT absolutely continuous (or discrete) r.v. However, the sum of INDEPENDENT absolutely continuous r.v. merit special consideration — its p.d.f. has a specific designation: *convolution of p.d.f.*.

Definition 3.51 — Convolution of p.d.f. (Karr, 1993, p. 77)
Let:

- X and Y be two INDEPENDENT absolutely continuous r.v.;
- f and g be the p.d.f. of X and Y, respectively.

Then the p.d.f. of $X + Y$ is termed the convolution of the p.d.f. f and g, represented by $f \star g$ and given by

$$(f \star g)(t) = \int_{-\infty}^{+\infty} f(t - s) \times g(s)\, ds. \tag{3.21}$$

•

Proposition 3.52 — Properties of the convolution of p.d.f. (Karr, 1993, p. 78)
The convolution of p.d.f. is:

- **commutative** — $f \star g = g \star f$, for all p.d.f. f and g;
- **associative** — $(f \star g) \star h = f \star (g \star h)$, for all p.d.f. f, g and h. •

Exercise 3.53 — Convolution of p.f.
How could we define the convolution of the p.f. of two INDEPENDENT discrete r.v.? •

Exercise 3.54 — Sum of independent binomial distributions
Let $X \sim \text{Binomial}(n_X, p)$ and $Y \sim \text{Binomial}(n_Y, p)$ be independent.

Prove that $X + Y \sim \text{Binomial}(n_X + n_Y, p)$ by using the Vandermonde's identity (http://en.wikipedia.org/wiki/Vandermonde's_identity).[6] •

Exercise 3.54 gives an example of a distribution family which is closed under convolution. There are several other families with the same property, as illustrated by the next proposition.

Proposition 3.55 — A few distribution families closed under convolution

R.v.	Convolution
$X_i \sim_{indep} \text{Binomial}(n_i, p)$, $i = 1, \ldots, k$	$\sum_{i=1}^k X_i \sim \text{Binomial}\left(\sum_{i=1}^k n_i, p\right)$
$X_i \sim_{indep} \text{NegativeBinomial}(r_i, p)$, $i = 1, \ldots, n$	$\sum_{i=1}^n X_i \sim \text{NegativeBinomial}\left(\sum_{i=1}^k r_i, p\right)$
$X_i \sim_{indep} \text{Poisson}(\lambda_i)$, $i = 1, \ldots, n$	$\sum_{i=1}^n X_i \sim \text{Poisson}\left(\sum_{i=1}^n \lambda_i\right)$
$X_i \sim_{indep.} \text{Normal}(\mu_i, \sigma_i^2)$, $i = 1, \ldots, n$	$\sum_{i=1}^n X_i \sim \text{Normal}\left(\sum_{i=1}^n \mu_i, \sum_{i=1}^n \sigma_i^2\right)$
	$\sum_{i=1}^n c_i X_i \sim \text{Normal}\left(\sum_{i=1}^n c_i \mu_i, \sum_{i=1}^n c_i^2 \sigma_i^2\right)$
$X_i \sim_{indep.} \text{Gamma}(\alpha_i, \lambda)$, $i = 1, \ldots, n$	$\sum_{i=1}^n X_i \sim \text{Gamma}\left(\sum_{i=1}^n \alpha_i, \lambda\right)$

•

Exercise 3.56 — Sum of (in)dependent normal distributions
Let (X, Y) have a (non-singular) bivariate normal distribution with mean vector and covariance matrix

$$\underline{\mu} = \begin{bmatrix} \mu_X \\ \mu_Y \end{bmatrix} \quad \text{and} \quad \mathbf{\Sigma} = \begin{bmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{bmatrix}, \tag{3.22}$$

respectively, that is, the joint p.d.f. is given by

$$f_{X,Y}(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x-\mu_X}{\sigma_X}\right)^2\right.\right. \tag{3.23}$$

$$\left.\left. -2\rho\left(\frac{x-\mu_X}{\sigma_X}\right)\left(\frac{y-\mu_Y}{\sigma_Y}\right) + \left(\frac{y-\mu_Y}{\sigma_Y}\right)^2\right]\right\}, \ (x,y) \in \mathbb{R}^2, \tag{3.24}$$

[6] In combinatorial mathematics, Vandermonde's identity, named after Alexandre-Théophile Vandermonde (1772), states that the equality $\binom{m+n}{r} = \sum_{k=0}^r \binom{m}{k}\binom{n}{r-k}$, $m, n, r \in \mathbb{N}_0$, for binomial coefficients holds. This identity was given already in 1303 by the Chinese mathematician Zhu Shijie (Chu Shi-Chieh).

for $|\rho| = |corr(X,Y)| < 1$.[7]

Prove that $X+Y$ is normally distributed with parameters $E(X+Y) = \mu_X + \mu_Y$ and $V(X+Y) = V(X) + 2cov(X,Y) + V(Y) = \sigma_X^2 + 2\rho\sigma_X\sigma_Y + \sigma_Y^2$. •

Exercise 3.57 — Distribution of the minimum of two exponentially distributed r.v. (Karr, 1993, p. 96, Exercise 3.7)

Let $X_i \sim$ Exponential(λ) and $Y \sim$ Exponential(μ) be two independent r.v.

Calculate the distribution of $Z = \min\{X, Y\}$. •

Exercise 3.58 — Distribution of the minimum of exponentially distributed r.v.

Let $X \overset{i.i.d.}{\sim}$ Exponential(λ_i) and $a_i > 0$, for $i = 1, \ldots, n$.

Prove that $\min_{i=1,\ldots,n}\{a_i\, X_i\} \sim$ Exponential $\left(\sum_{i=1}^n \frac{\lambda_i}{a_i}\right)$. •

Exercise 3.59 — Distribution of the minimum of Pareto distributed r.v.

The Pareto distribution, named after the Italian economist Vilfredo Pareto, was originally used to model the wealth of individuals, X.[8]

We say that $X \sim$ Pareto(b, α) if

$$f_X(x) = \frac{\alpha\, b^\alpha}{x^{\alpha+1}},\ x \geq b, \tag{3.25}$$

where $b > 0$ is the minimum possible value of X (it also represents the scale parameter) and $\alpha > 0$ is called the Pareto index (or the shape parameter)

Consider n individuals with wealths $X_i \overset{i.i.d.}{\sim} X$, $i = 1, \ldots, n$. Identify the survival function of the minimal wealth of these n individuals and comment on the result. •

Proposition 3.60 — A few distribution families closed under the minimum operation

R.v.	Minimum
$X_i \sim_{indep}$ Geometric(p_i), $i = 1, \ldots, n$	$\min_{i=1,\ldots,n} X_i \sim$ Geometric $\left(1 - \prod_{i=1}^n (1-p_i)\right)$
$X_i \sim_{indep}$ Exponential(λ_i), $a_i > 0$, $i = 1, \ldots, n$	$\min_{i=1,\ldots,n} a_i X_i \sim$ Exponential $\left(\sum_{i=1}^n \frac{\lambda_i}{a_i}\right)$
$X_i \sim_{indep}$ Pareto(b, α_i), $i = 1, \ldots, n$, $a > 0$	$\min_{i=1,\ldots,n} aX_i \sim$ Pareto $\left(ab, \sum_{i=1}^n \alpha_i\right)$

[7] The fact that two random variables X and Y both have a normal distribution does not imply that the pair (X, Y) has a joint normal distribution. A simple example is one in which $Y = X$ if $|X| > 1$ and $Y = -X$ if $|X| < 1$. This is also true for more than two random variables. (For more details see http://en.wikipedia.org/wiki/Multivariate_normal_distribution).

[8] The Pareto distribution seemed to show rather well the way that a larger portion of the wealth of any society is owned by a smaller percentage of the people in that society (http://en.wikipedia.org/wiki/Pareto_distribution).

R.v.	Minimum
$X_i \sim_{i.i.d.} \text{Weibull}(\alpha, \beta)$, $i = 1, \ldots, n$	$\min_{i=1,\ldots,n} X_i \sim \text{Weibull}\left(\alpha/n^{\frac{1}{\beta}}, \beta\right)$
$X_i \sim_{indep} \text{Weibull}(\alpha_i, \beta)$, $i = 1, \ldots, n$	$\min_{i=1,\ldots,n} X_i \sim \text{Weibull}\left(\left(\sum_{i=1}^{n} \alpha_i^{-\beta}\right)^{-\frac{1}{\beta}}, \beta\right)$

•

Exercise 3.61 — A few distribution families closed under the minimum operation

Prove Proposition 3.60 •

3.4 Order statistics

Algebraic operations on independent r.v., such as the minimum, the maximum and order statistics, are now further discussed because they play a major role in applied areas such as reliability.

Definition 3.62 — System reliability function (Barlow and Proschan, 1975, p. 82)
The system reliability function for the interval $[0, t]$ is the probability that the system functions successfully throughout the interval $[0, t]$.

If T represents the system lifetime then the system reliability function is the survival function of T,

$$S_T(t) = P(\{T > t\}) = 1 - F_T(t). \tag{3.26}$$

If the system has n components with INDEPENDENT lifetimes $X_1, \ldots, X_n$, with survival functions $S_{X_1}(t), \ldots, S_{X_n}(t)$, then system reliability function is a function of those n reliability functions, i.e,

$$S_T(t) = h\left[S_{X_1}(t), \ldots, S_{X_n}(t)\right]. \tag{3.27}$$

If they are not independent then $S_T(t)$ depends on more than the component marginal distributions at time t. •

Definition 3.63 — Order statistics
Given any r.v., $X_1, X_2, \ldots, X_n$,

- the 1st. order statistic is the minimum of $X_1, \ldots, X_n$, $X_{(1)} = \min_{i=1,\ldots,n} X_i$,
- nth. order statistic is the maximum of $X_1, \ldots, X_n$, $X_{(n)} = \max_{i=1,\ldots,n} X_i$, and
- the ith. order statistic corresponds to the ith.-smallest r.v. of $X_1, \ldots, X_n$, $X_{(i)}$, $i = 1, \ldots, n$.

Needless to say that the order statistics $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$ are also r.v., defined by sorting $X_1, X_2, \ldots, X_n$ in increasing order. Thus, $X_{(1)} \leq X_{(2)} \leq \ldots \leq X_{(n)}$. •

Motivation 3.64 — Importance of order statistics in reliabilty
A system lifetime T can be expressed as a function of order statistics of the components lifetimes, $X_1, \ldots, X_n$.

If we assume that $X_i \overset{i.i.d.}{\sim} X$, $i = 1, \ldots, n$, then the system reliability function $S_T(t) = P(\{T > t\})$ can be easily written in terms of the survival function (or reliability function) of X, $S_X(t) = P(\{X > t\})$, for some of the most usual reliability structures. •

Example 3.65 — Reliability function of a series system
A series system functions if all its components function. Therefore the system lifetime is given by

$$T = \min\{X_1, \ldots, X_n\} = X_{(1)}. \tag{3.28}$$

If $X_i \overset{i.i.d.}{\sim} X$, $i = 1, \ldots, n$, then the system reliability function is defined as

$$\begin{aligned} S_T(t) &= P\left(\bigcap_{i=1}^{n}\{X_i > t\}\right) \\ &= [S_X(t)]^n, \end{aligned} \tag{3.29}$$

where $S_X(t) = P(\{X > t\})$. •

Exercise 3.66 — Reliability function of a series system
A series system has two components with i.i.d. lifetimes with common failure rate function given by $\lambda_X(t) = \frac{f_X(t)}{S_X(t)} = 0.5t^{-0.5}$, $t \geq 0$, i.e., $S_X(t) = \exp\left[-\int_0^t \lambda_X(s)\, ds\right]$. (Prove this result!).

Derive the system reliability function. •

Example 3.67 — Reliability function of a parallel system
A parallel system functions if at least one of its components functions. Therefore the system lifetime is given by

$$T = \max\{X_1, \ldots, X_n\} = X_{(n)}. \tag{3.30}$$

If $X_i \overset{i.i.d.}{\sim} X$, $i = 1, \ldots, n$, then the system reliability function equals

$$\begin{aligned} S_T(t) &= 1 - F_T(t) \\ &= 1 - P\left(\bigcap_{i=1}^{n}\{X_i \leq t\}\right) \\ &= 1 - [1 - S_X(t)]^n, \end{aligned} \tag{3.31}$$

where $S_X(t) = P(\{X > t\})$. •

Exercise 3.68 — Reliability function of a parallel system
An obsolete electronic system has 6 valves set in parallel. Assume that the components lifetime (in years) are i.i.d. r.v. with common p.d.f. $f_X(t) = 50\, t\, e^{-25t^2}$, $t > 0$.

Obtain the system reliability for 2 months. •

Proposition 3.69 — Joint p.d.f. of the order statistics and more (Murteira, 1980, pp. 57, 55, 54)
Let $X_1, \ldots, X_n$ be absolutely continuous r.v. such that $X_i \overset{i.i.d.}{\sim} X$, $i = 1, \ldots, n$. Then:

$$f_{X_{(1)},\ldots,X_{(n)}}(x_1, \ldots, x_n) = n! \times \prod_{i=1}^{n} f_X(x_i), \tag{3.32}$$

for $x_1 \leq \ldots \leq x_n$;

$$\begin{aligned} F_{X_{(i)}}(x) &= \sum_{j=i}^{n} \binom{n}{j} \times [F_X(x)]^j \times [1 - F_X(x)]^{n-j} \\ &= 1 - F_{Binomial(n,F_X(x))}(i-1), \end{aligned} \tag{3.33}$$

for $i = 1, \ldots, n$;

$$f_{X_{(i)}}(x) = \frac{n!}{(i-1)!\,(n-i)!} \times [F_X(x)]^{i-1} \times [1 - F_X(x)]^{n-i} \times f_X(x), \tag{3.34}$$

for $i = 1, \ldots, n$;

$$\begin{aligned} f_{X_{(i)},X_{(j)}}(x_i, x_j) = &\frac{n!}{(i-1)!\,(j-i-1)!\,(n-j)!} \\ &\times [F_X(x_i)]^{i-1} \times [F_X(x_j) - F_X(x_i)]^{j-i-1} \times [1 - F_X(x_j)]^{n-j} \\ &\times f_X(x_i) \times f_X(x_j), \end{aligned} \tag{3.35}$$

for $x_i < x_j$, and $1 \leq i < j \leq n$. •

Exercise 3.70 — Joint p.d.f. of the order statistics and more
Prove Proposition 3.69 (http://en.wikipedia.org/wiki/Order_statistic). •

Example 3.71 — Reliability function of a k-out-of-n system
A k-out-of-n system functions if at least k out of its n components function. A series system corresponds to a n-out-of-n system, whereas a parallel system corresponds to a 1-out-of-n system. The lifetime of a k-out-of-n system is also associated to an order statistic:

$$T = X_{(n-k+1)}. \tag{3.36}$$

If $X_i \overset{i.i.d.}{\sim} X$, $i = 1, \ldots, n$, then the system reliability function can also be derived by using the auxiliary r.v.

$$Z_t = \text{number of } X_i's > t \ \sim \ \text{Binomial}(n, S_X(t)). \tag{3.37}$$

In fact,

$$\begin{aligned} S_T(t) &= P(Z_t \geq k) \\ &= 1 - P(Z_t \leq k-1) \\ &= 1 - F_{Binomial(n,S_X(t))}(k-1) \\ &= P(n - Z_t \leq n-k) \\ &= F_{Binomial(n,F_X(t))}(n-k). \end{aligned} \tag{3.38}$$

•

Exercise 3.72 — Reliability function of a k-out-of-n system
Admit a machine has 4 engines and it only functions if at least 3 of those engines are working. Moreover, suppose the lifetimes of the engines (in thousand hours) are i.i.d. r.v. with Exponential distribution with scale parameter $\lambda^{-1} = 2$.

Obtain the machine reliability for a period of 1000 h. •

3.5 Constructing independent r.v.

The following theorem is similar to Proposition 2.133 and guarantees that we can also construct independent r.v. with prescribed d.f.

Theorem 3.73 — Construction of a finite collection of independent r.v. with prescribed d.f. (Karr, 1993, p. 79)
Let $F_1, \ldots, F_n$ be d.f. on $\mathbb{R}$. Then there is a probability space $(\Omega, \mathcal{F}, P)$ and r.v. $X_1, \ldots, X_n$ defined on it such that $X_1, \ldots, X_n$ are independent r.v. and $F_{X_i} = F_i$ for each i. •

Exercise 3.74 — Construction of a finite collection of independent r.v. with prescribed d.f.
Prove Proposition 3.73 (Karr, 1993, p. 79). •

Theorem 3.75 — Construction of a sequence of independent r.v. with prescribed d.f. (Karr, 1993, p. 79)
Let $F_1, F_2, \ldots$ be d.f. on $\mathbb{R}$. Then there is a probability space $(\Omega, \mathcal{F}, P)$ and a r.v. $X_1, X_2, \ldots$ defined on it such that $X_1, X_2, \ldots$ are independent r.v. and $F_{X_i} = F_i$ for each i. •

Exercise 3.76 — Construction of a sequence of independent r.v. with prescribed d.f.
Prove Proposition 3.75 (Karr, 1993, pp. 79-80). •

3.6 Bernoulli process

Motivation 3.77 — **Bernoulli (counting) process** (Karr, 1993, p. 88)
Counting sucesses in repeated, independent trials, each of which has one of two possible outcomes (*success*[9] and *failure*). •

Definition 3.78 — **Bernoulli process** (Karr, 1993, p. 88)
A Bernoulli process with parameter p is a sequence $\{X_i,\ i \in \mathbb{N}\}$ of i.i.d. r.v. with Bernoulli distribution with parameter $p = P(success)$. •

Definition 3.79 — **Important r.v. in a Bernoulli process** (Karr, 1993, pp. 88–89)
In isolation a Bernoulli process is neither deep or interesting. However, we can identify three associated and very important r.v.:

- $S_n = \sum_{i=1}^{n} X_i$, the number of successes in the first n trials ($n \in \mathbb{N}$);
- $T_k = \min\{n : S_n = k\}$, the *time* (trial number) at which the kth. success occurs ($k \in \mathbb{N}$), that is, the number of trials needed to get k successes;
- $U_k = T_k - T_{k-1}$, the *time* (number of trials) between the kth. and $(k-1)$th. successes ($k \in \mathbb{N}$, $T_0 = 0$, $U_1 = T_1$). •

Definition 3.80 — **Bernoulli counting process** (Karr, 1993, p. 88)
The sequence $\{S_n,\ n \in \mathbb{N}\}$ is usually termed as Bernoulli counting process (or success counting process).[10] •

Exercise 3.81 — **Bernoulli counting process**
Simulate a Bernoulli process with parameter $p = \frac{1}{2}$ and consider $n = 100$ trials. Plot the realizations of both the Bernoulli process and the Bernoulli counting process. •

Definition 3.82 — **Bernoulli success time process** (Karr, 1993, p. 88)
The sequence $\{T_k,\ k \in \mathbb{N}\}$ is usually called the Bernoulli success time process. •

[9] Or *arrival*.

[10] S_n represents the total number of successes that have occurred up to trial n, thus $\{S_n,\ n \in \mathbb{N}\}$ satisfies: $S_n \in \mathbb{N}_0$, $n \in \mathbb{N}$; $S_m \leq S_n$, $m \leq n$, $m, n \in \mathbb{N}$; $S_n - S_m$ ($m \leq n$, $m, n \in \mathbb{N}$) corresponds to the number of successes that have occurred after trial m and up to tril n. (For the continuous time analogue, see Definition 3.96.)

Proposition 3.83 — Important distributions in a Bernoulli process (Karr, 1993, pp. 89–90)
In a Bernoulli process with parameter p ($p \in [0,1]$) we have:

- $S_n \sim \text{Binomial}(n,p)$, $n \in \mathbb{N}$;
- $T_k \sim \text{NegativeBinomial}(k,p)$, $k \in \mathbb{N}$;
- $U_k \overset{i.i.d.}{\sim} \text{Geometric}(p) \overset{d}{=} \text{NegativeBinomial}(1,p)$, $k \in \mathbb{N}$. •

Exercise 3.84 — Bernoulli counting process

(a) Prove that $T_k \sim \text{NegativeBinomial}(k,p)$ and $U_k \overset{i.i.d.}{\sim} \text{Geometric}(p)$, for $k \in \mathbb{N}$.

(b) Consider a Bernoulli process with parameter $p = 1/2$ and obtain the probability of having 57 successes between times 10 and 100. •

Exercise 3.85 — Relating the Bernoulli counting process and random walk (Karr, 1993, p. 97, Exercise 3.21)
Let S_n be a Bernoulli (counting) process with $p = \frac{1}{2}$.
Prove that the process $Z_n = 2S_n - n$ is a symmetric random walk. •

Proposition 3.86 — Properties of the Bernoulli counting process (Karr, 1993, p. 90)
The Bernoulli counting process $\{S_n, n \in \mathbb{N}\}$ has:

- independent increments — i.e., for $0 < n_1 < \ldots < n_k$, the r.v. S_{n_1}, $S_{n_2} - S_{n_1}$, $S_{n_3} - S_{n_2}$, …, $S_{n_k} - S_{n_{k-1}}$ are independent;
- stationary increments — that is, for fixed j, the distribution of $S_{k+j} - S_k$ is the same for all $k \in \mathbb{N}$. •

Exercise 3.87 — Properties of the Bernoulli counting process
Prove Proposition 3.86 (Karr, 1993, p. 90). •

Remark 3.88 — Bernoulli counting process (web.mit.edu/6.262/www/lectures/6.262.Lec1.pdf)
Some application areas for discrete stochastic processes such as the Bernoulli counting process (and the Poisson process, studied in the next section) are:

- **Operations Research**

  Queueing in any area, failures in manufacturing systems, finance, risk modelling, network models

- **Biology and Medicine**

  Epidemiology, genetics and DNA studies, cell modelling, bioinformatics, medical screening, neurophysiology

- **Computer Systems**

  Communication networks, intelligent control systems, data compression, detection of signals, job flow in computer systems, physics – statistical mechanics. •

Exercise 3.89 — Bernoulli process modelling of sexual HIV transmission (Pinkerton and Holtgrave (1998, pp. 13–14))
In the Bernoulli-process model of sexual HIV transmission, each act of sexual intercourse is treated as an independent stochastic trial that is associated to a probability α of HIV transmission. α is also known as the *infectivity* of HIV and a number of factors are believed to influence α.[11]

(a) Prove that the expression of the probability of HIV transmission in n multiple contacts with the same infected partner is $1-(1-\alpha)^n$.

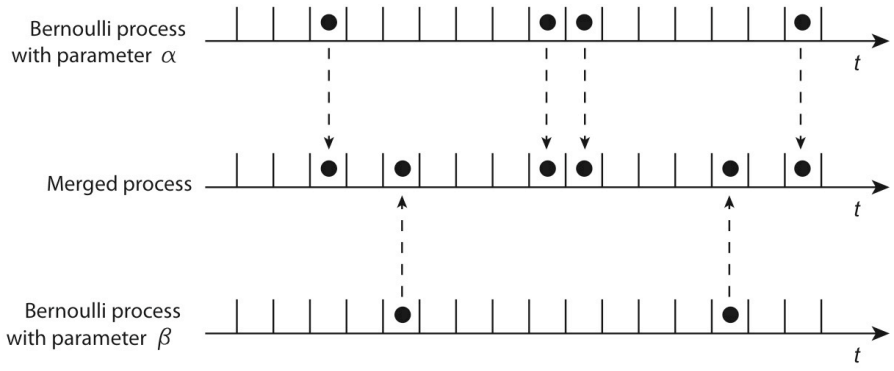
(b) Assume now that the consistent use of condoms reduce the infectivity from α to $\alpha' = (1-0.9)\times\alpha$.[12] Derive the relative change reduction in the probability defined in (a) due to the consistent use of condoms. Evaluate this reduction when $\alpha = 0.01$ and $n = 10$. •

Definition 3.90 — Independent Bernoulli processes
Two Bernoulli counting processes $\{S_n^{(1)},\, n \in I\!N\}$ and $\{S_n^{(2)},\, n \in I\!N\}$ are independent if for every positive integer k and all times $n_1, \ldots, n_k$, we have that the random vector $\left(S_{n_1}^{(1)}, \ldots, S_{n_k}^{(1)}\right)$ associated with the first process is independent of $\left(S_{n_1}^{(2)}, \ldots, S_{n_k}^{(2)}\right)$ associated with the second process. •

Proposition 3.91 — Merging independent Bernoulli processes
Let $\{S_n^{(1)},\, n \in I\!N\}$ and $\{S_n^{(2)},\, n \in I\!N\}$ be two independent Bernoulli counting processes with parameters α and β, respectively. Then the merged process $\{S_n^{(1)} \oplus S_n^{(2)},\, n \in I\!N\}$ is a Bernoulli counting process with parameter $\alpha + \beta - \alpha\beta$.

[11] Such as the type of sex act engaged, sex role, etc.

[12] α' is termed *reduced infectivity*; 0.9 represents a conservative estimate of condom effectiveness.



•

Exercise 3.92 — Merging independent Bernoulli processes

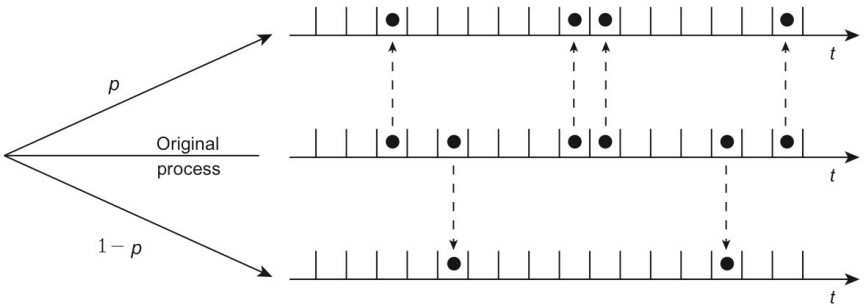
(a) Prove Proposition 3.91.

(b) Assume time is divided into consecutive fixed-length time-slots. Consider N sensors and assume that the i^{th} sensor triggers an alarm in any given fixed-length time-slot with probability α_i $(i = 1, \ldots, N)$, independently of the remaining sensors.

Obtain the probability that at least one alarm sounds in a given time-slot (Zukerman, 2000–2012, p. 60) and the p.f. of the number of times at least one alarm sounds in the first n time-slots. •

Proposition 3.93 — Splitting a Bernoulli process (or sampling a Bernoulli process)

Let $\{S_n,\, n \in \mathbb{N}\}$ be a Bernoulli counting process with parameter α. Splitting the original Bernoulli counting process based on a *selection* probability p yields two Bernoulli counting processes with parameters αp and $\alpha(1-p)$.



•

Exercise 3.94 — Splitting a Bernoulli process

(a) Prove Proposition 3.93.

Are the two resulting processes independent?[13]

(b) Assume once again that time is divided into consecutive fixed-length time-slots. Moreover, assume an alarm is triggered in a fixed-length time-slot with probability α and subsequently checked whether it is a false alarm (with probability p) or not.

Determine the p.f. of the number of time-slots between consecutive false alarms. •

[13]NO! If we try to merge the two splitting processes and assume they are independent we get a parameter $\alpha p + \alpha(1-p) - \alpha p \times \alpha(1-p)$ which is different form α.

3.7 Poisson process

In what follows we use the notation of Ross (1989, Chapter 5) which is slightly different from the one of Karr (1993, Chapter 3).

Motivation 3.95 — **Poisson process** (Karr, 1993, p. 91)
Is there a continuous analogue of the Bernoulli process? YES!

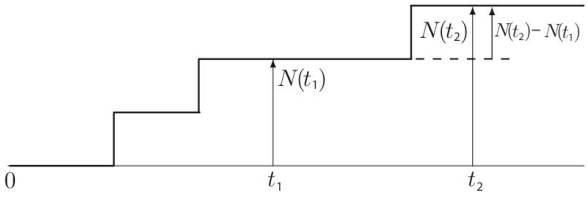
The Poisson process, named after the French mathematician Siméon-Denis Poisson (1781–1840), is the stochastic process in which events occur continuously and independently of one another. Examples that are well-modeled as Poisson processes include the radioactive decay of atoms, telephone calls arriving at a switchboard, and page view requests to a website.[14] •

Definition 3.96 — **Counting process (in continuous time)** (Ross, 1989, p. 209)
A stochastic process $\{N(t),\, t \geq 0\}$ is said to be a counting process if $N(t)$ represents the total number of events (e.g. arrivals) that have occurred up to time t. From this definition we can conclude that a counting process $\{N(t),\, t \geq 0\}$ must satisfy:

- $N(t) \in \mathbb{N}_0,\ \forall\, t \geq 0$;
- $N(s) \leq N(t),\ \forall\, 0 \leq s < t$;
- $N(t) - N(s)$ corresponds to the number of events that have occurred in the interval $(s, t],\ \forall\, 0 \leq s < t$. •

Definition 3.97 — **Counting process (in continuous time) with independent increments** (Ross, 1989, p. 209)
The counting process $\{N(t),\, t \geq 0\}$ is said to have independent increments if the number of events that occur in disjoint intervals are independent r.v., i.e.,

- for $0 < t_1 < \ldots < t_n$, $N(t_1)$, $N(t_2) - N(t_1)$, $N(t_3) - N(t_2)$, …, $N(t_n) - N(t_{n-1})$ are independent r.v.



•

[14] For more examples, check http://en.wikipedia.org/wiki/Poisson_process.

Definition 3.98 — Counting process (in continuous time) with stationary increments (Ross, 1989, p. 210)
The counting process $\{N(t),\, t \geq 0\}$ is said to have stationary increments if distribution of the number of events that occur in any interval of time depends only on the length of the interval,[15] that is,

- $N(t_2 + s) - N(t_1 + s) \overset{d}{=} N(t_2) - N(t_1)$, $\forall\, s > 0$, $0 \leq t_1 < t_2$. •

Definition 3.99 — Poisson process (Karr, 1993, p. 91)
A counting process $\{N(t),\, t \geq 0\}$ is said to be a Poisson process with rate λ ($\lambda > 0$) if:

- $\{N(t),\, t \geq 0\}$ has independent and stationary increments;
- $N(t) \sim \text{Poisson}(\lambda t)$. •

Remark 3.100 — Poisson process (Karr, 1993, p. 91)
Actually, $N(t) \sim \text{Poisson}(\lambda t)$ follows from the fact that $\{N(t),\, t \geq 0\}$ has independent and stationary increments, thus, redundant in Definition 3.99. •

Definition 3.101 — The definition of a Poisson process revisited (Ross, 1989, p. 212)
The counting process $\{N(t), t \geq 0\}$ is said to be a Poisson process with rate λ, if

- $N(0) = 0$;
- $\{N(t),\, t \geq 0\}$ has independent and stationary increments;
- $P(\{N(h) = 1\}) = \lambda h + o(h)$;[16]
- $P(\{N(h) \geq 2\}) = o(h)$. •

Exercise 3.102 — The definition of a Poisson process revisited
Prove that Definitions 3.99 and 3.101 are equivalent (Ross, 1989, pp. 212–214).

[15] The distributions do not depend on the origin of the time interval; they only depend on the length of the interval.

[16] The function f is said to be $o(h)$ if $\lim_{h\to 0} \frac{f(h)}{h} = 0$ (Ross, 1989, p. 211).
The function $f(x) = x^2$ is $o(h)$ since $\lim_{h\to 0} \frac{f(h)}{h} = \lim_{h\to 0} h = 0$.
The function $f(x) = x$ is not $o(h)$ since $\lim_{h\to 0} \frac{f(h)}{h} = \lim_{h\to 0} 1 = 1 \neq 0$.

Proposition 3.103 — Joint p.f. of $N(t_1), \ldots, N(t_n)$ in a Poisson process (Karr, 1993, p. 91)
For $0 < t_1 < \ldots < t_n$ and $0 \leq k_1 \leq \ldots \leq k_n$,

$$P(\{N(t_1) = k_1, \ldots, N(t_n) = k_n\}) = \prod_{j=1}^{n} \frac{e^{-\lambda(t_j - t_{j-1})} \left[\lambda(t_j - t_{j-1})\right]^{k_j - k_{j-1}}}{(k_j - k_{j-1})!}, \tag{3.39}$$

where $t_0 = 0$ and $k_0 = 0$. •

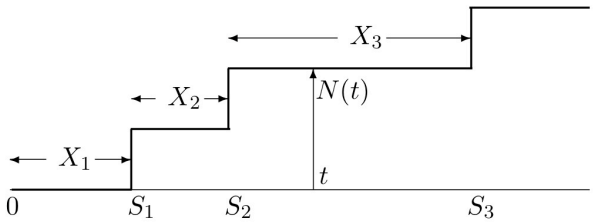
Exercise 3.104 — Joint p.f. of $N(t_1), \ldots, N(t_n)$ in a Poisson process
Prove Proposition 3.103 (Karr, 1993, p. 92) by taking advantage, namely, of the fact that a Poisson process has independent and stationary increments. •

Exercise 3.105 — Joint p.f. of $N(t_1), \ldots, N(t_n)$ in a Poisson process ("Stochastic Processes" — Test of 2002-11-09)
A machine produces electronic components according to a Poisson process with rate equal to 10 components per hour. Let $N(t)$ be the number of produced components up to time t.

Evaluate the probability of producing at least 8 components in the first hour given that exactly 20 components have been produced in the first two hours. •

Definition 3.106 — Important r.v. in a Poisson process (Karr, 1993, pp. 88–89)
Let $\{N(t),\, t \geq 0\}$ be a Poisson process with rate λ. Then:

- $S_n = \inf\{t : N(t) = n\}$ represents the time of the occurrence of the nth. event (e.g. arrival), $n \in \mathbb{N}$;
- $X_n = S_n - S_{n-1}$ corresponds to the time between the nth. and $(n-1)$th. events (e.g. interarrival time), $n \in \mathbb{N}$.



•

Proposition 3.107 — Important distributions in a Poisson process (Karr, 1993, pp. 92–93)
So far we know that $N(t) \sim \text{Poisson}(\lambda t)$, $t > 0$. We can also add that:

- $S_n \sim \text{Erlang}(n, \lambda)$, $n \in \mathbb{N}$;
- $X_n \overset{i.i.d.}{\sim} \text{Exponential}(\lambda)$, $n \in \mathbb{N}$. •

Remark 3.108 — Relating $N(t)$ and S_n in a Poisson process
We ought to note that:

$$\begin{aligned} N(t) \geq n &\Leftrightarrow S_n \leq t \qquad (3.40) \\ F_{S_n}(t) &= F_{Erlang(n,\lambda)}(t) \\ &= P(\{N(t) \geq n\}) \\ &= \sum_{j=n}^{+\infty} \frac{e^{-\lambda t}(\lambda t)^j}{j!} \\ &= 1 - F_{Poisson(\lambda t)}(n-1),\ n \in \mathbb{N}. \qquad (3.41) \end{aligned}$$

•

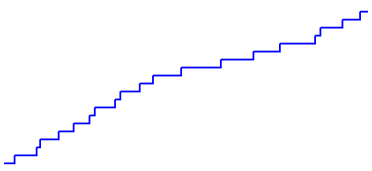
Exercise 3.109 — Important distributions in a Poisson process
Prove Proposition 3.107 (Karr, 1993, pp. 92–93). •

Exercise 3.110 — Time between events in a Poisson process
Suppose that people immigrate into a territory at a Poisson rate $\lambda = 1$ per day.

What is the probability that the elapsed time between the tenth and the eleventh arrival exceeds two days? (Ross, 1989, pp. 216–217). •

Exercise 3.111 — Poisson process
Simulate a Poisson process with rate $\lambda = 1$ considering the interval $[0, 100]$. Plot the realizations of the Poisson process.

The sample path of a Poisson process should look like this:



•

Motivation 3.112 — Conditional distribution of the first arrival time (Ross, 1989, p. 222)

Suppose we are told that exactly one event of a Poisson process has taken place by time t (i.e. $N(t) = 1$), and we are asked to determine the distribution of the time at which the event occurred (S_1). •

Proposition 3.113 — Conditional distribution of the first arrival time (Ross, 1989, p. 223)

Let $\{N(t),\, t \geq 0\}$ be a Poisson process with rate $\lambda > 0$. Then

$$S_1 | \{N(t) = 1\} \sim \text{Uniform}(0, t). \tag{3.42}$$

•

Exercise 3.114 — Conditional distribution of the first arrival time

Prove Proposition 3.113 (Ross, 1989, p. 223). •

Proposition 3.115 — Conditional distribution of the arrival times (Ross, 1989, p. 224)

Let $\{N(t),\, t \geq 0\}$ be a Poisson process with rate $\lambda > 0$. Then

$$f_{S_1,\ldots,S_n|\{N(t)=n\}}(s_1, \ldots, s_n) = \frac{n!}{t^n}, \tag{3.43}$$

for $0 < s_1 < \ldots < s_n < t$ and $n \in \mathbb{N}$. •

Remark 3.116 — Conditional distribution of the arrival times (Ross, 1989, p. 224)

Proposition 3.115 is usually paraphrased as stating that, under the condition that n events have occurred in $(0, t)$, the times $S_1, \ldots, S_n$ at which events occur, considered as unordered r.v., are i.i.d. and Uniform$(0, t)$.[17] •

Exercise 3.117 — Conditional distribution of the arrival times

Prove Proposition 3.115 (Ross, 1989, p. 224). •

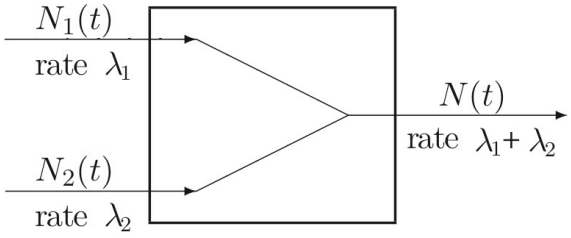
Proposition 3.118 — Merging independent Poisson processes

Let $\{N_1(t),\, t \geq 0\}$ and $\{N_2(t),\, t \geq 0\}$ be two INDEPENDENT Poisson processes[18] with rates λ_1 and λ_2, respectively.

[17] I.e., they behave as the order statistics $Y_{(1)}, \ldots, Y_{(n)}$, associated to $Y_i \overset{i.i.d.}{\sim} \text{Uniform}(0, t)$.

[18] How can one define two independent Poisson processes? As in Definition 3.90: two Poisson processes $\{N_1(t),\, t \geq 0\}$ and $\{N_2(t),\, t \geq 0\}$ are independent if for every positive integer k and all times $t_1, \ldots, t_k$, we have that the random vector $(N_1(t_1), \ldots, N_1(t_k))$ associated with the first process is independent of $(N_2(t_1), \ldots, N_2(t_k))$ associated with the second process.

Then the merged process $\{N_1(t)+N_2(t),\, t \geq 0\}$ is a Poisson process with rate $\lambda_1+\lambda_2$.



•

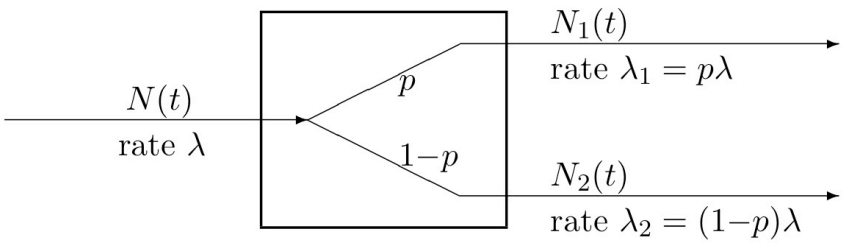
Exercise 3.119 — Merging independent Poisson processes
Prove Proposition 3.118. •

Exercise 3.120 — Merging independent Poisson processes
Men and women enter a supermarket according to independent Poisson processes having respective rates two and four per minute.

(a) Starting at an arbitrary time, compute the probability that at least two men arrive before three women arrive (Ross, 1989, p. 242, Exercise 20).

(b) What is the probability that the number of arrivals (men and women) exceeds ten in the first 20 minutes? •

Proposition 3.121 — Splitting a Poisson process (or sampling a Poisson process)
Let $\{N(t),\, t \geq 0\}$ be a Poisson process with rate λ. Splitting the original Poisson process based on a *selection* probability p yields two INDEPENDENT Poisson processes with rates λp and $\lambda(1-p)$.



Moreover, we can add that $N_1(t)|\{N(t)=n\} \sim \text{Binomial}(n,p)$ and $N_2(t)|\{N(t)=n\} \sim \text{Binomial}(n,1-p)$. •

Exercise 3.122 — Splitting a Poisson process
Prove Proposition 3.121 (Ross, 1989, pp. 218–219).

Why are the two resulting processes independent? •

Exercise 3.123 — Splitting a Poisson process
If immigrants to area A arrive at a Poisson rate of ten per week, and if each immigrant is of English descent with probability $\frac{1}{12}$, then what is the probability that no people of English descent will emigrate to area A during the month of February (Ross, 1989, p. 220). •

Exercise 3.124 — Splitting a Poisson process (Ross, 1989, p. 243, Exercise 23)
Cars pass a point on the highway at a Poisson rate of one per minute. If five percent of the cars on the road are Dodges, then:

(a) What is the probability that at least one Dodge passes during an hour?

(b) If 50 cars have passed by an hour, what is the probability that five of them were Dodges?

(c) Given that ten Dodges have passed by in an hour, obtain the expected value of the number of cars to have passed by in that time. •

3.8 Generalizations of the Poisson process

In this section we consider three generalizations of the Poisson process. The first of these is the non homogeneous Poisson process, which is obtained by allowing the arrival rate at time t to be a function of t.

Definition 3.125 — **Non homogeneous Poisson process** (Ross, 1989, p. 234)
The counting process $\{N(t), t \geq 0\}$ is said to be a non homogeneous Poisson process with intensity function $\lambda(t)$ $(t \geq 0)$ if

- $N(0) = 0$;
- $\{N(t), t \geq 0\}$ has independent increments;
- $P(\{N(t+h) - N(t) = 1\}) = \lambda(t) \times h + o(h)$, $t \geq 0$;
- $P(\{N(t+h) - N(t) \geq 2\}) = o(h)$, $t \geq 0$.

Moreover,

$$N(t+s) - N(s) \sim \text{Poisson}\left(\int_s^{t+s} \lambda(z)\, dz\right) \tag{3.44}$$

for $s \geq 0$ and $t > 0$. •

Exercise 3.126 — **Non homogeneous Poisson process** ("Stochastic Processes" test, 2003-01-14)
The number of arrivals to a shop is governed by a Poisson process with time dependent rate

$$\lambda(t) = \begin{cases} 4 + 2t, & 0 \leq t \leq 4 \\ 24 - 3t, & 4 < t \leq 8. \end{cases}$$

(a) Obtain the expression of the expected value of the number of arrivals until t $(0 \leq t \leq 8)$. Derive the probability of no arrivals in the interval [3,5].

