Appendix

A.1 Probability Factsheet

This section provides a snapshot of some of the main probabilistic concepts and properties that are made use of in the text. For a more detailed coverage, the reader is referred to Knight [14], Chaps. 1–3.

Events

A random experiment is a process whose outcome is uncertain. The possible outcomes, and combinations thereof, are described in the language of set theory. In principle, any statement that makes reference to the outcome of a random experiment should be expressible via this language. In detail:

- A possible outcome ω of a random experiment is called an elementary event.
- The set of all possible outcomes, say Ω is assumed non-empty, $\Omega \neq \emptyset$.
- An event is a subset $F \subset \Omega$ of Ω. An event F "is realised" (or "occurs") whenever the outcome of the experiment is an element of F.
- The union of two events F_1 and F_2, written $F_1 \cup F_2$ occurs if and only if either of F_1 or F_2 occurs. Equivalently, $\omega \in F_1 \cup F_2$ if and only if $\omega \in F_1$ or $\omega \in F_2$,

$$F_1 \cup F_2 = \{\omega \in \Omega : \omega \in F_1 \text{ or } \omega \in F_2\}$$

- The intersection of two events F_1 and F_2, written $F_1 \cap F_2$ occurs if and only if both F_1 and F_2 occur. Equivalently, $\omega \in F_1 \cap F_2$ if and only if $\omega \in F_1$ and $\omega \in F_2$,

$$F_1 \cap F_2 = \{\omega \in \Omega : \omega \in F_1 \text{ and } \omega \in F_2\}$$

- Unions and intersections of several events, $F_1 \cup \ldots \cup F_n$ and $F_1 \cap \ldots \cap F_n$ are defined iteratively from the definition for unions and intersections of pairs.

V.M. Panaretos, *Statistics for Mathematicians*, Compact Textbooks in Mathematics,
DOI 10.1007/978-3-319-28341-8

- The complement of an event F, denoted F^c, contains all the elements of Ω that are not contained in F,

$$F^c = \{\omega \in \Omega : \omega \notin F\}.$$

- Two events F_1 and F_2 are called disjoint if they contain no common elements, that is $F_1 \cap F_2 = \emptyset$.
- A partition $\{F_n\}_{n\geq 1}$ of Ω is a collection of events such that $F_i \cap F_j = \emptyset$ for all $i \neq j$, and $\cup_{n\geq 1} F_n = \Omega$.
- The difference of two events F_1 and F_2 is defined as $F_1 \backslash F_2 = F_1 \cap F_2^c$. It contains all the elements of F_1 that are not contained in F_2. Notice that the difference is not symmetric: $F_1 \setminus F_2 \neq F_2 \setminus F_1$.
- It can be checked that the following properties hold true
  (i) $(F_1 \cup F_2) \cup F_3 = F_1 \cup (F_2 \cup F_3) = F_1 \cup F_2 \cup F_3$
  (ii) $(F_1 \cap F_2) \cap F_3 = F_1 \cap (F_2 \cap F_3) = F_1 \cap F_2 \cap F_3$
  (iii) $F_1 \cap (F_2 \cup F_3) = (F_1 \cap F_2) \cup (F_1 \cap F_3)$
  (iv) $F_1 \cup (F_2 \cap F_3) = (F_1 \cup F_2) \cap (F_1 \cup F_3)$
  (v) $(F_1 \cup F_2)^c = F_1^c \cap F_2^c$ and $(F_1 \cap F_2)^c = F_1^c \cup F_2^c$

Probability Axioms

A probability measure $\mathbb{P}$ is a real function defined over the events of Ω, assigning a probability to any event. This can be interpreted as a measure of how certain we are that the event will occur. It is postulated to satisfy the following properties:

1. $\mathbb{P}(F) \geq 0$, for all events F.
2. $\mathbb{P}(\Omega) = 1$.
3. If $\{F_n\}_{n\geq 1}$ are disjoint events, and $F = \cup_{n\geq 1} F_n$ is an event given by their union, then

$$\mathbb{P}(F) = \sum_{n\geq 1} \mathbb{P}(F_n).$$

The following properties are immediate consequences of the probability axioms:

- $\mathbb{P}(F^c) = 1 - \mathbb{P}(F)$.
- $\mathbb{P}(F_1 \cap F_2) \leq \min\{\mathbb{P}(F_1), \mathbb{P}(F_2)\}$.
- $\mathbb{P}(F_1 \cup F_2) = \mathbb{P}(F_1) + \mathbb{P}(F_2) - \mathbb{P}(F_1 \cap F_2)$.
- Continuity from below: let $\{F_n\}_{n\geq 1}$ be nested events, such that $F_j \subseteq F_{j+1}$ for all j, and let F be an event given by $F = \cup_{n\geq 1} F_n$. Then $\mathbb{P}(F_n) \overset{n\to\infty}{\longrightarrow} \mathbb{P}(F)$.
- Continuity from above: let $\{F_n\}_{n\geq 1}$ be nested events, such that $F_j \supseteq F_{j+1}$ for all j, and let F be an event given by $F = \cap_{n\geq 1} F_n$. Then $\mathbb{P}(F_n) \overset{n\to\infty}{\longrightarrow} \mathbb{P}(F)$.
- If $\Omega = \{\omega_1, \ldots, \omega_K\}$, $K < \infty$, is a finite set, then for any event $F \subseteq \Omega$, we have $\mathbb{P}(F) = \sum_{j:\omega_j \in F} \mathbb{P}(\omega_j)$.

Conditional Probability and Independence

Suppose we do not know the precise outcome $\omega \in \Omega$ that has occurred, but we are told that $\omega \in F_2$ for some event F_2. If we are asked to now calculate the probability that $\omega \in F_1$ also, for some other event F_1, then we need to calculate the conditional probability of F_1 given F_2.

- For any pair of events F_1, F_2 such that $\mathbb{P}(F_2) > 0$, we define the conditional probability of F_1 given F_2 to be

$$\mathbb{P}(F_1|F_2) = \frac{\mathbb{P}(F_1 \cap F_2)}{\mathbb{P}(F_2)}.$$

- Let G be an event and $\{F_n\}_{n\geq 1}$ be a partition of Ω such that $\mathbb{P}(F_n) > 0$ for all n. We then have:
  - Law of total probability:

$$\mathbb{P}(G) = \sum_{n=1}^{\infty} \mathbb{P}(G|F_n)\mathbb{P}(F_n)$$

  - Bayes' theorem:

$$\mathbb{P}(F_j|G) = \frac{\mathbb{P}(F_j \cap G)}{\mathbb{P}(G)} = \frac{\mathbb{P}(G|F_j)\mathbb{P}(F_j)}{\sum_{n=1}^{\infty} \mathbb{P}(G|F_n)\mathbb{P}(F_n)}$$

- The events $\{G_n\}_{n\geq 1}$ are called independent if and only if for any finite sub-collection $\{G_{i_1}, \ldots, G_{i_K}\}$, $K < \infty$, we have:

$$\mathbb{P}(G_{i_1} \cap \cdots \cap G_{i_K}) = \mathbb{P}(G_{i_1}) \times \mathbb{P}(G_{i_2}) \times \ldots \times \mathbb{P}(G_{i_K})$$

Random Variables and Distribution Functions

Random variables are, simply stated, numerical summaries of the outcome of a random experiment. Since the result is random, such numerical summaries are random, too. They allow us to not worry too much about the precise structure of the outcome $\omega \in \Omega$, but concentrate on a numerical summary instead. If that numerical summary is all we really care about, we can concentrate on the range of a random variable X, rather than consider Ω itself.

- A random variable is a real function $X : \Omega \to \mathbb{R}$.
- We write $\{a \leq X \leq b\}$ to denote the event

$$\{\omega \in \Omega : a \leq X(\omega) \leq b\}.$$

More generally, if $A \subset \mathbb{R}$ is a more general subset, we write $\{X \in A\}$ to denote the event

$$\{\omega \in \Omega : X(\omega) \in A\}.$$

- If we have a probability measure defined on the events of Ω, then X induces a new probability measure on subsets of the real line. This is described by the distribution function (or cumulative distribution function) $F_X : \mathbb{R} \to [0, 1]$ of a random variable X (or the law of X). This is defined as

$$F_X(x) = \mathbb{P}(X \leq x).$$

- By its definition, a distribution function satisfies the following properties:
  (i) $x \leq y \Rightarrow F_X(x) \leq F_X(y)$
  (ii) $\lim_{x\to\infty} F_X(x) = 1$, $\lim_{x\to-\infty} F_X(x) = 0$
  (iii) $\lim_{y\downarrow x} F_X(y) = F_X(x)$, that is, F_X is right-continuous.
  (iv) $\lim_{y\uparrow x} F_X(y)$ exists, that is, F_X is left-limited.
  (v) $\mathbb{P}(a < X \leq b) = F_X(b) - F_X(a)$.
  (vi) $\mathbb{P}(X > a) = 1 - F(a)$.
  (vii) Let $D_X := \{x \in \mathbb{R} : F_X(x) - \lim_{y\uparrow x} F_X(y) > 0\}$ be the set of points where F_X is not continuous.
    – D_X is a countable set (Lemma A.11, p. 169).
    – If $\mathbb{P}(\{X