(b) Determine the expected value of the number of arrivals in the last 5 opening hours (interval $[3, 8]$) given that 15 customers have arrived in the last 3 opening hours (interval $[5, 8]$). •

Exercise 3.127 — The output process of an infinite server Poisson queue and the non homogeneous Poisson process
Prove that the output process of the $M/G/\infty$ queue — i.e., the number of customers who (by time t) have already left the infinite server queue with Poisson arrivals and general service d.f. G — is a non homogeneous Poisson process with intensity function $\lambda G(t)$. •

Definition 3.128 — Compound Poisson process (Ross, 1989, p. 237)
A stochastic process $\{X(t), t \geq 0\}$ is said to be a compound Poisson process if it can be represented as

$$X(t) = \sum_{i=1}^{N(t)} Y_i, \tag{3.45}$$

where

- $\{N(t),\, t \geq 0\}$ is a Poisson process with rate λ ($\lambda > 0$) and
- $Y_i \overset{i.i.d.}{\sim} Y$ and independent of $\{N(t),\, t \geq 0\}$. •

Proposition 3.129 — Compound Poisson process (Ross, 1989, pp. 238–239)
Let $\{X(t), t \geq 0\}$ be a compound Poisson process. Then

$$E[X(t)] = \lambda t \times E[Y] \tag{3.46}$$
$$V[X(t)] = \lambda t \times E[Y^2]. \tag{3.47}$$

•

Exercise 3.130 — Compound Poisson process
Prove Proposition 3.129 by noting that $E[X(t)] = E\{E[X(t)|N(t)]\}$ and $V[X(t)] = E\{V[X(t)|N(t)]\} + V\{E[X(t)|N(t)]\}$ (Ross, 1989, pp. 238–239). •

Exercise 3.131 — Compound Poisson process (Ross, 1989, p. 239)
Suppose that families migrate to an area at a Poisson rate $\lambda = 2$ per week. Assume that the number of people in each family is independent and takes values 1, 2, 3 and 4 with respective probabilities $\frac{1}{6}$, $\frac{1}{3}$, $\frac{1}{3}$ and $\frac{1}{6}$.

What is the expected value and variance of the number of individuals migrating to this area during a five-week period? •

Definition 3.132 — Conditional Poisson process (Ross, 1983, pp. 49–50)
Let:

- Λ be a positive r.v. having d.f. G; and
- $\{N(t),\, t \geq 0\}$ be a counting process such that, given that $\{\Lambda = \lambda\}$, $\{N(t),\, t \geq 0\}$ is a Poisson process with rate λ.

Then $\{N(t),\, t \geq 0\}$ is called a conditional Poisson process and

$$P(\{N(t+s) - N(s) = n\}) = \int_0^{+\infty} \frac{e^{-\lambda t}(\lambda t)^n}{n!}\, dG(\lambda). \tag{3.48}$$

•

Remark 3.133 — Conditional Poisson process (Ross, 1983, p. 50)
$\{N(t),\, t \geq 0\}$ is not a Poisson process. For instance, whereas it has stationary increments, it has not independent increments. •

Exercise 3.134 — Conditional Poisson process
Suppose that, depending on factors not at present understood, the rate at which seismic shocks occur in a certain region over a given season is either λ_1 or λ_2. Suppose also that the rate equals λ_1 for $p \times 100\%$ of the seasons and λ_2 in the remaining time.

A simple model would be to suppose that $\{N(t),\, t \geq 0\}$ is a conditional Poisson process such that Λ is either λ_1 or λ_2 with respective probabilities p and $1 - p$.

Prove that the probability that it is a λ_1-season, given n shocks in the first t units of a season, equals

$$\frac{p\, e^{-\lambda_1 t}(\lambda_1 t)^n}{p\, e^{-\lambda_1 t}(\lambda_1 t)^n + (1-p)\, e^{-\lambda_2 t}(\lambda_2 t)^n}, \tag{3.49}$$

by applying the Bayes' theorem (Ross, 1983, p. 50). •

Stochastic process	**Independent increments?**	**Stationary increments?**
Homogeneous PP	YES!!!	YES!!!
Non-homogeneous PP	YES!!!	NO!
Conditional PP	NO!	YES!!!
Compound PP	YES!!!	YES!!!

References

- Barlow, R.E. and Proschan, F. (1965/1996). *Mathematical Theory of Reliability.* SIAM (Classics in Applied Mathematics).
(TA169.BAR.64915)

- Barlow, R.E. and Proschan, F. (1975). *Reliability and Life Testing.* Holt, Rinehart and Winston, Inc.

- Grimmett, G.R. and Stirzaker, D.R. (2001). *Probability and Random Processes (3rd. edition).* Oxford University Press. (QA274.12-.76.GRI.30385 and QA274.12-.76.GRI.40695 refer to the library code of the 1st. and 2nd. editions from 1982 and 1992, respectively.)

- Karr, A.F. (1993). *Probability.* Springer-Verlag.

- Pinkerton, S.D. and Holtgrave, D.R. (1998). The Bernoulli-process model in HIV transmission: applications and implications. In *Handbook of economic evaluation of HIV prevention programs*, Holtgrave, D.R. (Ed.), pp. 13–32. Plenum Press, New York.

- Resnick, S.I. (1999). *A Probability Path.* Birkhäuser. (QA273.4-.67.RES.49925)

- Ross, S.M. (1983). *Stochastic Processes.* John Wiley & Sons. (QA274.12-.76.ROS.36921 and QA274.12-.76.ROS.37578)

- Ross, S.M. (1989). *Introduction to Probability Models* (fourth edition). Academic Press. (QA274.12-.76.ROS.43540 refers to the library code of the 5th. revised edition from 1993.)

- Zukerman, M. (2000–2012). *Introduction to Queueing Theory and Stochastic Teletraffic Models.* (arxiv.org/pdf/1307.2968)

Chapter 4

Expectation

One of the most fundamental concepts of probability theory and mathematical statistics is the expectation of a r.v. (Resnick, 1999, p. 117).

Motivation 4.1 — **Expectation** (Karr, 1993, p. 101)
The expectation represents the center of *gravity* of a r.v. and has a measure theory counterpart in integration theory.

Key computational formulas — not definitions of expectation — to obtain the expectation of

- a discrete r.v. X with values in a countable set $\mathcal{C}$ and p.f. $P(\{X = x\})$ and
- the one of an absolutely continuous r.v. Y with p.d.f. $f_Y(y)$

are

$$E(X) = \sum_{x \in \mathcal{C}} x \times P(\{X = x\}) \tag{4.1}$$

$$E(Y) = \int_{-\infty}^{+\infty} y \times f_Y(y)\, dy, \tag{4.2}$$

respectively.

When $X \geq 0$ it is permissible that $E(X) = +\infty$, but finiteness is mandatory when X can take both positive and negative (or null) values. •

Remark 4.2 — **Desired properties of expectation** (Karr, 1993, p. 101)

1. **Constant preserved**

   If $X \equiv c$ then $E(X) = c$.

2. **Monotonicity**

   If $X \leq Y$ [1] then $E(X) \leq E(Y)$.

3. **Linearity**

   For $a, b \in I\!R$, $E(aX + bY) = aE(X) + bE(Y)$.

4. **Continuity**[2]

   If $X_n \to X$ then $E(X_n) \to E(X)$.

5. **Relation to the probability**

   For each event A, $E(\mathbf{1}_A) = P(A)$.[3] •

Expectation is to r.v. as probability is to events so that properties of expectation extend those of probability.

4.1 Definition and fundamental properties

Many integration results are proved by first showing that they hold true for simple r.v. and then extending the result to more general r.v. (Resnick, 1999, p. 117).

4.1.1 Simple r.v.

Let $(\Omega, \mathcal{F}, P)$ be a probability space and let us remind the reader that X is said to be a simple r.v. if it assumes only finitely many values in which case

$$X = \sum_{i=1}^{n} a_i \times \mathbf{1}_{A_i}, \tag{4.3}$$

where:

- $a_1, \ldots, a_n$ are real numbers not necessarily distinct;
- $\{A_1, \ldots, A_n\}$ constitutes a partition of Ω;
- $\mathbf{1}_{A_i}$ is the indicator function of event A_i, $i = 1, \ldots, n$.

[1] I.e. $X(\omega) \leq Y(\omega), \forall\, \omega \in \Omega$.
[2] Continuity is not valid without restriction.
[3] Recall that $\mathbf{1}_{A_i}(\omega) = 1$, if $w \in A_i$, and $\mathbf{1}_{A_i}(\omega) = 0$, otherwise.

Consider, for example, $X \sim \text{Binomial}(2, p)$. In this case:

- $\Omega = \{FF, FS, SF, SS\}$ (where $F =$ fail and $S =$ success);
- $A_1 = FF$, $A_2 = FS$, $A_3 = SF$, $A_4 = SS$;
- $a_1 = 0$, $a_2 = 1$, $a_3 = 1$, $a_4 = 2$;
- $X = \sum_{i=1}^{n} a_i \times \mathbf{1}_{A_i}$.

Definition 4.3 — Expectation of a simple r.v. (Karr, 1993, p. 102)
The expectation of the simple r.v. $X = \sum_{i=1}^{n} a_i \times \mathbf{1}_{A_i}$ is given by

$$E(X) = \sum_{i=1}^{n} a_i \times P(A_i). \tag{4.4}$$

•

Remark 4.4 — Expectation of a simple r.v. (Resnick, 1999, p. 119; Karr, 1993, p. 102)

- Note that Definition 4.3 coincides with our knowledge of discrete probability from more elementary courses: the expectation is computed by taking a possible value, multiplying by the probability of the possible value and then summing over all possible values.
- $E(X)$ is well-defined in the sense that all representations of X yield the same value for $E(X)$: different representations of X, $X = \sum_{i=1}^{n} a_i \times \mathbf{1}_{A_i}$ and $X = \sum_{j=1}^{m} a'_j \times \mathbf{1}_{A'_j}$, lead to the same expected value $E(X) = \sum_{i=1}^{n} a_i \times P(A_i) = \sum_{j=1}^{m} a'_j \times P(A'_j)$.
- The expectation of an indicator function is indeed the probability of the associated event. •

Proposition 4.5 — Properties of the set of simple r.v. (Resnick, 1999, p. 118)
Let $\mathcal{E}$ be the set of all simple r.v. defined on $(\Omega, \mathcal{F}, P)$. We have the following properties of $\mathcal{E}$.

1. $\mathcal{E}$ is a vector space, i.e.:
   (a) if $X = \sum_{i=1}^{n} a_i \times \mathbf{1}_{A_i} \in \mathcal{E}$ and $\alpha \in \mathbb{R}$ then $\alpha X = \sum_{i=1}^{n} \alpha a_i \times \mathbf{1}_{A_i} \in \mathcal{E}$; and

(b) If $X = \sum_{i=1}^{n} a_i \times \mathbf{1}_{A_i} \in \mathcal{E}$ and $Y = \sum_{j=1}^{m} b_j \times \mathbf{1}_{B_j} \in \mathcal{E}$ then

$$X + Y = \sum_{i=1}^{n} \sum_{j=1}^{m} (a_i + b_j) \times \mathbf{1}_{A_i \cap B_j} \in \mathcal{E}. \tag{4.5}$$

2. If $X, Y \in \mathcal{E}$ then $XY \in \mathcal{E}$ since

$$XY = \sum_{i=1}^{n} \sum_{j=1}^{m} (a_i \times b_j) \times \mathbf{1}_{A_i \cap B_j}. \tag{4.6}$$

•

Proposition 4.6 — Expectation of a linear combination of simple r.v. (Karr, 1993, p. 103)
Let X and Y be two simple r.v. and $a, b \in \mathbb{R}$. Then $aX + bY$ is also a simple r.v. and

$$E(aX + bY) = aE(X) + bE(Y). \tag{4.7}$$

•

Exercise 4.7 — Expectation of a linear combination of simple r.v.
Prove Proposition 4.6 by capitalizing on Proposition 4.5 (Karr, 1993, p. 103). •

Exercise 4.8 — Expectation of a sum of discrete r.v. in a distribution problem (Walrand, 2004, p. 51, Example 4.10.6)
Suppose you put m balls randomly in n boxes. Each box can hold an arbitrarily large number of balls.

Prove that the expected number of empty boxes is equal to $n \times \left(\frac{n-1}{n}\right)^m$. •

Exercise 4.9 — Expectation of a sum of discrete r.v. in a selection problem (Walrand, 2004, p. 52, Example 4.10.7)
A cereal company is running a promotion for which it is giving a toy in every box of cereal. There are n different toys and each box is equally likely to contain any one of the n toys.

Prove that the expected number of boxes of cereal you have to purchase to collect all n toys is given by $n \times \sum_{m=1}^{n} \frac{1}{m}$. •

Remark 4.10 — Monotonicity of expectation for simple r.v. (Karr, 1993, p. 103)
The monotonicity of expectation for simple r.v. is a desired property which follows from

- linearity and
- positivity (or better said non negativity) — if $X \geq 0$ then $E(X) \geq 0$ —, a seemingly weaker property of expectation.

In fact, if $X \leq Y \Leftrightarrow Y - X \geq 0$ then

- $E(Y) - E(X) = E(Y - X) \geq 0$.

This argument is valid provided that $E(Y) - E(X)$ is not of the form $+\infty - \infty$. •

Proposition 4.11 — Monotonicity of expectation for simple r.v. (Karr, 1993, p. 103)
Let X and Y be two simple r.v. such that $X \leq Y$. Then $E(X) \leq E(Y)$. •

Example 4.12 — On the (dis)continuity of expectation of simple r.v. (Karr, 1993, pp. 103–104)
Continuity of expectation fails even for simple r.v. Let P be the uniform distribution on $[0, 1]$ and

$$X_n = n \times \mathbf{1}_{(0,\frac{1}{n})}. \tag{4.8}$$

(X_n takes values: n, if $\omega \in (0, \frac{1}{n})$; 0, otherwise.) Then $X_n(\omega) \to 0, \forall\, w \in \Omega$, but $E(X_n) = 1$, for each n.

Thus, we need additional conditions to guarantee continuity of expectation. •

4.1.2 Non negative r.v.

Before we proceed with the definition of the expectation of non negative r.v.,[4] we need to recall the measurability theorem. This theorem state that any non negative r.v. can be approximated by a simple r.v., and it is the reason why it is often the case that an integration result about non negative r.v. — such as the expectation and its properties — is proven first to simple r.v.

Theorem 4.13 — Measurability theorem (Resnick, 1999, p. 91; Karr, 1993, p. 50) Suppose $X(\omega) \geq 0$, for all ω. Then $X : \Omega \to I\!R$ is a Borel measurable function (i.e. a r.v.) iff there is an increasing sequence of simple and non negative r.v. $X_1, X_2, \ldots$ $(0 \leq X_1 \leq X_2 \leq \ldots)$ such that

$$X_n \uparrow X, \tag{4.9}$$

$(X_n(\omega) \uparrow X(\omega)$, for every $\omega)$. •

Exercise 4.14 — Measurability theorem
Prove Theorem 4.13 by considering

$$X_n = \sum_{k=1}^{n2^n} \frac{k-1}{2^n} \mathbf{1}_{\{\frac{k-1}{2^n} \leq X < \frac{k}{2^n}\}} + n \times \mathbf{1}_{\{X \geq n\}}, \tag{4.10}$$

for each n (Resnick, 1999, p. 118; Karr, 1993, p. 50). •

Motivation 4.15 — Expectation of a non negative r.v. (Karr, 1993, pp. 103–104) We now extend the definition of expectation to all non negative r.v. However, we have already seen that continuity of expectation fails even for simple r.v. and therefore we cannot define the expected value of a non negative r.v. simply as $E(X) = \lim_{n \to +\infty} E(X_n)$.

Unsurprisingly, if we apply the measurability theorem then the definition of expectation of a non negative r.v. virtually forces monotone continuity for increasing sequences of non negative r.v.:

- if $X_1, X_2, \ldots$ are simple and non negative r.v. and X is a non negative r.v. such that $X_n \uparrow X$ (pointwise) then $E(X_n) \uparrow E(X)$.

[4] Karr (1993) and Resnick (1999) call these r.v. *positive* when they are actually non negative.

It is convenient and useful to assume that these non negative r.v. can take values in the extended set of non negative real numbers, $\overline{I\!R}_0^+$.

Further on, we shall have to establish another restricted form of continuity: dominated continuity for integrable r.v.[5] •

Definition 4.16 — Expectation of a non negative r.v. (Karr, 1993, p. 104)
The expectation of a non negative r.v. X is

$$E(X) = \lim_{n \to +\infty} E(X_n) \leq +\infty, \tag{4.11}$$

where X_n are simple and non negative r.v. such that $X_n \uparrow X$.

The expectation of X over the event A is $E(X; A) = E(X \times \mathbf{1}_A)$. •

Remark 4.17 — Expectation of a non negative r.v. (Karr, 1993, p. 104)
The limit defining $E(X)$

- exists in the set of extended non negative real numbers $\overline{I\!R}_0^+$, and
- does not depend on the approximating sequence $\{X_n, n \in I\!N\}$, as stated in the next proposition. •

Proposition 4.18 — Expectation of a non negative r.v. (Karr, 1993, p. 104)
Let $\{X_n, n \in I\!N\}$ and $\{\tilde{X}_m, m \in I\!N\}$ be sequences of simple and non negative r.v. increasing to X. Then

$$\lim_{n \to +\infty} E(X_n) = \lim_{m \to +\infty} E(\tilde{X}_m). \tag{4.12}$$

•

Exercise 4.19 — Expectation of a non negative r.v.
Prove Proposition 4.18 (Karr, 1993, p. 104; Resnick, 1999, pp. 122–123). •

We now list some basic properties of the expectation operator applied to non negative r.v. For instance, linearity, monotonicity and monotone continuity/convergence. This last property describes how expectation and limits interact, and under which circunstances we are allowed to interchange expectation and limits.

[5] We shall soon define integrable r.v.

Proposition 4.20 — Expectation of a linear combination of non negative r.v. (Karr, 1993, p. 104; Resnick, 1999, p. 123)
Let X and Y be two non negative r.v. and $a, b \in \mathbb{R}^+$. Then

$$E(aX + bY) = aE(X) + bE(Y). \tag{4.13}$$

•

Exercise 4.21 — Expectation of a linear combination of non negative r.v.
Prove Proposition 4.20 by considering two sequences of simple and non negative r.v. $\{X_n, n \in \mathbb{N}\}$ and $\{Y_n, n \in \mathbb{N}\}$ such that $X_n \uparrow X$ and $Y_n \uparrow Y$ — and, thus, $(aX_n + bY_n) \uparrow (aX + bY)$ — (Karr, 1993, p. 104). •

Corollary 4.22 — Monotonicity of expectation for non negative r.v. (Karr, 1993, p. 105)
Let X and Y be two non negative r.v. such that $X \leq Y$. Then $E(X) \leq E(Y)$. •

Remark 4.23 — Monotonicity of expectation for non negative r.v. (Karr, 1993, p. 105)
Monotonicity of expectation follows, once again, from positivity and linearity. •

Motivation 4.24 — Fatou's lemma (Karr, 1993, p. 105)
The next result plays a vital role in the definition of monotone continuity/convergence. •

Theorem 4.25 — Fatou's lemma (Karr, 1993, p. 105; Resnick, 1999, p. 132)
Let $\{X_n, n \in \mathbb{N}\}$ be a sequence of non negative r.v. Then

$$E(\liminf X_n) \leq \liminf E(X_n). \tag{4.14}$$

•

Remark 4.26 — Fatou's lemma (Karr, 1993, p. 105)
The inequality (4.14) in Fatou's lemma can be strict. For instance, in Example 4.12 we are dealing with $E(\liminf X_n) = 0 < \liminf E(X_n) = 1$. •

Exercise 4.27 — Fatou's lemma
Prove Theorem 4.25 (Karr, 1993, p. 105; Resnick, 1999, p. 132). •

Exercise 4.28 — **Fatou's lemma and continuity of p.f.**
Verify that Theorem 4.25 could be used in a part of the proof of the continuity of p.f. if we considered $X_n = \mathbf{1}_{A_n}$ (Karr, 1993, p. 106). •

We now state another property of expectation of non negative r.v.: the monotone continuity/convergence of expectation.

Theorem 4.29 — **Monotone convergence theorem** (Karr, 1993, p. 106; Resnick, 1999, pp. 123–124)
Let $\{X_n, n \in \mathbb{N}\}$ be an increasing sequence of non negative r.v. and X a non negative r.v. If

$$X_n \uparrow X \tag{4.15}$$

then

$$E(X_n) \uparrow E(X). \tag{4.16}$$

•

Remark 4.30 — **Monotone convergence theorem** (Karr, 1993, p. 106)
The sequence of simple and non negative r.v. from Example 4.12, $X_n = n \times \mathbf{1}_{(0,\frac{1}{n})}$, does not violate the monotone convergence theorem because in that instance it is not true that $X_n \uparrow X$.[6] •

Exercise 4.31 — **Monotone convergence theorem**
Prove Theorem 4.29 (Karr, 1993, p. 106; Resnick, 1999, pp. 124–125, for a more sophisticated proof). •

Exercise 4.32 — **Monotone convergence theorem and monotone continuity of p.f.**
Verify that Theorem 4.29 could be used to prove the monotone continuity of p.f. if we considered $X_n = \mathbf{1}_{A_n}$ and $X = \mathbf{1}_A$, where $A_n \uparrow A$ (Karr, 1993, p. 106). •

One of the implications of the monotone convergence theorem is the linearity of expectation for convergent series, and is what Resnick (1999, p. 131) calls the series version of the monotone convergence theorem. This results refers under which circunstances we are allowed to interchange expectation and limits.

[6] Please note that the sequence is not even increasing: n increases but the sequence of sets $(0, \frac{1}{n})$ is a decreasing one.

Theorem 4.33 — Expectation of a linear convergent series of non negative r.v. (Karr, 1993, p. 106; Resnick, 1999, p. 131)
Let $\{Y_k, k \in I\!N\}$ be a collection of non negative r.v. such that $\sum_{k=1}^{+\infty} Y_k(\omega) < +\infty$, for every ω. Then

$$E\left(\sum_{k=1}^{+\infty} Y_k\right) = \sum_{k=1}^{+\infty} E(Y_k). \tag{4.17}$$

•

Exercise 4.34 — Expectation of a linear convergent series of non negative r.v.
Prove Theorem 4.33 by considering $X_n = \sum_{k=1}^{n} Y_k$ and applying the monotone convergence theorem (Karr, 1993, p. 106). •

Exercise 4.35 — Expectation of a linear convergent series of non negative r.v. and $\sigma-$additivity
Verify that Theorem 4.33 could be used to prove the $\sigma-$additivity of p.f. if we considered $Y_k = \mathbf{1}_{A_k}$, where A_k are disjoint, so that $\sum_{k=1}^{+\infty} Y_k = \mathbf{1}_{\bigcup_{k=1}^{+\infty} A_k}$ (Karr, 1993, p. 106). •

Proposition 4.36 – "Converse" of the positivity of expectation (Karr, 1993, p. 107)
Let X be a non negative r.v. If $E(X) = 0$ then $X \overset{a.s.}{=} 0$. •

Exercise 4.37 – "Converse" of the positivity of expectation
Prove Proposition 4.36 (Karr, 1993, p. 107). •

4.1.3 Integrable r.v.

It is time to extend the definition of expectation to r.v. X that can take both positive and negative (or null) values. But first recall that:

- $X^+ = \max\{X, 0\}$ represents the positive part of the r.v. X;
- $X^- = -\min\{X, 0\} = \max\{-X, 0\}$ represents the negative part of the r.v. X;
- $X = X^+ - X^-$;
- $|X| = X^+ + X^-$.

The definition of expectation of such a r.v. preserves linearity and is based on the fact that X can be written as a linear combination of two non negative r.v.: $X = X^+ - X^-$.

Definition 4.38 — Integrable r.v.; the set of integrable r.v. (Karr, 1993, p. 107; Resnick, 1999, p. 126)
Let X be a r.v., not necessarily non negative. Then X is said to be integrable if $E(|X|) < +\infty$.

The set of integrable r.v. is denoted by L^1 or $L^1(P)$ if the probability measure needs to be emphasized. •

Definition 4.39 — Expectation of an integrable r.v. (Karr, 1993, p. 107)
Let X be an integrable r.v. Then the expectation of X is given by

$$E(X) = E(X^+) - E(X^-). \tag{4.18}$$

For an event A, the expectation of X over A is $E(X; A) = E(X \times \mathbf{1}_A)$. •

Remark 4.40 — Expectation of an integrable r.v. (Karr, 1993, p. 107; Resnick, 1999, p. 126)

1. If X is an integrable r.v. then

$$E(X^+) + E(X^-) = E(X^+ + X^-) = E(|X|) < +\infty \tag{4.19}$$

so both $E(X^+)$ and $E(X^-)$ are finite, $E(X^+) - E(X^-)$ is not of the form $\infty - \infty$, thus, the definition of expectation of X is coherent.

2. Moreover, since $|X \times \mathbf{1}_A| \leq |X|$, $E(X; A) = E(X \times \mathbf{1}_A)$ is finite (i.e. exists!) as long as $E(|X|) < +\infty$, that is, as long as $E(X)$ exists.

3. Some conventions when X is **not integrable**...

   If $E(X^+) < +\infty$ but $E(X^-) = +\infty$ then we consider $E(X) = -\infty$.

   If $E(X^+) = +\infty$ but $E(X^-) < +\infty$ then we take $E(X) = +\infty$.

   If $E(X^+) = +\infty$ and $E(X^-) = +\infty$ then $E(X)$ does not exist. •

What follows refers to properties of the expectation operator.

Theorem 4.41 — Expectation of a linear combination of integrable r.v. (Karr, 1993, p. 107)
Let X and Y be two integrable r.v. — i.e., $X, Y \in L^1$ — and $a, b \in \mathbb{R}$. Then $aX + bY$ is also an integrable r.v.[7] and

$$E(aX + bY) = aE(X) + bE(Y). \tag{4.20}$$

•

Exercise 4.42 — Expectation of a linear combination of integrable r.v.
Prove Theorem 4.41 (Karr, 1993, p. 108). •

Corollary 4.43 — Modulus inequality (Karr, 1993, p. 108; Resnick, 1999, p. 128)
If $X \in L^1$ then

$$|E(X)| \leq E(|X|). \tag{4.21}$$

•

Exercise 4.44 — Modulus inequality
Prove Corollary 4.43 (Karr, 1993, p. 108). •

Corollary 4.45 — Monotonicity of expectation for integrable r.v. (Karr, 1993, p. 108; Resnick, 1999, p. 127)
If $X, Y \in L^1$ and $X \leq Y$ then

$$E(X) \leq E(Y). \tag{4.22}$$

•

Exercise 4.46 — Monotonicity of expectation for integrable r.v.
Prove Corollary 4.45 (Resnick, 1999, pp. 127–128). •

[7] That is, $aX + bY \in L^1$. In fact, L^1 is a vector space.

The continuity of expectation for integrable r.v. can be finally stated.

Theorem 4.47 — Dominated convergence theorem (Karr, 1993, p. 108; Resnick, 1999, p. 133)
Let $X_1, X_2, \ldots \in L^1$ and $X \in L^1$ with

$$X_n \to X. \tag{4.23}$$

If there is a dominating r.v. $Y \in L^1$ such that

$$|X_n| \leq Y, \tag{4.24}$$

for each n, then

$$\lim_{n \to +\infty} E(X_n) = E(X). \tag{4.25}$$

•

Remark 4.48 — Dominated convergence theorem (Karr, 1993, p. 109)
The sequence of simple r.v., $X_n = n \times \mathbf{1}_{(0, \frac{1}{n})}$, from Example 4.12 does not violate the dominated convergence theorem because any r.v. Y dominating X_n for each n must satisfy $Y \geq \sum_{n=1}^{+\infty} n \times \mathbf{1}_{(\frac{1}{n+1}, \frac{1}{n})}$, which implies that $E(Y) = +\infty$, thus $Y \notin L^1$ and we cannot apply Theorem 4.47. •

Exercise 4.49 — Dominated convergence theorem
Prove Theorem 4.47 (Karr, 1993, p. 109; Resnick, 1999, p. 133, for a detailed proof). •

Exercise 4.50 — Dominated convergence theorem and continuity of p.f.
Verify that Theorem 4.47 could be used to prove the continuity of p.f. if we considered $X_n = \mathbf{1}_{A_n}$, where $A_n \to A$, and $Y \equiv 1$ as the dominating integrable r.v. (Karr, 1993, p. 109). •

4.1.4 Complex r.v.

Definition 4.51 — Integrable complex r.v.; expectation of a complex r.v. (Karr, 1993, p. 109)
A complex r.v. $Z = X + iY \in L^1$ if $E(|Z|) = E(\sqrt{X^2 + Y^2}) < +\infty$, and in this case the expectation Z is $E(Z) = E(X) + iE(Y)$. •

4.2 Integrals with respect to distribution functions

Integrals (of Borel measurable functions) with respect to d.f. are known as Lebesgue–Stieltjes integrals. Moreover, they are really expectations with respect to probabilities on $\mathbb{R}$ and are reduced to sums and Riemann (more generally, Lebesgue) integrals.

4.2.1 On integration

Remark 4.52 — **Riemann integral** (http://en.wikipedia.org/wiki/Riemann_integral) In the branch of mathematics known as real analysis, the Riemann integral, created by Bernhard Riemann (1826–1866), was the first rigorous definition of the integral of a function on an interval.

- **Overview**

  Let g be a non-negative real-valued function of the interval $[a, b]$, and let $S = \{(x, y) : 0 < y < g(x)\}$ be the region of the plane under the graph of the function g and above the interval $[a, b]$.

  The basic idea of the Riemann integral is to use very simple approximations for the area of S, denoted by $\int_a^b g(x)\,dx$, namely by taking better and better approximations — we can say that "in the limit" we get exactly the area of S under the curve.

- **Riemann sums**

  Choose a real-valued function f which is defined on the interval $[a, b]$. The Riemann sum of f with respect to the tagged partition $a = x_0 < x_1 < x_2 < \ldots < x_n = b$ together with $t_0, \ldots, t_{n-1}$ (where $x_i \leq t_i \leq x_{i+1}$) is

  $$\sum_{i=0}^{n-1} g(t_i)(x_{i+1} - x_i), \tag{4.26}$$

  where each term represents the area of a rectangle with height $g(t_i)$ and length $x_{i+1} - x_i$. Thus, the Riemann sum is the signed area under all the rectangles.

- **Riemann integral**

  Loosely speaking, the Riemann integral is the limit of the Riemann sums of a function as the partitions get finer.

  If the limit exists then the function is said to be integrable (or more specifically Riemann-integrable).

- **Limitations of the Riemann integral**

  With the advent of Fourier series, many analytical problems involving integrals came up whose satisfactory solution required interchanging limit processes and integral signs.

  Failure of monotone convergence — The indicator function $\mathbf{1}_{\mathbb{Q}}$ on the rationals is not Riemann integrable. No matter how the set $[0, 1]$ is partitioned into subintervals, each partition will contain at least one rational and at least one irrational number, since rationals and irrationals are both dense in the reals. Thus, the upper Darboux sums[8] will all be one, and the lower Darboux sums[9] will all be zero.

  Unsuitability for unbounded intervals — The Riemann integral can only integrate functions on a bounded interval. It can however be extended to unbounded intervals by taking limits, so long as this does not yield an answer such as $+\infty - \infty$.

  What about integrating on structures other than Euclidean space? — The Riemann integral is inextricably linked to the order structure of the line. How do we free ourselves of this limitation? •

Remark 4.53 — **Lebesgue integral** (http://en.wikipedia.org/wiki/Lebesgue_integral; http://en.wikipedia.org/wiki/Henri_Lebesgue)
Lebesgue integration plays an important role in real analysis, the axiomatic theory of probability, and many other fields in the mathematical sciences. The Lebesgue integral is a construction that extends the integral to a larger class of functions defined over spaces more general than the real line.

- **Lebesgue's theory of integration**

  Henri Léon Lebesgue (1875–1941) invented a new method of integration to solve this problem. Instead of using the areas of rectangles, which put the focus on the domain of the function, Lebesgue looked at the codomain of the function for his fundamental unit of area. Lebesgue's idea was to first build the integral for what he called simple functions, measurable functions that take only finitely many values. Then he defined it for more complicated functions as the least upper bound of all the integrals of simple functions smaller than the function in question.

[8]The upper Darboux sum of g with respect to the partition is $\sum_{i=0}^{n-1}(x_{i+1} - x_i)M_{i+1}$, where $M_{i+1} = \sup_{x\in[x_i,x_{i+1}]} g(x)$.

[9]The lower Darboux sum of g with respect to the partition is $\sum_{i=0}^{n}(x_{i+1} - x_i)m_{i+1}$, where $m_{i+1} = \inf_{x\in[x_i,x_{i+1}]} g(x)$.

Lebesgue integration has the beautiful property that every bounded function defined over a bounded interval with a Riemann integral also has a Lebesgue integral, and for those functions the two integrals agree. But there are many functions with a Lebesgue integral that have no Riemann integral.

As part of the development of Lebesgue integration, Lebesgue invented the concept of Lebesgue measure, which extends the idea of length from intervals to a very large class of sets, called measurable sets.

- **Integration**

  We start with a measure space $(\Omega, \mathcal{F}, \mu)$ where Ω is a set, $\mathcal{F}$ is a $\sigma - algebra$ of subsets of Ω and μ is a (non-negative) measure on $\mathcal{F}$ of subsets of Ω.

  In the mathematical theory of probability, we confine our study to a probability measure μ, which satisfies $\mu(\Omega) = 1$.

  In Lebesgue's theory, integrals are defined for a class of functions called measurable functions.

  We build up an integral $\int_\Omega g\, d\mu$ for measurable real-valued functions g defined on Ω in stages:

  - **Indicator functions**. To assign a value to the integral of the indicator function of a measurable set S consistent with the given measure μ, the only reasonable choice is to set
    $$\int_\Omega \mathbf{1}_S \, d\mu = \mu(S).$$
  - **Simple functions**. A finite linear combination of indicator functions $\sum_k a_k \mathbf{1}_{S_k}$. When the coefficients a_k are non-negative and S_k are disjoint sets of Ω, we set
    $$\int_\Omega (\sum_k a_k \mathbf{1}_{S_k})\, d\mu = \sum_k a_k \int_\Omega 1_{S_k}\, d\mu = \sum_k a_k\, \mu(S_k).$$
  - **Non-negative functions**. We define
    $$\int_\Omega g\, d\mu = \sup \left\{ \int_\Omega s\, d\mu : 0 \le s \le g,\ s \text{ simple} \right\}.$$
  - **Signed functions**. $g = g^+ - g^-$... And it makes sense to define
    $$\int_\Omega g d\mu = \int_\Omega g^+ d\mu - \int_\Omega g^- d\mu.$$

-

Remark 4.54 — Lebesgue/Riemann–Stieltjes integration
(http://en.wikipedia.org/wiki/Lebesgue-Stieltjes_integration)
The Lebesgue–Stieltjes integral is the ordinary Lebesgue integral with respect to a measure known as the Lebesgue–Stieltjes measure, which may be associated to any function of bounded variation on the real line.

- **Definition**

  The Lebesgue–Stieltjes integral $\int_a^b g(x)\,dF(x)$ is defined when $g : [a, b] \to \mathbb{R}$ is Borel-measurable and bounded and $F : [a, b] \to \mathbb{R}$ is of bounded variation in $[a, b]$ and right-continuous, or when g is non-negative and F is monotone and right-continuous.

- **Riemann–Stieltjes integration and probability theory**

  When g is a continuous real-valued function of a real variable and F is a non-decreasing real function, the Lebesgue–Stieltjes integral is equivalent to the Riemann–Stieltjes integral, in which case we often write $\int_a^b g(x)\,dF(x)$ for the Lebesgue–Stieltjes integral, letting the measure P_F remain implicit.

  This is particularly common in probability theory when F is the cumulative distribution function of a real-valued random variable X, in which case

  $$\int_{-\infty}^{\infty} g(x)\,dF(x) = E_F[g(X)].$$

•

4.2.2 Generalities

First of all, we should recall that given a d.f. F on $\mathbb{R}$, there is a unique p.f. on $\mathbb{R}$ such that $P_F((a, b]) = F(b) - F(a)$.

Moreover, all functions g appearing below are assumed to be Borel measurable.

Definition 4.55 — Integral of a nonnegative g with respect to a d.f. (Karr, 1993, p. 110)
Let F be a d.f. on $\mathbb{R}$ and g a non negative function. Then the integral of g with respect to F is given by

$$\int_{\mathbb{R}} g(x)\,dF(x) = E_F(g) \leq +\infty, \tag{4.27}$$

where the expectation is that of $g(X)$ as a Borel measurable function of the r.v. X defined on the probability space $(\mathbb{R}, \mathcal{B}(\mathbb{R}), P_F)$. •

Definition 4.56 — Integrable of a function with respect to a d.f. (Karr, 1993, p. 110)
Let F be a d.f. on $\mathbb{R}$ and g a signed function. Then g is said to be integrable with respect to F if $\int_{\mathbb{R}} g(x)\,dF(x) < +\infty$, and in this case, the integral of g with respect to F equals

$$\int_{\mathbb{R}} g(x)\,dF(x) = \int_{\mathbb{R}} g^+(x)\,dF(x) - \int_{\mathbb{R}} g^-(x)\,dF(x). \tag{4.28}$$

•

Definition 4.57 — Integral of a function over a set with respect to a d.f. (Karr, 1993, p. 110)
Let F be a d.f. on $\mathbb{R}$ and g either non negative or integrable and $B \in \mathcal{B}(\mathbb{R})$. The integral of g over B with respect to F is equal to

$$\int_B g(x)\,dF(x) = \int_{\mathbb{R}} g(x) \times \mathbf{1}_B(x)\,dF(x). \tag{4.29}$$

•

The properties of the integral of a function with respect to a d.f. are those of expectation:

1. Constant preserved
2. Monotonicity
3. Linearity
4. Relation to P_F
5. Fatou's lemma
6. Monotone convergence theorem
7. Dominated convergence theorem

4.2.3 Discrete distribution functions

Keep in mind that integrals with respect to discrete d.f. are sums.

Theorem 4.58 — Integral with respect to a discrete d.f. (Karr, 1993, p. 111)
Consider a d.f. $F(t)$ that can be written as

$$F(x) = \sum_i p_i \times \mathbf{1}_{[x_i, +\infty)}(x). \tag{4.30}$$

Then, for each $g \geq 0$,

$$\int_{\mathbb{R}} g(x)\, dF(x) = \sum_i g(x_i) \times p_i. \tag{4.31}$$

•

Exercise 4.59 — Integral with respect to a discrete d.f.
Prove Theorem 4.58 (Karr, 1993, p. 111). •

Corollary 4.60 — Integrable function with respect to a discrete d.f. (Karr, 1993, p. 111)
The function g is said to be integrable with respect to the discrete d.f. F iff

$$\sum_i |g(x_i)| \times p_i < +\infty, \tag{4.32}$$

and in this case

$$\int_{\mathbb{R}} g(x)\, dF(x) = \sum_i g(x_i) \times p_i. \tag{4.33}$$

•

4.2.4 Absolutely continuous distribution functions

Now note that integrals with respect to absolutely continuous d.f. are Riemann integrals.

Theorem 4.61 — Integral with respect to an absolutely continuous d.f. (Karr, 1993, p. 112)
Suppose that the d.f. F is absolutely continuous and is associated to a piecewise continuous p.d.f. f. If g is a non negative function and piecewise continuous then

$$\int_{\mathbb{R}} g(x)\, dF(x) = \int_{-\infty}^{+\infty} g(x) \times f(x)\, dx, \tag{4.34}$$

where the integral on the right-hand side is an improper Riemann integral. •

Exercise 4.62 — **Integral with respect to an absolutely continuous d.f.**
Prove Theorem 4.61 (Karr, 1993, p. 112). •

Corollary 4.63 — **Integral with respect to an absolutely continuous d.f.** (Karr, 1993, p. 112)
A piecewise continuous function g is said to be integrable with respect to the d.f. F iff

$$\int_{-\infty}^{+\infty} |g(x)| f(x)\, dx < +\infty, \tag{4.35}$$

and in this case

$$\int_{I\!R} g(x)\, dF(x) = \int_{-\infty}^{+\infty} g(x) \times f(x)\, dx. \tag{4.36}$$

•

4.2.5 Mixed distribution functions

Recall that a mixed d.f. F is a convex combination of a discrete d.f.

$$F_d(x) = \sum_i p_i \times \mathbf{1}_{[x_i,+\infty)}(x) \tag{4.37}$$

and an absolutely continuous d.f.

$$F_a(x) = \int_{-\infty}^{x} f_a(s)\, ds. \tag{4.38}$$

Thus,

$$F(x) = \alpha \times F_d(x) + (1-\alpha) \times F_a(x), \tag{4.39}$$

where $\alpha \in (0,1)$.

Corollary 4.64 — **Integral with respect to a mixed d.f.** (Karr, 1993, p. 112)
The integral of g with respect to the mixed d.f. F is a corresponding combination of integrals with respect to F_d and F_a:

$$\begin{aligned} \int_{I\!R} g(x)\, dF(x) &= \alpha \times \int_{I\!R} g(x)\, dF_d(x) + (1-\alpha) \times \int_{I\!R} g(x)\, dF_a(x) \\ &= \alpha \times \sum_i g(x_i) \times p_i + (1-\alpha) \times \int_{-\infty}^{+\infty} g(x) \times f_a(x)\, dx. \end{aligned} \tag{4.40}$$

In order that the integral with respect to a mixed d.f. exists, g must be piecewise continuous and either non negative or integrable with respect to both F_d and F_a. •

4.3 Computation of expectations

So far we have defined the expectation for simple r.v.

The expectations of other types of r.v. — such as non negative, integrable and mixed r.v. — naturally involve integrals with respect to distribution functions.

4.3.1 Non negative r.v.

The second equality in the next formula is quite convenient because it allows us to obtain the expectation of a non negative r.v. — be it a discrete, absolutely continuous or mixed — in terms of an improper Riemann integral.

Theorem 4.65 — Expected value of a non negative r.v. (Karr, 1993, p. 113)
If $X \geq 0$ then

$$E(X) = \int_0^{+\infty} x \, dF_X(x) = \int_0^{+\infty} [1 - F_X(x)] \, dx. \tag{4.41}$$

•

Exercise 4.66 — Expected value of a non negative r.v.
Prove Theorem 4.65 (Karr, 1993, pp. 113–114). •

Corollary 4.67 — Expected value of a non negative integer-valued r.v. (Karr, 1993, p. 114)
Let X be a non negative integer-valued r.v. Then

$$E(X) = \sum_{n=1}^{+\infty} n \times P(\{X = n\}) = \sum_{n=1}^{+\infty} P(\{X \geq n\}) = \sum_{n=0}^{+\infty} P(\{X > n\}). \tag{4.42}$$

•

Exercise 4.68 — Expected value of a non negative integer-valued r.v.
Prove Corollary 4.67 (Karr, 1993, p. 114). •

Corollary 4.69 — Expected value of a non negative absolutely continuous r.v.
Let X be a non negative absolutely continuous r.v. with p.d.f. $f_X(x)$.

$$E(X) = \int_0^{+\infty} x \times f_X(x) \, dx = \int_0^{+\infty} [1 - F_X(x)] \, dx. \tag{4.43}$$

•

Exercise 4.70 — A nonnegative r.v. with infinite expectation
Let $X \sim \text{Pareto}(b = 1, \alpha = 1)$. i.e.

$$f_X(x) = \begin{cases} \frac{\alpha\, b^\alpha}{x^{\alpha+1}} = \frac{1}{x^2}, & x \geq b = 1 \\ 0, & \text{otherwise.} \end{cases} \tag{4.44}$$

Prove that $E(X)$ exists and $E(X) = +\infty$ (Resnick, 1999, p. 126, Example 5.2.1). •

4.3.2 Integrable r.v.

Let us remind the reader that X is said to be an integrable r.v. if $E(|X|) = E(X^+) + E(X^-) < +\infty$.

Theorem 4.71 — Expected value of an integrable r.v. (Karr, 1993, p. 114)
If X is an integrable r.v. then

$$E(X) = \int_{-\infty}^{+\infty} x\, dF_X(x). \tag{4.45}$$

•

Exercise 4.72 — Expected value of an integrable r.v.
Prove Theorem 4.71 (Karr, 1993, p. 114). •

Corollary 4.73 — Expected value of an integrable discrete or absolutely continuous r.v.
Let X be an integrable discrete or absolutely continuous r.v. with p.f. $P(X = x)$ or p.d.f. $f_X(x)$ then

$$E(X) = \sum_x x \times P(X = x)\, dx \tag{4.46}$$

$$E(X) = \int_{-\infty}^{+\infty} x \times f_X(x)\, dx, \tag{4.47}$$

respectively. •

Exercise 4.74 — Real r.v. without expectation
After having derived the c.d.f. of X^+ and X^-, use Theorem 4.65 to prove that $E(X^+) = E(X^-) = +\infty$ — and therefore $E(X)$ does not exist — if X has p.d.f. equal to:

(a) $$f_X(x) = \begin{cases} \frac{1}{2x^2}, & |x| > 1 \\ 0, & \text{otherwise;} \end{cases}$$

(b) $f_X(x) = \dfrac{1}{\pi(1+x^2)}$, $x \in \mathbb{R}$.

(Resnick, 1999, p. 126, Example 5.2.1.)[10] •

4.3.3 Mixed r.v.

When dealing with mixed r.v. X we take advantage of the fact that $F_X(x)$ is a convex combination of the d.f. of a discrete r.v. X_d and the d.f. of an absolutely continuous r.v. X_a.

Corollary 4.75 — Expectation of a mixed r.v.
The expected value of the mixed r.v. X with d.f. $F_X(x) = \alpha \times F_{X_d}(x) + (1-\alpha) \times F_{X_a}(x)$, where $\alpha \in (0,1)$, is given by

$$\begin{aligned}
\int_{\mathbb{R}} x\, dF_X(x) &= \alpha \times \int_{\mathbb{R}} x\, dF_{X_d}(x) + (1-\alpha) \times \int_{\mathbb{R}} x\, dF_{X_a}(x) \\
&= \alpha \times \sum_i x_i \times P(\{X_d = x_i\}) \\
&\qquad + (1-\alpha) \times \int_{-\infty}^{+\infty} x \times f_{X_a}(x)\, dx \qquad (4.48) \\
&= \alpha\, E(X_d) + (1-\alpha)\, E(X_a). \qquad (4.49)
\end{aligned}$$

•

Exercise 4.76 — Expectation of a mixed r.v.
A random variable X has the following d.f.:[11]

$$F_X(x) = \begin{cases} 0, & x < 0 \\ 0.3, & 0 \leq x < 2 \\ 0.3 + 0.2x, & 2 \leq x < 3 \\ 1, & x \geq 3. \end{cases} \qquad (4.50)$$

(a) Why is X a mixed r.v.?

(b) Write $F_X(x)$ as a linear combination of the d.f. of two r.v.: a discrete and an absolutely continuous r.v.

(c) Obtain the expected value of X, by using the fact that X is non negative, thus, $E(X) = \int_0^{+\infty} [1 - F_X(x)]\, dx$.

Compare this value with the one you would obtain using Corollary 4.75. •

[10] There is a typo in the definition of the first p.d.f. in Resnick (1999): $x > 1$ should read as $|x| > 1$. The second p.d.f. corresponds to the one of a r.v. with (standard) Cauchy distribution.

[11] Adapted from Walrand (2004, pp. 53–55, Example 4.10.9).

Exercise 4.77 — Expectation of a mixed r.v. in a queueing setting
Consider a $M/M/1$ system.[12] Let:

- L_s be the number of customers an arriving customer finds in the system in equilibrium;[13]
- W_q be the waiting time in queue of this arriving customer.[14]

Under these conditions, we can state that

$$P(L_s = k) = (1-\rho) \times \rho^k,\ k \in \mathbb{N}_0; \tag{4.51}$$

thus, $L_s \sim \text{geometric}^*(1-\rho)$, where $\rho = \frac{\lambda}{\mu} \in (0,1)$ and $E(L_s) = \frac{\rho}{1-\rho}$.

(a) Argue that $W_q|\{L_s = k\} \sim \text{Gamma}(k, \mu)$, for $k \in \mathbb{N}$.

(b) Prove that $W_q|\{W_q > 0\} \sim \text{Exponential}(\mu(1-\rho))$.

(c) Demonstrate that W_q is a mixed r.v. with d.f. given by:

$$F_{W_q}(w) = \begin{cases} 0, & w < 0 \\ (1-\rho) + \rho \times F_{Exp(\mu(1-\rho))}(w), & w \geq 0. \end{cases} \tag{4.52}$$

(d) Verify that $E(W_q) = \frac{\rho}{\mu(1-\rho)}$. •

[12] The arrivals to the system are governed by a Poisson process with rate λ, i.e. the time between arrivals has an exponential distribution with parameter λ; needless to say, M stands for memoryless. The service times are not only i.i.d. with exponential distribution with parameter μ, but also independent from the arrival process. There is only one server, and the service policy is FCFS (first come first served). $\rho = \frac{\lambda}{\mu}$ represents the traffic intensity and we assume that $\rho \in (0,1)$.

[13] Equilibrium roughly means that a lot of time has elapsed since the system has been operating and therefore the initial conditions no longer influence the state of system.

[14] W_q is the time elapsed from the moment the customer arrives until his/her service starts in the system in equilibrium.

4.3.4 Functions of r.v.

Unsurprisingly, we are surely able to derive expressions for the expectation of a Borel measurable function g of the r.v. X, $E[g(X)]$. Obtaining this expectation does not require the derivation of the d.f. of $g(X)$ and follows from section 4.2.

In the two sections we shall discuss the expectation of specific functions of r.v.: $g(X) = X^k$, $k \in \mathbb{N}$.

Theorem 4.78 — Expected value of a function of a r.v. (Karr, 1993, p. 115)
Let X be a r.v., and g be a Borel measurable function either non negative or integrable. Then

$$E[g(X)] = \int_{\mathbb{R}} g(x)\, dF_X(x). \tag{4.53}$$

•

Exercise 4.79 — Expected value of a function of a r.v.
Prove Theorem 4.78 (Karr, 1993, p. 115). •

Corollary 4.80 — Expected value of a function of a discrete r.v. (Karr, 1993, p. 115)
Let X be a discrete r.v., and g be a Borel measurable function either non negative or integrable. Then

$$E[g(X)] = \sum_{x_i} g(x_i) \times P(\{X = x_i\}). \tag{4.54}$$

•

Corollary 4.81 — Expected value of a function of an absolutely continuous r.v. (Karr, 1993, p. 115)
Let X be an absolutely continuous r.v., and g be a Borel measurable function either non negative or integrable. Then

$$E[g(X)] = \int_{-\infty}^{+\infty} g(x) \times f_X(x)\, dx. \tag{4.55}$$

•

4.3.5 Functions of random vectors

When dealing with functions of random vectors, the only useful formulas are those referring to the expectation of functions of discrete and absolutely continuous random vectors.

These formulas will be used to obtain, for instance, what we shall call measures of (linear) association between r.v.

Theorem 4.82 — Expectation of a function of a discrete random vector (Karr, 1993, p. 116)
Let:

- $X_1, \ldots, X_d$ be a discrete r.v., with values in the countable sets $C_1, \ldots, C_d$, respectively;
- $g : \mathbb{R}^d \to \mathbb{R}$ be a Borel measurable function either non negative or integrable (i.e. $g(X_1, \ldots, X_d) \in L^1$).

Then

$$E[g(X_1, \ldots, X_d)] = \sum_{x_1 \in C_1} \ldots \sum_{x_d \in C_d} g(x_1, \ldots, x_d) \times P(\{X_1 = x_1, \ldots, X_d = x_d\}). \quad (4.56)$$

•

Theorem 4.83 — Expectation of a function of an absolutely continuous random vector (Karr, 1993, p. 116)
Let:

- $(X_1, \ldots, X_d)$ be an absolutely continuous random vector with joint p.d.f. $f_{X_1,\ldots,X_d}(x_1, \ldots, x_d)$;
- $g : \mathbb{R}^d \to \mathbb{R}$ be a Borel measurable function either non negative or integrable.

Then

$$E[g(X_1, \ldots, X_d)] = \int_{-\infty}^{+\infty} \ldots \int_{-\infty}^{+\infty} g(x_1, \ldots, x_d) \times f_{X_1,\ldots,X_d}(x_1, \ldots, x_d)\, dx_1 \ldots dx_d. \quad (4.57)$$

•

Exercise 4.84 — Expectation of a function of an absolutely continuous random vector
Prove Theorem 4.83 (Karr, 1993, pp. 116–117). •

4.3.6 Functions of independent r.v.

When all the components of the random vector $(X_1, \ldots, X_d)$ are independent, the formula of $E[g(X_1, \ldots, X_d)]$ can be simplified.

The next results refer to two independent random variables ($d = 2$). The generalization for $d > 2$ is straightforward.

Theorem 4.85 — Expectation of a function of two independent r.v. (Karr, 1993, p. 117)
Let:

- X and Y be two independent r.v.
- $g : \mathbb{R}^2 \to \mathbb{R}^+$ be a Borel measurable non negative function.

Then

$$\begin{aligned} E[g(X,Y)] &= \int_{\mathbb{R}} \left[\int_{\mathbb{R}} g(x,y)\, dF_X(x) \right] dF_Y(y) \\ &= \int_{\mathbb{R}} \left[\int_{\mathbb{R}} g(x,y)\, dF_Y(y) \right] dF_X(x). \end{aligned} \tag{4.58}$$

•

Moreover, the expectation of the product of functions of independent r.v. is the product of their expectations. Also note that the product of two integrable r.v. need not be integrable.

Corollary 4.86 — Expectation of a function of two independent r.v. (Karr, 1993, p. 118)
Let:

- X and Y be two independent r.v.
- $g_1, g_2 : \mathbb{R} \to \mathbb{R}$ be two Borel measurable functions either non negative or integrable.

Then $g_1(X) \times g_2(Y)$ is integrable and

$$E[g_1(X) \times g_2(Y)] = E[g_1(X)] \times E[g_2(Y)]. \tag{4.59}$$

•

Exercise 4.87 — Expectation of a function of two independent r.v.
Prove Theorem 4.85 (Karr, 1993, p. 117) and Corollary 4.86 (Karr, 1993, p. 118). •

4.3.7 Sum of independent r.v.

We are certainly not going to state that $E(X+Y) = E(X) + E(Y)$ when X and Y are simple or non negative or integrable independent r.v.[15]

Instead, we are going to write the d.f. of the sum of two INDEPENDENT r.v. in terms of integrals with respect to the d.f. and define a convolution of d.f.[16]

Theorem 4.88 — D.f. of a sum of two independent r.v. (Karr, 1993, p. 118)
Let X and Y be two INDEPENDENT r.v. Then

$$F_{X+Y}(t) = \int_{\mathbb{R}} F_X(t-y)\, dF_Y(y) = \int_{\mathbb{R}} F_Y(t-x)\, dF_X(x). \tag{4.60}$$

•

Exercise 4.89 — D.f. of a sum of two independent r.v.
Prove Theorem 4.88 (Karr, 1993, p. 118). •

Corollary 4.90 — D.f. of a sum of two independent discrete r.v.
Let X and Y be two INDEPENDENT discrete r.v. Then

$$F_{X+Y}(t) = \sum_y F_X(t-y) \times P(\{Y=y\}) = \sum_x F_Y(t-x) \times P(\{X=x\}). \tag{4.61}$$

•

Remark 4.91 — D.f. of a sum of two independent discrete r.v.
The previous formula is not preferable to the one we derived for the p.f. of $X+Y$ in Chapter 2 because it depends in fact on two sums... •

Corollary 4.92 — D.f. of a sum of two independent absolutely continuous r.v.
Let X and Y be two INDEPENDENT absolutely continuous r.v. Then

$$F_{X+Y}(t) = \int_{-\infty}^{+\infty} F_X(t-y) \times f_Y(y)\, dy = \int_{-\infty}^{+\infty} F_Y(t-x) \times f_X(x)\, dx. \tag{4.62}$$

•

Let us revisit an exercise from Chapter 2 to illustrate the use of Corollary 4.92.

[15] This result follows from the linearity of expectation.

[16] Recall that in Chapter 2 we derived expressions for the p.f. and the p.d.f. of the sum of two independent r.v.

Exercise 4.93 — D.f. of the sum of two independent absolutely continuous r.v. Let X and Y be the durations of two independent system components set in what is called a *stand by connection*.[17] In this case the system duration is given by $X + Y$.

(a) Derive the d.f. of $X + Y$, assuming that $X \sim$ Exponencial(α) and $Y \sim$ Exponencial(β), where $\alpha, \beta > 0$ and $\alpha \neq \beta$, and using Corollary 4.92.

(b) Prove that the associated p.d.f. equals $f_{X+Y}(z) = \frac{\alpha\beta\left(e^{-\beta z} - e^{-\alpha z}\right)}{\alpha - \beta}$, $z > 0$. •

Definition 4.94 — Convolution of d.f. (Karr, 1993, p. 119)
Let X and Y be independent r.v. Then

$$(F_X \star F_Y)(t) = F_{X+Y}(t) = \int_{\mathbb{R}} F_X(t - y)\, dF_Y(y) = \int_{\mathbb{R}} F_Y(t - x)\, dF_X(x) \tag{4.63}$$

is said to be the convolution of the d.f. F_X and F_Y. •

[17] At time 0, only the component with duration X is on. The component with duration Y replaces the other one as soon as it fails.

4.4 L^p spaces

Motivation 4.95 — **L^p spaces** (Karr, 1993, p. 119)
While describing a r.v. in a partial way, we tend to deal with $E(X^p)$, $p \in [1, +\infty)$, or a function of several such expected values. Needless to say that we have to guarantee that $E(|X|^p)$ is finite. •

Definition 4.96 — **L^p spaces** (Karr, 1993, p. 119)
The space L^p, for a fixed $p \in [1, +\infty)$, consists of all r.v. X whose pth absolute power is integrable, that is,

$$E(|X|^p) < +\infty. \tag{4.64}$$

•

Exercise 4.97 — **Exponential distributions and L^p spaces**
Let $X \sim \text{Exponential}(\lambda)$, $\lambda > 0$.

Prove that $X \in L^p$, for any $p \in [1, +\infty)$. •

Exercise 4.98 — **Pareto distributions and L^p spaces**
Let $X \sim \text{Pareto}(b, \alpha)$ i.e.

$$f_X(x) = \begin{cases} \frac{\alpha\, b^\alpha}{x^{\alpha+1}}, & x \geq b \\ 0, & \text{otherwise} \end{cases} \tag{4.65}$$

where $b > 0$ is the minimum possible value of X and $\alpha > 0$ is called the Pareto index.

For which values of $p \in [1, +\infty)$ we have $X \in L^p$? •

4.5 Key inequalities

What immediately follows is a table with an overview of a few extremely useful inequalities involving expectations.

Some of these inequalities are essential to prove certain types of convergence of sequences of r.v. in L^p and uniform integrability (Resnick, 1999, p. 189)[18] and provide answers to a few questions we compiled after the table.

Finally, we state and treat each inequality separately.

Proposition 4.99 — A few (moment) inequalities (Karr, 1993, p. 123; http://en.wikipedia.org)

(Moment) inequality	**Conditions**	**Statement of the inequality**
Young	$h : \mathbb{R}_0^+ \to \mathbb{R}_0^+$ continuous, strictly increasing, $h(0) = 0$, $h(+\infty) = +\infty$, $H(x) = \int_0^x h(y)\,dy$ k pointwise inverse of h, $K(x) = \int_0^x k(y)\,dy$ $a, b \in \mathbb{R}^+$	$a \times b \leq H(a) + K(b)$
Hölder	$X \in L^p$, $Y \in L^q$, where $p, q \in [1, +\infty) : \frac{1}{p} + \frac{1}{q} = 1$	$E(\lvert X \times Y\rvert) \leq E^{\frac{1}{p}}(\lvert X\rvert^p) \times E^{\frac{1}{q}}(\lvert Y\rvert^q)$ $(X \times Y \in L^1)$
Cauchy-Schwarz	$X, Y \in L^2$	$E(\lvert X \times Y\rvert) \leq \sqrt{E(X^2) \times E(Y^2)}$ $(X \times Y \in L^1)$
Liapunov	$X \in L^s$, $1 \leq r \leq s$	$E^{\frac{1}{r}}(\lvert X\rvert^r) \leq E^{\frac{1}{s}}(\lvert X\rvert^s)$ $(L^s \subseteq L^r)$
Minkowski	$X, Y \in L^p$, $p \in [1, +\infty)$	$E^{\frac{1}{p}}(\lvert X + Y\rvert^p) \leq E^{\frac{1}{p}}(\lvert X\rvert^p) + E^{\frac{1}{p}}(\lvert Y\rvert^p)$ $(X + Y \in L^p)$
Jensen	g convex; $X, g(X) \in L^1$	$g[E(X)] \leq E[g(X)]$
	g concave; $X, g(X) \in L^1$	$g[E(X)] \geq E[g(X)]$
Chebyshev	$X \geq 0$, g non negative and increasing, $a > 0$	$P(\{X \geq a\}) \leq \frac{E[g(X)]}{g(a)}$
(Chernoff)	$X \geq 0$, $a, t > 0$	$P(\{X \geq a\}) \leq \frac{E(e^{tX})}{e^{ta}}$
(Markov)	$X \in L^1$, $a > 0$	$P(\{\lvert X\rvert \geq a\}) \leq \frac{E[\lvert X\rvert]}{a}$;
	$X \in L^p$, $a > 0$	$P(\{\lvert X\rvert \geq a\}) \leq \frac{E[\lvert X\rvert^p]}{a^p}$
(Chebyshev-Bienaymé)	$X \in L^2$, $a > 0$	$P(\{\lvert X - E(X)\rvert \geq a\}) \leq \frac{V(X)}{a^2}$
	$X \in L^2$, $a > 0$	$P\left(\left\{\lvert X - E(X)\rvert \geq a\sqrt{V(X)}\right\}\right) \leq \frac{1}{a^2}$
(Cantelli)	$X \in L^2$, $a > 0$	$P(\{\lvert X - E(X)\rvert \geq a\}) \leq \frac{2V(X)}{a^2 + V(X)}$
(one-sided Chebyshev)	$X \in L^2$, $a > 0$	$P\left(\left\{X - E(X) \geq a\sqrt{V(X)}\right\}\right) \leq \frac{1}{1+a^2}$

•

[18] For a definition of uniform integrability see http://en.wikipedia.org/wiki/Uniform_integrability.

Motivation 4.100 — A few (moment) inequalities

- **Young** — How can we relate the areas under (resp. above) an increasing function h in the interval $[0, a]$ (resp. in the interval of $[0, h^{-1}(b)]$) with the area of the rectangle with vertices $(0, 0)$, $(0, b)$, $(a, 0)$ and (a, b), where $b \in (0, \max_{x\in[0,a]} h(x)]$?
- **Hölder/Cauchy-Schwarz** — What are the sufficient conditions on r.v. X and Y to be dealing with an integrable product XY?
- **Liapunov** — What happens to the spaces L^p when p increases in $[1, +\infty)$? Is it a decreasing (increasing) sequence of sets?

  What happens to the norm of a r.v. in L^p, $||X||_p = E^{\frac{1}{p}}(|X|^p)$? Is it an increasing (decreasing) function of $p \in [1, +\infty)$?
- **Minkowski** — What are the sufficient conditions on r.v. X and Y to be dealing with a sum $X + Y \in L^p$? Is L^p a vector space?
- **Jensen** — Under what conditions we can relate $g[E(X)]$ and $E[g(X)]$?
- **Chebyshev** — When can we provide non trivial upper bounds for the tail probability $P(\{X \geq a\})$ •

4.5.1 Young's inequality

The first inequality (not a moment inequality) is named after William Henry Young (1863–1942), an English mathematician, and can be used to prove Hölder inequality.

Lemma 4.101 — Young's inequality (Karr, 1993, p. 119)
Let:

- $h : \mathbb{R}_0^+ \to \mathbb{R}_0^+$ be a continuous and strictly increasing function such that $h(0) = 0$, $h(+\infty) = +\infty$;
- k be the pointwise inverse of h;
- $H(x) = \int_0^x h(y)\, dy$ be the area under h in the interval $[0, x]$;
- $K(x) = \int_0^x k(y)\, dy$ be the area above h in the interval $[0, h^{-1}(x)] = [0, k(x)]$;
- $a, b \in \mathbb{R}^+$.

Then

$$a \times b \leq H(a) + K(b). \tag{4.66}$$

•

Exercise 4.102 — **Young's inequality** (Karr, 1993, p. 119)
Prove Lemma 4.101, by using a graphical argument (Karr, 1993, p. 119). •

Remark 4.103 — **A special case of Young's inequality**
If we apply Young's inequality to $h(x) = x^{p-1}$, $p \in [1, +\infty)$ and consider

- a and b non negative real numbers,
- $q = 1 + \frac{1}{p-1} \in [1, +\infty)$ i.e. $\frac{1}{p} + \frac{1}{q} = 1$,

then

$$a \times b \leq \frac{a^p}{p} + \frac{b^q}{q}. \tag{4.67}$$

For the proof of this result see http://en.wikipedia.org/wiki/Young's_inequality, which states (4.67) as Young's inequality. See also Karr (1993, p. 120) for a reference to (4.67) as a consequence of Young's inequality as stated in (4.66). •

4.5.2 Hölder's moment inequality

In mathematical analysis Hölder's inequality, named after the German mathematician Otto Hölder (1859–1937), is a fundamental inequality between integrals, an indispensable tool for the study of L^p spaces and essential to prove Liapunov's and Minkowski's inequalities.

Interestingly enough, Hölder's inequality was first found by the British mathematician L.J. Rogers (1862–1933) in 1888, and discovered independently by Hölder in 1889.

Theorem 4.104 — **Hölder's moment inequality** (Karr, 1993, p. 120)
Let

- $X \in L^p$, $Y \in L^q$, where $p, q \in [1, +\infty) : \frac{1}{p} + \frac{1}{q} = 1$.

Then

$$X \times Y \in L^1 \tag{4.68}$$
$$E(|XY|) \leq E^{\frac{1}{p}}(|X|^p) \times E^{\frac{1}{q}}(|Y|^q). \tag{4.69}$$

•

Remarks 4.105 — Hölder's (moment) inequality

- The numbers p and q above are said to be Hölder conjugates of each other.
- For the detailed statement of Hölder inequality in measure spaces, check http://en.wikipedia.org/wiki/Hölder's_inequality. Two notable special cases follow...
- In case we are dealing with S, a measurable subset of $\mathbb{R}$ with the Lebesgue measure, and f and g are measurable real-valued functions on S then Hölder's inequality reads as follows:

$$\int_S |f(x) \times g(x)|\, dx \leq \left(\int_S |f(x)|^p\, dx \right)^{\frac{1}{p}} \times \left(\int_S |g(x)|^q\, dx \right)^{\frac{1}{q}}. \tag{4.70}$$

- When we are dealing with $n-$dimensional Euclidean space and the counting measure, we have

$$\sum_{k=1}^{n} |x_k \times y_k| \leq \left(\sum_{k=1}^{n} |x_k|^p \right)^{\frac{1}{p}} \times \left(\sum_{k=1}^{n} |y_k|^q \right)^{\frac{1}{q}}, \tag{4.71}$$

  for all $(x_1, \ldots, x_n), (y_1, \ldots, y_n) \in \mathbb{R}^n$.
- For a generalization of Hölder's inequality involving n (instead of 2) Hölder conjugates, see http://en.wikipedia.org/wiki/Hölder's_inequality. •

Exercise 4.106 — Hölder's moment inequality

Prove Theorem 4.104, by using the special case of Young's inequality (4.67), considering $a = \frac{|X|}{E^{\frac{1}{p}}(|X|^p)}$ and $b = \frac{|Y|}{E^{\frac{1}{q}}(|Y|^q)}$, taking expectations to (4.67), and applying the result $\frac{1}{p} + \frac{1}{q} = 1$ (Karr, 1993, p. 120). •

4.5.3 Cauchy-Schwarz's moment inequality

A special case of Hölder's moment inequality — $p = q = 2$ — is nothing but the Cauchy-Schwarz's moment inequality.

In mathematics, the Cauchy-Schwarz inequality[19] is a useful inequality encountered in many different settings, such as linear algebra applied to vectors, in analysis applied to infinite series and integration of products, and in probability theory, applied to variances and covariances.

The inequality for sums was published by Augustin Cauchy in 1821, while the corresponding inequality for integrals was first stated by Viktor Yakovlevich Bunyakovsky in 1859 and rediscovered by Hermann Amandus Schwarz in 1888 (often misspelled "Schwartz").

Corollary 4.107 — Cauchy-Schwarz's moment inequality (Karr, 1993, p. 120)
Let $X, Y \in L^2$. Then

$$X \times Y \in L^1 \tag{4.72}$$

$$E(|X \times Y|) \leq \sqrt{E(|X|^2) \times E(|Y|^2)}. \tag{4.73}$$

•

Remarks 4.108 — Cauchy-Schwarz's moment inequality
(http://en.wikipedia.org/wiki/Cauchy-Schwarz_inequality)

- In the Euclidean space $\mathbb{R}^n$ with the standard inner product, the Cauchy-Schwarz's inequality is

$$\left(\sum_{i=1}^{n} x_i \times y_i\right)^2 \leq \left(\sum_{i=1}^{n} x_i^2\right) \times \left(\sum_{i=1}^{n} y_i^2\right). \tag{4.74}$$

- The triangle inequality for the inner product is often shown as a consequence of the Cauchy-Schwarz inequality, as follows: given vectors $\underline{x}$ and $\underline{y}$, we have

$$\begin{aligned} \|x + y\|^2 &= \langle x + y, x + y \rangle \\ &\leq (\|x\| + \|y\|)^2 . \end{aligned} \tag{4.75}$$

•

[19] Also known as the Bunyakovsky inequality, the Schwarz inequality, or the Cauchy-Bunyakovsky-Schwarz inequality (http://en.wikipedia.org/wiki/Cauchy-Schwarz_inequality).

Exercise 4.109 — Confronting the squared covariance and the product of the variance of two r.v.
Prove that

$$|cov(X,Y)|^2 \leq V(X) \times V(Y), \tag{4.76}$$

where $X, Y \in L^2$ (http://en.wikipedia.org/wiki/Cauchy-Schwarz's_inequality). •

4.5.4 Lyapunov's moment inequality

The spaces L^p decrease as p increases in $[1, +\infty)$. Moreover, $E(|X|^p)^{\frac{1}{p}}$ is an increasing function of $p \in [1, +\infty)$.

The following moment inequality is a special case of Hölder's inequality and is due to Aleksandr Mikhailovich Lyapunov (1857–1918), a Russian mathematician, mechanician and physicist.[20]

Corollary 4.110 — Lyapunov's moment inequality (Karr, 1993, p. 120)
Let $X \in L^s$, where $1 \leq r \leq s$. Then

$$L^s \subseteq L^r \tag{4.77}$$

$$E^{\frac{1}{r}}(|X|^r) \leq E^{\frac{1}{s}}(|X|^s). \tag{4.78}$$

•

Remarks 4.111 — Lyapunov's moment inequality

- Taking $s = 2$ and $r = 1$ we can conclude that

$$E^2(|X|) \leq E(X^2). \tag{4.79}$$

  This result is not correctly stated in Karr (1993, p. 121) and can be also deduced from the Cauchy-Schwarz's inequality, as well as from Jensen's inequality, stated below.

- Rohatgi (1976, p. 103) states Liapunov's inequality in a slightly different way. It can be put as follows: for $X \in L^k$, $k \in [1, +\infty)$,

$$E^{\frac{1}{k}}(|X|^k) \leq E^{\frac{1}{k+1}}(|X|^{k+1}). \tag{4.80}$$

[20] Lyapunov is known for his development of the stability theory of a dynamical system, as well as for his many contributions to mathematical physics and probability theory (http://en.wikipedia.org/wiki/Aleksandr_Lyapunov).

- The equality in (4.78) holds iff X is a degenerate r.v., i.e. $X \stackrel{d}{=} c$, where c is a real constant (Rohatgi, 1976, p. 103). •

Exercise 4.112 — Lyapunov's moment inequality
Prove Corollary 4.110, by applying Hölder's inequality to $X' = X^r$, where $X \in L^r$, and to $Y \stackrel{d}{=} 1$, and considering $p = \frac{s}{r}$ (Karr, 1993, pp. 120–121).[21] •

4.5.5 Minkowski's moment inequality

The Minkowski's moment inequality establishes that the L^p spaces are vector spaces.

Theorem 4.113 — Minkowski's moment inequality (Karr, 1993, p. 121)
Let $X, Y \in L^p$, $p \in [1, +\infty)$. Then

$$X + Y \in L^p \tag{4.81}$$

$$E^{\frac{1}{p}}(|X+Y|^p) \leq E^{\frac{1}{p}}(|X|^p) + E^{\frac{1}{p}}(|Y|^p). \tag{4.82}$$

•

Remarks 4.114 — Minkowski's moment inequality
(http://en.wikipedia.org/wiki/Minkowski_inequality)

- The Minkowski inequality is the triangle inequality[22] in L^p.
- Like Hölder's inequality, the Minkowski's inequality can be specialized to (sequences and) vectors by using the counting measure:

$$\left(\sum_{k=1}^{n} |x_k + y_k|^p\right)^{\frac{1}{p}} \leq \left(\sum_{k=1}^{n} |x_k|^p\right)^{\frac{1}{p}} + \left(\sum_{k=1}^{n} |y_k|^p\right)^{\frac{1}{p}}, \tag{4.83}$$

  for all $(x_1, \ldots, x_n), (y_1, \ldots, y_n) \in \mathbb{R}^n$. •

Exercise 4.115 — Minkowski's moment inequality
Prove Theorem 4.113, by applying the triangle inequality followed by Hölder's inequality and the fact that $q(p-1) = p$ and $1 - \frac{1}{q} = \frac{1}{p}$ (Karr, 1993, p. 121). •

[21] Rohatgi (1976, p. 103) provides an alternative proof.

[22] The real line is a normed vector space with the absolute value as the norm, and so the triangle inequality states that $|x + y| \leq |x| + |y|$, for any real numbers x and y. The triangle inequality is useful in mathematical analysis for determining the best upper estimate on the size of the sum of two numbers, in terms of the sizes of the individual numbers (http://en.wikipedia.org/wiki/Triangle_inequality).

4.5.6 Jensen's moment inequality

Jensen's inequality, named after the Danish mathematician and engineer Johan Jensen (1859–1925), relates the value of a convex function of an integral to the integral of the convex function. It was proved by Jensen in 1906.

Given its generality, the inequality appears in many forms depending on the context. In its simplest form the inequality states, that

- the convex transformation of a mean is less than or equal to the mean after convex transformation.

It is a simple corollary that the opposite is true of concave transformations (http://en.wikipedia.org/wiki/Jensen's_inequality).

Theorem 4.116 — Jensen's moment inequality (Karr, 1993, p. 121)
Let g convex and assume that $X, g(X) \in L^1$. Then

$$g[E(X)] \leq E[g(X)] \tag{4.84}$$

•

Corollary 4.117 — Jensen's moment inequality for concave functions
Let g concave and assume that $X, g(X) \in L^1$. Then

$$g[E(X)] \geq E[g(X)] \tag{4.85}$$

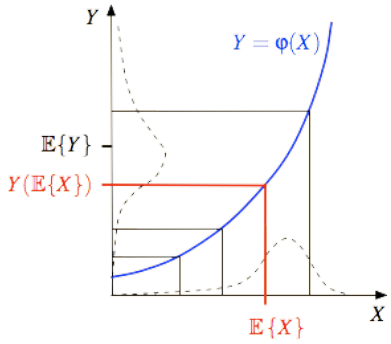
•

Remarks 4.118 — Jensen's (moment) inequality
(http://en.wikipedia.org/wiki/Jensen's_inequality)

- A proof of Jensen's inequality can be provided in several ways. However, it is worth analyzing an intuitive graphical argument based on the probabilistic case where X is a real r.v.

  Assuming a hypothetical distribution of X values, one can immediately identify the position of $E(X)$ and its image $g[E(X)] = \varphi[E(X)]$ in the graph.

  Noticing that for convex mappings $Y = g(X) = \varphi(X)$ the corresponding distribution of Y values is increasingly "stretched out" for increasing values of X, the expectation of $Y = g(X)$ will always shift upwards with respect to the position of $g[E(X)] = \varphi[E(X)]$, and this "proves" the inequality.



- For a real convex function g, numbers $x_1, x_2, \ldots, x_n$ in its domain, and positive weights a_i, $i = 1, \ldots, n$, Jensen's inequality can be stated as:

$$g\left(\frac{\sum_{i=1}^{n} a_i \times x_i}{\sum_{i=1}^{n} a_i}\right) \leq \frac{\sum_{i=1}^{n} a_i \times g(x_i)}{\sum_{i=1}^{n} a_i}. \tag{4.86}$$

  The inequality is reversed if g is concave.

- As a particular case, if the weights $a_i = 1$, $i = 1, \ldots, n$, then

$$g\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) \leq \frac{1}{n}\sum_{i=1}^{n} g(x_i) \Leftrightarrow g(\bar{x}) \leq \overline{g(x)}. \tag{4.87}$$

- For instance, considering $g(x) = \log(x)$, which is a concave function, we can establish the arithmetic mean-geometric mean inequality:[23] for any list of n non negative real numbers $x_1, x_2, \ldots, x_n$,

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} \geq \sqrt[n]{x_1 \times x_2 \times \ldots \times x_n} = mg. \tag{4.88}$$

  Moreover, equality in (4.88) holds iff $x_1 = x_2 = \ldots = x_n$. •

Exercise 4.119 — Jensen's moment inequality (for concave functions)
Prove Theorem 4.116 (Karr, 1993, pp. 121-122), Corollary 4.117 and Equation (4.88). •

Exercise 4.120 — Jensen's inequality and the distance between the mean and the median
Prove that for any r.v. having an expected value and a median, the mean and the median can never differ from each other by more than one standard deviation:

$$|E(X) - med(X)| \leq \sqrt{V(X)}, \tag{4.89}$$

by using Jensen's inequality twice — applied to the absolute value function and to the square root function[24] (http://en.wikipedia.org/wiki/Chebyshev's_inequality). •

[23] See http://en.wikipedia.org/wiki/AM-GM_inequality.
[24] In this last case we should apply the concave version of Jensen's inequality.

4.5.7 Chebyshev's inequality

Curiously, Chebyshev's inequality is named after the Russian mathematician Pafnuty Lvovich Chebyshev (1821–1894), although it was first formulated by his friend and French colleague Irénée-Jules Bienaymé (1796–1878), according to http://en.wikipedia.org/wiki/Chebyshev's_inequality.

In probability theory, the Chebyshev's inequality,[25] in the most usual version — what Karr (1993, p. 122) calls the Bienaymé-Chebyshev's inequality —, can be ultimately stated as follows:

- no more than $\frac{1}{k^2} \times 100\%$ of the values of the r.v. X are more than k standard deviations away from the expected value of X.

Theorem 4.121 — Chebyshev's inequality (Karr, 1993, p. 122)
Let:

- X be a non negative r.v.;
- g non negative and increasing function on $\mathbb{R}^+$;
- $a > 0$.

Then

$$P(\{X \geq a\}) \leq \frac{E[g(X)]}{g(a)}. \tag{4.90}$$

•

Exercise 4.122 — Chebyshev's inequality
Prove Theorem 4.121 (Karr, 1993, p. 122). •

Remarks 4.123 — Several cases of Chebyshev's inequality (Karr, 1993, p. 122)

- **Chernoff's inequality**[26]
  $X \geq 0,\ a, t > 0 \Rightarrow P(\{X \geq a\}) \leq \frac{E(e^{tX})}{e^{ta}}$
- **Markov's inequalities**
  $X \in L^1,\ a > 0 \Rightarrow P(\{|X| \geq a\}) \leq \frac{E[|X|]}{a}$
  $X \in L^p,\ a > 0 \Rightarrow P(\{|X| \geq a\}) \leq \frac{E[|X|^p]}{a^p}$

[25] Also known as Tchebysheff's inequality, Chebyshev's theorem, or the Bienaymé-Chebyshev's inequality (http://en.wikipedia.org/wiki/Chebyshev's_inequality).

[26] Karr (1993) does not mention this inequality. For more details see http://en.wikipedia.org/wiki/Chernoff's_inequality.

- **Chebyshev-Bienaymé's inequalities**

  $X \in L^2$, $a > 0 \Rightarrow P(\{|X - E(X)| \geq a\}) \leq \frac{V(X)}{a^2}$

  $X \in L^2$, $a > 0 \Rightarrow P\left(\left\{|X - E(X)| \geq a\sqrt{V(X)}\right\}\right) \leq \frac{1}{a^2}$

- **Cantelli's inequality**

  $X \in L^2$, $a > 0 \Rightarrow P(\{|X - E(X)| \geq a\}) \leq \frac{2V(X)}{a^2+V(X)}$

- **One-sided Chebyshev's inequality**

  $X \in L^2$, $a > 0 \Rightarrow P\left(\left\{X - E(X) \geq a\sqrt{V(X)}\right\}\right) \leq \frac{1}{1+a^2}$

  According to http://en.wikipedia.org/wiki/Chebyshev's_inequality, the one-sided version of the Chebyshev inequality is also called Cantelli's inequality, and is due to the Italian mathematician Francesco Paolo Cantelli (1875–1966). •

Remark 4.124 — Chebyshev(-Bienaymé)'s inequality
(http://en.wikipedia.org/wiki/Chebyshev's_inequality)
The Chebyshev(-Bienaymé)'s inequality can be useful despite loose bounds because it applies to random variables of any distribution, and because these bounds can be calculated knowing no more about the distribution than the mean and variance. •

Exercise 4.125 — Chebyshev(-Bienaymé)'s inequality
Assume that we have a large body of text, for example articles from a publication and that we know that the articles are on average 1000 characters long with a standard deviation of 200 characters.

(a) Prove that from the Chebyshev(-Bienaymé)'s inequality we can then infer that the chance that a given article is between 600 and 1400 characters would be at least 75%.

(b) The inequality is coarse: a more accurate guess would be possible if the distribution of the length of the articles is known.

  Demonstrate that, for example, a normal distribution would yield a 75% chance of an article being between 770 and 1230 characters long.

(http://en.wikipedia.org/wiki/Chebyshev's_inequality.) •

Exercise 4.126 — Chebyshev(-Bienaymé)'s inequality
Let $X \sim \text{Uniform}(0, 1)$.

(a) Obtain $P\left(\left\{|X - \frac{1}{2}| < 2\sqrt{1/12}\right\}\right)$.

(b) Obtain a lower bound for $P\left(\left\{|X - \frac{1}{2}| < 2\sqrt{1/12}\right\}\right)$, by noting that $E(X) = \frac{1}{2}$ and $V(X) = \frac{1}{12}$. Compare this bound with the value you obtained in (a).

(Rohatgi, 1976, p. 101.) •

Exercise 4.127 — Meeting the Chebyshev(-Bienaymé)'s bounds exactly
Typically, the Chebyshev(-Bienaymé)'s inequality will provide rather loose bounds.

(a) Prove that these bounds cannot be improved upon for the r.v. X with p.f.

$$P(\{X = x\}) = \begin{cases} P(\{X = -1\}) = \frac{1}{2k^2}, & x = -1 \\ P(\{X = 0\}) = 1 - \frac{1}{k^2}, & x = 0 \\ P(\{X = 1\}) = \frac{1}{2k^2}, & x = 1 \\ 0, & \text{otherwise,} \end{cases} \tag{4.91}$$

where $k > 1$, that is, $P\left(|X - E(X)| \geq k\sqrt{V(X)}\right) = \frac{1}{k^2}$. (For more details see http://en.wikipedia.org/wiki/Chebyshev's_inequality.)[27]

(b) Prove that equality holds exactly for any r.v. Y that is a linear transformation of X.[28] •

Remark 4.128 — Chebyshev(-Bienaymé)'s inequality and the weak law of large numbers (http://en.wikipedia.org/ wiki/Chebyshev's_inequality)
Chebyshev(-Bienaymé)'s inequality is used for proving the following version of the weak law of large numbers: when dealing with a sequence of i.i.d. r.v., $X_1, X_2, \ldots$, with finite expected value and variance ($\mu, \sigma^2 < +\infty$),

$$\lim_{n \to +\infty} P\left(\left\{\left|\bar{X}_n - \mu\right| < \epsilon\right\}\right) = 1, \tag{4.92}$$

where $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$. That is, $\bar{X}_n \xrightarrow{P} \mu$ as $n \to +\infty$. •

[27] This is the answer to Exercise 4.36 from Karr (1993, p. 133).

[28] Inequality holds for any r.v. that is not a linear transformation of X (http://en.wikipedia.org/wiki/Chebyshev's_inequality).

Exercise 4.129 — Chebyshev(-Bienaymé)'s inequality and the weak law of large numbers

Use Chebyshev(-Bienaymé)'s inequality to prove the weak law of large numbers stated in Remark 4.128. •

Exercise 4.130 — Cantelli's inequality (Karr, 1993, p. 132, Exercise 4.30)

When does $P(\{|X - E(X)| \geq a\}) \leq \frac{2V(X)}{a^2+V(X)}$ give a better bound than Chebyshev(-Bienaymé)'s inequality? •

Exercise 4.131 — One-sided Chebyshev's inequality and the distance between the mean and the median

Use the one-sided Chebyshev's inequality to prove that for any r.v. having an expected value and a median, the mean and the median can never differ from each other by more than one standard deviation, i.e.

$$|E(X) - med(X)| \leq \sqrt{V(X)} \tag{4.93}$$

(http://en.wikipedia.org/wiki/Chebyshev's_inequality). •

4.6 Moments

Motivation 4.132 — Moments of r.v.
The nature of a r.v. can be partial described in terms of a number of features — such as the expected value, the variance, the skewness, kurtosis, etc. — that can written in terms of expectation of powers of X, the moments of a r.v. •

4.6.1 Moments of r.v.

Definition 4.133 — kth. moment and kth. central moment of a r.v. (Karr, 1993, p. 123)
Let

- X be a r.v. such that $X \in L^k$, for some $k \in \mathbb{N}$.

Then:

- the kth. moment of X is given by the Riemann-Stieltjes integral

$$E(X^k) = \int_{-\infty}^{\infty} x^k \, dF_X(x); \tag{4.94}$$

- similarly, the kth. central moment of X equals

$$E\left\{[X - E(X)]^k\right\} = \int_{-\infty}^{+\infty} [x - E(X)]^k \, dF_X(x). \tag{4.95}$$

•

Remarks 4.134 — kth. moment and kth. central moment of a r.v. (Karr, 1993, p. 123; http://en.wikipedia.org/wiki/Moment_(mathematics))

- The kth. central moment exists under the assumption that $X \in L^k$ because $L^k \subseteq L^1$, for any $k \in \mathbb{N}$ (a consequence of Lyapunov's inequality).
- If the kth. (central) moment exists[29] so does the $(k-1)$th. (central) moment, and all lower-order moments. This is another consequence of Lyapunov's inequality.
- If $X \in L^1$ the first moment is the expectation of X; the first central moment is thus null. In higher orders, the central moments are more interesting than the moments about zero. •

[29] Or the kth. moment about any point exists. Note that the kth. central moment is nothing but the kth. moment about $E(X)$.

Proposition 4.135 — Computing the kth. moment of a non negative r.v.
If X is a non negative r.v. and $X \in L^k$, for $k \in \mathbb{N}$, we can write the kth. moment of X in terms of the following Riemann integral:

$$E(X^k) = \int_0^\infty k\, x^{k-1} \times [1 - F_X(x)]\, dx. \tag{4.96}$$

•

Exercise 4.136 — Computing the kth. moment of a non negative r.v.
Prove Proposition 4.135. •

Exercise 4.137 — Computing the kth. moment of a non negative r.v.
Let $X \sim \text{Exponential}(\lambda)$. Use Proposition 4.135 to prove that $E(X^k) = \frac{\Gamma(k+1)}{\lambda^k}$, for any $k \in \mathbb{N}$. •

Exercise 4.138 — The median of a r.v. and the minimization of the expected absolute deviation (Karr, 1993, p. 130, Exercise 4.2)
The median of the r.v. X, $med(X)$, is such that $P(\{X \leq med(X)\}) \geq \frac{1}{2}$ and $P(\{X \geq med(X)\}) \geq \frac{1}{2}$.

Prove that if $X \in L^1$ then

$$E(|X - med(X)|) \leq E(|X - a|), \tag{4.97}$$

for all $a \in \mathbb{R}$. •

Exercise 4.139 — Minimizing the mean squared error (Karr, 1993, p. 131, Exercise 4.12)
Let $\{A_1, \ldots, A_n\}$ be a finite partition of Ω. Suppose that we know which of $A_1, \ldots, A_n$ has occurred, and wish to predict whether some other event B has occurred. Since we know the values of the indicator functions $\mathbf{1}_{A_1}, \ldots, \mathbf{1}_{A_n}$, it make sense to use a predictor that is a function of them, namely linear predictors of the form $Y = \sum_{i=1}^n a_i \times \mathbf{1}_{A_i}$, whose accuracy is assessed via the mean squared error:

$$\text{MSE}(Y) = E[(\mathbf{1}_B - Y)^2]. \tag{4.98}$$

Prove that the values $a_i = P(B|A_i)$, $i = 1, \ldots, n$ minimize $\text{MSE}(Y)$. •

Exercise 4.140 — Expectation of a r.v. with respect to a conditional probability function (Karr, 1993, p. 132, Exercise 4.20)
Let A be an event such that $P(A) > 0$.

Show that if X is positive or integrable then $E(X|A)$, the expectation of X with respect to the conditional probability function $P_A(B) = P(B|A)$, is given by

$$E(X|A) \stackrel{def}{=} \frac{E(X;A)}{P(A)}, \tag{4.99}$$

where $E(X;A) = E(X \times \mathbf{1}_A)$ represents the expectation of X over the event A. •

4.6.2 Variance and standard deviation

Definition 4.141 — Variance and standard deviation of a r.v. (Karr, 1993, p. 124)
Let $X \in L^2$. Then:

- the 2nd. central moment is the variance of X,

$$V(X) = E\left\{[X - E(X)]^2\right\}; \tag{4.100}$$

- the positive square root of the variance is the standard deviation of X,

$$SD(X) = +\sqrt{V(X)}. \tag{4.101}$$

•

Remark 4.142 — Computing the variance of a r.v. (Karr, 1993, p. 124)
The variance of a r.v. $X \in L^2$ can also be expressed as

$$V(X) = E(X^2) - E^2(X), \tag{4.102}$$

which is more convenient than (4.100) for computational purposes. •

Exercise 4.143 — The meaning of a null variance (Karr, 1993, p. 131, Exercise 4.19)
Prove that if $V(X) = 0$ then $X \stackrel{a.s.}{=} E(X)$. •

Exercise 4.144 — Comparing the variance of X and $\min\{X, c\}$ (Karr, 1993, p. 133, Exercise 4.32)
Let X be a r.v. such that $E(X^2) < +\infty$ and c a real constant. Prove that

$$V(\min\{X, c\}) \leq V(X). \tag{4.103}$$

•

Proposition 4.145 — Variance of the sum (or difference) of two independent r.v. (Karr, 1993, p. 124)
If $X, Y \in L^2$ and are two INDEPENDENT r.v. then

$$V(X+Y) = V(X-Y) = V(X) + V(Y). \tag{4.104}$$

•

Exercise 4.146 — Expected values and variances of some important r.v. (Karr, 1993, pp. 125 and 130, Exercise 4.1)
Verify the entries of the following table.

Distribution	**Parameters**	**Expected value**	**Variance**
Discrete Uniform$(\{x_1, x_2, \ldots, x_n\})$	$\{x_1, x_2, \ldots, x_n\}$	$\frac{1}{n}\sum_{i=1}^{n} x_i$	$\left(\frac{1}{n}\sum_{i=1}^{n} x_i^2\right) - \left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)^2$
Bernoulli(p)	$p \in [0,1]$	p	$p\,(1-p)$
Binomial(n,p)	$n \in \mathbb{N}$; $p \in [0,1]$	$n\,p$	$n\,p\,(1-p)$
Hipergeometric(N,M,n)	$N \in \mathbb{N}$ $M \in \mathbb{N}$, $M \leq N$ $n \in \mathbb{N}$, $n \leq N$	$n\,\frac{M}{N}$	$n\,\frac{M}{N}\,\left(1-\frac{M}{N}\right)$
Geometric(p)	$p \in [0,1]$	$\frac{1}{p}$	$\frac{1-p}{p^2}$
Poisson(λ)	$\lambda \in \mathbb{R}^+$	λ	λ
NegativeBinomial(r,p)	$r \in \mathbb{N}$; $p \in [0,1]$	$\frac{r}{p}$	$\frac{r(1-p)}{p^2}$
Uniform(a,b)	$a, b \in \mathbb{R}$, $a < b$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
Normal(μ,σ^2)	$\mu \in \mathbb{R}$; $\sigma^2 \in \mathbb{R}^+$	μ	σ^2
Lognormal(μ,σ^2)	$\mu \in \mathbb{R}$; $\sigma^2 \in \mathbb{R}^+$	$e^{\mu+\frac{1}{2}\sigma^2}$	$(e^{\sigma^2}-1)e^{2\mu+\sigma^2}$
Exponential(λ)	$\lambda \in \mathbb{R}^+$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$
Gamma(α,β)	$\alpha, \beta \in \mathbb{R}^+$	$\frac{\alpha}{\beta}$	$\frac{\alpha}{\beta^2}$
Beta(α,β)	$\alpha, \beta \in \mathbb{R}^+$	$\frac{\alpha}{\alpha+\beta}$	$\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$
Weibull(α,β)	$\alpha, \beta \in \mathbb{R}^+$	$\alpha\,\Gamma\left(1+\frac{1}{\beta}\right)$	$\alpha^2\left[\Gamma\left(1+\frac{2}{\beta}\right) - \Gamma^2\left(1+\frac{1}{\beta}\right)\right]$

•

Definition 4.147 — Normalized (central) moments of a r.v.
(http://en.wikipedia.org/wiki/Moment_(mathematics))
Let X be a r.v. such that $X \in L^k$, for some $k \in \mathbb{N}$. Then:

- the normalized kth. moment of the X is the kth. moment divided by $[SD(X)]^k$,

$$\frac{E(X^k)}{[SD(X)]^k}; \tag{4.105}$$

- the normalized kth. central moment of X is given by

$$\frac{E\{[X - E(X)]^k\}}{[SD(X)]^k}; \tag{4.106}$$

These normalized central moments are dimensionless quantities, which represent the distribution independently of any linear change of scale. •

4.6.3 Skewness and kurtosis

Motivation 4.148 — Skewness of a r.v.
(http://en.wikipedia.org/wiki/Moment_(mathematics))
Any r.v. $X \in L^3$ with a symmetric p.(d.)f. will have a NULL 3rd. central moment. Thus, the 3rd. central moment is a measure of the lopsidedness of the distribution. •

Definition 4.149 — Skewness of a r.v.
(http://en.wikipedia.org/wiki/Moment_(mathematics))
Let $X \in L^3$ be a r.v. Then the normalized 3rd. central moment is called the skewness — or skewness coefficient (SC) —,

$$SC(X) = \frac{E\{[X - E(X)]^3\}}{[SD(X)]^3}. \tag{4.107}$$

•

Remark 4.150 — Skewness of a r.v.
(http://en.wikipedia.org/wiki/Moment_(mathematics))

- A r.v. X that is skewed to the left (the tail of the p.(d.)f. is heavier on the left) will have a negative skewness.
- A r.v. that is skewed to the right (the tail of the p.(d.)f. is heavier on the right), will have a positive skewness. •

Exercise 4.151 — Skewness of a r.v.
Prove that the skewness of:

(a) $X \sim \text{Exponential}(\lambda)$ equals $SC(X) = 2$
(http://en.wikipedia.org/wiki/Exponential_distribution);

(b) $X \sim \text{Pareto}(b, \alpha)$ is given by $SC(X) = \frac{2(1+\alpha)}{\alpha-3}\sqrt{\frac{\alpha-2}{\alpha}}$, for $\alpha > 3$
(http://en.wikipedia.org/wiki/Pareto_distribution). •

Motivation 4.152 — Kurtosis of a r.v.
(http://en.wikipedia.org/wiki/Moment_(mathematics))
The normalized 4th. central moment of any normal distribution is 3. Unsurprisingly, the normalized 4th. central moment is a measure of whether the distribution is tall and skinny or short and squat, compared to the normal distribution of the same variance. •

Definition 4.153 — Kurtosis of a r.v.
(http://en.wikipedia.org/wiki/Moment_(mathematics))
Let $X \in L^4$ be a r.v. Then the kurtosis — or kurtosis coefficient (KC) — is defined to be the normalized 4th. central moment minus 3,[30]

$$KC(X) = \frac{E\{[X - E(X)]^4\}}{[SD(X)]^4} - 3. \tag{4.108}$$

•

Remarks 4.154 — Kurtosis of a r.v.
(http://en.wikipedia.org/wiki/Moment_(mathematics))

- If the p.(d.)f. of the r.v. X has a peak at the expected value and long tails, the 4th. moment will be high and the kurtosis positive. Bounded distributions tend to have low kurtosis.
- $KC(X)$ must be greater than or equal to $[SC(X)]^2 - 2$; equality only holds for Bernoulli distributions (prove!).
- For unbounded skew distributions not too far from normal, $KC(X)$ tends to be somewhere between $[SC(X)]^2$ and $2 \times [SC(X)]^2$. •

[30] Some authors do not subtract three.

Exercise 4.155 — Kurtosis of a r.v.
Prove that the kurtosis of:

(a) $X \sim \text{Exponential}(\lambda)$ equals $KC(X) = 6$
(http://en.wikipedia.org/wiki/Exponential_distribution);

(b) $X \sim \text{Pareto}(b, \alpha)$ is given by $\frac{6(\alpha^3+\alpha^2-6\alpha-2)}{\alpha(\alpha-3)(\alpha-4)}$ for $\alpha > 4$
(http://en.wikipedia.org/wiki/Pareto_distribution). •

4.6.4 Covariance

Motivation 4.156 — Covariance (and correlation) between two r.v.
It is crucial to obtain measures of how much two variables change together, namely absolute and relative measures of (linear) association between pairs of r.v. •

Definition 4.157 — Covariance between two r.v. (Karr, 1993, p. 125)
Let $X, Y \in L^2$ be two r.v. Then the covariance between X and Y is equal to

$$\begin{aligned} cov(X,Y) &= E\{[X - E(X)] \times [Y - E(Y)]\} \\ &= E(XY) - E(X)E(Y) \end{aligned} \tag{4.109}$$

(this last formula is more useful for computational purposes, prove it!) •

Remark 4.158 — Covariance between two r.v.
(http://en.wikipedia.org/wiki/Covariance)
The units of measurement of the covariance between the r.v. X and Y are those of X times those of Y. •

Proposition 4.159 — Properties of the covariance
Let $X, Y, Z \in L^2$, $X_1, \ldots, X_n \in L^2$, $Y_1, \ldots, Y_n \in L^2$, and $a, b \in \mathbb{R}$. Then:

1. $X \perp\!\!\!\perp Y \Rightarrow cov(X,Y) = 0$
2. $cov(X,Y) = 0 \not\Rightarrow X \perp\!\!\!\perp Y$
3. $cov(X,Y) \neq 0 \Rightarrow X \not\perp\!\!\!\perp Y$

4. $cov(X, Y) = cov(Y, X)$ (symmetric operator!)
5. $cov(X, X) = V(X) \geq 0$ and $V(X) = 0 \Rightarrow X \overset{a.s.}{=} E(X)$ (positive semi-definite operator!)
6. $cov(aX, bY) = a\, b\, cov(X, Y)$
7. $cov(X + a, Y + b) = cov(X, Y)$
8. $cov(aX + bY, Z) = a\, cov(X, Z) + b\, cov(Y, Z)$ (bilinear operator!)
9. $cov\left(\sum_{i=1}^{n} X_i, \sum_{j=1}^{n} Y_j\right) = \sum_{i=1}^{n} \sum_{j=1}^{n} cov(X_i, Y_j)$
10. $cov\left(\sum_{i=1}^{n} X_i, \sum_{j=1}^{n} X_j\right) = \sum_{i=1}^{n} V(X_i) + 2 \times \sum_{i=1}^{n} \sum_{j=i+1}^{n} cov(X_i, X_j)$. •

Exercise 4.160 — Covariance
Prove properties 6 through 10 from Proposition 4.159. •

Proposition 4.161 — Variance of some linear combinations of r.v.
Let $X_1, \ldots, X_n \in L^2$. Then:

$$V\left(c_1 X_1 + c_2 X_2\right) = c_1^2 V(X_1) + c_2^2 V(X_2) + 2 c_1 c_2 cov(X_1, X_2); \tag{4.110}$$

$$V\left(X_1 + X_2\right) = V(X_1) + V(X_2) + 2 cov(X_1, X_2); \tag{4.111}$$

$$V\left(X_1 - X_2\right) = V(X_1) + V(X_2) - 2 cov(X_1, X_2); \tag{4.112}$$

$$V\left(\sum_{i=1}^{n} c_i X_i\right) = \sum_{i=1}^{n} c_i^2 V(X_i) + 2 \sum_{i=1}^{n} \sum_{j=i+1}^{n} c_i\, c_j\, cov(X_i, X_j). \tag{4.113}$$

When we deal with uncorrelated r.v. — i.e., if $cov(X_i, X_j) = 0$, $\forall i \neq j$ — or with pairwise independent r.v. — that is, $X_i \perp\!\!\!\perp X_j$, $\forall i \neq j$ —, we have:

$$V\left(\sum_{i=1}^{n} c_i X_i\right) = \sum_{i=1}^{n} c_i^2 V(X_i). \tag{4.114}$$

And if, besides being uncorrelated or (pairwise) independent r.v., we have $c_i = 1$, for $i = 1, \ldots, n$, we get:

$$V\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} V(X_i), \tag{4.115}$$

i.e. the variance of the sum of uncorrelated or (pairwise) independent r.v. is the sum of the individual variances. •

4.6.5 Correlation

Motivation 4.162 — Correlation between two r.v.
(http://en.wikipedia.org/wiki/Correlation_and_dependence)
The most familiar measure of dependence between two r.v. is (Pearson's) correlation. It is obtained by dividing the covariance between two variables by the product of their standard deviations.

Correlations are useful because they can indicate a predictive relationship that can be exploited in practice. For example, an electrical utility may produce less power on a mild day based on the correlation between electricity demand and weather. Moreover, correlations can also suggest possible causal, or mechanistic relationships. •

Definition 4.163 — Correlation between two r.v. (Karr, 1993, p. 125)
Let $X, Y \in L^2$ be two r.v. Then the correlation[31] between X and Y is given by

$$corr(X,Y) = \frac{cov(X,Y)}{\sqrt{V(X)\,V(Y)}}. \tag{4.116}$$

•

Remark 4.164 — Correlation between two r.v.
(http://en.wikipedia.org/wiki/Covariance)
Correlation is a dimensionless measure of linear dependence. •

Definition 4.165 — Uncorrelated r.v. (Karr, 1993, p. 125)
Let $X, Y \in L^2$. Then if

$$corr(X,Y) = 0 \tag{4.117}$$

X and Y are said to be uncorrelated r.v.[32] •

Exercise 4.166 — Uncorrelated r.v. (Karr, 1993, p. 131, Exercise 4.14)
Give an example of r.v. X and Y that are uncorrelated but for which there is a function g such that $Y = g(X)$. •

Exercise 4.167 — Uncorrelated r.v. (Karr, 1993, p. 131, Exercise 4.18)
Prove that if $V, W \in L^2$ and $(V, W) \stackrel{d}{=} (-V, W)$ then V and W are uncorrelated. •

[31] Also know as the Pearson's correlation coefficient (http://en.wikipedia.org/wiki/Correlation_and_dependence).

[32] $X, Y \in L^2$ are said to be correlated r.v. if $corr(X,Y) \neq 0$.

Exercise 4.168 — Sufficient conditions to deal with uncorrelated sample mean and variance (Karr, 1993, p. 132, Exercise 4.28)
Let $X_i \overset{i.i.d.}{\sim} X$, $i = 1, \ldots, n$, such that $E(X) = E(X^3) = 0$.

Prove that the sample mean $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ and the sample variance $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$ are uncorrelated r.v. •

Proposition 4.169 — Properties of the correlation
Let $X, Y \in L^2$, and $a, b \in I\!R$. Then:

1. $X \perp\!\!\!\perp Y \Rightarrow corr(X, Y) = 0$
2. $corr(X, Y) = 0 \not\Rightarrow X \perp\!\!\!\perp Y$
3. $corr(X, Y) \neq 0 \Rightarrow X \not\perp\!\!\!\perp Y$
4. $corr(X, Y) = corr(Y, X)$
5. $corr(X, X) = 1$
6. $corr(aX, bY) = corr(X, Y)$
7. $-1 \leq corr(X, Y) \leq 1$, for any pair of r.v.[33]
8. $corr(X, Y) = -1 \Leftrightarrow Y \overset{a.s.}{=} aX + b$, $a < 0$
9. $corr(X, Y) = 1 \Leftrightarrow Y \overset{a.s.}{=} aX + b$, $a > 0$. •

Exercise 4.170 — Properties of the correlation
Prove properties 7 through 9 from Proposition 4.169. •

Exercise 4.171 — Negative linear association between three r.v. (Karr, 1993, p. 131, Exercise 4.17)
Prove that there are no r.v. X, Y and Z such that $corr(X, Y) = corr(Y, Z) = corr(Z, X) = -1$. •

Remark 4.172 — Interpretation of the sign of a correlation
The correlation sign entre X e Y should be interpreted as follows:

- if $corr(X, Y)$ is "considerably" larger than zero (resp. smaller than zero) we can cautiously add that if X increases then Y "tends" to increase (resp. decrease). •

[33] A consequence of Cauchy-Schwarz's inequality.

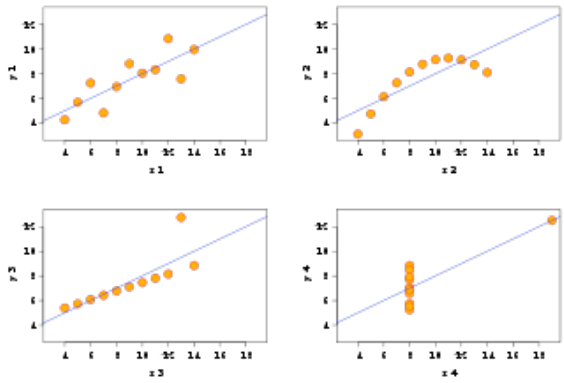
Remark 4.173 — Interpretation of the size of a correlation
(http://en.wikipedia.org/wiki/Pearson_product-moment_correlation_coefficient)
Several authors have offered guidelines for the interpretation of a correlation coefficient. Others have observed, however, that all such criteria are in some ways arbitrary and should not be observed too strictly.

The interpretation of a correlation coefficient depends on the context and purposes. A correlation of 0.9 may be very low if one is verifying a physical law using high-quality instruments, but may be regarded as very high in the social sciences where there may be a greater contribution from complicating factors. •

Remark 4.174 — Correlation and linearity
(http://en.wikipedia.org/wiki/Correlation_and_dependence)
Properties 8 and 9 from Proposition 4.169 suggest that

- correlation "quantifies" the linear association between X e Y.

Thus, if the absolute value of $corr(X, Y)$ is very close to the unit we are tempted to add that the association between X and Y is "likely" to be linear.



However, the Pearson's correlation coefficient indicates the strength of a linear relationship between two variables, but its value generally does not completely characterize their relationship.

The image on the right shows scatterplots of Anscombe's quartet, a set of four different pairs of variables created by Francis Anscombe. The four y variables have the same mean (7.5), standard deviation (4.12), correlation (0.816) and regression line ($y = 3 + 0.5x$). However, as can be seen on the plots, the distribution of the variables is very different.

The first one (top left) seems to be distributed normally, and corresponds to what one would expect when considering two variables correlated and following the assumption of normality.

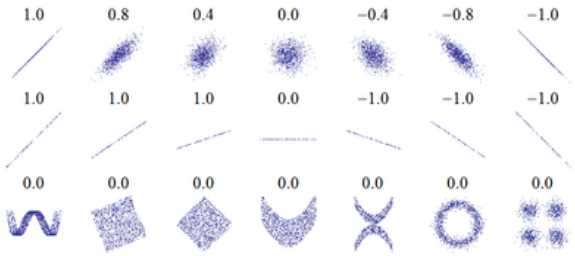
The second one (top right) is not distributed normally; while an obvious relationship between the two variables can be observed, it is not linear, and the Pearson correlation coefficient is not relevant.

In the third case (bottom left), the linear relationship is perfect, except for one outlier which exerts enough influence to lower the correlation coefficient from 1 to 0.816.

Finally, the fourth example (bottom right) shows another example when one outlier is enough to produce a high correlation coefficient, even though the relationship between the two variables is not linear. •

Remark 4.175 — Correlation and causality
(http://en.wikipedia.org/wiki/Correlation_and_dependence)
The conventional dictum that "correlation does not imply causation" means that correlation cannot be used to infer a causal relationship between the variables.[34]

This dictum should not be taken to mean that correlations cannot indicate the potential existence of causal relations. However, the causes underlying the correlation, if any, may be indirect and unknown, and high correlations also overlap with identity relations, where no causal process exists. Consequently, establishing a correlation between two variables is not a sufficient condition to establish a causal relationship (in either direction).



Several sets of (x, y) points, with the correlation coefficient of x and y for each set. Note that the correlation reflects the noisiness and direction of a linear relationship (top row), but not the slope of that relationship (middle), nor many aspects of nonlinear relationships (bottom). The figure in the center has a slope of 0 but in that case the correlation coefficient is undefined because the variance of y is zero. •

[34] A correlation between age and height in children is fairly causally transparent, but a correlation between mood and health in people is less so. Does improved mood lead to improved health; or does good health lead to good mood; or both? Or does some other factor underlie both? In other words, a correlation can be taken as evidence for a possible causal relationship, but cannot indicate what the causal relationship, if any, might be.

4.6.6 Moments of random vectors

Moments of random vectors are defined component, pairwise, etc. Instead of an expected value (resp. variance) we shall deal with a mean vector (resp. covariance matrix).

Definition 4.176 — Mean vector and covariance matrix of a random vector (Karr, 1993, p. 126)
Let $\underline{X} = (X_1, \ldots, X_d)$ be a $d-$dimensional random vector. Then provided that:

- $X_i \in L^1$, $i = 1, \ldots, d$, the mean vector of $\underline{X}$ is the $d-$vector of the individual means, $\underline{\mu} = (E(X_1), \ldots, E(X_d))$;
- $X_i \in L^2$, $i = 1, \ldots, d$, the covariance matrix of $\underline{X}$ is a $d \times d$ matrix given by $\mathbf{\Sigma} = [cov(X_i, X_j)]_{i,j=1,\ldots,d}$. •

Proposition 4.177 — Properties of the covariance matrix of a random vector (Karr, 1993, p. 126)
Let $\underline{X} = (X_1, \ldots, X_d)$ be a $d-$dimensional random vector with covariance matrix $\mathbf{\Sigma}$. Then:

- the diagonal of $\mathbf{\Sigma}$ has entries equal to $cov(X_i, X_i) = V(X_i)$, $i = 1, \ldots, d$;
- $\mathbf{\Sigma}$ is a symmetric matrix since $cov(X_i, X_j) = cov(X_j, X_i)$, $i, j = 1, \ldots, d$;
- $\mathbf{\Sigma}$ is a positive-definite matrix, that is, $\sum_{i=1}^{d} \sum_{j=1}^{d} c_i \times cov(X_i, X_j) \times c_j > 0$, for every $d-$vector $\underline{c} = (c_1, \ldots, c_d)$. •

Exercise 4.178 — Mean vector and covariance matrix of a linear combination of r.v. (matrix notation)
Let:

- $\underline{X} = (X_1, \ldots, X_d)$ a $d-$dimensional random vector;
- $\underline{\mu} = (E(X_1), \ldots, E(X_d))$ the mean vector of $\underline{X}$;
- $\mathbf{\Sigma} = [cov(X_i, X_j)]_{i,j=1,\ldots,d}$ the covariance matrix of $\underline{X}$;
- $\underline{c} = (c_1, \ldots, c_d)$ a vector of weights.

By noting that $\sum_{i=1}^{d} c_i X_i = \underline{c}^\top \underline{X}$, verify that:

- $E\left(\sum_{i=1}^{d} c_i X_i\right) = \underline{c}^\top \underline{\mu}$;
- $V\left(\sum_{i=1}^{d} c_i X_i\right) = \underline{c}^\top \mathbf{\Sigma}\, \underline{c}$. •

4.6.7 Multivariate normal distributions

Motivation 4.179 — Multivariate normal distribution (Tong, 1990, p. 1)
There are many reasons for the predominance of the multivariate normal distribution:

- it represents a natural extension of the univariate normal distribution and provides a suitable model for many real-life problems concerning vector-valued data;
- even if the original data cannot be fitted satisfactorily with a multivariate normal distribution, by the central limit theorem the distribution of the sample mean vector is asymptotically normal;
- the p.d.f. of a multivariate normal distribution is uniquely determined by the mean vector and the covariance matrix;
- zero correlation imply independence between two components of the random vector with multivariate normal distribution;
- the family of multivariate normal distributions is closed under linear transformations or linear combinations;
- the marginal distribution of any subset of components of a random vector with multivariate normal distribution is also multivariate normal;
- the conditional distribution in a multivariate normal distribution is also multivariate normal. •

Remark 4.180 — Multivariate normal distribution (Tong, 1990, p. 2)
Studies of the bivariate normal distribution seem to begin in the middle of the XIX century, and moved forward in 1888 with F. Galton's (1822–1911) work on the applications of correlations analysis in genetics. In 1896, K. Pearson (1857–1936) gave a definitive mathematical formulation of the bivariate normal distribution.

The multivariate normal distribution was treated comprehensively for the first time in 1892 by F.Y. Edgeworth (1845–1926). •

A random vector has a multivariate normal distribution if it is a linear transformation of a random vector with i.i.d. components with standard normal distribution (Karr, 1993, p. 126).

Definition 4.181 — Multivariate normal distribution (Karr, 1993, p. 126)
Let:

- $\underline{\mu} = (\mu_1, \ldots, \mu_d) \in I\!R^d$;
- $\mathbf{\Sigma} = [\sigma_{ij}]_{i,j=1,\ldots,d}$ be a symmetric, positive-definite, non-singular $d \times d$ matrix;[35]

Then the random vector $\underline{X} = (X_1, \ldots, X_d)$ has a multivariate normal distribution with mean vector $\underline{\mu}$ and covariance matrix $\mathbf{\Sigma}$ if

$$\underline{X} = \mathbf{\Sigma}^{\frac{1}{2}} \underline{Y} + \underline{\mu}, \tag{4.118}$$

where:

- $\underline{Y} = (Y_1, \ldots, Y_d)$ with $Y_i \overset{i.i.d.}{\sim} \text{Normal}(0, 1)$, $i = 1, \ldots, d$;
- $\mathbf{\Sigma}^{\frac{1}{2}}$ is the unique matrix satisfying $\left(\mathbf{\Sigma}^{\frac{1}{2}}\right)^\top \times \mathbf{\Sigma}^{\frac{1}{2}} = \mathbf{\Sigma}$.

In this case we write $\underline{X} \sim \text{Normal}_d(\underline{\mu}, \mathbf{\Sigma})$. •

We can use Definition 4.181 to simulate a multivariate normal distribution as mentioned below.

Remark 4.182 — Simulating a multivariate normal distribution (Gentle, 1998, pp. 105–106)
Since $Y_i \overset{i.i.d.}{\sim} \text{Normal}(0, 1)$, $i = 1, \ldots, d$, implies that $\underline{X} = \mathbf{\Sigma}^{\frac{1}{2}} \underline{Y} + \underline{\mu} \sim \text{Normal}_d(\underline{\mu}, \mathbf{\Sigma})$ we can obtain a $d-$dimensional pseudo-random vector from this multivariate normal distribution if we:

1. generate d independent pseudo-random numbers, $y_1, \ldots, y_d$, from the standard normal distribution;
2. assign $\underline{x} = \mathbf{\Sigma}^{\frac{1}{2}} \underline{y} + \underline{\mu}$, where $\underline{y} = (y_1, \ldots, y_d)$.

Gentle (1998, p. 106) refers other procedures to generate pseudo-random numbers with multivariate normal distribution. •

[35] A $d \times d$ matrix $\mathbf{A}$ is called invertible or non-singular or non-degenerate if there exists an $d \times d$ matrix $\mathbf{B}$ such that $\mathbf{AB} = \mathbf{BA} = \mathbf{I}_d$, where $\mathbf{I}_d$ denotes the $d \times d$ identity matrix. (http://en.wikipedia.org/wiki/Invertible_matrix) .

Proposition 4.183 — Characterization of the multivariate normal distribution (Karr, 1993, pp. 126–127)
Let $\underline{X} \sim \text{Normal}_d(\underline{\mu}, \mathbf{\Sigma})$ where $\underline{\mu} = (\mu_1, \ldots, \mu_d)$ and $\mathbf{\Sigma} = [\sigma_{i\,j}]_{i,j=1,\ldots,d}$. Then:

$$E(X_i) = \mu_i,\ i = 1, \ldots, d; \tag{4.119}$$

$$cov(X_i, X_j) = \sigma_{ij},\ i, j = 1, \ldots, d; \tag{4.120}$$

$$f_{\underline{X}}(\underline{x}) = (2\pi)^{-\frac{d}{2}} |\mathbf{\Sigma}|^{-\frac{1}{2}} \times \exp\left[-\frac{1}{2}\,(\underline{x} - \underline{\mu})^{\top} \mathbf{\Sigma}^{-1} (\underline{x} - \underline{\mu})\right], \tag{4.121}$$

for $\underline{x} = (x_1, \ldots, x_d) \in I\!R^d$. •

Exercise 4.184 — P.d.f. of a bivariate normal distribution
Let (X_1, X_2) have a (non-singular) bivariate normal distribution with mean vector and covariance matrix

$$\underline{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} \quad \text{and} \quad \mathbf{\Sigma} = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}, \tag{4.122}$$

respectively, where $|\rho| = |corr(X, Y)| < 1$.

(a) Verify that the joint p.d.f. is given by

$$f_{X_1,X_2}(x_1, x_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2 - 2\rho\left(\frac{x_1-\mu_1}{\sigma_1}\right)\left(\frac{x_2-\mu_2}{\sigma_2}\right) + \left(\frac{x_2-\mu_2}{\sigma_2}\right)^2\right]\right\},\ (x_1, x_2) \in I\!R^2. \tag{4.123}$$

(b) Use *Mathematica* to plot this joint p.d.f. for $\mu_1 = \mu_2 = 0$ and $\sigma_1^2 = \sigma_2^2 = 1$, and at least five different values of the correlation coefficient ρ. •

Exercise 4.185 — Normally distributed r.v. with a non bivariate normal distribution
We have already mentioned that if two r.v. X_1 and X_2 both have a standard normal distribution this does not imply that the random vector (X_1, X_2) has a joint normal distribution.[36]

Prove that $X_2 = X_1$ if $|X_1| > c$ and $X_2 = -X_1$ if $|X_1| < c$, where $c > 0$, illustrates this fact. •

[36] See http://en.wikipedia.org/wiki/Multivariate_normal_distribution.

In what follows we describe a few distributional properties of bivariate normal distributions and, more generally, multivariate normal distributions.

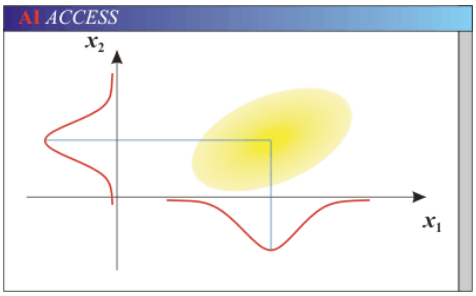
Proposition 4.186 — Marginal distributions/moments in the bivariate normal setting (Tong, 1990, p. 8, Theorem 2.1.1)
Let $\underline{X} = (X_1, X_2)$ be distributed according to a bivariate normal distribution with parameters

$$\underline{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} \quad \text{and} \quad \mathbf{\Sigma} = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}. \tag{4.124}$$

Then the marginal distribution of X_i, $i = 1, 2$, is normal. In fact,

$$X_i \sim \text{Normal}(\mu_i, \sigma_i^2),\ i = 1, 2. \tag{4.125}$$

The following figure[37] shows the two marginal distributions of a bivariate normal distribution:



•

Consider the partitions of $\underline{X}$, $\underline{\mu}$ and $\mathbf{\Sigma}$ given below,

$$\underline{X} = \begin{bmatrix} \underline{X}_1 \\ \underline{X}_2 \end{bmatrix}, \quad \underline{\mu} = \begin{bmatrix} \underline{\mu}_1 \\ \underline{\mu}_2 \end{bmatrix} \quad \text{and} \quad \mathbf{\Sigma} = \begin{bmatrix} \mathbf{\Sigma}_{11} & \mathbf{\Sigma}_{12} \\ \mathbf{\Sigma}_{21} & \mathbf{\Sigma}_{22} \end{bmatrix}, \tag{4.126}$$

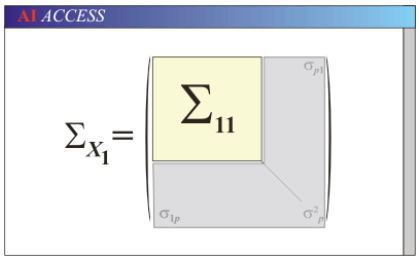
where:

- $\underline{X}_1 = (X_1, \ldots, X_k)$ is made up of the first $k < d$ components of $\underline{X}$;
- $\underline{X}_2 = (X_{k+1}, \ldots, X_d)$ is made up of the remaining components of $\underline{X}$;
- $\underline{\mu}_1 = (\mu_1, \ldots, \mu_k)$;
- $\underline{\mu}_2 = (\mu_{k+1}, \ldots, \mu_d)$;

[37] Taken from http://www.aiaccess.net/English/Glossaries/GlosMod/e_gm_multinormal_distri.htm.

- $\mathbf{\Sigma}_{11} = [\sigma_{ij}]_{i,j=1,\ldots,k}$; $\mathbf{\Sigma}_{12} = [\sigma_{ij}]_{1 \leq i < j \leq k}$;
- $\mathbf{\Sigma}_{21} = \mathbf{\Sigma}_{12}^{\top}$;
- $\mathbf{\Sigma}_{22} = [\sigma_{ij}]_{i,j=k+1,\ldots,d}$.

The following figure (where $d = p$)[38] represents the covariance matrix of $\underline{X}_1$, $\mathbf{\Sigma}_{11}$, which is just the upper left corner square submatrix of order k of the original covariance matrix:



Theorem 4.187 — Marginal distributions/moments in the multivariate normal setting (Tong, 1990, p. 30, Theorem 3.3.1)
Let $\underline{X} \sim \text{Normal}_d(\underline{\mu}, \mathbf{\Sigma})$. Then for every $k < d$ the marginal distributions of $\underline{X}_1$ and $\underline{X}_2$ are also multivariate normal:

$$\underline{X}_1 \sim \text{Normal}_k(\underline{\mu}_1, \mathbf{\Sigma}_{11}) \tag{4.127}$$
$$\underline{X}_2 \sim \text{Normal}_{d-k}(\underline{\mu}_2, \mathbf{\Sigma}_{22}), \tag{4.128}$$

respectively. •

The family of multivariate normal distributions is closed under linear transformations, as stated below.

Theorem 4.188 — Distribution/moments of a linear transformation of a bivariate normal random vector (Tong, 1990, p. 10, Theorem 2.1.2)
Let:

- $\underline{X} \sim \text{Normal}_2(\underline{\mu}, \mathbf{\Sigma})$;
- $\mathbf{C} = [c_{ij}]$ be a 2×2 real matrix;
- $\underline{b} = (b_1, b_2)$ be a real vector.

Then

$$\underline{Y} = \mathbf{C}\,\underline{X} + \underline{b} \sim \text{Normal}_2(\mathbf{C}\,\underline{\mu} + \underline{b}, \mathbf{C}\,\mathbf{\Sigma}\,\mathbf{C}^{\top}). \tag{4.129}$$

•

[38] Also taken from http://www.aiaccess.net/English/Glossaries/GlosMod/e_gm_multinormal_distri.htm.

Exercise 4.189 — Distribution/moments of a linear transformation of a bivariate normal random vector

(a) Prove that if in Theorem 4.188 we choose

$$\mathbf{C} = \begin{bmatrix} \sigma_1^{-1} & 0 \\ 0 & \sigma_2^{-1} \end{bmatrix} \tag{4.130}$$

and $\underline{b} = -\mathbf{C}\,\underline{\mu}$, then $\underline{Y}$ is a bivariate normal variable with zero means, unit variances and correlation coefficient ρ (Tong, 1990, p. 10).

(b) Now consider a linear transformation of $\underline{Y}$, $\underline{Y}^*$, by rotating the xy axes by 45 degrees counterclockwise:

$$\underline{Y}^* = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix} \underline{Y}. \tag{4.131}$$

Verify that $\underline{Y}^*$ is a bivariate normal variable with zero means, variances $1-\rho$ and $1+\rho$ and null correlation coefficient (Tong, 1990, p. 10). Comment.

(c) Conclude that if $\underline{X} \sim \text{Normal}_2(\underline{\mu}, \mathbf{\Sigma})$ such that $|\rho| < 1$ then

$$\begin{bmatrix} \frac{1}{\sqrt{1-\rho}} & 0 \\ 0 & \frac{1}{\sqrt{1+\rho}} \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix} \begin{bmatrix} \sigma_1^{-1} & 0 \\ 0 & \sigma_2^{-1} \end{bmatrix} \begin{bmatrix} X_1 - \mu_1 \\ X_2 - \mu_2 \end{bmatrix} \sim \text{Normal}_2(\underline{0}, \mathbf{I}_2), \tag{4.132}$$

where $\underline{0} = (0, 0)$ and $\mathbf{I}_2$ is the 2×2 identity matrix (Tong, 1990, p. 11).

(d) Prove that if $Z_i \overset{i.i.d.}{\sim} \text{Normal}(0, 1)$ then

$$\begin{aligned} &\begin{bmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix} \begin{bmatrix} \sqrt{1-\rho} & 0 \\ 0 & \sqrt{1+\rho} \end{bmatrix} \begin{bmatrix} Z_1 \\ Z_2 \end{bmatrix} + \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} \\ &\overset{st}{=} \begin{bmatrix} \sigma_1 & 0 \\ \sigma_2\rho & \sigma_2\sqrt{1-\rho^2} \end{bmatrix} \begin{bmatrix} Z_1 \\ Z_2 \end{bmatrix} + \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} \sim \text{Normal}_2(\underline{\mu}, \mathbf{\Sigma}), \end{aligned} \tag{4.133}$$

i.e. we can obtain a bivariate normal distribution with any mean vector and (non-singular, semi-definite positive) covariance matrix through a transformation of two independent standard normal r.v. (Tong, 1990, p. 11; http://xbeta.org/wiki/show/Bivariate+normal+distribution). •

Theorem 4.190 — Distribution/moments of a linear transformation of a multivariate normal distribution (Tong, 1990, p. 32, Theorem 3.3.3)
Let:

- $\underline{X} \sim \text{Normal}_d(\underline{\mu}, \mathbf{\Sigma})$;
- $\mathbf{C} = [c_{ij}]$ be any given $k \times d$ real matrix;
- $\underline{b}$ is any $k \times 1$ real vector.

Then

$$\underline{Y} = \mathbf{C}\,\underline{X} + \underline{b} \sim \text{Normal}_k(\mathbf{C}\,\underline{\mu} + \underline{b}, \mathbf{C}\,\mathbf{\Sigma}\,\mathbf{C}^\top). \tag{4.134}$$

•

The family of multivariate normal distributions is closed not only under linear transformations, as stated in the previous theorem, but also under linear combinations.

Corollary 4.191 — Distribution/moments of a linear combination of the components of a multivariate normal random vector (Tong, 1990, p. 33, Corollary 3.3.3)
Let:

- $\underline{X} \sim \text{Normal}_d(\underline{\mu}, \mathbf{\Sigma})$ partitioned as in (4.126);
- $\mathbf{C}_1$ and $\mathbf{C}_2$ be two $m \times k$ and $m \times (d-k)$ real matrices, respectively.

Then

$$\underline{Y} = \mathbf{C}_1\,\underline{X}_1 + \mathbf{C}_2\,\underline{X}_2 \sim \text{Normal}_m(\underline{\mu}_{\underline{Y}}, \mathbf{\Sigma}_{\underline{Y}}), \tag{4.135}$$

where the mean vector and the covariance matrix of $\underline{Y}$ are given by

$$\underline{\mu}_{\underline{Y}} = \mathbf{C}_1\,\underline{\mu}_1 + \mathbf{C}_2\,\underline{\mu}_2 \tag{4.136}$$

$$\mathbf{\Sigma}_{\underline{Y}} = \mathbf{C}_1\,\mathbf{\Sigma}_{11}\,\mathbf{C}_1^\top + \mathbf{C}_2\,\mathbf{\Sigma}_{22}\,\mathbf{C}_2^\top + \mathbf{C}_1\,\mathbf{\Sigma}_{12}\,\mathbf{C}_2^\top + \mathbf{C}_2\,\mathbf{\Sigma}_{21}\,\mathbf{C}_1^\top, \tag{4.137}$$

respectively. •

The result that follows has already been proved in Chapter 3 and is a particular case of Theorem 4.193.

Corollary 4.192 — Correlation and independence in a bivariate normal setting (Tong, 1990, p. 8, Theorem 2.1.1)
Let $\underline{X} = (X_1, X_2) \sim \text{Normal}_2(\underline{\mu}, \mathbf{\Sigma})$. Then X_1 and X_2 are independent iff $\rho = 0$. •

In general, r.v. may be uncorrelated but highly dependent. But if a random vector has a multivariate normal distribution then any two or more of its components that are uncorrelated are independent.

Theorem 4.193 — Correlation and independence in a multivariate normal setting (Tong, 1990, p. 31, Theorem 3.3.2)
Let $\underline{X} \sim \text{Normal}_d(\underline{\mu}, \mathbf{\Sigma})$ partitioned as in (4.126). Then $\underline{X}_1$ and $\underline{X}_2$ are independent random vectors iff $\mathbf{\Sigma}_{12} = \mathbf{\Sigma}_{12}^\top = \mathbf{0}_{k\times(d-k)}$. •

Corollary 4.194 — Linear combination of independent multivariate normal random vectors (Tong, 1990, p. 33, Corollary 3.3.4)
Let $\underline{X}_1, \ldots, \underline{X}_N$ be independent $\text{Normal}_d(\underline{\mu}_i, \mathbf{\Sigma}_i)$, $i = 1, \ldots, N$, random vectors. Then

$$\underline{Y} = \sum_{i=1}^{N} c_i \, \underline{X}_i \sim \text{Normal}_d \left(\sum_{i=1}^{N} c_i \, \underline{\mu}_i, \sum_{i=1}^{N} c_i^2 \, \mathbf{\Sigma}_i \right). \tag{4.138}$$

•

Proposition 4.195 — Independence between the sample mean vector and covariance matrix (Tong, 1990, pp. 47–48)
Let:

- N be a positive integer;
- $\underline{X}_1, \ldots, \underline{X}_N$ be i.i.d. random vectors with a common $\text{Normal}_d(\underline{\mu}, \mathbf{\Sigma})$ distribution, such that $\mathbf{\Sigma}$ is positive definite;
- $\bar{\underline{X}}_N = \frac{1}{N}\sum_{t=1}^{N} \underline{X}_t = (\bar{X}_1, \ldots, \bar{X}_d)$ denote the sample mean vector, where $\bar{X}_i = \frac{1}{N}\sum_{t=1}^{N} X_{it}$ and X_{it} the ith component of $\underline{X}_t$;
- $\mathbf{S}_N = [S_{ij}]_{i,j=1,\ldots,d}$ denote the sample covariance matrix, where $S_{ij} = \frac{1}{N-1}\sum_{t=1}^{N}(X_{it} - \bar{X}_i)(X_{jt} - \bar{X}_j)$.

Then $\bar{\underline{X}}_N$ and $\mathbf{S}_N$ are independent. •

Definition 4.196 — Mixed (central) moments
Let:

- $\underline{X} = (X_1, \ldots, X_d)$ be a random $d-$vector;
- $r_1, \ldots, r_d \in I\!N$.

Then:

- the mixed moment of order $(r_1, \ldots, r_d)$ of $\underline{X}$ is given by

$$E\left[X_1^{r_1} \times \ldots \times X_d^{r_d}\right], \tag{4.139}$$

  and is also called a $(\sum_{i=1}^{d} r_i)$th order moment of $\underline{X}$;[39]

- the mixed central moment of order $(r_1, \ldots, r_d)$ of $\underline{X}$ is defined as

$$E\left\{[X_1 - E(X_1)]^{r_1} \times \ldots [X_d - E(X_d)]^{r_d}\right\}. \tag{4.140}$$

•

The Isserlis' theorem is a formula that allows one to compute mixed moments of the multivariate normal distribution with null mean vector[40] in terms of the entries of its covariance matrix.

Remarks 4.197 — Isserlis' theorem (http://en.wikipedia.org/wiki/Isserlis'_theorem)

- In his original paper from 1918, Isserlis considered only the fourth-order moments, in which case the formula takes appearance

$$\begin{aligned} E(X_1X_2X_3X_4) &= E(X_1X_2) \times E(X_3X_4) + E(X_1X_3) \times E(X_2X_4) \\ &\quad + E(X_1X_4) \times E(X_2X_3), \end{aligned} \tag{4.141}$$

  which can be written in terms of the covariances: $\sigma_{12} \times \sigma_{34} + \sigma_{13} \times \sigma_{24} + \sigma_{14} \times \sigma_{23}$.

  It also added that if $(X_1, \ldots, X_{2n})$ is a zero mean multivariate normal random vector, then

$$E(X_1 \ldots X_{2n-1}) = 0 \tag{4.142}$$

$$E(X_1 \ldots X_{2n}) = \sum \prod E(X_i X_j), \tag{4.143}$$

[39] See for instance http://en.wikipedia.org/wiki/Multivariate_normal_distribution.

[40] Or mixed moments of the difference between a multivariate normal random vector $\underline{X}$ and its mean vector.

where the notation $\sum\prod$ means summing over all distinct ways of partitioning $X_1, \ldots, X_{2n}$ into pairs.

- This theorem is particularly important in particle physics, where it is known as Wick's theorem.
- Another applications include the analysis of portfolio returns, quantum field theory, generation of colored noise, etc. •

Theorem 4.198 — Isserlis' theorem
(http://en.wikipedia.org/wiki/Multivariate_normal_distribution)
Let:

- $\underline{X} = (X_1, \ldots, X_d) \sim \text{Normal}_d(\underline{\mu}, \boldsymbol{\Sigma})$;
- $E\left[\prod_{i=1}^{d}(X_i - \mu_i)^{r_i}\right]$ be the mixed central moment of order $(r_1, \ldots, r_d)$ of $\underline{X}$;
- $k = \sum_{i=1}^{d} r_i$.

Then:

- if k is odd (i.e. $k = 2n - 1$, $n \in \mathbb{N}$)

$$E\left[\prod_{i=1}^{d}(X_i - \mu_i)^{r_i}\right] = 0; \tag{4.144}$$

- if k is even (i.e. $k = 2n$, $n \in \mathbb{N}$)

$$E\left[\prod_{i=1}^{d}(X_i - \mu_i)^{r_i}\right] = \sum\prod \sigma_{ij}, \tag{4.145}$$

where the $\sum\prod$ is taken over all allocations of the set $\{1, 2, \ldots, n\}$ into n (unordered) pairs, that is, if you have a $k = 2n = 6$th order central moment, you will be summing the products of $n = 3$ covariances. •

Exercise 4.199 — **Isserlis' theorem**
Let $\underline{X} = (X_1, \ldots, X_4) \sim \text{Normal}_4(\underline{0}, \boldsymbol{\Sigma} = [\sigma_{ij}])$. Prove that

(a) $E(X_i^4) = 3\sigma_{ii}^2$, $i = 1, \ldots, 4$,

(b) $E(X_i^3 X_j) = 3\sigma_{ii}\sigma_{ij}$, $i, j = 1, \ldots, 4$,

(c) $E(X_i^2 X_j^2) = \sigma_{ii}\sigma_{jj} + 2\sigma_{ij}^2$, $i, j = 1, \ldots, 4$,

(d) $E(X_i^2 X_j X_l) = \sigma_{ii}\sigma_{jl} + 2\sigma_{ij}\sigma_{il}$, $i, j, l = 1, \ldots, 4$,

(e) $E(X_i X_j X_l X_n) = \sigma_{ij}\sigma_{lm} + \sigma_{il}\sigma_{jn} + \sigma_{in}\sigma_{jl}$, $i, j, l, n = 1, \ldots, 4$

(http://en.wikipedia.org/wiki/Isserlis'_theorem). •

In passing from univariate to multivariate distributions, some essentially new features require our attention: these features are connected not only with relations among sets of variables including covariance and correlation, but also regressions (conditional expected values) and, generally, conditional distributions (Johnson and Kotz, 1969, p. 280).

Theorem 4.200 — **Conditional distributions and regressions in the bivariate normal setting** (Tong, 1990, p. 8, Theorem 2.1.1)
Let $\underline{X} = (X_1, X_2)$ be distributed according to a bivariate normal distribution with parameters

$$\underline{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} \quad \text{and} \quad \boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix} = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{bmatrix}. \tag{4.146}$$

If $|\rho| < 1$ then

$$X_1|\{X_2 = x_2\} \sim \text{Normal}\left(\mu_1 + \frac{\rho\sigma_1}{\sigma_2}(x_2 - \mu_2), \sigma_1^2(1 - \rho^2)\right), \tag{4.147}$$

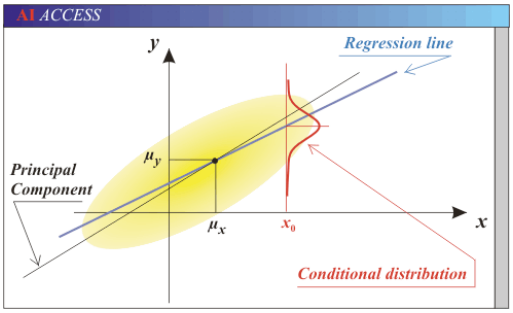
i.e.

$$X_1|\{X_2 = x_2\} \sim \text{Normal}\left(\mu_1 + \sigma_{12}\,\sigma_{22}^{-1}\,(x_2 - \mu_2), \sigma_{11} - \sigma_{12}\,\sigma_{22}^{-1}\,\sigma_{21}\right). \tag{4.148}$$

•

Exercise 4.201 — **Conditional distributions and regressions in the bivariate normal setting**
Prove results (4.147) and (4.148), and show that they are equivalent. •

The following figure[41] shows the conditional distribution of $Y|\{X = x_0\}$ of a random vector (X, Y) with a bivariate normal distribution:



The inverse Mills ratio is the ratio of the probability density function over the cumulative distribution function of a distribution and corresponds to a specific conditional expectation, as stated below.

Definition 4.202 — Inverse Mills' ratio
(http://en.wikipedia.org/wiki/Inverse_Mills_ratio; Tong, 1990, p. 174)
Let $\underline{X} = (X_1, X_2)$ be a bivariate normal random vector with **zero means**, **unit variances** and correlation coefficient ρ. Then the conditional expectation

$$E(X_1|\{X_2 > x_2\}) = \rho \frac{\phi(x_2)}{\Phi(-x_2)} \tag{4.149}$$

is often called the inverse Mills' ratio. •

Remark 4.203 — Inverse Mills' ratio
(http://en.wikipedia.org/wiki/Inverse_Mills_ratio)
A common application of the inverse Mills' ratio arises in regression analysis to take account of a possible selection bias. •

Exercise 4.204 — Conditional distributions and the inverse Mills' ratio in the bivariate normal setting
Assume that X_1 represents the log-dose of insuline that has been administrated and X_2 the decrease in blood sugar after a fixed amount of time. Also assume that (X_1, X_2) has a bivariate normal distribution with mean vector and covariance matrix

$$\underline{\mu} = \begin{bmatrix} 0.56 \\ 53 \end{bmatrix} \quad \text{e} \quad \mathbf{\Sigma} = \begin{bmatrix} 0.027 & 2.417 \\ 2.417 & 407.833 \end{bmatrix}. \tag{4.150}$$

[41] Once again taken from http://www.aiaccess.net/English/Glossaries/GlosMod/e_gm_multinormal_distri.htm.

(a) Obtain the probability that the decrease in blood sugar exceeds 70, given that log-dose of insuline that has been administrated is equal to 0.5.

(b) Determine the log-dose of insuline that has to be administrated so that the expected value of the decrease in blood sugar equals 70.

(c) Obtain the expected value of the decrease in blood sugar, given that log-dose of insuline that has been administrated exceeds 0.5. •

Theorem 4.205 — Conditional distributions and regressions in the multivariate normal setting (Tong, 1990, p. 35, Theorem 3.3.4)
Let $\underline{X} \sim \text{Normal}_d(\underline{\mu}, \mathbf{\Sigma})$ partitioned as in (4.126). Then

$$\underline{X}_1 | \{\underline{X}_2 = \underline{x}_2\} \sim \text{Normal}_k \left(\underline{\mu}_1 + \mathbf{\Sigma}_{12}\, \mathbf{\Sigma}_{22}^{-1}\, (\underline{x}_2 - \underline{\mu}_2),\ \mathbf{\Sigma}_{11} - \mathbf{\Sigma}_{12}\, \mathbf{\Sigma}_{22}^{-1}\, \mathbf{\Sigma}_{21} \right). \quad (4.151)$$

•

Exercise 4.206 — Conditional distributions and regressions in the multivariate normal setting
Derive (4.148) from (4.151). •

4.6.8 Multinomial distributions

The genesis and the definition of multinomial distributions follow.

Motivation 4.207 — Multinomial distribution
(http://en.wikipedia.org/wiki/Multinomial_distribution)
The multinomial distribution is a generalization of the binomial distribution.

The binomial distribution is the probability distribution of the number of "successes" in n independent Bernoulli trials, with the same probability of "success" on each trial.

In a multinomial distribution, the analog of the Bernoulli distribution is the categorical distribution, where each trial results in exactly one of some fixed finite number d of possible outcomes, with probabilities $p_1, \ldots, p_d$ ($p_i \in [0,1]$, $i = 1, \ldots, d$, and $\sum_{i=1}^{d} p_i = 1$), and there are n independent trials. •

Definition 4.208 — Multinomial distribution (Johnson and Kotz, 1969, p. 281)
Consider a series of n independent trials, in each of which just one of d mutually exclusive events $E_1, \ldots, E_d$ must be observed, and in which the probability of occurrence of event E_i is equal to p_i for each trial, with, of course, $p_i \in [0,1]$, $i = 1, \ldots, d$, and $\sum_{i=1}^{d} p_i = 1$. Then the joint distribution of the r.v. $N_1, \ldots, N_d$, representing the numbers of occurrences of the events $E_1, \ldots, E_d$ (respectively) in n trials, is defined by

$$P(\{N_1 = n_1, \ldots, N_d = n_d\}) = \frac{n!}{\prod_{i=1}^{d} n_i!} \times \prod_{i=1}^{d} p_i^{n_i}, \tag{4.152}$$

for $n_i \in \mathbb{N}_0$, $i = 1, \ldots, d$, such that $\sum_{i=1}^{d} n_i = n$. The random $d-$vector $\underline{N} = (N_1, \ldots, N_d)$ is said to have a multinomial distribution with parameters n and $\underline{p} = (p_1, \ldots, p_d)$) — in short $\underline{N} \sim \text{Multinomial}_{d-1}(n, \underline{p} = (p_1, \ldots, p_d))$.[42] •

Remark 4.209 — Special case of the multinomial distribution (Johnson and Kotz, 1969, p. 281)
Needless to say, we deal with the binomial distribution when $d = 2$, i.e.,

$$\text{Multinomial}_{2-1}(n, \underline{p} = (p, 1-p)) \stackrel{d}{=} \text{Binomial}(n, p). \tag{4.153}$$

Curiously, J. Bernoulli, who worked with the binomial distribution, also used the multinomial distribution. •

[42] The index $d-1$ follows from the fact that the r.v. N_d (or any other component of $\underline{N}$) is redundant: $N_d = n - \sum_{i=1}^{d-1} N_i$.

Remark 4.210 — Applications of multinomial distribution
(http://controls.engin.umich.edu/wiki/index.php/Multinomial_distributions)
Multinomial systems are a useful analysis tool when a "success-failure" description is insufficient to understand the system. For instance, in chemical engineering applications, multinomial distributions are relevant to situations where there are more than two possible outcomes (temperature = high, med, low).

A continuous form of the multinomial distribution is the Dirichlet distribution (http://en.wikipedia.org/wiki/Dirichlet_distribution).[43] •

Exercise 4.211 — Multinomial distribution (p.f.)
In a recent three-way election for a large country, candidate A received 20% of the votes, candidate B received 30% of the votes, and candidate C received 50% of the votes.

If six voters are selected randomly, what is the probability that there will be exactly one supporter for candidate A, two supporters for candidate B and three supporters for candidate C in the sample? (http://en.wikipedia.org/wiki/Multinomial_distribution) •

Exercise 4.212 — Multinomial distribution
A runaway reaction occurs when the heat generation from an exothermic reaction exceeds the heat loss. Elevated temperature increases reaction rate, further increasing heat generation and pressure buildup inside the reactor. Together, the uncontrolled escalation of temperature and pressure inside a reactor may cause an explosion. The precursors to a runaway reaction — high temperature and pressure — can be detected by the installation of reliable temperature and pressure sensors inside the reactor. Runaway reactions can be prevented by lowering the temperature and/or pressure inside the reactor before they reach dangerous levels. This task can be accomplished by sending a cold inert stream into the reactor or venting the reactor.

Les is a process engineer at the Miles Reactor Company that has been assigned to work on a new reaction process. Using historical data from all the similar reactions that have been run before, Les has estimated the probabilities of each outcome occurring during the new process. The potential outcomes of the process include all permutations of the possible reaction temperatures (low and high) and pressures (low and high). He has combined this information into the table below:

[43] The Dirichlet distribution is in turn the multivariate generalization of the beta distribution.

Outcome	Temperature	Pressure	Probability
1	high	high	0.013
2	high	low	0.267
3	low	high	0.031
4	low	low	0.689

Worried about risk of runaway reactions, the Miles Reactor Company is implementing a new program to assess the safety of their reaction processes. The program consists of running each reaction process 100 times over the next year and recording the reactor conditions during the process every time. In order for the process to be considered safe, the process outcomes must be within the following limits:

Outcome	Temperature	Pressure	Frequency
1	high	high	$n_1 = 0$
2	high	low	$n_2 \leq 20$
3	low	high	$n_3 \leq 2$
4	low	low	$n_4 = 100 - n1 - n2 - n3$

Help Les predict whether or not the new process is safe by answering the following question: "What is the probability that the new process will meet the specifications of the new safety program?" (http://controls.engin.umich.edu/wiki/index.php/Multinomial_distributions). •

Remark 4.213 — **Multinomial expansion** (Johnson and Kotz, 1969, p. 281)
If we recall the multinomial theorem[44] then the expression of $P(\{N_1 = n_1, \ldots, N_d = n_d\})$ can be regarded as the coefficient of $\prod_{i=1}^{d} t_i^{n_i}$ in the multinomial expansion of

$$(t_1\, p_1 + \ldots + t_d\, p_d)^n = \sum_{(n_1,\ldots,n_d)} P(\{N_1 = n_1, \ldots, N_d = n_d\}) \times \prod_{i=1}^{d} t_i^{n_i}, \tag{4.155}$$

where the summation is taken over all $(n_1, \ldots, n_d) \in \{(m_1, \ldots, m_d) \in I\!N_0^k : \sum_{i=1}^{d} m_i = n\}$ and $\underline{N} \sim \text{Multinomial}_{d-1}(n, \underline{p} = (p_1, \ldots, p_d))$. •

[44] For any positive integer d and any nonnegative integer n, we have

$$(x_1 + \ldots + x_d)^n = \sum_{(n_1,\ldots,n_d)} \frac{n!}{\prod_{i=1}^{d} n_i!} \times \prod_{i=1}^{d} x_i^{n_i}, \tag{4.154}$$

where the summation is taken over all $d-$vectors of nonnegative integer indices $n_1, \ldots, n_d$ such that the sum of all n_i is n. As with the binomial theorem, quantities of the form 0^0 which appear are taken to be equal 1. See http://en.wikipedia.org/wiki/Multinomial_theorem for more details.

Definition 4.214 — Mixed factorial moments
Let:

- $\underline{X} = (X_1, \ldots, X_d)$ be a random $d-$vector;
- $r_1, \ldots, r_d \in \mathbb{N}_0$.

Then the mixed factorial moment of order $(r_1, \ldots, r_d)$ of $\underline{X}$ is equal to

$$\begin{aligned} & E\left[X_1^{(r_1)} \times \ldots \times X_d^{(r_d)}\right] \\ & = E\left\{[X_1(X_1-1)\ldots(X_1-r_1+1)] \times \ldots \times [X_d(X_d-1)\ldots(X_d-r_d+1)]\right\}. \end{aligned} \tag{4.156}$$

•

Marginal moments and marginal central moments, and covariances and correlations between the components of a random vector can be written in terms of mixed (central/factorial) moments. This is particularly useful when we are dealing with the multinomial distribution.

Exercise 4.215 — Writing the variance and covariance in terms of mixed (central/factorial) moments
Write

(a) the marginal variance of X_i and

(b) $cov(X_i, X_j)$

in terms of mixed factorial moments. •

Proposition 4.216 — Mixed factorial moments of a multinomial distribution (Johnson and Kotz, 1969, p. 284)
Let:

- $\underline{N} = (N_1, \ldots, N_d) \sim \text{Multinomial}_{d-1}(n, \underline{p} = (p_1, \ldots, p_d))$;
- $r_1, \ldots, r_d \in \mathbb{N}_0$.

Then the mixed factorial moment of order $(r_1, \ldots, r_d)$ is equal to

$$E\left[N_1^{(r_1)} \times \ldots \times N_d^{(r_d)}\right] = n^{(\sum_{i=1}^d r_i)} \times \prod_{i=1}^d p_i^{r_i}, \tag{4.157}$$

where $n^{(\sum_{i=1}^d r_i)} = \frac{n!}{(n-\sum_{i=1}^d r_i))!}$. •

From the general formula (4.157), we find the expected value of N_i, and the covariances and correlations between N_i and N_j.

Corollary 4.217 — Mean vector, covariance and correlation matrix of a multinomial distribution (Johnson and Kotz, 1969, p. 284; http://en.wikipedia.org/wiki/Multinomial_distribution)
The expected number of times the event E_i was observed over n trials, N_i, is

$$E(N_i) = n\,p_i, \tag{4.158}$$

for $i = 1, \ldots, d$.

The covariance matrix is as follows. Each diagonal entry is the variance

$$V(N_i) = n\,p_i\,(1 - p_i), \tag{4.159}$$

for $i = 1, \ldots, d$. The off-diagonal entries are the covariances

$$cov(N_i, N_j) = -n\,p_i\,p_j, \tag{4.160}$$

for $i, j = 1, \ldots, d$, $i \neq j$. All covariances are negative because, for fixed n, an increase in one component of a multinomial vector requires a decrease in another component. The covariance matrix is a $d \times d$ positive-semidefinite matrix of rank $d - 1$.

The off-diagonal entries of the corresponding correlation matrix are given by

$$corr(N_i, N_j) = -\sqrt{\frac{p_i\,p_j}{(1 - p_i)\,(1 - p_j)}}, \tag{4.161}$$

for $i, j = 1, \ldots, d$, $i \neq j$.[45] Note that the number of trials (n) drops out of the expression of $corr(N_i, N_j)$. •

Exercise 4.218 — Mean vector, covariance and correlation matrices of a multinomial distribution
Use (4.157) to derive the entries of the mean vector, and the covariance and correlation matrices of a multinomial distribution. •

Exercise 4.219 — Mean vector and correlation matrix of a multinomial distribution
Resume Exercise 4.211 and calculate the mean vector and the correlation matrix. Comment the values you have obtained for the off-diagonal entries of the correlation matrix. •

[45] The diagonal entries of the correlation matrix are obviously equal to 1.

Proposition 4.220 — Marginal distributions in a multinomial setting (Johnson and Kotz, 1969, p. 281)
The marginal distribution of any N_i, $i = 1, \ldots, d$, is Binomial with parameters n and p_i. I.e.

$$N_i \sim \text{Binomial}(n, p_i). \tag{4.162}$$

for $i = 1, \ldots, d$. •

(4.162) is a special case of a more general result.

Proposition 4.221 — Joint distribution of a subset of r.v. from a multinomial distribution (Johnson and Kotz, 1969, p. 281)
The joint distribution of any subset of s ($s = 1, \ldots, d-1$) r.v., say $N_{a_1}, \ldots, N_{a_s}$ of the N_j's, is also multinomial with an $(s+1)^{th}$ r.v. equal to $N_{a_{s+1}} = n - \sum_{i=1}^{s} N_{a_i}$. In fact

$$\begin{aligned} &P\left(\left\{N_{a_1} = n_{a_1}, \ldots, N_{a_s} = n_{a_s}, N_{a_{s+1}} = n - \textstyle\sum_{i=1}^{s} n_{a_i}\right\}\right) \\ &= \frac{n!}{\prod_{i=1}^{s} n_{a_i}! \times (n - \sum_{i=1}^{s} n_{a_i})!} \times \prod_{i=1}^{s} \left[p_{a_i}^{n_{a_i}} \times \left(1 - \textstyle\sum_{j=1}^{s} p_{a_j}\right)^{n - \sum_{j=1}^{s} n_{a_j}} \right], \end{aligned} \tag{4.163}$$

for $n_{a_i} \in I\!N_0$, $i = 1, \ldots, s$ such that $\sum_{i=1}^{s} n_{a_i} \leq n$. •

Proposition 4.222 — Some regressions and conditional distributions in the multinomial distribution setting (Johnson and Kotz, 1969, p. 284)

- The regression of N_i on N_j ($j \neq i$) is linear:

$$E(N_i | N_j) = (n - N_j) \times \frac{p_i}{1 - p_j}. \tag{4.164}$$

- The multiple regression of N_i on $N_{b_1}, \ldots, N_{b_r}$ ($b_j \neq i$, $j = 1, \ldots, r$) is also linear:

$$E(N_i | \{N_{b_1}, \ldots, N_{b_r}\}) = \left(n - \sum_{j=1}^{r} N_{b_j}\right) \times \frac{p_i}{1 - \sum_{j=1}^{r} p_{b_j}}. \tag{4.165}$$

- The random vector $(N_{a_1}, \ldots N_{a_s})$ conditional on a event referring to any subset of the remaining N_j's, say $\{N_{b_1} = n_{b_1}, \ldots, N_{b_r} = n_{b_r}\}$, has also a multinomial distribution. Its p.f. can be found in Johnson and Kotz (1969, p. 284). •

Remark 4.223 — Conditional distributions and the simulation of a multinomial distribution (Gentle, 1998, p. 106)
The following conditional distributions taken from Gentle (1998, p. 106) suggest a procedure to generate pseudo-random vectors with a multinomial distribution:

- $N_1 \sim \text{Binomial}(n, p_1)$;
- $N_2|\{N_1 = n_1\} \sim \text{Binomial}\left(n - n_1, \frac{p_2}{1-p_1}\right)$;
- $N_3|\{N_1 = n_1, N_2 = n_2\} \sim \text{Binomial}\left(n - n_1 - n_2, \frac{p_3}{1-p_1-p_2}\right)$;
- ...
- $N_{d-1}|\{N_1 = n_1, \ldots, N_{d-2} = n_{d-2}\} \sim \text{Binomial}\left(n - \sum_{i=1}^{d-2} n_i, \ \frac{p_{d-1}}{1-\sum_{i=1}^{d-2} p_i}\right)$;
- $N_d|\{N_1 = n_1, \ldots, N_{d-1} = n_{d-1}\} \stackrel{d}{=} n - \sum_{i=1}^{d-1} n_i$.

Thus, we can generate a pseudo-random vector from a multinomial distribution by sequentially generating independent pseudo-random numbers from the binomial conditional distributions stated above.

Gentle (1998, p. 106) refers other ways of generating pseudo-random vectors from a multinomial distribution. •

Remark 4.224 — Speeding up the simulation of a multinomial distribution (Gentle, 1998, p. 106)
To speed up the generation process, Gentle (1998, p. 106) recommends that we previously order the probabilities $p_1, \ldots, p_d$ in descending order — thus, getting the vector of probabilities $(p_{(d)}, \ldots, p_{(1)})$, where $p_{(d)} = \max_{i=1,\ldots,d} p_i$, ..., and $p_{(1)} = \min_{i=1,\ldots,d} p_i$. Then we generate d pseudo-random numbers with the following binomial distributions with parameters

1. n and the largest probability of "success" $p_{(d)}$, say $n_{(d)}$,
2. $n - n_{(d)}$ and $\frac{p_{(d-1)}}{1-p_{(d-1)}}$, say $n_{(d-1)}$,
3. $n - n_{(d)} - n_{(d-1)}$ and $\frac{p_{(d-2)}}{1-p_{(d-1)}-p_{(d-2)}}$, say $n_{(d-1)}$,
4. ...
5. $n - \sum_{i=1}^{d-2} n_{(d+1-i)}$ and $\frac{p_{(2)}}{1-\sum_{i=1}^{d-2} p_{(d+1-i)}}$, say $n_{(2)}$,

and finally

6. assign $n_{(1)} = n - \sum_{i=1}^{d-1} n_{(d+1-i)}$. •

Remark 4.225 — Speeding up the simulation of a multinomial distribution (http://en.wikipedia.org/wiki/Multinomial_distribution)
Assume the parameters $p_1, \dots p_d$ are already sorted descendingly (this is only to speed up computation and not strictly necessary). Now, for each trial, generate a pseudo-random number from $U \sim \text{Uniform}(0,1)$, say u. The resulting outcome is the event E_j where

$$\begin{aligned} j &= \arg \min_{j'=1,\dots,k} \left(\sum_{i=1}^{j'} p_i \geq u \right) \\ &= F_Z^{-1}(u), \end{aligned} \tag{4.166}$$

with Z an integer r.v. that takes values $1, \dots, d$, with probabilities $p_1, \dots p_d$, respectively. This is a sample for the multinomial distribution with $n = 1$.

The absolute frequencies of events $E_1, \dots, E_d$, resulting from n independent repetitions of the procedure we just described, constitutes a pseudo-random vector from a multinomial distribution with parameters n and $\underline{p} = (p_1, \dots p_d)$. •

References

- Gentle, J.E. (1998). *Random Number Generation and Monte Carlo Methods.* Springer-Verlag. (QA298.GEN.50103)
- Johnson, N.L. and Kotz, S. (1969). *Discrete distributions* John Wiley & Sons. (QA273-280/1.JOH.36178)
- Karr, A.F. (1993). *Probability.* Springer-Verlag.
- Resnick, S.I. (1999). *A Probability Path.* Birkhäuser. (QA273.4-.67.RES.49925)
- Rohatgi, V.K. (1976). *An Introduction to Probability Theory and Mathematical Statistics.* John Wiley & Sons. (QA273-280/4.ROH.34909)
- Tong, Y.L. (1990). *The Multivariate Normal Distribution.* Springer-Verlag. (QA278.5-.65.TON.39685)
- Walrand, J. (2004). *Lecture Notes on Probability Theory and Random Processes.* Department of Electrical Engineering and Computer Sciences, University of California, Berkeley.

Chapter 5

Convergence concepts and classical limit theorems

Throughout this chapter we assume that $\{X_1, X_2, \ldots\}$ is a sequence of r.v. and X is a r.v., and all of them are defined on the same probability space $(\Omega, \mathcal{F}, P)$.

Stochastic convergence formalizes the idea that a sequence of r.v. sometimes is expected to settle into a pattern.[1] The pattern may for instance be that:

- there is a convergence of $X_n(\omega)$ in the classical sense to a fixed value $X(\omega)$, for each and every event ω;
- the probability that the distance between X_n from a particular r.v. X exceeds any prescribed positive value decreases and converges to zero;
- the sequence formed by calculating the expected value of the (absolute or quadratic) distance between X_n and X converges to zero;
- the distribution of X_n may "grow" increasingly similar to the distribution of a particular r.v. X.

Just as in real analysis, we can distinguish among several types of convergence (Rohatgi, 1976, p. 240). Thus, in this chapter we investigate modes of convergence of sequences of r.v.:

- almost sure convergence ($\overset{a.s.}{\rightarrow}$);
- convergence in probability ($\overset{P}{\rightarrow}$);

[1] See http://en.wikipedia.org/wiki/Convergence_of_random_variables.

- convergence in quadratic mean or in L^2 ($\xrightarrow{q.m.}$);
- convergence in L^1 or in mean ($\xrightarrow{L^1}$);
- convergence in distribution ($\xrightarrow{d}$).

It is important for the reader to be familiarized with all these modes of convergence, the way they can be related and with the applications of such results and understand their considerable significance in probability, statistics and stochastic processes.

5.1 Modes of convergence

The first four modes of convergence ($\xrightarrow{*}$, where $* = a.s., P, q.m., L^1$) pertain to the sequence of r.v. and to X as functions of Ω, while the fifth ($\xrightarrow{d}$) is related to the convergence of d.f. (Karr, 1993, p. 135).

5.1.1 Convergence of r.v. as functions on Ω

Motivation 5.1 — Almost sure convergence (Karr, 1993, p. 135)
Almost sure convergence — or convergence with probability one — is the probabilistic version of pointwise convergence known from elementary real analysis. •

Definition 5.2 — Almost sure convergence (Karr, 1993, p. 135; Rohatgi, 1976, p. 249)
The sequence of r.v. $\{X_1, X_2, \ldots\}$ is said to converge almost surely to a r.v. X if

$$P\left(\left\{w : \lim_{n\to+\infty} X_n(\omega) = X(\omega)\right\}\right) = 1. \tag{5.1}$$

In this case we write $X_n \xrightarrow{a.s.} X$ (or $X_n \to X$ with probability 1). •

Remark 5.3 — Almost sure convergence
Equation (5.1) does not mean that $\lim_{n\to+\infty} P(\{w : X_n(\omega) = X(\omega)\}) = 1$. •

Exercise 5.4 — Almost sure convergence
Let $\{X_1, X_2, \ldots\}$ be a sequence of independent r.v. such that $X_n \sim \text{Bernoulli}(\frac{1}{n})$, $n \in \mathbb{N}$.

Prove that $X_n \not\xrightarrow{a.s.} 0$, by deriving $P(\{X_n = 0, \text{for every } m \leq n \leq n_0\})$ and observing that this probability does not converge to 1 as $n_0 \to +\infty$ for all values of m (Rohatgi, 1976, p. 252, Example 9). •

Motivation 5.5 — Convergence in probability (Karr, 1993, p. 135; http://en.wikipedia.org/wiki/Convergence_of_random_variables)
Convergence in probability essentially means that the probability that $|X_n - X|$ exceeds any prescribed, strictly positive value converges to zero.

The basic idea behind this type of convergence is that the probability of an "unusual" outcome becomes smaller and smaller as the sequence progresses. •

Definition 5.6 — Convergence in probability (Karr, 1993, p. 136; Rohatgi, 1976, p. 243)
The sequence of r.v. $\{X_1, X_2, \ldots\}$ is said to converge in probability to a r.v. X — denoted by $X_n \xrightarrow{P} X$ — if

$$\lim_{n \to +\infty} P\left(\{|X_n - X| > \epsilon\}\right) = 0, \tag{5.2}$$

for every $\epsilon > 0$. •

Remarks 5.7 — Convergence in probability (Rohatgi, 1976, p. 243; http://en.wikipedia.org/wiki/Convergence_of_random_variables)

- The definition of convergence in probability says nothing about the convergence of r.v. X_n to r.v. X in the sense in which it is understood in real analysis. Thus, $X_n \xrightarrow{P} X$ does not imply that, given $\epsilon > 0$, we can find an N such that $|X_n - X| < \epsilon$, for $n \geq N$.

  Definition 5.6 speaks only of the convergence of the sequence of probabilities $P\left(|X_n - X| > \epsilon\right)$ to zero.

- Formally, Definition 5.6 means that

$$\forall \epsilon, \delta > 0,\ \exists N_\delta : P(\{|X_n - X| > \epsilon\}) < \delta,\ \forall n \geq N_\delta. \tag{5.3}$$

- The concept of convergence in probability is used very often in statistics. For example, an estimator is called consistent if it converges in probability to the parameter being estimated.

- Convergence in probability is also the type of convergence established by the weak law of large numbers. •

Example 5.8 — Convergence in probability
Let $\{X_1, X_2, \ldots\}$ be a sequence of i.i.d. r.v. such that Uniform$(0, \theta)$, where $\theta > 0$.

(a) Check if $X_{(n)} = \max_{i=1,\ldots,n} X_i \xrightarrow{P} \theta$.

- **R.v.**
  $X_i \overset{i.i.d.}{\sim} X$, $i \in I\!N$
  $X \sim$ Uniform$(0, \theta)$
- **D.f. of X**
  $$F_X(x) = \begin{cases} 0, & x < 0 \\ \frac{x}{\theta}, & 0 \leq x \leq \theta \\ 1, & x > \theta \end{cases}$$
- **New r.v.**
  $X_{(n)} = \max_{i=1,\ldots,n} X_i$
- **D.f. of $X_{(n)}$**
  $$\begin{aligned} F_{X_{(n)}}(x) &= [F_X(x)]^n \\ &= \begin{cases} 0, & x < 0 \\ \left(\frac{x}{\theta}\right)^n, & 0 \leq x \leq \theta \\ 1, & x > \theta \end{cases} \end{aligned}$$
- **Checking the convergence in probability $X_{(n)} \xrightarrow{P} \theta$**
  Making use of the definition of this type of convergence and capitalizing on the d.f. of $X_{(n)}$, we get, for every $\epsilon > 0$:
  $$\begin{aligned} \lim_{n\to+\infty} P\left(|X_{(n)} - \theta| > \epsilon\right) &= 1 - \lim_{n\to+\infty} P\left(\theta - \epsilon \leq X_{(n)} \leq \theta + \epsilon\right) \\ &= 1 - \lim_{n\to+\infty} \left[F_{X_{(n)}}(\theta + \epsilon) - F_{X_{(n)}}(\theta - \epsilon)\right] \\ &= \begin{cases} 1 - \lim_{n\to+\infty} \left[F_{X_{(n)}}(\theta) - F_{X_{(n)}}(\theta - \epsilon)\right], \\ \qquad\qquad 0 < \epsilon < \theta \\ 1 - \lim_{n\to+\infty} F_{X_{(n)}}(\theta), \ \epsilon \geq \theta \end{cases} \\ &= \begin{cases} 1 - \lim_{n\to+\infty} \left[1 - \left(\frac{\theta-\epsilon}{\theta}\right)^n\right] \\ = 1 - (1 - 0), \ 0 < \epsilon < \theta \\ 1 - 1, \ \epsilon \geq \theta \end{cases} \\ &= 0. \end{aligned}$$

- **Conclusion**

  $X_{(n)} \xrightarrow{P} \theta$.

  Interestingly enough, $X_{(n)}$ is the ML estimator of θ and also a consistent estimator of θ ($X_{(n)} \xrightarrow{P} \theta$). However, $E[X_{(n)}] = n\theta/(n+1) \neq \theta$, i.e. $X_{(n)}$ is a biased estimator of θ.

(b) Prove that $X_{(1:n)} = \min_{i=1,\ldots,n} X_i \xrightarrow{P} 0$.

- **New r.v.**

  $X_{(1:n)} = \min_{i=1,\ldots,n} X_i$

- **D.f. of $X_{(1:n)}$**

$$\begin{aligned} F_{X_{(1:n)}}(x) &= 1 - [1 - F_X(x)]^n \\ &= \begin{cases} 0, & x < 0 \\ 1 - \left(1 - \frac{x}{\theta}\right)^n, & 0 \leq x \leq \theta \\ 1, & x > \theta \end{cases} \end{aligned}$$

- **Checking the convergence in probability $X_{(1:n)} \xrightarrow{P} 0$**

  For every $\epsilon > 0$, we have

$$\begin{aligned} \lim_{n\to+\infty} P\left(|X_{(1:n)} - 0| > \epsilon\right) &= 1 - \lim_{n\to+\infty} \left[F_{X_{(1:n)}}(\epsilon) - F_{X_{(1:n)}}(-\epsilon)\right] \\ &= 1 - \lim_{n\to+\infty} F_{X_{(1:n)}}(\epsilon) \\ &= \begin{cases} 1 - \lim_{n\to+\infty} \left[1 - \left(1 - \frac{\epsilon}{\theta}\right)^n\right] \\ = 1 - (1 - 0),\ 0 < \epsilon < \theta \\ 1 - \lim_{n\to+\infty} F_{X_{(1:n)}}(\theta) = 1 - 1,\ \epsilon \geq \theta \end{cases} \\ &= 0. \end{aligned}$$

- **Conclusion**

  $X_{(1:n)} \xrightarrow{P} 0$. •

Remark 5.9 — Chebyshev(-Bienaymé)'s inequality and convergence in probability

Chebyshev(-Bienaymé)'s inequality can be useful to prove that some sequences of r.v. converge in probability to a degenerate r.v. (i.e., a constant). •

Example 5.10 — Chebyshev(-Bienaymé)'s inequality and convergence in probability

Let $\{X_1, X_2, \ldots\}$ be a sequence of r.v. such that $X_n \sim \text{Gamma}(n, n)$, $n \in \mathbb{N}$. Prove that $X_n \xrightarrow{P} 1$, by making use of Chebyshev(-Bienaymé)'s inequality.

- **R.v.**

  $X_n \sim \text{Gamma}(n, n)$, $n \in \mathbb{N}$

  $E(X_n) = \frac{n}{n} = 1$

  $V(X_n) = \frac{n}{n^2} = \frac{1}{n}$

- **Checking the convergence in probability** $X_n \xrightarrow{P} 1$

  The application of the definition of this type of convergence and Chebyshev(-Bienaymé)'s inequality leads to

$$\begin{aligned}
\lim_{n\to+\infty} P\left(|X_n - 1| > \epsilon\right) &= \lim_{n\to+\infty} P\left(|X_n - E(X_n)| \geq \frac{\epsilon}{\sqrt{V(X_n)}}\sqrt{V(X_n)}\right) \\
&\leq \lim_{n\to+\infty} \frac{1}{\left(\frac{\epsilon}{\sqrt{\frac{1}{n}}}\right)^2} \\
&= \frac{1}{\epsilon^2} \lim_{n\to+\infty} \frac{1}{n} \\
&= 0,
\end{aligned}$$

  for every $\epsilon > 0$.

- **Conclusion**

  $X_n \xrightarrow{P} 1$. •

Exercise 5.11 — Chebyshev(-Bienaymé)'s inequality and convergence in probability

Prove that $X_{(n)} = \max_{i=1,\ldots,n} X_i$, where $X_i \sim_{i.i.d.} \text{Uniform}(0, \theta)$, is a consistent estimator of $\theta > 0$, by using Chebyshev(-Bienaymé)'s inequality and the fact that $E[X_{(n)}] = \frac{n}{n+1}\theta$ and $V[X_{(n)}] = \frac{n}{(n+2)(n+1)^2}\theta^2$. •

Exercise 5.12 — Convergence in probability
Let $\{X_1, X_2, \ldots\}$ be a sequence of r.v. such that $X_n \sim \text{Bernoulli}(\frac{1}{n})$, $n \in \mathbb{N}$.

(a) Show that $X_n \xrightarrow{P} 0$, by obtaining $P(\{|X_n| > \epsilon\})$, for $0 < \epsilon < 1$ and $\epsilon \geq 1$ (Rohatgi, 1976, pp. 243–244, Example 5).

(b) Verify that $E(X_n^k) \to E(X^k)$, where $k \in \mathbb{N}$ and $X \stackrel{d}{=} 0$. •

Exercise 5.13 — Convergence in probability does not imply convergence of kth. moments
Let $\{X_1, X_2, \ldots\}$ be a sequence of r.v. such that $X_n \stackrel{d}{=} n \times \text{Bernoulli}(\frac{1}{n})$, $n \in \mathbb{N}$, i.e.

$$P(\{X_n = x\}) = \begin{cases} 1 - \frac{1}{n}, & x = 0 \\ \frac{1}{n}, & x = n \\ 0, & \text{otherwise.} \end{cases} \tag{5.4}$$

Prove that $X_n \xrightarrow{P} 0$, however $E(X_n^k) \not\to E(X^k)$, where $k \in \mathbb{N}$ and the r.v. X is degenerate at 0 (Rohatgi, 1976, p. 247, Remark 3). •

Motivation 5.14 — Convergence in quadratic mean and in L^1
We have just seen that convergence in probability does not imply the convergence of moments, namely of orders 2 or 1. •

Definition 5.15 — Convergence in quadratic mean or in L^2 (Karr, 1993, p. 136)
Let $X, X_1, X_2, \ldots$ belong to L^2. Then the sequence of r.v. $\{X_1, X_2, \ldots\}$ is said to converge to X in quadratic mean (or in L^2) — denoted by $X_n \xrightarrow{q.m.} X$ (or $X_n \xrightarrow{L^2} X$) — if

$$\lim_{n \to +\infty} E\left[(X_n - X)^2\right] = 0. \tag{5.5}$$

•

Exercise 5.16 — Convergence in quadratic mean
Let $\{X_1, X_2, \ldots\}$ be a sequence of r.v. such that $X_n \sim \text{Bernoulli}\left(\frac{1}{n}\right)$.

Prove that $X_n \xrightarrow{q.m.} X$, where the r.v. X is degenerate at 0 (Rohatgi, 1976, p. 247, Example 6). •

Exercise 5.17 — Convergence in quadratic mean (bis)
Let $\{X_1, X_2, \ldots\}$ be a sequence of r.v. with $P\left(\left\{X_n = \pm\frac{1}{n}\right\}\right) = \frac{1}{2}$.

Prove that $X_n \xrightarrow{q.m.} X$, where the r.v. X is degenerate at 0 (Rohatgi, 1976, p. 252, Example 11). •

Exercise 5.18 — **Convergence in quadratic mean implies convergence of 2nd. moments** (Karr, 1993, p. 158, Exercise 5.6(a))
Show that $X_n \overset{q.m.}{\to} X \Rightarrow E(X_n^2) \to E(X^2)$ (Rohatgi, 1976, p. 248, proof of Theorem 8). •

Exercise 5.19 — **Convergence in quadratic mean of partial sums** (Karr, 1993, p. 159, Exercise 5.11)
Let $X_1, X_2, \ldots$ be pairwise uncorrelated r.v. with mean zero and partial sums $S_n = \sum_{i=1}^{n} X_i$.

Prove that if there is a constant c such that $V(X_i) \leq c$, for every i, then $\frac{S_n}{n^\alpha} \overset{q.m.}{\to} 0$ for all $\alpha > \frac{1}{2}$. •

Definition 5.20 — **Convergence in mean or in L^1** (Karr, 1993, p. 136)
Let $X, X_1, X_2, \ldots$ belong to L^1. Then the sequence of r.v. $\{X_1, X_2, \ldots\}$ is said to converge to X in mean (or in L^1) — denoted by $X_n \overset{L^1}{\to} X$ — if

$$\lim_{n \to +\infty} E\left(|X_n - X|\right) = 0. \tag{5.6}$$

•

Exercise 5.21 — **Convergence in mean implies convergence of 1st. moments** (Karr, 1993, p. 158, Exercise 5.6(b))
Prove that $X_n \overset{L^1}{\to} X \Rightarrow E(X_n) \to E(X)$ (Rohatgi, 1976, p. 248, proof Theorem 8). •

5.1.2 Convergence in distribution

Motivation 5.22 — Convergence in distribution
(http://en.wikipedia.org/wiki/Convergence_of_random_variables)
Convergence in distribution is very frequently used in practice, most often it arises from the application of the central limit theorem. •

Definition 5.23 — Convergence in distribution (Karr, 1993, p. 136; Rohatgi, 1976, pp. 240–1)
The sequence of r.v. $\{X_1, X_2, \ldots\}$ converges to X in distribution — denoted by $X_n \xrightarrow{d} X$ — if

$$\lim_{n \to +\infty} F_{X_n}(x) = F_X(x), \tag{5.7}$$

for all x at which F_X is continuous. •

Remarks 5.24 — Convergence in distribution
(http://en.wikipedia.org/wiki/Convergence_of_random_variables; Karr, 1993, p. 136; Rohatgi, 1976, p. 242)

- With this mode of convergence, we increasingly expect to see the next r.v. in a sequence of r.v. becoming better and better modeled by a given d.f., as seen in exercises 5.25 and 5.26.
- It must be noted that it is quite possible for a given sequence of d.f. to converge to a function that is not a d.f., as shown in Example 5.27 and Exercise 5.28.
- The requirement that only the continuity points of F_X should be considered is essential, as we shall see in exercises 5.29 and 5.30.
- The convergence in distribution does not imply the convergence of corresponding p.(d.)f., as shown in Exercise 5.32. Sequences of absolutely continuous r.v. that converge in distribution to discrete r.v. (and vice-versa) are obvious illustrations, as shown in examples 5.31 and 5.33. •

Exercise 5.25 — Convergence in distribution
Let $X_1, X_2, \ldots, X_n$ be i.i.d. r.v. with common p.d.f.

$$f(x) = \begin{cases} \frac{1}{\theta}, & 0 < x < \theta \\ 0, & \text{otherwise,} \end{cases} \tag{5.8}$$

where $0 < \theta < +\infty$, and $X_{(n)} = \max_{1,\ldots,n} X_i$.

Show that $X_{(n)} \xrightarrow{d} \theta$ (Rohatgi, 1976, p. 241, Example 2). •

Exercise 5.26 — Convergence in distribution (bis)
Let:

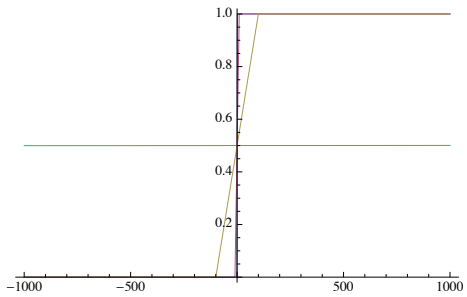
- $\{X_1, X_2, \ldots\}$ be a sequence of r.v. such that $X_n \sim \text{Bernoulli}(p_n)$, $n = 1, 2, \ldots$;
- $X \sim \text{Bernoulli}(p)$.

Prove that $X_n \xrightarrow{d} X$ iff $p_n \to p$. •

Example 5.27 — A sequence of d.f. converging to a non d.f. (Murteira, 1979, pp. 330–331)
Let $\{X_1, X_2, \ldots\}$ be a sequence of r.v. with d.f.

$$F_{X_n}(x) = \begin{cases} 0, & x < -n \\ \frac{x+n}{2n}, & -n \leq x < n \\ 1, & x \geq n. \end{cases} \tag{5.9}$$

Please note that $\lim_{n\to+\infty} F_{X_n}(x) = \frac{1}{2}$, $x \in I\!R$, as suggested by the graph below with some terms of the sequence of d.f., for $n = 1, 10^3, 10^6$ (from top to bottom):



Consequently, the limit of the sequence of d.f. is not itself a d.f. •

Exercise 5.28 — A sequence of d.f. converging to a non d.f.
Consider the sequence of d.f.

$$F_{X_n}(x) = \begin{cases} 0, & x < n \\ 1, & x \geq n, \end{cases} \tag{5.10}$$

where $F_{X_n}(x)$ is the d.f. of the r.v. X_n degenerate at $x = n$.

Verify that $F_{X_n}(x)$ converges to a function (that is identically equal to 0!!!) which is not a d.f. (Rohatgi, 1976, p. 241, Example 1). •

Exercise 5.29 — The requirement that only the continuity points of F_X should be considered is essential
Let $X_n \sim \text{Uniform}\left(\frac{1}{2} - \frac{1}{n}, \frac{1}{2} + \frac{1}{n}\right)$ and X be a r.v. degenerate at $\frac{1}{2}$.

(a) Prove that $X_n \xrightarrow{d} X$ (Karr, 1993, p. 142).

(b) Verify that $F_{X_n}\left(\frac{1}{2}\right) = \frac{1}{2}$ for each n, and these values do not converge to $F_X\left(\frac{1}{2}\right) = 1$. Is there any contradiction with the convergence in distribution previously proved? (Karr, 1993, p. 142.) •

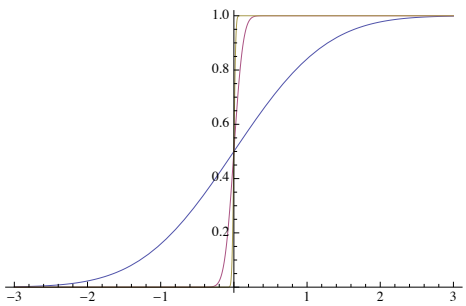
Exercise 5.30 — The requirement that only the continuity points of F_X should be considered is essential (bis)
Let $X_n \sim \text{Uniform}\left(0, \frac{1}{n}\right)$ and X a r.v. degenerate at 0.

Prove that $X_n \xrightarrow{d} X$, even though $F_{X_n}(0) = 0$, for all n, and $F_X(0) = 1$, that is, the convergence of d.f. fails at the point $x = 0$ where F_X is discontinuous (http://en.wikipedia.org/wiki/Convergence_of_random_variables). •

Example 5.31 — Convergence in distribution does not imply convergence of corresponding p.(d.)f. (Murteira, 1979, p. 331)
Let $\{X_1, X_2, \ldots\}$ be a sequence of r.v. such that $X_n \sim \text{Normal}\left(0, \frac{1}{n^2}\right)$.

An analysis of the representation of some terms of the sequence of d.f. (e.g. $n = 1, 10, 50$, from left to right in the following graph) and the notion of convergence in distribution leads us to conclude that $X_n \xrightarrow{d} X$, where $X \overset{d}{=} 0$, even though

$$\lim_{n \to +\infty} F_{X_n}(0) = \lim_{n \to +\infty} \Phi\left(\frac{0-0}{\sqrt{\frac{1}{n^2}}}\right) = \Phi(0) = \frac{1}{2}$$



and

$$\lim_{n\to+\infty} F_{X_n}(x) = \begin{cases} 0, & x < 0 \\ \frac{1}{2}, & x = 0 \\ 1, & x > 0 \end{cases}$$

is not a a d.f. (it is not left- or right-continuous).

Note that $X \overset{d}{=} 0$, therefore the d.f. of the limit of the sequence of r.v. $\{X_1, X_2, \ldots\}$ is the Heaviside function, i.e. $F_X(x) = I_{[0,+\infty)}(x)$. •

Exercise 5.32 — Convergence in distribution does not imply convergence of corresponding p.(d.)f.
Let $\{X_1, X_2, \ldots\}$ be a sequence of r.v. with p.f.

$$P(\{X_n = x\}) = \begin{cases} 1, & x = 2 + \frac{1}{n} \\ 0, & \text{otherwise.} \end{cases} \tag{5.11}$$

(a) Prove that $X_n \xrightarrow{d} X$, where X a r.v. degenerate at 2.

(b) Verify that none of the p.f. $P(\{X_n = x\})$ assigns any probability to the point $x = 2$, for all n, and that $P(\{X_n = x\}) \to 0$ for all x (Rohatgi, 1976, p. 242, Example 4). •

Example 5.33 — A sequence of discrete r.v. that converges in distribution to an absolutely continuous r.v. (Rohatgi, 1976, p. 256, Exercise 10)
Let:

- $\{X_1, X_2, \ldots\}$ be a sequence of r.v. such that $X_n \sim \text{Geometric}\left(\frac{\lambda}{n}\right)$, where $n > \lambda > 0$;
- $\{Y_n,\ n \in I\!N\}$ a sequence of r.v. such that $Y_n = \frac{X_n}{n}$.

Show that $Y_n \xrightarrow{d} \text{Exponential}\,(\lambda)$.

- **R.v.**

  $X_n \sim \text{Geometric}\left(\frac{\lambda}{n}\right),\ n \in I\!N$

- **P.f. of X_n and Y_n**

  $P(X_n = x) = \left(1 - \frac{\lambda}{n}\right)^{x-1} \times \frac{\lambda}{n},\ x = 1, 2, \ldots$

  $P(Y_n = y) = P(X_n = ny) = \left(1 - \frac{\lambda}{n}\right)^{ny-1} \times \frac{\lambda}{n},\ y = \frac{1}{n}, \frac{2}{n}, \ldots$

- **D.f. of Y_n**

$$\begin{aligned} F_{Y_n}(y) &= F_{X_n}(ny) \\ &= \begin{cases} 0, & y < \frac{1}{n} \\ \sum_{x=1}^{[ny]} P(X_n = x), & y \geq \frac{1}{n}, \end{cases} \end{aligned}$$

where $[ny]$ represents the integer part of the real number ny and

$$\begin{aligned} \sum_{x=1}^{[ny]} P(X_n = x) &= \sum_{x=0}^{[ny]-1} \left(1 - \frac{\lambda}{n}\right)^x \times \frac{\lambda}{n} \\ &= 1 - \left(1 - \frac{\lambda}{n}\right)^{[ny]}. \end{aligned}$$

- **Checking the convergence in distribution**

Let us remind the reader that $[ny] = ny - \epsilon$, for some $\epsilon \in [0, 1)$. Thus:

$$\begin{aligned} \lim_{n \to +\infty} F_{Y_n}(y) &= 1 - \lim_{n \to +\infty} \left(1 - \frac{\lambda}{n}\right)^{[ny]} \\ &= 1 - \lim_{n \to +\infty} \left(1 - \frac{\lambda}{n}\right)^{ny} \times \lim_{n \to +\infty} \left(1 - \frac{\lambda}{n}\right)^{-\epsilon} \\ &= 1 - \left[\lim_{n \to +\infty} \left(1 - \frac{\lambda}{n}\right)^{n}\right]^{y} \times 1 \\ &= 1 - e^{-\lambda y} \\ &= F_{Exponential(\lambda)}(y). \end{aligned}$$

- **Conclusion**

$Y_n \xrightarrow{d}$ Exponential(λ). •

Exercise 5.34 — A sequence of discrete r.v. that converges in distribution to an absolutely continuous r.v. (bis)

Let $\{X_1, X_2, \ldots\}$ be a sequence of discrete r.v. such that $X_n \sim \text{Uniform}\{0, 1, \ldots, n\}$.

Prove that $Y_n = \frac{X_n}{n} \xrightarrow{d}$ Uniform$(0, 1)$.[2] •

[2] This result is very important in the generation of pseudo-random numbers from the Uniform$(0, 1)$ distribution by using computers since these machines "deal" with discrete mathematics.

The following table condenses the definitions of convergence of sequences of r.v.

Mode of convergence	**Assumption**	**Defining condition**
$X_n \overset{a.s.}{\to} X$ (almost sure)	—	$P(\{w : X_n(\omega) \to X(\omega)\}) = 1$
$X_n \overset{P}{\to} X$ (in probability)	—	$P(\{\lvert X_n - X\rvert > \epsilon\}) \to 0$, for all $\epsilon > 0$
$X_n \overset{q.m}{\to} X$ (in quadratic mean)	$X, X_1, X_2, \ldots \in L^2$	$E\left[(X_n - X)^2\right] \to 0$
$X_n \overset{L^1}{\to} X$ (in L^1)	$X, X_1, X_2, \ldots \in L^1$	$E(\lvert X_n - X\rvert) \to 0$
$X_n \overset{d}{\to} X$ (in distribution)	—	$F_{X_n}(x) \to F_X(x)$, at continuity points x of F_X

Exercise 5.35 — **Modes of convergence and uniqueness of limit** (Karr, 1993, p. 158, Exercise 5.1)
Prove that for all five forms of convergence the limit is unique. In particular:

(a) if $X_n \overset{*}{\to} X$ and $X_n \overset{*}{\to} Y$, where $* = a.s., P, q.m., L^1$, then $X \overset{a.s.}{=} Y$;

(b) if $X_n \overset{d}{\to} X$ and $X_n \overset{d}{\to} Y$, then $X \overset{d}{=} Y$; •

Exercise 5.36 — **Modes of convergence and the vector space structure of the family of r.v.** (Karr, 1993, p. 158, Exercise 5.2)
Prove that, for $* = a.s., P, q.m., L^1$,

$$X_n \overset{*}{\to} X \Leftrightarrow X_n - X \overset{*}{\to} 0, \tag{5.12}$$

i.e. the four function-based forms of convergence are compatible with the vector space structure of the family of r.v. •

5.1.3 Alternative criteria

The definition of almost sure convergence and its verification are far from trivial. More tractable criteria have to be stated...

Proposition 5.37 — Relating almost sure convergence and convergence in probability (Karr, 1993, p. 137; Rohatgi, 1976, p. 249)
$X_n \overset{a.s.}{\to} X$ iff

$$\forall \epsilon > 0, \lim_{n \to +\infty} P\left(\left\{ \sup_{k \geq n} |X_k - X| > \epsilon \right\} \right) = 0, \tag{5.13}$$

i.e.

$$X_n \overset{a.s.}{\to} X \quad \Leftrightarrow \quad Y_n = \sup_{k \geq n} |X_k - X| \overset{P}{\to} 0. \tag{5.14}$$

•

Remarks 5.38 — Relating almost sure convergence and convergence in probability (Karr, 1993, p. 137; Rohatgi, 1976, p. 250, Remark 6)

- Proposition 5.37 states an equivalent form of almost sure convergence that illuminates its relationship to convergence in probability.
- $X_n \overset{a.s.}{\to} 0$ means that,

$$\forall \epsilon, \eta > 0, \exists n_0 \in I\!N : P\left(\left\{ \sup_{k \geq n_0} |X_k| > \epsilon \right\} \right) < \eta. \tag{5.15}$$

  Indeed, we can write, equivalently, that

$$\lim_{n \to +\infty} P\left(\bigcup_{k \geq n} \{|X_k| > \epsilon\} \right) = 0, \tag{5.16}$$

  for $\epsilon > 0$ arbitrary. •

Exercise 5.39 — Relating almost sure convergence and convergence in probability
Prove Proposition 5.37 (Karr, 1993, p. 137; Rohatgi, 1976, p. 250). •

Exercise 5.40 — Relating almost sure convergence and convergence in probability (bis)
Let $\{X_1, X_2, \ldots\}$ be a sequence of r.v. with $P\left(\left\{X_n = \pm\frac{1}{n}\right\}\right) = \frac{1}{2}$.

Prove that $X_n \overset{a.s.}{\to} X$, where the r.v. X is degenerate at 0, by using (5.16) (Rohatgi, 1976, p. 252). •

Theorem 5.41 — Cauchy criterion (Rohatgi, 1976, p. 270)

$$X_n \overset{a.s.}{\to} X \quad \Leftrightarrow \quad \lim_{n\to+\infty} P\left(\left\{\sup_m |X_{n+m} - X_n| \leq \epsilon\right\}\right) = 1, \forall\epsilon > 0. \tag{5.17}$$

•

Exercise 5.42 — Cauchy criterion
Prove Theorem 5.41 (Rohatgi, 1976, pp. 270–2). •

Definition 5.43 — Complete convergence (Karr, 1993, p. 138)
The sequence of r.v. $\{X_1, X_2, \ldots\}$ is said to converge completely to X if

$$\sum_{n=1}^{+\infty} P\left(\{|X_n - X| > \epsilon\}\right) < +\infty, \tag{5.18}$$

for every $\epsilon > 0$. •

The next results relate complete convergence, which is stronger than almost sure convergence, and sometimes more convenient to establish (Karr, 1993, p. 137).

Proposition 5.44 — Relating almost sure convergence and complete convergence (Karr, 1993, p. 138)

$$\sum_{n=1}^{+\infty} P\left(\{|X_n - X| > \epsilon\}\right) < +\infty, \forall\epsilon > 0 \quad \Rightarrow \quad X_n \overset{a.s.}{\to} X. \tag{5.19}$$

•

Remark 5.45 — Relating almost sure convergence and complete convergence (Karr, 1993, p. 138)
$X_n \overset{P}{\to} X$ iff the probabilities $P\left(\{|X_n - X| > \epsilon\}\right)$ converge to zero, while $X_n \overset{a.s.}{\to} X$ if (but not only if) the convergence of probabilities $P\left(\{|X_n - X| > \epsilon\}\right)$ is fast enough that their sum, $\sum_{n=1}^{+\infty} P\left(\{|X_n - X| > \epsilon\}\right)$, is finite. •

Exercise 5.46 — **Relating almost sure convergence and complete convergence**
Show Proposition 5.44, by using the (1st.) Borel-Cantelli lemma (Karr, 1993, p. 138). •

Theorem 5.47 — **Almost sure convergence of a sequence of independent r.v.** (Rohatgi, 1976, p. 265)
Let $\{X_1, X_2, \ldots\}$ be a sequence of independent r.v. Then

$$X_n \overset{a.s.}{\to} 0 \quad \Leftrightarrow \quad \sum_{n=1}^{+\infty} P\left(\{|X_n| > \epsilon\}\right) < +\infty, \forall \epsilon > 0. \tag{5.20}$$

•

Exercise 5.48 — **Almost sure convergence of a sequence of independent r.v.**

(a) Prove Theorem 5.47 (Rohatgi, 1976, pp. 265–6).

(b) Use Theorem 5.47 to solve Exercise 5.40. •

The definition of convergence in distribution is cumbersome because of the proviso regarding continuity points of the limit d.f. F_X. An alternative criterion follows.

Theorem 5.49 — **Alternative criterion for convergence in distribution** (Karr, 1993, p. 138)
Let $\mathbf{C}$ be the set of bounded, continuous functions $f : \mathbb{R} \to \mathbb{R}$. Then

$$X_n \overset{d}{\to} X \quad \Leftrightarrow \quad E[f(X_n)] \to E[f(X)], \forall\, f \in \mathbf{C}. \tag{5.21}$$

•

Remark 5.50 — **Alternative criterion for convergence in distribution** (Karr, 1993, p. 138)
Theorem 5.49 provides a criterion for convergence in distribution which is superior to the definition of convergence in distribution in that one needs not to deal with continuity points of the limit d.f. •

Exercise 5.51 — **Alternative criterion for convergence in distribution**
Prove Theorem 5.49 (Karr, 1993, pp. 138–139). •

Since in the proof of Theorem 5.49 the continuous functions used to approximate indicator functions can be taken to be arbitrarily smooth we can add a sufficient condition that guarantees convergence in distribution.

Corollary 5.52 — Sufficient condition for convergence in distribution (Karr, 1993, p. 139)
Let:

- k be a fixed non-negative integer;
- $\mathbf{C}^{(k)}$ be the space of bounded, $k-$times uniformly continuously differentiable functions $f : I\!R \to I\!R$.

Then

$$E[f(X_n)] \to E[f(X)], \forall\, f \in \mathbf{C}^{(k)} \quad \Rightarrow \quad X_n \xrightarrow{d} X. \tag{5.22}$$

•

The next table summarizes the alternative criteria and sufficient conditions for almost sure convergence and convergence in distribution of sequences of r.v.

Alternative criterion or sufficient condition		**Mode of convergence**
$\forall \epsilon > 0, \lim_{n\to+\infty} P\left(\left\{\sup_{k\geq n} \lvert X_k - X\rvert > \epsilon\right\}\right) = 0$	$\Leftrightarrow$	$X_n \xrightarrow{a.s.} X$
$Y_n = \sup_{k\geq n} \lvert X_k - X\rvert \xrightarrow{P} 0$	$\Leftrightarrow$	$X_n \xrightarrow{a.s.} X$
$\lim_{n\to+\infty} P\left(\{\sup_m \lvert X_{n+m} - X_n\rvert \leq \epsilon\}\right) = 1, \forall \epsilon > 0$	$\Leftrightarrow$	$X_n \xrightarrow{a.s.} X$
$\sum_{n=1}^{+\infty} P\left(\{\lvert X_n - X\rvert > \epsilon\}\right) < +\infty, \forall\, \epsilon > 0$	$\Rightarrow$	$X_n \xrightarrow{a.s.} X$
$\sum_{n=1}^{+\infty} P\left(\{\lvert X_n\rvert > \epsilon\}\right) < +\infty, \forall\, \epsilon > 0$	$\Leftrightarrow$	$X_n \xrightarrow{a.s.} 0$
$E[f(X_n)] \to E[f(X)], \forall\, f \in \mathbf{C}$	$\Leftrightarrow$	$X_n \xrightarrow{d} X$
$E[f(X_n)] \to E[f(X)], \forall\, f \in \mathbf{C}^{(k)}$ for a fixed $k \in I\!N_0$	$\Rightarrow$	$X_n \xrightarrow{d} X$

We should also add that Grimmett and Stirzaker (2001, p. 310) state that if $X_n \xrightarrow{P} X$ and $P(\{|X_n| \leq k\}) = 1$, for all n and some k, then $X_n \xrightarrow{L^r} X$, for all $r \geq 1$,[3] namely $X_n \xrightarrow{q.m.} X$ (which in turn implies $X_n \xrightarrow{L^1} X$).

[3]Let $X, X_1, X_2, \ldots$ belong to L^r $(r \geq 1)$. Then the sequence of r.v. $\{X_1, X_2, \ldots\}$ is said to converge to X in L^r) — denoted by $X_n \xrightarrow{L^r} X$ — if $\lim_{n\to+\infty} E\left(|X_n - X|^r\right) = 0$ (Grimmett and Stirzaker, 2001, p. 308).

5.2 Relationships among the modes of convergence

Given the plethora of modes of convergence, it is natural to inquire how they can be always related or hold true in the presence of additional assumptions (Karr, 1993, pp. 140 and 142).

5.2.1 Implications always valid

Proposition 5.53 — Almost sure convergence implies convergence in probability (Karr, 1993, p. 140; Rohatgi, 1976, p. 250)

$$X_n \xrightarrow{a.s.} X \quad \Rightarrow \quad X_n \xrightarrow{P} X. \tag{5.23}$$

•

Exercise 5.54 — Almost sure convergence implies convergence in probability Show Proposition 5.53 (Karr, 1993, p. 140; Rohatgi, 1976, p. 251). •

Proposition 5.55 — Convergence in quadratic mean implies convergence in L^1 (Karr, 1993, p. 140)

$$X_n \xrightarrow{q.m.} X \quad \Rightarrow \quad X_n \xrightarrow{L^1} X. \tag{5.24}$$

•

Exercise 5.56 — Convergence in quadratic mean implies convergence in L^1 Prove Proposition 5.55, by applying Cauchy-Schwarz's inequality (Karr, 1993, p. 140). •

Proposition 5.57 — Convergence in L^1 implies convergence in probability (Karr, 1993, p. 141)

$$X_n \xrightarrow{L^1} X \quad \Rightarrow \quad X_n \xrightarrow{P} X. \tag{5.25}$$

•

Exercise 5.58 — Convergence in L^1 implies convergence in probability Prove Proposition 5.57, by using Chebyshev's inequality (Karr, 1993, p. 141). •

Proposition 5.59 — Convergence in probability implies convergence in distribution (Karr, 1993, p. 141)

$$X_n \xrightarrow{P} X \quad \Rightarrow \quad X_n \xrightarrow{d} X. \tag{5.26}$$

•

Exercise 5.60 — Convergence in probability implies convergence in distribution
Show Proposition 5.59, (Karr, 1993, p. 141). •

Figure 5.1 shows that convergence in distribution is the weakest form of convergence, since it is implied by all other types of convergence studied so far.

$$\begin{array}{ccccc} X_n \overset{q.m.}{\to} X & \Rightarrow & X_n \overset{L^1}{\to} X & & \\ & & \Rightarrow & & \\ & & & X_n \overset{P}{\to} X & \Rightarrow \quad X_n \overset{d}{\to} X \\ & & \Rightarrow & & \\ & & X_n \overset{a.s.}{\to} X & & \end{array}$$

Figure 5.1: Implications always valid between modes of convergence.

Grimmett and Stirzaker (2001, p. 314) refer that convergence in distribution is the weakest form of convergence for two reasons: it only involves d.f. and makes no reference to an underlying probability space.[4] However, convergence in distribution has an useful representation in terms of almost sure convergence, as stated in the next theorem.

Theorem 5.61 — Skorokhod's representation theorem (Grimmett and Stirzaker, 2001, p. 314)
Let $\{X_1, X_2, \ldots\}$ be a sequence of r.v., $\{F_1, F_2, \ldots\}$ the associated sequence of d.f. and X be a r.v. with d.f. F. If $X_n \overset{d}{\to} X$ then there is a probability space $(\Omega', \mathcal{F}', P')$ and r.v. $\{Y_1, Y_2, \ldots\}$ and Y mapping Ω' into $\mathbb{R}$ such that $\{Y_1, Y_2, \ldots\}$ and Y have d.f. $\{F_1, F_2, \ldots\}$ and F and $Y_n \overset{a.s.}{\to} Y$. •

Remark 5.62 — Skorokhod's representation theorem (Grimmett and Stirzaker, 2001, p. 315)
Although X_n may fail to converge to X in any mode than in distribution, there is a sequence of r.v. $\{Y_1, Y_2, \ldots\}$ such that Y_n is identically distributed to X_n, for every n, which converges almost surely to a "copy" of X. •

[4] Let us remind the reader that that there is an equivalent formulation of convergence in distribution which involves d.f. alone: the sequence of d.f. $\{F_1, F_2, \ldots\}$ converges to the d.f. F, if $\lim_{n\to+\infty} F_n(x) = F(x)$ at each point x where F is continuous (Grimmett and Stirzaker, 2001, p. 190).

5.2.2 Counterexamples

Counterexamples to all implications among the modes of convergence (and more!) are condensed in Figure 5.2 and presented by means of several exercises.

$$\begin{array}{ccccc} X_n \overset{q.m.}{\to} X & \not\Leftarrow & X_n \overset{L^1}{\to} X & & \\ & & \not\Leftarrow & & \\ & & & X_n \overset{P}{\to} X & \not\Leftarrow \quad X_n \overset{d}{\to} X \\ & & \not\Leftarrow & & \\ & & X_n \overset{a.s.}{\to} X & & \end{array}$$

Figure 5.2: Counterexamples to implications among the modes of convergence.

Before proceeding with exercises, recall exercises 5.4 and 5.12 which pertain to the sequence of r.v. $\{X_1, X_2, \ldots\}$, where $X_n \sim \text{Bernoulli}(\frac{1}{n})$, $n \in \mathbb{N}$. In the first exercise we proved that $X_n \overset{a.s.}{\not\to} 0$, whereas in the second one we concluded that $X_n \overset{P}{\to} 0$. Thus, combining these results we can state that $X_n \overset{P}{\to} 0 \not\Rightarrow X_n \overset{a.s.}{\to} 0$.

Exercise 5.63 — Almost sure convergence does not imply convergence in quadratic mean
Let $\{X_1, X_2, \ldots\}$ be a sequence of r.v. such that

$$P(\{X_n = x\}) = \begin{cases} 1 - \frac{1}{n}, & x = 0 \\ \frac{1}{n}, & x = n \\ 0, & \text{otherwise.} \end{cases} \tag{5.27}$$

Prove that $X_n \overset{a.s.}{\to} 0$, and, hence, $X_n \overset{P}{\to} 0$ and $X_n \overset{d}{\to} 0$, but $X_n \overset{L^1}{\not\to} 0$ and $X_n \overset{q.m.}{\not\to} 0$ (Karr, 1993, p. 141, Counterexample a)). •

Exercise 5.64 — Almost sure convergence does not imply convergence in quadratic mean (bis)
Let $\{X_1, X_2, \ldots\}$ be a sequence of r.v. such that

$$P(\{X_n = x\}) = \begin{cases} 1 - \frac{1}{n^r}, & x = 0 \\ \frac{1}{n^r}, & x = n \\ 0, & \text{otherwise,} \end{cases} \tag{5.28}$$

where $r \geq 2$.

Prove that $X_n \overset{a.s.}{\to} 0$, but $X_n \overset{q.m.}{\not\to} 0$ for $r = 2$ (Rohatgi, 1976, p. 252, Example 10). •

Exercise 5.65 — Convergence in quadratic mean does not imply almost sure convergence

Let $X_n \sim \text{Bernoulli}\left(\frac{1}{n}\right)$.

Prove that $X_n \overset{q.m.}{\to} 0$, but $X_n \overset{a.s.}{\not\to} 0$ (Rohatgi, 1976, p. 252, Example 9). •

Exercise 5.66 — Convergence in L^1 does not imply convergence in quadratic mean

Let $\{X_1, X_2, \ldots\}$ be a sequence of r.v. such that

$$P(\{X_n = x\}) = \begin{cases} 1 - \frac{1}{n}, & x = 0 \\ \frac{1}{n}, & x = \sqrt{n} \\ 0, & \text{otherwise.} \end{cases} \tag{5.29}$$

Show that $X_n \overset{a.s.}{\to} 0$ and $X_n \overset{L^1}{\to} 0$, however $X_n \overset{q.m.}{\not\to} 0$ (Karr, 1993, p. 141, Counterexample b)). •

Exercise 5.67 — Convergence in probability does not imply almost sure convergence

For each positive integer n there exists integers m and k (uniquely determined) such that

$$n = 2^k + m,\ m = 0, 1, \ldots, 2^k - 1,\ k = 0, 1, 2, \ldots \tag{5.30}$$

Thus, for $n = 1$, $k = m = 0$; for $n = 5$, $k = 2$, $m = 1$; and so on.

Define r.v. X_n, for $n = 1, 2, \ldots$, on $\Omega = [0, 1]$ by

$$X_n(\omega) = \begin{cases} 2^k, & \frac{m}{2^k} \leq w < \frac{m+1}{2^k} \\ 0, & \text{otherwise.} \end{cases} \tag{5.31}$$

Let the probability distribution of X_n be given by $P(\{I\}) =$ length of the interval $I \subset \Omega$. Thus,

$$P(\{X_n = x\}) = \begin{cases} 1 - \frac{1}{2^k}, & x = 0 \\ \frac{1}{2^k}, & x = 2^k \\ 0, & \text{otherwise.} \end{cases} \tag{5.32}$$

Prove that $X_n \overset{P}{\to} 0$, but $X_n \overset{a.s.}{\not\to} 0$ (Rohatgi, 1976, pp. 251–2, Example 8). •

Exercise 5.68 — Convergence in distribution does not imply convergence in probability

Let $\{X_2, X_3, \ldots\}$ be a sequence of r.v. such that

$$F_{X_n}(x) = \begin{cases} 0, & x < 0 \\ \frac{1}{2} - \frac{1}{n}, & 0 \leq x < 1 \\ 1, & x \geq 1, \end{cases} \tag{5.33}$$

i.e. $X_n \sim \text{Bernoulli}\left(\frac{1}{2} + \frac{1}{n}\right)$, $n = 2, 3, \ldots$

Prove that $X_n \xrightarrow{d} X$, where $X \sim \text{Bernoulli}\left(\frac{1}{2}\right)$ and independent of any X_n, but $X_n \not\xrightarrow{P} X$ (Karr, 1993, p. 142, Counterexample d)). •

Exercise 5.69 — Convergence in distribution does not imply convergence in probability (bis)

Let $X, X_1, X_2, \ldots$ be identically distributed r.v. and let the joint p.f. of (X, X_n) be $P(\{X = 0, X_n = 1\}) = P(\{X = 1, X_n = 0\}) = \frac{1}{2}$.

Prove that $X_n \xrightarrow{d} X$, but $X_n \not\xrightarrow{P} X$ (Rohatgi, 1976, p. 247, Remark 2). •

5.2.3 Implications of restricted validity

Proposition 5.70 — Convergence in distribution to a constant implies convergence in probability (Karr, 1993, p. 140; Rohatgi, 1976, p. 246)
Let $\{X_1, X_2, \ldots\}$ be a sequence of r.v. and $c \in \mathbb{R}$. Then

$$X_n \xrightarrow{d} c \quad \Rightarrow \quad X_n \xrightarrow{P} c. \tag{5.34}$$

•

Remark 5.71 — Convergence in distribution to a constant is equivalent to convergence in probability (Rohatgi, 1976, p. 246)
If we add to the previous result the fact that $X_n \xrightarrow{P} c \Rightarrow X_n \xrightarrow{d} c$, we can conclude that

$$X_n \xrightarrow{P} c \quad \Leftrightarrow \quad X_n \xrightarrow{d} c. \tag{5.35}$$

•

Exercise 5.72 — Convergence in distribution to a constant implies convergence in probability
Show Proposition 5.70 (Karr, 1993, p. 142). •

Exercise 5.73 — Convergence in distribution to a constant implies convergence in probability (bis)
Let $(X_1, \ldots, X_n)$ be a random vector where X_i are i.i.d. r.v. with common p.d.f.

$$f_X(x) = \theta x^{-2} \times I_{[\theta, +\infty)}(x),$$

where $\theta \in \mathbb{R}^+$.

(a) After having proved that

$$F_{X_{(1:n)}}(x) = P\left(\min_{i=1,\ldots,n} X_i \leq x\right) = [1 - (\theta/x)^n] \times I_{[\theta,+\infty)}(x), \tag{5.36}$$

derive the following result: $X_{(1:n)} \xrightarrow{d} \theta$.

(b) Is $X_{(1:n)}$ a consistent estimator of θ? •

Definition 5.74 — Uniform integrability (Karr, 1993, p. 142)
A sequence of r.v. $\{X_1, X_2, \ldots\}$ is uniformly integrable if $X_n \in L^1$ for each $n \in \mathbb{N}$ and if

$$\lim_{a \to +\infty} \sup_n E(|X_n|; \{|X_n| > a\}) = 0. \tag{5.37}$$

Recall that the expected value of a r.v. X over an event A is given by $E(X; A) = E(X \times \mathbf{1}_A)$. •

Proposition 5.75 — Alternative criterion for uniform integrability (Karr, 1993, p. 143)
A sequence of r.v. $\{X_1, X_2, \ldots\}$ is uniformly integrable iff

- $\sup_n E(|X_n|) < +\infty$ and
- $\{X_1, X_2, \ldots\}$ is uniformly absolutely continuous: for each $\epsilon > 0$ there is $\delta > 0$ such that $\sup_n E(|X_n|; A) < \epsilon$ whenever $P(A) > \delta$. •

Proposition 5.76 — Combining convergence in probability and uniform integrability is equivalent to convergence in L^1 (Karr, 1993, p. 144)
Let $X, X_1, X_2, \ldots \in L^1$. Then

$$X_n \xrightarrow{P} X \text{ and } \{X_1, X_2, \ldots\} \text{ is uniformly integrable } \Leftrightarrow \; X_n \xrightarrow{L^1} X. \tag{5.38}$$

•

Exercise 5.77 — Combining convergence in probability and uniform integrability is equivalent to convergence in L^1
Prove Proposition 5.76 (Karr, 1993, p. 144). •

Exercise 5.78 — Combining convergence in probability of the sequence of r.v. and convergence of sequence of the means implies convergence in L^1 (Karr, 1993, p. 160, Exercise 5.16)
Let $X, X_1, X_2, \ldots$ be positive r.v.
Prove that if $X_n \xrightarrow{P} X$ and $E(X_n) \to E(X)$, then $X_n \xrightarrow{L^1} X$. •

Exercise 5.79 — Increasing character and convergence in probability combined imply almost sure convergence (Karr, 1993, p. 160, Exercise 5.15)
Show that if $X_1 \leq X_2 \leq \ldots$ and $X_n \xrightarrow{P} X$, then $X_n \xrightarrow{a.s.} X$. •

Exercise 5.80 — Strictly decreasing and positive character and convergence in probability combined imply almost sure convergence (Rohatgi, 1976, p. 252, Theorem 13)
Let $\{X_1, X_2, \ldots\}$ be a strictly decreasing sequence of positive r.v.
Prove that if $X_n \xrightarrow{P} 0$ then $X_n \xrightarrow{a.s.} 0$. •

5.3 Convergence under transformations

Since the original sequence(s) of r.v. is (are) bound to be transformed, it is natural to inquire whether the modes of convergence are preserved under continuous mappings and algebraic operations of the r.v.

5.3.1 Continuous mappings

Only convergence almost surely, in probability and in distribution are preserved under continuous mappings (Karr, 1993, p. 145).

Theorem 5.81 — Preservation of $\{a.s., P, d\}-$convergence under continuous mappings (Karr, 1993, p. 148)
Let:

- $\{X_1, X_2, \ldots\}$ be a sequence of r.v. and X a r.v.;
- $g : I\!R \to I\!R$ be a continuous function.

Then

$$X_n \xrightarrow{*} X \quad \Rightarrow \quad g(X_n) \xrightarrow{*} g(X), \ * = a.s., P, d. \tag{5.39}$$

•

Exercise 5.82 — Preservation of $\{a.s., P, d\}-$convergence under continuous mappings
Show Theorem 5.81 (Karr, 1993, p. 148). •

5.3.2 Algebraic operations

With the exception of the convergence in distribution, addition is preserved by the modes of convergence of r.v. as functions on Ω, as stated in the next theorem.

Theorem 5.83 — Preservation of $\{a.s., P, q.m, L^1\}-$convergence under addition (Karr, 1993, p. 145)
Let $X_n \xrightarrow{*} X$ and $Y_n \xrightarrow{*} Y$, where $* = a.s., P, q.m, L^1$. Then

$$X_n + Y_n \xrightarrow{*} X + Y, \ * = a.s., P, q.m, L^1. \tag{5.40}$$

•

Remark 5.84 — Preservation of $\{a.s., P, q.m, L^1\}-$convergence under addition
Under the conditions of Theorem 5.83,

- $X_n \pm Y_n \xrightarrow{*} X \pm Y$, $* = a.s., P, q.m, L^1$. •

Exercise 5.85 — Preservation of $\{a.s., P, q.m, L^1\}-$convergence under addition
Prove Theorem 5.83 (Karr, 1993, pp. 145–6). •

Convergence in distribution is only preserved under addition if one of the limits is constant.

Theorem 5.86 — Slutsky's theorem or preservation of $d-$convergence under (restricted) addition (Karr, 1993, p. 146)
Let:

- $X_n \xrightarrow{d} X$;
- $Y_n \xrightarrow{d} c$, $c \in I\!R$.

Then

$$X_n + Y_n \xrightarrow{d} X + c. \tag{5.41}$$

•

Remarks 5.87 — Slutsky's theorem or preservation of $d-$convergence under (restricted) addition and subtraction
(http://en.wikipedia.org/wiki/Slutsky's_theorem; Rohatgi, 1976, p. 253)

- The requirement that $\{Y_n\}$ converges in distribution to a constant is important — if it were to converge to a non-degenerate random variable, Theorem 5.86 would be no longer valid.
- Theorem 5.86 remains valid if we replace all convergences in distribution with convergences in probability because it implies the convergence in distribution.
- Moreover, Theorem 15 (Rohatgi, 1976, p. 253) reads as follows:

$$X_n \xrightarrow{d} X,\ Y_n \xrightarrow{P} c,\ c \in I\!R \quad \Rightarrow \quad X_n \pm Y_n \xrightarrow{d} X \pm c. \tag{5.42}$$

In this statement, the condition of $Y_n \xrightarrow{d} c$, $c \in I\!R$ in Theorem 5.86 was replaced with $Y_n \xrightarrow{P} c$, $c \in I\!R$. This by no means a contradiction because these two conditions are equivalent, according to Proposition 5.70. •

Exercise 5.88 — Slutsky's theorem or preservation of $d-$convergence under (restricted) addition
Prove Theorem 5.86 (Karr, 1993, p. 146; Rohatgi, 1976, pp. 253–4). •

As for the product, almost sure convergence and convergence in probability are preserved.

Theorem 5.89 — Preservation of $\{a.s., P\}-$convergence under product (Karr, 1993, p. 147)
Let $X_n \xrightarrow{*} X$ and $Y_n \xrightarrow{*} Y$, where $* = a.s., P$. Then

$$X_n \times Y_n \xrightarrow{*} X \times Y, * = a.s., P. \tag{5.43}$$

•

Exercise 5.90 — Preservation of $\{a.s., P\}-$convergence under product
Show Theorem 5.89 (Karr, 1993, p. 147).[5] •

Theorem 5.91 — (Non)preservation of $q.m.-$convergence under product (Karr, 1993, p. 147)
Let $X_n \xrightarrow{q.m.} X$ and $Y_n \xrightarrow{q.m.} Y$. Then

$$X_n \times Y_n \xrightarrow{L^1} X \times Y. \tag{5.44}$$

•

Remark 5.92 — (Non)preservation of $q.m.-$convergence under product (Karr, 1993, pp. 146–7)
Quadratic mean convergence of products does not hold in general, since $X \times Y$ need not belong to L^2 when X and Y do:

$$X_n \xrightarrow{q.m.} X,\ Y_n \xrightarrow{q.m.} Y \quad \not\Rightarrow \quad X_n \times Y_n \xrightarrow{q.m.} X \times Y. \tag{5.45}$$

However, the product of r.v. in L^2 belongs to L^1, and L^2 convergence of factors implies L^1 convergence of products. •

[5] Proposition 5.18 of Karr (1993, p. 144) may come handy to prove the result. This proposition reads as follows: the sequence of r.v. $\{X_1, X_2, \ldots\}$ converges in probability to X iff each subsequence $\{X_{1'}, X_{2'}, \ldots\}$ contains a further subsequence $\{X_{1''}, X_{2''}, \ldots\}$ such that $X_n \xrightarrow{a.s.} X$.

Exercise 5.93 — (Non)preservation of $q.m.-$convergence under product
Prove Theorem 5.91 (Karr, 1993, p. 147; Rohatgi, 1976, p. 254). •

Convergence in distribution is preserved under product, provided that one limit factor is constant (Karr, 1993, p. 146).

Theorem 5.94 — Slutsky's theorem (bis) or preservation of $d-$convergence under (restricted) product (Karr, 1993, p. 147)
Let:

- $X_n \xrightarrow{d} X$;
- $Y_n \xrightarrow{d} c$, $c \in \mathbb{R}$.

Then

$$X_n \times Y_n \xrightarrow{d} X \times c. \tag{5.46}$$

•

Remark 5.95 — Slutsky's theorem or preservation of $d-$convergence under (restricted) product (Rohatgi, 1976, p. 253)
Rohatgi (1976, p. 253, Theorem 15) also states that

$$X_n \xrightarrow{d} X,\ Y_n \xrightarrow{P} c,\ c \in \mathbb{R} \ \Rightarrow\ X_n \times Y_n \xrightarrow{d} X \times c \tag{5.47}$$

$$X_n \xrightarrow{d} X,\ Y_n \xrightarrow{P} c,\ c \in \mathbb{R}\backslash\{0\} \ \Rightarrow\ \frac{X_n}{Y_n} \xrightarrow{d} \frac{X}{c}. \tag{5.48}$$

(Discuss the validity of both results.) •

	Preservation under...		
Mode of convergence	**Continuous mapping**	**Addition & Subtraction**	**Product**
$\xrightarrow{a.s.}$ (almost sure)	YES	YES	YES
$\xrightarrow{P}$ (in probability)	YES	YES	YES
$\xrightarrow{q.m.}$ (in quadratic mean)	NO	YES	$\xrightarrow{L^1}$
$\xrightarrow{L^1}$ (in L^1)	NO	YES	YES
$\xrightarrow{d}$ (in distribution)	YES	RV*	RV*

* Restricted validity (RV): one of the summands/factors has to converge in distribution to a constant

Exercise 5.96 — Slutsky's theorem or preservation of $d-$convergence under (restricted) product

Prove Theorem 5.94 (Karr, 1993, pp. 147–8). •

Example/Exercise 5.97 — Slutsky's theorem or preservation of $d-$convergence under (restricted) product

Consider the sequence of r.v. $\{X_1, X_2, \ldots\}$, where $X_n \overset{i.i.d.}{\sim} X$ and let $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ and $S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$ be the sample mean and the variance of the first n r.v.

(a) Show that

$$\frac{\bar{X}_n - \mu}{S_n/\sqrt{n}} \xrightarrow{d} \text{Normal}(0,1), \tag{5.49}$$

for any $X \in L^4$.

- **R.v.**

  $X_i \overset{i.i.d.}{\sim} X$, $i \in I\!N$

  $X : E(X) = \mu, V(X) = \sigma^2 = \mu_2, E\left[(X-\mu)^4\right] = \mu_4$, which are finite moments since $X \in L^4$.

- **Auxiliary results**

  $E(\bar{X}_n) = \mu$

  $V(\bar{X}_n) = \frac{\sigma^2}{n} = \frac{\mu_2}{n}$

  $E(S_n^2) = \sigma^2 = \mu_2$

  $V(S_n^2) = \left(\frac{n}{n-1}\right)^2 \left[\frac{\mu_4 - \mu_2^2}{n} - \frac{2(\mu_4 - 2\mu_2^2)}{n^2} + \frac{2(\mu_4 - 3\mu_2^2)}{n^3}\right]$ (Murteira, 1980, p. 46).

- **Asymptotic sample distribution of $\frac{\bar{X}_n - \mu}{S_n/\sqrt{n}}$**

  To show that $\frac{\bar{X}_n - \mu}{S_n/\sqrt{n}} \xrightarrow{d} \text{Normal}(0,1)$ it suffices to note that

  $$\frac{\bar{X}_n - \mu}{S_n/\sqrt{n}} = \frac{\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}}{\sqrt{\frac{S_n^2}{\sigma^2}}}, \tag{5.50}$$

  prove that $\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \xrightarrow{d} \text{Normal}(0,1)$ and $\sqrt{\frac{S_n^2}{\sigma^2}} \xrightarrow{P} 1$, and then apply Slutsky's theorem as stated in (5.48).

- **Convergence in distribution of the numerator**

  It follows immediately from the Central Limit Theorem.[6]

[6] This well known theorem is thoroughly discussed by Karr (1993, pp. 190–196) and also in Section 5.9.

- **Convergence in probability of the denominator**

  By using the definition of convergence in probability and the Chebyshev(-Bienaymé)'s inequality, we get, for any $\epsilon > 0$:

$$\begin{aligned}
\lim_{n\to+\infty} P\left(|S_n^2 - \sigma^2| > \epsilon\right) &= \lim_{n\to+\infty} P\left(\left|S_n^2 - E(S_n^2)\right| \geq \frac{\epsilon}{\sqrt{V(S_n^2)}}\sqrt{V(S_n^2)}\right) \\
&\leq \lim_{n\to+\infty} \frac{1}{\left(\frac{\epsilon}{\sqrt{V(S_n^2)}}\right)^2} \\
&= \frac{1}{\epsilon^2} \lim_{n\to+\infty} V(S_n^2) \\
&= 0, \qquad (5.51)
\end{aligned}$$

  i.e. $S_n^2 \xrightarrow{P} \sigma^2$.

  Finally, note that convergence in probability is preserved under continuous mappings such as $g(x) = \sqrt{\frac{x}{\sigma}}$, hence

$$S_n^2 \xrightarrow{P} \sigma^2 \quad \Rightarrow \quad \sqrt{\frac{S_n^2}{\sigma^2}} \xrightarrow{P} \sqrt{\frac{\sigma^2}{\sigma^2}} = 1. \qquad (5.52)$$

- **Conclusion**

  $\frac{\bar{X}-\mu}{S/\sqrt{n}} \xrightarrow{d} \text{Normal}(0,1)$.

(b) Discuss the utility of this result. •

Exercise 5.98 — Slutsky's theorem or preservation of $d-$convergence under (restricted) division

Let $X_i \overset{i.i.d.}{\sim} \text{Normal}(0,1)$, $i \in I\!N$.

Determine the limiting distribution of $W_n = \frac{U_n}{V_n}$, where

$$U_n = \frac{\sum_{i=1}^n X_i}{\sqrt{n}} \qquad (5.53)$$

$$V_n = \frac{\sum_{i=1}^n X_i^2}{n}, \qquad (5.54)$$

by proving that

$$U_n \xrightarrow{d} \text{Normal}(0,1) \qquad (5.55)$$

$$V_n \xrightarrow{d} 1 \qquad (5.56)$$

(Rohatgi, 1976, pp. 254, Example 12). •

Exercise 5.99 — Slutsky's theorem or preservation of $d-$convergence under (restricted) division (bis)
Let $\{X_1, X_2, \ldots\}$ a sequence of i.i.d. r.v. with common distribution Bernoulli(p) and $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ the maximum likelihood estimator of p.

(a) Prove that

$$\frac{\bar{X}_n - p}{\sqrt{\frac{\bar{X}_n(1-\bar{X}_n)}{n}}} \xrightarrow{d} \text{Normal}(0,1). \tag{5.57}$$

(b) Discuss the relevance of this convergence in distribution. •

5.4 Convergence of random vectors

Before defining modes of convergence of a sequence of random $d-$vectors we need two recall the definition of norm of a vector.

Definition 5.100 — **L^2 (or Euclidean) and L^1 norms of $\underline{x}$** (Karr, 1993, p. 149)
Let $\underline{x} \in I\!R^d$ and $\underline{x}(i)$ its ith component. Then

$$||\underline{x}||_{L^2} = \sqrt{\sum_{i=1}^{d} \underline{x}(i)^2} \tag{5.58}$$

$$||\underline{x}||_{L^1} = \sum_{i=1}^{d} |\underline{x}(i)| \tag{5.59}$$

denote the L^2 norm (or Euclidean norm) and the L^1 norm of $\underline{x}$, respectively. •

Remark 5.101 — **L^2 (or Euclidean) and L^1 norms of $\underline{x}$**
(http://en.wikipedia.org/wiki/Norm_mathematics♯Definition)
On $I\!R^d$, the intuitive notion of length of the vector $\underline{x}$ is captured by its L^2 or Euclidean norm: this gives the ordinary distance from the origin to the point $\underline{x}$, a consequence of the Pythagorean theorem.

The Euclidean norm is by far the most commonly used norm on $I\!R^d$, but there are other norms, such as the L^1 norm on this vector space. •

Definition 5.102 — **Four modes of convergence (as functions of Ω) of sequences of random vectors** (Karr, 1993, p. 149)
Let $\underline{X}$, $\underline{X}_1$, $\underline{X}_2$, ... be random $d-$vectors. Then the four modes of convergence $\underline{X}_n \xrightarrow{*} \underline{X}$, $* = a.s., P, q.m., L^1$ are natural extensions of their counterparts in the univariate case:

- $\underline{X}_n \xrightarrow{a.s.} \underline{X}$ if $P(\{\omega : \lim_{n\to+\infty} ||\underline{X}_n(\omega) - \underline{X}(\omega)||_{L^1} = 0\}) = 1$;
- $\underline{X}_n \xrightarrow{P} \underline{X}$ if $\lim_{n\to+\infty} P\left(\{||\underline{X}_n - \underline{X}||_{L^1} > \epsilon\}\right) = 0$, for every $\epsilon > 0$;
- $\underline{X}_n \xrightarrow{q.m.} \underline{X}$ if $\lim_{n\to+\infty} E\left(||\underline{X}_n - \underline{X}||_{L^2}\right) = 0$;
- $\underline{X}_n \xrightarrow{L^1} \underline{X}$ if $\lim_{n\to+\infty} E\left(||\underline{X}_n - \underline{X}||_{L^1}\right) = 0$. •

Proposition 5.103 — **Alternative criteria for the four modes of convergence of sequences of random vectors** (Karr, 1993, p. 149)
$\underline{X}_n \xrightarrow{*} \underline{X}$, $* = a.s., P, q.m., L^1$ iff the same kind of stochastic convergence holds for each component, i.e. $\underline{X}_n(i) \xrightarrow{*} \underline{X}(i)$, $* = a.s., P, q.m., L^1$, $i = 1, \ldots, d$. •

Remark 5.104 — Convergence in distribution of a sequence of random vectors (Karr, 1993, p. 149)
Due to the intractability of multi-dimension d.f., convergence in distribution — unlike the four previous modes of convergence — has to be defined by taking advantage of the alternative criterion for convergence in distribution stated in Theorem 5.49. •

Definition 5.105 — Convergence in distribution of a sequence of random vectors
(Karr, 1993, p. 149)
Let $\underline{X}$, $\underline{X}_1$, $\underline{X}_2$, ... be random $d-$vectors. Then:

- $\underline{X}_n \xrightarrow{d} \underline{X}$ if $E[f(\underline{X}_n)] \to E[f(\underline{X})]$, for all bounded, continuous functions $f : \mathbb{R}^d \to \mathbb{R}$. •

Proposition 5.106 — A sufficient condition for the convergence in distribution of the components of a sequence of random vectors (Karr, 1993, p. 149)
Unlike the four previous modes of convergence, convergence in distribution of the components of a sequence of random vectors is implied, but need not imply, convergence in distribution of the sequence of random vectors:

$$\underline{X}_n \xrightarrow{d} \underline{X} \Rightarrow (\not\Leftarrow)\ \underline{X}_n(i) \xrightarrow{d} \underline{X}(i), \tag{5.60}$$

for each i. •

A sequence of random vectors converges in distribution iff every linear combination of their components converges in distribution; this result constitutes the Cramér-Wold device.

Theorem 5.107 (Cramér-Wold device) — An alternative criterion for the convergence in distribution of a sequence of random vectors (Karr, 1993, p. 150)
Let $\underline{X}$, $\underline{X}_1$, $\underline{X}_2$, ... be random $d-$vectors. Then

$$\underline{X}_n \xrightarrow{d} \underline{X} \Leftrightarrow \underline{a}^\top \underline{X}_n = \sum_{i=1}^{d} \underline{a}(i) \times \underline{X}_n(i) \xrightarrow{d} \sum_{i=1}^{d} \underline{a}(i) \times \underline{X}(i) = \underline{a}^\top \underline{X}, \tag{5.61}$$

for all $\underline{a} \in \mathbb{R}^d$. •

Exercise 5.108 — An alternative criterion for the convergence in distribution of a sequence of random vectors
Show Theorem 5.107 (Karr, 1993, p. 150). •

As with sequences of r.v., convergence almost surely, in probability and in distribution are preserved under continuous mappings of sequences of random vectors.

Theorem 5.109 — Preservation of $\{a.s., P, d\}-$convergence under continuous mappings of random vectors (Karr, 1993, p. 148)
Let:

- $\{\underline{X}_1, \underline{X}_2, \ldots\}$ be a sequence of random $d-$vectors and $\underline{X}$ a random $d-$vector;
- $g : I\!R^d \to I\!R^m$ be a continuous mapping of $I\!R^d$ into $I\!R^m$.

Then

$$\underline{X}_n \xrightarrow{*} \underline{X} \quad \Rightarrow \quad g(\underline{X}_n) \xrightarrow{*} g(\underline{X}), \ * = a.s., P, d. \tag{5.62}$$

•

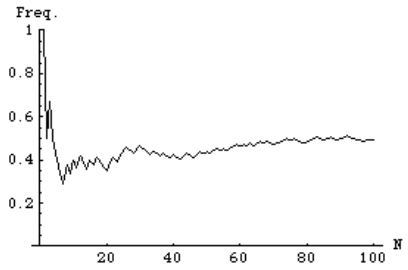
5.5 Limit theorems for Bernoulli summands

Let $\{X_1, X_2, \ldots\}$ be a Bernoulli process with parameter $p \in (0,1)$. In this section we study the asymptotic behavior of the Bernoulli counting process $\{S_1, S_2, \ldots\}$, where $S_n = \sum_{i=1}^{n} X_i \sim \text{Binomial}(n,p)$.

5.5.1 Laws of large numbers for Bernoulli summands

Motivation 5.110 — Laws of large numbers
(http://en.wikipedia.org/wiki/Law_of_large_numbers; Murteira, 1979, p. 313)
In probability theory, the law of large numbers (LLN) is a theorem that describes the result of performing the same experiment a large number of times. According to the law, the average of the results obtained from a large number of trials (e.g. Bernoulli trials) should be close to the expected value, and will tend to become closer as more trials are performed.

For instance, when a FAIR coin is flipped once, the expected value of the number of heads is equal to one half. Therefore, according to the law of large numbers, the proportion of heads in a large number of coin flips should be roughly one half, as depicted by the next figure (where N stands for n).



This illustration suggests the following statement: $\frac{S_n}{n} = \bar{X}_n$ converges, in some sense, to $p = \frac{1}{2}$. In fact, if we use Chebyshev(-Bienaymé)'s inequality we can prove that

$$\lim_{n \to +\infty} P\left(\left\{\left|\frac{S_n}{n} - p\right| > \varepsilon\right\}\right) = 0, \tag{5.63}$$

that is, $\frac{S_n}{n} \xrightarrow{P} p = \frac{1}{2}$. (Show this result!) In addition, we can also prove that the proportion of heads after n flips will almost surely converge to one half as n approaches infinity, i.e., $\frac{S_n}{n} \xrightarrow{a.s.} p = \frac{1}{2}$. Similar convergences can be devised for the mean of n i.i.d. r.v.

The Indian mathematician Brahmagupta (598–668) and later the Italian mathematician Gerolamo Cardano (1501–1576) stated without proof that the accuracies of empirical statistics tend to improve with the number of trials. This was then formalized as a law of large numbers (LLN).

The LLN was first proved by Jacob Bernoulli. It took him over 20 years to develop a sufficiently rigorous mathematical proof which was published in his Ars Conjectandi (The Art of Conjecturing) in 1713. He named this his *Golden Theorem* but it became generally known as "Bernoulli's Theorem". In 1835, S.D. Poisson further described it under the name *La loi des grands nombres* (*The law of large numbers*). Thereafter, it was known under both names, but the *Law of large numbers* is most frequently used.

Other mathematicians also contributed to refinement of the law, including Chebyshev, Markov, Borel, Cantelli and Kolmogorov. These further studies have given rise to two prominent forms of the LLN:

- the weak law of large numbers (WLLN);
- the strong law of large numbers (SLLN);

These forms do not describe different laws but instead refer to different ways of describing the mode of convergence of the cumulative sample means to the expected value:

- the WLLN refers to a convergence in probability;
- the SLLN is concerned with an almost sure convergence;

Needless to say that the SLLN implies the WLLN. •

Theorem 5.111 — Weak law of large numbers for Bernoulli summands
(Karr, 1993, p. 151)
Let:

- $\{X_1, X_2, \ldots\}$ be a Bernoulli process with parameter $p \in (0, 1)$;
- $\frac{S_n}{n} = \bar{X}_n$ be the proportion of successes in the first n Bernoulli trials.

Then

$$\frac{S_n}{n} \overset{q.m.}{\to} p, \tag{5.64}$$

therefore

$$\frac{S_n}{n} \overset{P}{\to} p. \tag{5.65}$$

•

Exercise 5.112 — Weak law of large numbers for Bernoulli summands
Show Theorem 5.111, by calculating the limit of $E\left[\left(\frac{S_n}{n}-p\right)^2\right]$ (thus proving the convergence in quadratic mean) and combining Proposition 5.55 (which states that convergence in quadratic mean implies convergence in L^1) and Proposition 5.57 (it says that convergence in L^1 implies convergence in probability) (Karr, 1993, p. 151). •

Theorem 5.113 — Strong law of large numbers for Bernoulli summands or Borel's SLLN (Karr, 1993, p. 151; Rohatgi, 1976, p. 273, Corollary 3)
Let:

- $\{X_1, X_2, \ldots\}$ be a Bernoulli process with parameter $p \in (0,1)$;
- $\frac{S_n}{n} = \bar{X}_n$ be the proportion of successes in the first n Bernoulli trials.

Then

$$\frac{S_n}{n} \overset{a.s.}{\to} p. \tag{5.66}$$

•

Exercise 5.114 — Strong law of large numbers for Bernoulli summands or Borel's SLLN
Prove Theorem 5.113, by: using Theorem 4.121 (Chebyshev's inequality) with $g(x) = x^4$ to set an upper limit to $P\left(\{|S_n - np| > n\epsilon\}\right)$, which is smaller than $O(n^{-2})$,[7] thus, proving that $\sum_{i=1}^{+\infty} P\left(\left\{\left|\frac{S_n}{n} - p\right| > \epsilon\right\}\right) < \infty$, i.e., that the sequence $\left\{\frac{S_1}{1}, \frac{S_2}{2}, \ldots\right\}$ completely converges to p; finally applying Proposition 5.44 which relates almost sure convergence and complete convergence (Karr, 1993, pp. 151–152).[8] •

Remark 5.115 — Weak and strong laws of large numbers for Bernoulli summands (http://en.wikipedia.org/wiki/Law_of_large_numbers; Karr, 1993, p. 152)

- Theorem 5.113 can be invoked to support the frequency interpretation of probability.
- The WLLN for Bernoulli summands states that for a specified large n, $\frac{S_n}{n}$ is likely to be near p. Thus, it leaves open the possibility that the event $\{|\frac{S_n}{n} - p| > \epsilon\}$, for any $\epsilon > 0$, happens an infinite number of times, although at infrequent intervals.

  The SLLN for Bernoulli summands shows that this almost surely will not occur. In particular, it implies that with probability 1, we have that, for any $\epsilon > 0$, the inequality $|\frac{S_n}{n} - p| > \epsilon$ holds for all large enough n.

[7] Let $f(x)$ and $g(x)$ be two functions defined on some subset of the real numbers. One writes $f(x) = O(g(x))$ as $x \to \infty$ iff there exists a positive real number M and a real number x_0 such that $|f(x)| \leq M|g(x)|$ for all $x > x_0$ (http://en.wikipedia.org/wiki/Big_O_notation).

[8] A simple proof (using the 2nd. Borel-Cantelli lemma) can be found in Rohatgi (1976, p. 265).

- Finally, the proofs of theorems 5.111 and 5.113 only involve the moments of X_i. Unsurprisingly, these two theorems can be reproduced for other sequences of i.i.d. r.v., namely those in L^2 (in the case of the WLLN) and in L^4 (for the SLLN), as we shall see in sections 5.6 and 5.7. •

5.5.2 Central limit theorems for Bernoulli summands

Motivation 5.116 — Central limit theorems for Bernoulli summands (Karr, 1993, p. 152)
They essentially state that, in the Bernoulli summands case and for large n, $S_n \sim \text{Binomial}(n,p)$ is such that $\frac{S_n - E(S_n)}{\sqrt{V(S_n)}} = \frac{S_n - np}{\sqrt{np(1-p)}}$ has approximately a standard normal distribution.

The local (resp. global) central limit theorem — also known as the DeMoivre-Laplace local (resp. global) limit theorem — provides an approximation to the p.f. (resp. d.f.) of S_n in terms of the standard normal p.d.f. (resp. d.f.). •

Theorem 5.117 — DeMoivre-Laplace local limit theorem (Karr, 1993, p. 153)
Let:

- $k_n = 0, 1, \ldots, n$;
- $x_n = \frac{k_n - np}{\sqrt{np(1-p)}} = o(n^{1/6})$;[9]
- $\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$ be the standard normal p.d.f.

Then

$$\lim_{n \to +\infty} \frac{P(\{S_n = k_n\})}{\frac{\phi(x_n)}{\sqrt{np(1-p)}}} = 1. \tag{5.67}$$

•

Remark 5.118 — DeMoivre-Laplace local limit theorem (Karr, 1993, p. 153)
The proof of Theorem 5.117 shows that the convergence in (5.67) is uniform in values of k satisfying $|k - np| = o(n^{2/3})$. As a consequence, for large values of n and values of k_n not to different from np,

$$P(\{S_n = k_n\}) \simeq \frac{1}{\sqrt{np(1-p)}} \times \phi\left[\frac{k_n - np}{\sqrt{np(1-p)}}\right], \tag{5.68}$$

[9] The relation $f(x) = o(g(x))$ is read as "$f(x)$ is little-o of $g(x)$". Intuitively, it means that $g(x)$ grows much faster than $f(x)$. Formally, it states $\lim_{x \to \infty} \frac{f(x)}{g(x)} = 0$ (http://en.wikipedia.org/wiki/Big_O_notation#Little-o_notation).

that is, the p.f. of $S_n \sim \text{Binomial}(n, p)$ evaluated at k_n can be properly approximated by the p.d.f. of a normal distribution, with mean $E(S_n) = np$ and variance $V(S_n) = np(1-p)$, evaluated at $\frac{k_n - np}{\sqrt{np(1-p)}}$. •

Exercise 5.119 — **DeMoivre-Laplace local limit theorem**

(a) Show Theorem 5.117 (Karr, 1993, p. 153).

(b) What is the probability that exactly 20 heads result when you flip a fair coin 40 times? (www.maths.bris.ac.uk/∼mb13434/Stirling_DeMoivre_Laplace.pdf) •

Theorem 5.120 — **DeMoivre-Laplace global limit theorem** (Karr, 1993, p. 154; Murteira, 1979, p. 347)
Let $S_n \sim \text{Binomial}(n, p)$, $n \in I\!N$. Then

$$\frac{S_n - np}{\sqrt{np(1-p)}} \xrightarrow{d} \text{Normal}(0, 1). \tag{5.69}$$

•

Remark 5.121 — **DeMoivre-Laplace global limit theorem** (Karr, 1993, p. 155; Murteira, 1979, p. 347)

- Theorem 5.120 justifies the following approximation:

$$P(S_n \leq x) \simeq \Phi\left[\frac{x - np}{\sqrt{np(1-p)}}\right]. \tag{5.70}$$

- According to Murteira (1979, p. 348), the well known continuity correction was proposed by Feller in 1968 to improve the normal approximation to the binomial distribution,[10] and can be written as:

$$P(a \leq S_n \leq b) \simeq \Phi\left[\frac{b + \frac{1}{2} - np}{\sqrt{np(1-p)}}\right] - \Phi\left[\frac{a - \frac{1}{2} - np}{\sqrt{np(1-p)}}\right]. \tag{5.71}$$

[10] However, http://en.wikipedia.org/wiki/Continuity_correction suggests that continuity correction dates back from Feller, W. (1945). On the normal approximation to the binomial distribution. *The Annals of Mathematical Statistics* **16**, pp. 319–329.

- The proof of the central limit theorem for summands (other than Bernoulli ones) involves a Taylor series expansion[11] and requires dealing with the notion of characteristic function of a r.v.[12] Such proof can be found in Murteira (1979, pp. 354–355); Karr (1993, pp. 190–196) devotes a whole section to this theorem. •

Exercise 5.122 — DeMoivre-Laplace global limit theorem

(a) Show Theorem 5.120 (Karr, 1993, pp. 154–155).

(b) The ideal size of a course is 150 students. On average 30% of those accepted will enroll, therefore the organisers accept 450 students.

What is the probability that more than 150 students enroll? (www.maths.bris.ac.uk/∼mb13434/Stirling_DeMoivre_Laplace.pdf) •

5.5.3 The Poisson limit theorem

Motivation 5.123 — Poisson limit theorem
(Karr, 1993, p. 155; http://en.wikipedia.org/wiki/Poisson_limit_theorem)

- In the two central limit theorems for Bernoulli summands, although $n \to +\infty$, the parameter p remained fixed. These theorems provide useful approximations to binomial probabilities, as long as the values of p are close to neither zero or one, and inaccurate ones, otherwise.

- The Poisson limit theorem gives a Poisson approximation to the binomial distribution, under certain conditions, namely, it considers the effect of simultaneously allowing $n \to +\infty$ and $p = p_n \to 0$ with the proviso that $n \times p_n \to \lambda$, where $\lambda \in \mathbb{R}^+$. This theorem was obviously named after Siméon-Denis Poisson (1781–1840). •

[11] The Taylor series of a real or complex function $f(x)$ that is infinitely differentiable in a neighborhood of a real (or complex number) a is the power series written in the more compact sigma notation as $\sum_{n=0}^{+\infty} \frac{f^{(n)}(a)}{n!} (x-a)^n$, where $f^{(n)}(a)$ denotes the nth derivative of f evaluated at the point a. In the case that a = 0, the series is also called a Maclaurin series (http://en.wikipedia.org/wiki/Taylor_series).

[12] For a scalar random variable X the characteristic function is defined as the expected value of e^{itX}, $E(e^{itX})$, where i is the imaginary unit, and $t \in \mathbb{R}$ is the argument of the characteristic function (http://en.wikipedia.org/wiki/Characteristic_function_(probability_theory)).

Theorem 5.124 — Poisson limit theorem (Karr, 1993, p. 155)
Let:

- $\{X_1, X_2, \ldots\}$ be a sequence of r.v. such that $X_n \sim \text{Binomial}(n, p_n)$, for each n;
- $n \times p_n \to \lambda$, where $\lambda \in \mathbb{R}^+$.

Then

$$X_n \xrightarrow{d} \text{Poisson}(\lambda). \tag{5.72}$$

•

Example/Exercise 5.125 — Poisson limit theorem

(a) Consider $0 < \lambda < n$ and let us verify that

$$\lim_{\substack{n \to +\infty \\ p_n \to 0 \\ np_n = \lambda \text{ fix}}} \binom{n}{x} p_n^x (1 - p_n)^{n-x} = e^{-\lambda} \frac{\lambda^x}{x!}.$$

- **R.v.**

  $X_n \sim \text{Binomial}(n, p_n)$

- **Parameters**

  n

  $p_n = \frac{\lambda}{n}$ $(0 < \lambda < n)$

- **P.f.**

  $P(X_n = x) = \binom{n}{x} \left(\frac{\lambda}{n}\right)^x \left(1 - \frac{\lambda}{n}\right)^{n-x}$, $x = 0, 1, \ldots, n$

- **Limit p.f.**

  For any $x \in \{0, 1, \ldots, n\}$, we get

$$\begin{aligned}
\lim_{n \to +\infty} P(X_n = x) &= \frac{\lambda^x}{x!} \times \lim_{n \to +\infty} \frac{n(n-1)\ldots(n-x+1)}{n^x} \\
&\quad \times \lim_{n \to +\infty} \left(1 + \frac{-\lambda}{n}\right)^n \times \lim_{n \to +\infty} \left(1 - \frac{\lambda}{n}\right)^{-x} \\
&= \frac{\lambda^x}{x!} \times 1 \times e^{-\lambda} \times 1 \\
&= e^{-\lambda} \frac{\lambda^x}{x!}.
\end{aligned}$$

- **Conclusion**

  If the limit p.f. of X_n coincides with p.f. of $X \sim \text{Poisson}(\lambda)$ then the same holds for the limit d.f. of X_n and the d.f. of X. Hence

  $$X_n \xrightarrow{d} \text{Poisson}(\lambda).$$

(b) Now, prove Theorem 5.124 (Karr, 1993, p. 155).

(c) Suppose that in an interval of length 1000, 500 points are placed randomly.

Use the Poisson limit theorem to prove that we can approximate the p.f. of the number points that will be placed in a sub-interval of length 10 by

$$e^{-5}\frac{5^k}{k!} \tag{5.73}$$

(http://en.wikipedia.org/wiki/Poisson_limit_theorem). •

5.6 Weak law of large numbers

Motivation 5.126 — Weak law of large numbers (Rohatgi, 1976, p. 257)
Let:

- $\{X_1, X_2, \ldots\}$ be a sequence of r.v. in L^2;
- $S_n = \sum_{i=1}^{n} X_i$ be the sum of the first n terms of such a sequence.

In this section we are going to answer the next question in the affirmative:

- Are there constants a_n and b_n ($b_n > 0$) such that $\frac{S_n - a_n}{b_n} \xrightarrow{P} 0$?

In other words, what follows are extensions of the WLLN for Bernoulli summands (Theorem 5.111), to other sequences of:

- i.i.d. r.v. in L^2;
- pairwise uncorrelated and identically distributed r.v. in L^2;
- pairwise uncorrelated r.v. in L^2;
- r.v. in L^2 with a specific variance behavior;
- i.i.d. r.v. in L^1. •

Definition 5.127 — Obeying the weak law of large numbers (Rohatgi, 1976, p. 257)
Let:

- $\{X_1, X_2, \ldots\}$ be a sequence of r.v.;
- $S_n = \sum_{i=1}^{n} X_i$, $n = 1, 2, \ldots$;

Then $\{X_1, X_2, \ldots\}$ is said to obey the weak law of large numbers (WLLN) with respect to the sequence of constants $\{b_1, b_2, \ldots\}$ ($b_n > 0$, $b_n \uparrow +\infty$) if there is a sequence of real constants $\{a_1, a_2, \ldots\}$ such that

$$\frac{S_n - a_n}{b_n} \xrightarrow{P} 0. \tag{5.74}$$

a_n and b_n are called centering and norming constants, respectively. •

Remark 5.128 — Obeying the weak law of large numbers (Rohatgi, 1976, p. 257)
The definition in Murteira (1979, p. 319) is a particular case of Definition 5.127 with $a_n = \sum_{i=1}^{n} E(X_i)$ and $b_n = n$.

- Let $\{X_1, X_2, \ldots\}$ be a sequence of r.v., $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$, and $\{Z_1, Z_2, \ldots\}$ be another sequence of r.v. such that $Z_n = \frac{S_n - a_n}{b_n} = \bar{X}_n - E(\bar{X}_n)$, $n = 1, 2, \ldots$

  Then $\{X_1, X_2, \ldots\}$ is said to obey the WLLN if $Z_n \xrightarrow{P} 0$.

Hereafter the convergence results are stated either in terms of S_n or $\bar{X}_n$. •

Theorem 5.129 — Weak law of large numbers, i.i.d. r.v. in L^2 (Karr, 1993, p. 152)
Let $\{X_1, X_2, \ldots\}$ be a sequence of i.i.d. r.v. in L^2 with common expected value μ and variance σ^2. Then

$$\bar{X}_n \xrightarrow{q.m.} \mu, \tag{5.75}$$

therefore $\{X_1, X_2, \ldots\}$ obeys the WLLN:

$$\bar{X}_n \xrightarrow{P} \mu, \tag{5.76}$$

i.e., $\frac{S_n - n\mu}{n} \xrightarrow{P} 0$.[13] •

Exercise 5.130 — Weak law of large numbers, i.i.d. r.v. in L^2
Prove Theorem 5.129, by mimicking the proof of the WLLN for Bernoulli summands. •

Exercise 5.131 — Weak law of large numbers, i.i.d. r.v. in L^2 (bis)
Let $\{X_1, X_2, \ldots\}$ a sequence of i.i.d. r.v. with common p.d.f.

$$f(x) = \begin{cases} e^{-x+q}, & x > q \\ 0, & \text{otherwise} \end{cases} \tag{5.77}$$

(a) Prove that $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{P} 1 + q$.

(b) Show that $X_{(1:n)} = \min_{i=1,\ldots,n} X_i \xrightarrow{P} q$.[14] •

A closer look to the proof Theorem 5.129 leads to the conclusion that the r.v. need only be pairwise uncorrelated and identically distributed in L^2, since in this case we still have $V(\bar{X}_n) = \frac{\sigma^2}{n}$ (Karr, 1993, p. 152).

[13] As suggested by Rohatgi (1976, p. 258, Corollary 3).
[14] Use the d.f. of $X_{(1:n)}$.

Theorem 5.132 — Weak law of large numbers, pairwise uncorrelated and identically distributed r.v. in L^2 (Karr, 1993, p. 152; Rohatgi, 1976, p. 258)
Let $\{X_1, X_2, \ldots\}$ be a sequence of pairwise uncorrelated and identically distributed r.v. in L^2 with common expected value μ and variance σ^2. Thus, $\bar{X}_n \overset{q.m.}{\to} \mu$ and[15] hence $\{X_1, X_2, \ldots\}$ obeys the WLLN:

$$\bar{X}_n \overset{P}{\to} \mu. \tag{5.78}$$

•

We can also drop the assumption that we are dealing with identically distributed r.v. as suggested by the following theorem.

Theorem 5.133 — Weak law of large numbers, pairwise uncorrelated r.v. in L^2 (Rohatgi, 1976, p. 258, Theorem 1)
Let:

- $\{X_1, X_2, \ldots\}$ be a sequence of pairwise uncorrelated r.v. in L^2 with $E(X_i) = \mu_i$ and $V(X_i) = \sigma_i^2$;
- $a_n = \sum_{i=1}^n \mu_i$;
- $b_n = \sum_{i=1}^n \sigma_i^2$.

If $b_n \to +\infty$ then

$$\frac{S_n - a_n}{b_n} = \frac{\sum_{i=1}^n X_i - \sum_{i=1}^n \mu_i}{\sum_{i=1}^n \sigma_i^2} \overset{P}{\to} 0, \tag{5.79}$$

i.e., $\{X_1, X_2, \ldots\}$ obeys the WLLN with respect to b_n. •

Exercise 5.134 — Weak law of large numbers, pairwise uncorrelated r.v. in L^2

(a) Show Theorem 5.132.

(b) Prove Theorem 5.133 by applying Chebyshev(-Bienaymé)s inequality (Rohatgi, 1976, p. 258). •

Remark 5.135 — Weak law of large numbers, pairwise uncorrelated r.v. in L^2
A careful inspection of the proof of Theorem 5.133 (Rohatgi, 1976, p. 258) leads us to restate it as follows:

- Let $\{X_1, X_2, \ldots\}$ be a sequence of pairwise uncorrelated r.v. in L^2 with $E(X_i) = \mu_i$ and $V(X_i) = \sigma_i^2$, $a_n = \sum_{i=1}^n \mu_i$, and

[15] Rohatgi (1976, p. 258, Corollary 1) does not refer this convergence in quadratic mean.

$$b_n = \sqrt{\sum_{i=1}^{n} \sigma_i^2}. \tag{5.80}$$

If $b_n \to +\infty$ then

$$\frac{S_n - a_n}{b_n} = \frac{S_n - E(S_n)}{\sqrt{V(S_n)}} \xrightarrow{P} 0. \tag{5.81}$$

Theorem 5.133 can be further generalized: the sequence of r.v. need only have the mean of its first n terms, $\bar{X}_n$, with a specific variance behavior, as stated below. •

Theorem 5.136 — WLLN and Markov's theorem (Murteira, 1979, p. 320)
Let $\{X_1, X_2, \ldots\}$ be a sequence of r.v. in L^2. If

$$\lim_{n\to+\infty} V(\bar{X}_n) = \lim_{n\to+\infty} \frac{1}{n^2} V\left(\sum_{i=1}^{n} X_i\right) = 0, \tag{5.82}$$

then

$$\bar{X}_n - E(\bar{X}_n) \xrightarrow{P} 0, \tag{5.83}$$

that is, $\{X_1, X_2, \ldots\}$ obeys the WLLN with respect to $b_n = n$ ($a_n = \sum_{i=1}^{n} E(X_i)$). •

Exercise 5.137 — WLLN and Markov's theorem
Show Theorem 5.136, by simply applying Chebyshev(-Bienaymé)'s inequality. •

Remark 5.138 — (Special cases of) Markov's theorem (Murteira, 1979, pp. 320–321; Rohatgi, 1979, p. 258)

- The WLLN holds for a sequence of pairwise uncorrelated r.v., with common expected value μ and uniformly limited variance $V(X_n) < k$, $n = 1, 2, \ldots; k \in \mathbb{R}^+$.[16]
- The WLLN also holds for a sequence of pairwise uncorrelated and identically distributed r.v. in L^2, with common expected value μ and σ^2 (Theorem 5.132).

[16]This corollary of Markov's theorem is due to Chebyshev. Please note that when we dealing with pairwise uncorrelated r.v., the condition (5.82) in Markov's theorem still reads: $\lim_{n\to+\infty} V(\bar{X}_n) = \lim_{n\to+\infty} \frac{1}{n^2} \sum_{i=1}^{n} V(X_i) = 0$.

- Needless to say that the WLLN holds for any sequence of i.i.d. r.v. in L^2 (Theorem 5.129) and therefore $\bar{X}_n$ is a consistent estimator of μ.

  Moreover, according to http://en.wikipedia.org/wiki/Law_of_large_numbers, the assumption of finite variances ($V(X_i) = \sigma^2 < +\infty$) is not necessary; large or infinite variance will make the convergence slower, but the WLLN holds anyway, as stated in Theorem 5.143. This assumption is often used because it makes the proofs easier and shorter. •

The next theorem provides a necessary and sufficient condition for a sequence of r.v. $\{X_1, X_2, \ldots\}$ to obey the WLLN.

Theorem 5.139 — An alternative criterion for the WLLN (Rohatgi, 1976, p. 258, Theorem 2)
Let:

- $\{X_1, X_2, \ldots\}$ be a sequence of r.v. (in L^2);
- $Y_n = \bar{X}_n$, $n = 1, 2, \ldots$.

Then $\{X_1, X_2, \ldots\}$ satisfies the WLLN with respect to $b_n = n$ ($a_n = \sum_{i=1}^{n} E(X_i)$), i.e.

$$\bar{X}_n - E(\bar{X}_n) \xrightarrow{P} 0, \tag{5.84}$$

iff

$$\lim_{n \to +\infty} E\left(\frac{Y_n^2}{1 + Y_n^2}\right) = 0. \tag{5.85}$$

•

Remark 5.140 — An alternative criterion for the WLLN (Rohatgi, 1976, p. 259)
Since condition (5.85) does not apply to the individual r.v. X_i Theorem 5.139 is of limited use. •

Exercise 5.141 — An alternative criterion for the WLLN
Show Theorem 5.139 (Rohatgi, 1976, pp. 258–259). •

Exercise 5.142 — An alternative criterion for the WLLN (bis)
Let $(X_1, \ldots, X_n)$ be jointly normal and such that: $E(X_i) = 0$ and $V(X_i) = 1$ $(i = 1, 2, \ldots)$; and,

$$cov(X_i, X_j) = \begin{cases} 1, & i = j \\ \rho \in (-1, 1), & |j - i| = 1 \\ 0, & |j - i| > 1. \end{cases} \tag{5.86}$$

Use Theorem 5.139 to prove that $\bar{X}_n \xrightarrow{P} 0$ (Rohatgi, 1976, pp. 259–260, Example 2). •

Finally, the assumption that the r.v. belong to L^2 is dropped and we state a theorem due to Soviet mathematician Aleksandr Yakovlevich Khinchin (1894–1959).

Theorem 5.143 — Weak law of large numbers, i.i.d. r.v. in L^1 (Rohatgi, 1976, p. 261)
Let $\{X_1, X_2, \ldots\}$ be a sequence of i.i.d. r.v. in L^1 with common finite mean μ.[17] Then $\{X_1, X_2, \ldots\}$ satisfies the WLLN with respect to $b_n = n$ ($a_n = n\mu$), i.e.

$$\bar{X}_n \xrightarrow{P} \mu. \tag{5.87}$$

•

Exercise 5.144 — Weak law of large numbers, i.i.d. r.v. in L^1
Prove Theorem 5.143 (Rohatgi, 1976, p. 261). •

Exercise 5.145 — Weak law of large numbers, i.i.d. r.v. in L^1 (bis)
Let $\{X_1, X_2, \ldots\}$ be a sequence of i.i.d. r.v. to $X \in L^k$, for some positive integer k, and common kth. moment $E(X^k)$. Apply Theorem 5.143 to prove that:

(a) $\frac{1}{n}\sum_{i=1}^{n} X_i^k \xrightarrow{P} E(X^k)$;[18]

(b) if $k = 2$ then $\frac{1}{n}\sum_{i=1}^{n} X_i^2 - (\bar{X}_n)^2 \xrightarrow{P} V(X)$ (Rohatgi, 1976, p. 261, Example 4).[19] •

Exercise 5.146 — Weak law of large numbers, i.i.d. r.v. in L^1 (bis bis)
Let $\{X_1, X_2, \ldots\}$ be a sequence of i.i.d. r.v. with common p.d.f.

$$f_X(x) = \begin{cases} \frac{1+\delta}{x^{2+\delta}}, & x \geq 1 \\ 0, & \text{otherwise,} \end{cases} \tag{5.88}$$

where $\delta > 0$.[20] Show that $\bar{X}_n \xrightarrow{P} E(X) = \frac{1+\delta}{\delta}$ (Rohatgi, 1976, p. 262, Example 5). •

[17] Please note that nothing is said about the variance, it need not to be finite.

[18] This means that the kth. sample moment, $\frac{1}{n}\sum_{i=1}^{n} X_i^k$, is a consistent estimator of $E(X^k)$ if the i.i.d. r.v. belong to L^k.

[19] I.e., the sample variance, $\frac{1}{n}\sum_{i=1}^{n} X_i^2 - (\bar{X}_n)^2$, is a consistent estimator of $V(X)$ if we are dealing with i.i.d. r.v. in L^2.

[20] This is the p.d.f. of a Pareto$(1, 1+\delta)$ r.v.

5.7 Strong law of large numbers

This section is devoted to a few extensions of the SLLN for Bernoulli summands (or Borel's SLLN), Theorem 5.113. They refer to sequences of:

- i.i.d. r.v. in L^4;
- dominated i.i.d. r.v.;
- independent r.v. in L^2 with a specific variance behavior;
- i.i.d. r.v. in L^1.

Theorem 5.147 — Strong law of large numbers, i.i.d. r.v. in L^4 (Karr, 1993, p. 152; Rohatgi, 1976, pp. 264–265, Theorem 1)
Let $\{X_1, X_2, \ldots\}$ be a sequence of i.i.d. r.v. in L^4, with common expected value μ. Then

$$\bar{X}_n \overset{a.s.}{\to} \mu. \tag{5.89}$$

•

Exercise 5.148 — Strong law of large numbers, i.i.d. r.v. in L^4
Prove Theorem 5.147, by following the same steps as in the proof of the SLLN for Bernoulli summands (Rohatgi, 1976, p. 265). •

The proviso of a common finite fourth moment can be dropped if there is a degenerate r.v. that dominates the i.i.d. r.v. $X_1, X_2, \ldots$

Corollary 5.149 — Strong law of large numbers, dominated i.i.d. r.v. (Rohatgi, 1976, p. 265)
Let $\{X_1, X_2, \ldots\}$ be a sequence of i.i.d. r.v., with common expected value μ and such that

$$P(\{|X_n| < k\}) = 1, \text{ for all } n, \tag{5.90}$$

where k is a positive constant. Then

$$\bar{X}_n \overset{a.s.}{\to} \mu. \tag{5.91}$$

•

The next lemma is essential to prove yet another extension of Borel's SLLN (Theorem 5.113).

Lemma 5.150 — Kolmogorov's inequality (Rohatgi, 1976, p. 268)
Let $\{X_1, X_2, \ldots\}$ be a sequence of independent r.v., with common null mean and variances σ_i^2, $i = 1, 2, \ldots$ Then, for any $\epsilon > 0$,

$$P\left(\left\{\max_{k=1,\ldots,n} |S_k| > \epsilon\right\}\right) \leq \frac{\sum_{i=1}^{n} \sigma_i^2}{\epsilon^2}. \tag{5.92}$$

•

Exercise 5.151 — Kolmogorov's inequality
Show Lemma 5.150 (Rohatgi, 1976, pp. 268–269). •

Remark 5.152 — Kolmogorov's inequality (Rohatgi, 1976, p. 269)
If we take $n = 1$ then Lemma 5.150 can be written as $P(\{|X_1| > \epsilon\}) \leq \frac{\sigma_1^2}{\epsilon^2}$, which is Chebyshev's inequality. •

The condition of dealing with i.i.d. r.v. in L^4 can be further relaxed as long as the r.v. are still independent and the variances of $X_1, X_2, \ldots$ have a specific behavior, as stated below.

Theorem 5.153 — Strong law of large numbers, independent r.v. in L^2 (Rohatgi, 1976, p. 272)
Let $\{X_1, X_2, \ldots\}$ be a sequence of independent r.v. in L^2 with variances σ_i^2, $i = 1, 2, \ldots$, such that

$$\sum_{i=1}^{+\infty} V(X_i) < +\infty. \tag{5.93}$$

Then

$$S_n - E(S_n) \overset{a.s.}{\to} 0. \tag{5.94}$$

•

Exercise 5.154 — Strong law of large numbers, independent r.v. in L^2
Prove Theorem 5.153 by making use of Kolmogorov's inequality and Cauchy's criterion (Rohatgi, 1976, p. 272). •

Theorem 5.155 — Strong law of large numbers, i.i.d. r.v. in L^1, or Kolmogorov's SLLN (Karr, 1993, p. 188; Rohatgi, 1976, p. 274, Theorem 7)
Let $\{X_1, X_2, \ldots\}$ be a sequence of i.i.d. r.v. to X. Then

$$\bar{X}_n \overset{a.s.}{\to} \mu \tag{5.95}$$

iff $X \in L^1$, and then $\mu = E(X)$. •

Exercise 5.156 — Strong law of large numbers, i.i.d. r.v. in L^1, or Kolmogorov's SLLN

Show Theorem 5.155 (Karr, 1993, pp. 188–189; Rohatgi, 1976, pp. 274–275). •

5.8 Characteristic functions

In probability theory, the characteristic function of any real-valued r.v. X:

- uniquely defines its probability distribution (Karr, 1993, p. 163);
- always exists when treated as a function of a real-valued argument, unlike the moment-generating function (http://en.wikipedia.org/wiki/Characteristic_function_(probability_theory)).

Furthermore:

1. the obtention of the characteristic function of a sum of independent r.v. is converted to the simpler operation of pointwise multiplication (Karr, 1993, p. 163) of the individual characteristic functions;
2. a sequence of r.v. converges in distribution iff the corresponding characteristic functions converge pointwise (Karr, 1993, p. 163).

Result 1. proves to be absolutely necessary to tackle the fairly complex problem of determining the distribution of a sum of independent r.v. (Resnick, 1999, p. 293). Result 2. plays an essential role in the rigorous proof of the Central Limit Theorem (Resnick, 1999, p. 295) and is the main reason to study the characteristic function in this chapter.

Before we proceed, let us remind the reader that, for a given complex number $z = x + iy$:

- the real part of z is $Re(z) = x$;
- the imaginary part of z is $Im(z) = y$;
- the complex conjugate of z is $\bar{z} = x - iy$;
- z is real iff $\bar{z} = z$;
- the modulus of z is $|z| = \sqrt{x^2 + y^2}$;
- $e^{it} = \cos(t) + i \sin(t)$ (Euler's formula).

Definition 5.157 — Characteristic function (Karr, 1993, p. 163; Resnick, 1999, p. 295)
The characteristic function of the real-valued r.v. X, with c.d.f. $F_X(x)$, is the complex valued function of a real variable t,

$$\varphi_X : I\!R \to \mathbb{C}, \tag{5.96}$$

defined as the expected value of e^{itX}:

$$\begin{aligned} \varphi_X(t) &= E(e^{itX}) \\ &= \int_{-\infty}^{\infty} e^{itx}\, dF_X(x). \end{aligned} \tag{5.97}$$

•

Remark 5.158 — Characteristic function (Resnick, 1999, p. 295)
By using Euler's formula, the characteristic function can be rewritten as

$$\varphi_X(t) = \int_{-\infty}^{\infty} \cos(tx)\, dF_X(x) + i \int_{-\infty}^{\infty} \sin(tx)\, dF_X(x). \tag{5.98}$$

•

Example 5.159 — Characteristic function (Karr, 1993, p. 164)
The characteristic functions of a few key discrete and absolutely continuous distributions:

Distribution of X	Characteristic function $\varphi_X(t)$
Bernoulli(p)	$1 - p + p\, e^{it}$
Binomial(n, p)	$(1 - p + p\, e^{it})^n$
c	e^{itc}
Geometric(p)	$\frac{p\, e^{it}}{1-(1-p)\, e^{it}}$
NegativeBinomial(r, p)	$\left[\frac{p\, e^{it}}{1-(1-p)\, e^{it}}\right]^r$
Poisson(λ)	$e^{\lambda(e^{it}-1)}$
Exponential(λ)	$\frac{\lambda}{\lambda - it}$
Gamma(α, λ)	$\left(\frac{\lambda}{\lambda - it}\right)^{\alpha}$
Normal(μ, σ^2)	$e^{\mu it - \frac{\sigma^2 t^2}{2}}$
Uniform(a, b)	$\frac{e^{itb} - e^{ita}}{it(b-a)}$

•

Exercise 5.160 — Characteristic function
Obtain the characteristic function of at least two discrete (resp. three absolutely continuous) distributions. •

The characteristic function of a real-valued r.v. X always exists because it is an integral of a bounded continuous function over a space whose measure is finite (http://en.wikipedia.org/wiki/Characteristic_function_(probability_theory)). In fact, by successively using Jensen's inequality, Euler's formula and the Pythagorean trigonometric identity, we get

$$\begin{aligned} |\varphi_X(t)| &= |E(e^{itX})| \\ &\leq E(|e^{itX}|) \\ &= E\left[\sqrt{\sin^2(tX)+\cos^2(tX)}\right] \\ &= 1. \end{aligned}$$

We now list other elementary properties of characteristic functions.

Proposition 5.161 — Elementary properties of characteristic functions (Karr, 1993, pp. 164–165; Resnick, 1999, pp. 296–297; http://en.wikipedia.org/wiki/Characteristic_function_(probability_theory))

1. The characteristic function $\varphi_X(t)$ is uniformly continuous on $\mathbb{R}$.
2. $\varphi_X(t)$ satisfies:
   (a) $\varphi_X(0) = 1$ (i.e., it is non-vanishing in a region around zero);
   (b) $\varphi_X(-t) = \overline{\varphi_X(t)}$ (that is, it is Hermitian).
3. The effect on $\varphi_X(t)$ of an affine transformation on X is given by
$$\varphi_{aX+b}(t) = \varphi_X(at) \times e^{ibt}, \tag{5.99}$$
   where $a, b \in \mathbb{R}$.
4. Let $\overline{\varphi_X(t)}$ be the complex conjugate of $\varphi_X(t)$. Then
$$\varphi_X(-t) = \overline{\varphi_X(t)} = \varphi_{-X}(t). \tag{5.100}$$
5. The real part of $\varphi_X(t)$, $Re[\varphi_X(t)] = \int_{-\infty}^{\infty} \cos(tx)\, dF_X(x)$, is an even function, i.e., $Re[\varphi_X(t)] = Re[\varphi_X(-t)]$.

6. The imaginary part of $\varphi_X(t)$, $Im[\varphi_X(t)] = \int_{-\infty}^{\infty} \sin(tx)\, dF_X(x)$, is an odd function, that is, $Im[\varphi_X(t)] = -Im[\varphi_X(-t)]$.

7. If X and Y are independent r.v. then

$$\varphi_{X+Y}(t) = \varphi_X(t) \times \varphi_Y(t). \tag{5.101}$$

8. The previous result can be generalise as follows. Let $X_1, \ldots, X_n$ be independent r.v. and $a_1, \ldots, a_n$ be real constants. Then the characteristic function of the linear combination of X_i's is equal to

$$\varphi_{a_1X_1+\ldots+a_nX_n}(t) = \varphi_{X_1}(a_1t) \times \cdots \times \varphi_{X_n}(a_nt). \tag{5.102}$$

9. Let: $X_1, \ldots, X_n$ be i.i.d. r.v. with common characteristic function $\varphi_X(t)$; $S_n = \sum_{i=1}^{n} X_i$; $a_n \neq 0$ and $b_n \in I\!R$. Then

$$\varphi_{a_n^{-1}S_n - nb_n}(t) = e^{-itnb_n} \times \left[\varphi_X(a_n^{-1}t)\right]^n. \tag{5.103}$$

   This property is a generalisation of Result 7. in a direction useful for the Central Limit Theorem.

10. If a r.v. X has a moment-generating function $M_X(t) = E(e^{tX})$, then the domain of the characteristic function can be extended to the complex plane, and $\varphi_X(-it) = M_X(t)$. •

Exercise 5.162 — Elementary properties of characteristic functions
Prove results:

(a) 1. (Karr, 1993, p. 164);

(b) 3.;

(c) 7. (Karr, 1993, p. 165);

(d) 9. •

For a brief account on criteria for characteristic functions the reader is referred to http://en.wikipedia.org/wiki/Characteristic_function_(probability_theory)#Criteria_for_characteristic_functions

The c.d.f. of the r.v. X can be obtained in terms of the characteristic function $\varphi_X(t)$, as stated in the next two results.

Theorem 5.163 — Inversion theorem (Karr, 1993, p. 166)
Let $a < b$ be two continuity points of the c.d.f. of the r.v. X. Then

$$\begin{aligned} P(a < X < b) &= F_X(b) - F_X(a) \\ &= \lim_{T \to +\infty} \frac{1}{2\pi} \int_{-T}^{T} \frac{e^{-ita} - e^{-itb}}{it} \times \varphi_X(t)\, dt. \end{aligned} \tag{5.104}$$

•

Remark 5.164 — Inversion theorem
(http://en.wikipedia.org/wiki/Characteristic_function_(probability_theory))
Let x be a continuity point of the c.d.f. of the r.v. X. Then

$$F_X(x) = \frac{1}{2} - \frac{1}{\pi} \int_0^{+\infty} \frac{Im[e^{-itx} \times \varphi_X(t)]}{t}\, dt. \tag{5.105}$$

•

Exercise 5.165 — Inversion theorem
Prove Theorem 5.163 by making use of the trigonometric identity

$$\int_0^{+\infty} \frac{\sin(\alpha x)}{x}\, dx = \text{sign}(\alpha) \times \frac{\pi}{2}$$

(Karr, 1993, pp. 166–167). •

We can recover not only the p.d.f. of an absolutely continuous r.v., but also the individual probabilities $P(X = x)$ using the characteristic function of X.

Theorem 5.166 — Fourier inversion theorem (Karr, 1993, p. 167; Resnick, 1999, p. 303)
If $\int_{-\infty}^{+\infty} |\varphi_X(t)|\, dt < \infty$ then X is an absolutely continuous r.v. with p.d.f. given by

$$f_X(x) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} e^{-itx} \times \varphi_X(t)\, dt \tag{5.106}$$

•

Exercise 5.167 — Fourier inversion theorem

(a) Prove Theorem 5.166 (Karr, 1993, p. 168).

(b) Derive the p.d.f. associated to the characteristic function $e^{-\frac{t^2}{2}}$ by applying Theorem 5.166 (Karr, 1993, p. 168). •

Proposition 5.168 — Inversion theorem (Karr, 1993, p. 168)
Let X be a real discrete r.v. and $\varphi_X(t)$ its characteristic function. Then

$$P(X = x) = \lim_{T \to +\infty} \frac{1}{2T} \int_{-T}^{T} e^{-itx} \times \varphi_X(t)\, dt, \tag{5.107}$$

for $x \in \mathbb{R}$. •

Exercise 5.169 — Inversion theorem
Prove Proposition 5.168 (Karr, 1993, p. 168). •

Interestingly enough, the p.f. of an integer-valued r.v. X can be also written in terms of $\varphi_X(t)$, as mentioned below.

Corollary 5.170 — Inversion theorem: integer-valued r.v. (Karr, 1993, p. 169)
Let X be an integer-valued r.v. Then

$$P(X = n) = \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{-int} \times \varphi_X(t)\, dt, \tag{5.108}$$

for $n \in \mathbb{Z}$. •

Exercise 5.171 — Inversion theorem: integer-valued r.v.

(a) Prove Corollary 5.170 (Karr, 1993, p. 169).

(b) Derive the p.f. of a Bernoulli(p) r.v., by using Corollary 5.170.

(c) Use *Mathematica* to obtain the p.f. of a Poisson(1) r.v., by using Corollary 5.170. •

Characteristic functions can also be used to find moments of a r.v. X provided that they exist. Furthermore, by verifying a simple condition, characteristic functions establish that the moments of X exist.

Theorem 5.172 — Calculation of moments known to exist (Karr, 1993, p. 169; Resnick, 1999, pp. 301–302)
Consider that the k^{th} absolute moment of a r.v. X exists, i.e., $E(|X|^k) < \infty$. Then $E(X^k)$ can be computed by taking $k-$fold derivatives of the characteristic function of X:

$$E(X^k) = i^{-k}\, \varphi_X^{(k)}(0) \tag{5.109}$$

$$= i^{-k} \left[\frac{d^k\, \varphi_X(t)}{dt^k} \right]_{t=0}, \tag{5.110}$$

for $k \in \mathbb{N}$. •

Exercise 5.173 — Calculation of moments known to exist

(a) Prove Theorem 5.172 (Karr, 1993, p. 169).

(b) Use Theorem 5.172 to derive the first and second moments of $X \sim \text{Normal}(0, 1)$ (Karr, 1993, p. 171). •

Theorem 5.174 — Establishing the existence of moments (Karr, 1993, p. 170)
Let k be an even positive integer and suppose $\varphi_X^{(k)}(0)$ exists. Then $E(|X|^k) < \infty$.[21] •

Remark 5.175 — Establishing the existence of moments (http://en.wikipedia.org/wiki/Characteristic_function_(probability_theory))
Let k be an odd positive integer. Then if a characteristic function φ_X has a k^{th} derivative at zero, then the r.v. X has all moments only up to $k - 1$. •

Exercise 5.176 — Establishing the existence of moments
Prove Theorem 5.174 (Karr, 1993, p. 170). •

The Taylor expansion of characteristic functions is crucial to prove some limit theorems (Karr, 1993, p. 171).

Theorem 5.177 — Taylor expansions of characteristic functions (Karr, 1993, p. 171)
If $E(|X|^k) < \infty$, for some integer $k \in \mathbb{N}$, then

$$\varphi_X(t) = \sum_{j=0}^{k} \frac{(it)^j}{j!} E(X^j) + o(|t|^k), \tag{5.111}$$

as $t \to 0$.[22] •

Remark 5.178 — Taylor expansions of characteristic functions (Resnick, 1999, p. 300)
If $E(|X|^k) < \infty$, for all $k \in \mathbb{N}$, then

$$\varphi_X(t) = \sum_{j=0}^{+\infty} \frac{(it)^j}{j!} E(X^j). \tag{5.112}$$

•

[21] Thus, all moments $E(X^j)$, $j = 1, \ldots, k$, exist.

[22] Recall that the relation $f(x) \in o(g(x))$ is read as "$f(x)$ is little-o of $g(x)$". Intuitively, it means that $g(x)$ grows much faster than $f(x)$, or similarly, the growth of $f(x)$ is nothing compared to that of $g(x)$ and $\lim_{x\to\infty} \frac{f(x)}{g(x)} = 0$.

The next theorem states that the characteristic function of a r.v. uniquely determines its distribution (Resnick, 1999, p. 302).

Theorem 5.179 — Uniqueness theorem (Karr, 1993, p. 167; Resnick, 1999, p. 302)
If $\varphi_X(t) = \varphi_Y(t)$, for all t, then $X \stackrel{d}{=} Y$. •

Exercise 5.180 — Uniqueness theorem
Use Theorem 5.163 to prove Theorem 5.179 (Karr, 1993, p. 167; Resnick, 1999, pp. 302–303). •

The next theorem allows us to conclude the convergence in distribution of a sequence of r.v. from the pointwise convergence of their characteristic functions and vice versa.

Theorem 5.181 — Continuity theorem (Karr, 1993, p. 171)
$X_n \stackrel{d}{\to} X$ iff

$$\varphi_{X_n}(t) \to \varphi_X(t), \text{ for each } t \in I\!R. \tag{5.113}$$

•

Exercise 5.182 — Continuity theorem
Prove Theorem 5.181 (Karr, 1993, pp. 171–172). •

The following result — the Lévy continuity theorem — establishes that the pointwise limit of a sequence of characteristic functions is a characteristic function, provided that it is continuous at zero (Karr, 1993, p. 172).

The Lévy continuity theorem is frequently used to prove the law of large numbers and the Central Limit Theorem.

Theorem 5.183 — Lévy continuity theorem (Karr, 1993, p. 172; Resnick, 1999, pp. 304–305)
Let $\{X_1, X_2, \ldots\}$ be a sequence of r.v. and $\varphi_{X_1}(t), \varphi_{X_2}(t), \ldots$ the corresponding characteristic functions. If

(i) $\varphi(t) = \lim_{n \to +\infty} \varphi_{X_n}(t)$ for every $t \in I\!R$

(ii) φ is continuous at zero

then there is a r.v. X such that

$$\varphi_X = \varphi \tag{5.114}$$

$$X_n \stackrel{d}{\to} X. \tag{5.115}$$

•

Exercise 5.184 — Lévy continuity theorem
Prove Theorem 5.183 by using the following result, stated and proved by Resnick (1999, p. 311): there is $K \in \mathbb{R}$ such that for each X,

$$P(|X| \geq a^{-1}) \leq \frac{K}{a} \int_0^a \{1 - Re[\varphi_X(t)]\}\, dt, \tag{5.116}$$

for all $a > 0$ (Karr, 1993, pp. 172–173; Resnick, 1999, 311–312). •

Exercise 5.185 — Continuity theorems
Use the continuity theorems and other results you may see fit to prove:

(a) the weak law of large numbers stated in Theorem 5.143 (Karr, 1993, pp. 173–174);

(b) the Poisson limit theorem stated in Theorem 5.124 (Karr, 1993, p. 174);

(c) $\sum_{i=1}^{+\infty} 2^{-i} X_i \stackrel{d}{=} \text{Uniform}(-1, 1)$, when the X_i are i.i.d. r.v. with common p.f. $P(X_i = -1) = P(X_i = 1) = \frac{1}{2}$ (Karr, 1993, p. 173). •

5.9 The Central Limit Theorem

The Central Limit Theorem (CLT) is probably the most notable case of convergence in distribution. It states that, given certain conditions, the sum (or the arithmetic mean) of a sufficiently large number of i.i.d. r.v., each with a well-defined expected value and well-defined variance, will be approximately normally distributed (http://en.wikipedia.org/wiki/Central_limit_theorem#Classical_CLT). This result is particularly important because, unlike the Binomial, Poisson and Normal distributions, most distributions are not closed under convolution and it is crucial to provide an approximate distribution for sums (or means) of r.v.

The CLT has several variants. The version we state below:

- refers to i.i.d. r.v.;
- is sometimes referred to as the Lindeberg-Lévy CLT (Murteira, 1979, p. 354);
- extends the DeMoivre-Laplace global limit theorem (Theorem 5.120), in which the S_n have binomial distributions (Karr, 1993, p. 174).

Theorem 5.186 — Lindeberg-Lévy Central Limit Theorem (or CLT for i.i.d. r.v.) (Resnick, 1999, p. 313; Karr, 1993, p. 174; Murteira, 1979, p. 354)
Let:

- $\{X_1, X_2, \ldots\}$ be a sequence of i.i.d. r.v. such that $E(X_i) = \mu$ and $V(X_i) = \sigma^2 \in \mathbb{R}^+$, for $i = 1, 2, \ldots$;
- $S_n = \sum_{i=1}^{n} X_i$ be the sum of the first n terms of that sequence of i.i.d. r.v.;
- $\{Z_1, Z_2, \ldots\}$ be the sequence of the standardized partial sums, where

$$\begin{aligned} Z_n &= \frac{S_n - E(S_n)}{\sqrt{V(S_n)}} \\ &= \frac{S_n - n\mu}{\sqrt{n\sigma^2}}. \end{aligned} \tag{5.117}$$

Then

$$Z_n \xrightarrow{d} \text{Normal}(0, 1). \tag{5.118}$$

•

Remark 5.187 — Lindeberg-Lévy Central Limit Theorem (or CLT for i.i.d. r.v.)

- This variant of the CLT allows us to add that, when we deal with a sufficiently large number n of i.i.d. r.v. $X_1, \ldots, X_n$, with common mean μ and common positive and finite variance σ^2, the c.d.f. of the sum of these r.v. can be approximate as follows:

$$P(S_n \leq s) = P\left(\frac{S_n - n\mu}{\sqrt{n\sigma^2}} \leq \frac{s - n\mu}{\sqrt{n\sigma^2}}\right) \overset{CLT}{\simeq} \Phi\left(\frac{s - n\mu}{\sqrt{n\sigma^2}}\right). \tag{5.119}$$

- Because of the continuity theorem, characteristic functions[23] are used in the most frequently seen proof of this version of the CLT. •

Exercise 5.188 — Lindeberg-Lévy Central Limit Theorem (or CLT for i.i.d. r.v.)
Prove Theorem 5.186 (Karr, 1993, p. 174; Murteira, 1979, pp. 354–355; Resnick, 1999, pp. 313–314). •

[23] And their Taylor expansions omitting terms of higher order than the 2^{nd} degree.

In the classical form of the CLT, the r.v. must be identically distributed. However, the CLT can be generalized to the case where the summands are independent r.v. but not identically distributed (Resnick, 1999, p. 314), given that they comply with certain conditions.

Interestingly, the next variant of the CLT is due to Lyapunov and was proved before the Lindeberg-Lévy CLT (Murteira, 1979, p. 359). The Lyapunov CLT requires that the r.v. $|X_i|$ have finite moments of some order $(2+\delta)$, $\delta > 0$, and that the rate of growth of these moments is limited by the Lyapunov condition given below.

Theorem 5.189 — Lyapunov Central Limit Theorem (Murteira, 1979, p. 359; Resnick, 1999, p. 319; Karr, 1993, p. 191)
Let:

- $\{X_1, X_2, \ldots\}$ be a sequence of independent r.v. such that $E(X_i) = \mu_i$ and $V(X_i) = \sigma_i^2$,[24] $i = 1, 2, \ldots$;
- $S_n = \sum_{i=1}^{n} X_i$ be the partial sum of the first n terms of that sequence of independent r.v.;
- $\{Z_1, Z_2, \ldots\}$ be the sequence of the standardized partial sums, where

$$\begin{aligned} Z_n &= \frac{S_n - E(S_n)}{\sqrt{V(S_n)}} \\ &= \frac{\sum_{i=1}^{n} X_i - \sum_{i=1}^{n} \mu_i}{\sqrt{\sum_{i=1}^{n} \sigma_i^2}}. \end{aligned} \tag{5.120}$$

Then

$$Z_n \xrightarrow{d} \text{Normal}(0,1) \tag{5.121}$$

if $\{X_1, X_2, \ldots\}$ satisfies the *Lyapunov condition*, i.e., if

$$\exists\, \delta > 0 \;\; : \;\; \begin{cases} E(|X_n|^{2+\delta}) < +\infty,\ n = 1, 2, \ldots \\ \lim_{n \to +\infty} \frac{1}{\left(\sum_{i=1}^{n} \sigma_i^2\right)^{2+\delta}} \sum_{i=1}^{n} E\left[|X_i - \mu_i|^{2+\delta}\right] = 0. \end{cases} \tag{5.122}$$

•

Exercise 5.190 — Lyapunov Central Limit Theorem
Prove Theorem 5.189 (Karr, 1993, p. 192). •

[24] These variances are all finite because the sequence of of r.v. satisfies the Lyapunov condition. Moreover, Murteira (1979, p. 359) mentions that $\sigma_1 \neq 0$; we strongly believe this condition should read as follows: at least one of the variances should be non null.

Theorem 5.191 — Lindeberg-Feller Central Limit Theorem (Karr, 1993, p. 194; Murteira, 1979, p. 360; Resnick, 1999, p. 315)
Let:

- $\{X_1, X_2, \ldots\}$ be a sequence of independent r.v. such that $E(X_i) = \mu_i$ and $V(X_i) = \sigma_i^2$,[25] $i = 1, 2, \ldots$;
- $S_n = \sum_{i=1}^{n} X_i$ be the partial sum of the first n terms of that sequence of independent r.v.;
- $\{Z_1, Z_2, \ldots\}$ be the sequence of the standardized partial sums, where $Z_n = \frac{S_n - E(S_n)}{\sqrt{V(S_n)}} = \frac{\sum_{i=1}^{n} X_i - \sum_{i=1}^{n} \mu_i}{\sqrt{\sum_{i=1}^{n} \sigma_i^2}}$.

Then

$$Z_n \xrightarrow{d} \text{Normal}(0, 1) \tag{5.123}$$

and

$$\lim_{n \to +\infty} \max_{k=1,\ldots,n} \frac{\sigma_k^2}{V(S_n)} = 0, \tag{5.124}$$

iff $\{X_1, X_2, \ldots\}$ satisfies the *Lindeberg condition*, that is, if

$$\lim_{n \to +\infty} \frac{1}{V(S_n)} \sum_{k=1}^{n} \int_{|x - \mu_k| > \epsilon V(S_n)} (x - \mu_k)^2 \, dF_{X_k}(x) = 0. \tag{5.125}$$

•

Remark 5.192 — Lindeberg-Feller Central Limit Theorem

- The Lindeberg condition is not by itself a necessary condition for the validity of the CLT (Karr, 1993, p. 196).[26]
- Lindeberg (resp. Feller) proved the necessary (resp. sufficient) part of the Lindeberg-Feller CLT (Murteira, 1979, p. 360).

[25] Once again these variances are all finite (Murteira, 1979, p. 360) and at least one of them should be non null. Curiously, Resnick (1999, p. 314) does not mention these conditions on the variances.

[26] For instance, if $X_i \sim \text{Normal}(0, 2^i)$, then $V(S_n) = 2^{n+1} - 1 \simeq 2^{n+1}$ and $\lim_{n \to +\infty} \max_{k=1,\ldots,n} \frac{\sigma_k^2}{V(S_n)} = 2 \neq 0$. In this case, (5.124) fails and so does the Lindeberg condition, even though $S_n / \sqrt{V(S_n)} \sim \text{Normal}(0, 1)$, for all n (Karr, 1993, p. 196). However, once we stipulate that $\lim_{n \to +\infty} \max_{k=1,\ldots,n} \frac{\sigma_k^2}{V(S_n)} = 0$ the Lindeberg conditions is necessary: if $X_1, X_2, \ldots$ are independent r.v., with $\lim_{n \to +\infty} \max_{k=1,\ldots,n} \frac{\sigma_k^2}{V(S_n)} = 0$, and if $Z_n \xrightarrow{d} \text{Normal}(0, 1)$, then $\{X_1, X_2, \ldots\}$ satisfies the Lindeberg condition (Karr, 1993, Theorem 7.18, p. 196).

- The Lindeberg condition essentially means that, for each k, most of the mass of X_k is centered in an interval about the mean μ_k and this interval is small when compared to $V(S_n)$ (Resnick, 1999, p. 315).

- If the sequence of r.v. $\{X_1, X_2, \ldots\}$ satisfies the Lyapunov condition then it also satisfies the Lindeberg condition (Karr, 1993, p. 193). •

Exercise 5.193 — Lindeberg-Feller Central Limit Theorem
Prove Theorem 5.191 (Karr, 1993, pp. 194–196). •

Finally, note that characteristic functions can be extended to random vectors (http://en.wikipedia.org/wiki/Characteristic_function_(probability_theory)#Generalizations) and, unsurprisingly, the CLT has a multivariate variant. In fact, when we deal with a sequence of i.i.d. random vectors in $\mathbb{R}^k$, $\{\underline{X}_1, \underline{X}_2, \ldots\}$, with mean vector $\underline{\mu} = [E(X_i)]_{i=1,\ldots,k}$ and covariance matrix $\underline{\Sigma} = [cov(X_i, X_j)]_{i,j=1,\ldots,k}$, and take componentwise summations of these vectors, then the multidimensional CLT states that when scaled, the sequence of the resulting vectors converges to a multivariate normal distribution (http://en.wikipedia.org/wiki/Central_limit_theorem#Multidimensional_CLT).

5.10 The law of the iterated logarithm

It is important to determine the growth rate of the partial sums S_n: that rate is $\sqrt{2n\ \sigma^2\ \ln[\ln(n)]}$, thus the name "law of the iterated logarithm" (Karr, 1993, p. 196). According to http://en.wikipedia.org/wiki/Law_of_the_iterated_logarithm, the original statement of the law of the iterated logarithm is due to A.Y. Khinchin (1924); another statement was given by A.N. Kolmogorov in 1929.

Karr (1993, pp. 197–200) only proves this result when the summands are i.i.d. and have standard normal distribution.

Theorem 5.194 — Law of the iterated logarithm, i.i.d. summands with standard normal distribution (Karr, 1993, p. 198)
Let:

- $\{X_1, X_2, \ldots\}$ be a sequence of i.i.d. r.v. with Normal$(0, 1)$ distribution;

- $S_n = \sum_{i=1}^{n} X_i$ be the partial sum of the first n terms of that sequence of i.i.d. r.v.

Then

$$\limsup_{n\to+\infty} \frac{S_n}{\sqrt{2n\ \ln[\ln(n)]}} = \lim_{n\to+\infty} \sup_{m\geq n} \frac{S_m}{\sqrt{2m\ \ln[\ln(m)]}} \overset{a.s.}{=} 1. \tag{5.126}$$

•

Exercise 5.195 — Law of the iterated logarithm, standard normal and i.i.d. summands

Prove Theorem 5.194 Karr (1993, pp. 198–200). •

Theorem 5.196 — Law of the iterated logarithm, i.i.d. summands (Karr, 1993, p. 200)

Let:

- $\{X_1, X_2, \ldots\}$ be a sequence of i.i.d. r.v. such that $E(X_i) = \mu$ and $V(X_i) = \sigma^2 \in \mathbb{R}^+$, $i = 1, 2, \ldots$;
- $S_n = \sum_{i=1}^{n} X_i$ be the partial sum of the first n terms of that sequence of i.i.d. r.v.

Then

$$\limsup_{n\to+\infty} \frac{S_n - n\mu}{\sqrt{2n\ \sigma^2\ \ln[\ln(n)]}} \overset{a.s.}{=} 1. \tag{5.127}$$

•

Remark 5.197 — Law of the iterated logarithm, i.i.d. summands (http://en.wikipedia.org/wiki/Law_of_the_iterated_logarithm)

Let $\{X_1, X_2, \ldots\}$ be a sequence of i.i.d. r.v. such that $E(X_i) = 0$ and $V(X_i) = 1$, $i = 1, 2, \ldots$, and S_n the associated partial sum.

On one hand,

$$\bar{X}_n = \frac{S_n}{n} \overset{P}{\to} 0 \quad (\text{resp. } \overset{a.s.}{\to} 0), \tag{5.128}$$

according to the weak (resp. strong) law of large numbers. On the other hand,

$$\frac{S_n}{\sqrt{n}} \overset{d}{\to} \text{Normal}(0, 1), \tag{5.129}$$

by the CLT. Thus, the law of iterated logarithm operates "in between" the law of large numbers and the central limit theorem. •

5.11 Applications of the limit theorems

Monte Carlo integration, the characterisation of maximum likelihood estimators (MLE) and empirical distribution functions benefit from the strong law of large numbers, central limit theorem and the law of the iterated logarithm (Karr, 1993, pp. 200–207).

Theorem 5.198 — Monte Carlo integration and the strong law of large numbers (Karr, 1993, p. 201)
Let:

- h be a continuous (or even just Borel measurable) function on $[0,1]$ and such that $\int_0^1 |h(x)|\,dx < \infty$;
- $\{U_1, U_2, \ldots\}$ be a sequence of i.i.d. r.v. with the same distribution as $U \sim \text{Uniform}(0,1)$.

Then $\frac{1}{n}\sum_{i=1}^n h(U_i)$, the Monte Carlo estimator of $E[h(U)] = \int_0^1 h(x)\,dx$, satisfies

$$\frac{1}{n}\sum_{i=1}^{n} h(U_i) \overset{a.s.}{\to} \int_0^1 h(x)\,dx, \tag{5.130}$$

that is, $\frac{1}{n}\sum_{i=1}^n h(U_i)$ is a strongly consistent estimator of $\int_0^1 h(x)\,dx$ •

Since $\overset{a.s.}{\to} \Rightarrow \overset{P}{\to}$, we can add that $\frac{1}{n}\sum_{i=1}^n h(U_i)$ is a (weakly) consistent estimator of $\int_0^1 h(x)\,dx$.

Exercise 5.199 — Monte Carlo integration and the strong law of large numbers
Prove Theorem 5.198 (Karr, 1993, p. 201). •

Under widely satisfied conditions, maximum likelihood estimators are not only consisted, but also asymptotically normal (Karr, 1003, pp. 201–202).

Theorem 5.200 — Maximum likelihood estimation and the weak law of large numbers (Karr, 1993, pp. 201–202)
Let:

- $\{X_1, X_2, \ldots\}$ be a sequence of i.i.d. r.v. with the same p.d.f. (or p.f.) $f(.,\theta)$ as the r.v. X;
- $\theta \in \mathbb{R}$ is an unknown parameter we wish to estimate;
- $\hat{\theta}_n = \hat{\theta}_n(X_1, \ldots, X_n)$ be the MLE of θ based on the random sample of size n, $(X_1, \ldots, X_n)$.

Suppose:

- the mapping $\theta \to f(x, \theta)$ is continuous for (almost) every $x \in \mathbb{R}$;
- for each θ and $\gamma > 0$,

$$k_\theta(\gamma) = \inf_{|\theta' - \theta| > \gamma} \int_{-\infty}^{+\infty} \left[\sqrt{f(x, \theta)} - \sqrt{f(x, \theta')} \right]^2 dx > 0; \tag{5.131}$$

- for each θ,

$$\lim_{\delta \to 0} \left\{ \int_{-\infty}^{+\infty} \sup_{|h| \leq \delta} \left[\sqrt{f(x, \theta)} - \sqrt{f(x, \theta + h)} \right]^2 dx \right\}^{\frac{1}{2}} = 0; \tag{5.132}$$

- for each θ,

$$\lim_{c \to +\infty} \int_{-\infty}^{+\infty} \sup_{|u| \geq c} \left[\sqrt{f(x, \theta)} \times \sqrt{f(x, \theta + u)} \right]^2 dx < 1. \tag{5.133}$$

Then $\hat{\theta}_n$ is a consistent estimator of θ, i.e.,

$$\hat{\theta}_n \xrightarrow{P} \theta. \tag{5.134}$$

•

Exercise 5.201 — **Maximum likelihood estimation and limit theorems**
Prove Theorem 5.200 (Karr, 1993, pp. 202–204). •

Theorem 5.202 — **Maximum likelihood estimation and the CLT** (Karr, 1993, p. 204)
Under the conditions of Theorem 5.200 and the finiteness of the *Fisher information*,

$$I(\theta) = E\left[\left(\frac{\partial \ln f(X, \theta)}{\partial \theta} \right)^2 \right], \tag{5.135}$$

we get the asymptotic normality of the standardised estimation error:

$$\sqrt{n\, I(\theta)} [\hat{\theta}_n - \theta] \xrightarrow{d} \text{Normal}(0, 1). \tag{5.136}$$

•

Exercise 5.203 — **Maximum likelihood estimation and the CLT**
Prove Theorem 5.202 (Karr, 1993, pp. 204–205). •

Proposition 5.204 — Empirical distribution functions and the strong law of large numbers
Let:

- $\{X_1, X_2, \ldots\}$ be a sequence of i.i.d. r.v. with the same *entirely unknown* c.d.f. F as the r.v. X;
- $F_n(x, \underline{X}) = \frac{1}{n}\sum_{i=1}^{n} I_{(-\infty,x]}(X_i)$, $x \in \mathbb{R}$, be the empirical distribution function for the random sample $\underline{X} = (X_1, \ldots, X_n)$;[27]

Not only

$$P[F_n(x, \underline{X}) = s] = \frac{n!}{(ns)!\,(n-ns)!} \times [F(x)]^{ns} \times [1 - F(x)]^{n-ns}, \tag{5.137}$$

for $s = 0, \frac{1}{n}, \frac{2}{n}, \ldots, \frac{n-1}{n}, 1$,

$$E[F_n(x, \underline{X})] = F(x), \tag{5.138}$$

$$V[F_n(x, \underline{X})] = \frac{F(x)\,[1 - F(x)]}{n}, \tag{5.139}$$

but more important

$$F_n(x, \underline{X}) \xrightarrow{a.s.} F(x), \tag{5.140}$$

that is, $F_n(x, \underline{X})$ is a strongly consistent estimator of $F(x)$. •

This convergence is also uniform. This result is also known as the Glivenko-Cantelli theorem.

Theorem 5.205 — Glivenko-Cantelli theorem
Under the conditions of Proposition 5.204, we have

$$\forall \epsilon > 0, \lim_{n \to +\infty} P[\sup_{x \in \mathbb{R}} |F_n(x, \underline{X}) - F(x)| < \epsilon] = 1, \tag{5.141}$$

i.e.,

$$\sup_{x \in \mathbb{R}} |F_n(x, \underline{X}) - F(x)| \xrightarrow{a.s.} 0. \tag{5.142}$$

•

Suffice to say that we could have applied the CLT and conclude that

$$\frac{F_n(x, \underline{X}) - F(x)}{\sqrt{\frac{F(x)[1-F(x)]}{n}}} \xrightarrow{d} \text{Normal}(0, 1). \tag{5.143}$$

[27] $F_n(x, \underline{X})$ corresponds to the proportion X_i's smaller than or equal to x in the random sample $(X_1, \ldots, X_n)$.

Expectedly, $F_n(x, \underline{X})$ is used in the statistic of the Kolmogorov-Smirnov goodness of fit of test, $\sup_{x \in I\!R} |F_n(x, \underline{X}) - F_0(x)|$, where F_0 represents the conjectured (and known) distribution. Interestingly enough, for any absolutely continuous c.d.f. F, it is possible to:

- provide an asymptotic distribution for $\sqrt{n} \sup_{x \in I\!R} |F_n(x, \underline{X}) - F(x)|$ — this result constitutes the Kolmogorov-Smirnov theorem;
- state a law of the iterated logarithm for empirical distribution functions.

Theorem 5.206 — **Kolmogorov-Smirnov theorem** (Karr, 1993, pp. 206–207)
Under the conditions of Proposition 5.204 and an absolutely continuous c.d.f. F, we have

$$\sqrt{n} \sup_{x \in I\!R} |F_n(x, \underline{X}) - F(x)| \xrightarrow{d} Y, \tag{5.144}$$

where the c.d.f. of Y is given by

$$F_Y(y) = 1 - 2 \sum_{i=1}^{\infty} (-1)^{i+1} e^{-2i^2 y^2}, \; y > 0. \tag{5.145}$$

•

Theorem 5.207 — **Law of iterated logarithm for empirical distribution functions** (Karr, 1993, p. 207)
Under the conditions of Proposition 5.206

$$\limsup_{n \to +\infty} \frac{\sqrt{n} \sup_{x \in I\!R} |F_n(x, \underline{X}) - F(x)|}{\sqrt{2 \times \{\sup_{x \in I\!R} F(x)[1 - F(x)]\} \times \ln[\ln(n)]}} \overset{a.s.}{=} 1. \tag{5.146}$$

•

References

- Grimmett, G.R. and Stirzaker, D.R. (2001). *Probability and Random Processes* (3rd. edition). Oxford University Press. (QA274.12-.76.GRI.30385 and QA274.12-.76.GRI.40695 refer to the library code of the 1st. and 2nd. editions from 1982 and 1992, respectively.)

- Karr, A.F. (1993). *Probability*. Springer-Verlag.

- Murteira, B.J.F. (1979). *Probabilidades e Estatística*, Vol. 1. Editora McGraw-Hill de Portugal, Lda.

- Murteira, B.J.F. (1980). *Probabilidades e Estatística*, Vol. 2. Editora McGraw-Hill de Portugal, Lda. (QA273-280/3.MUR.34472, QA273-280/3.MUR.34475)

- Resnick, S.I. (1999). *A Probability Path*. Birkhäuser. (QA273.4-.67.RES.49925)

- Rohatgi, V.K. (1976). *An Introduction to Probability Theory and Mathematical Statistics*. John Wiley & Sons. (QA273-280/4.ROH.34909)

- Walrand, J. (2004). *Lecture Notes on Probability Theory and Random Processes*. Department of Electrical Engineering and Computer Sciences, University of California, Berkeley.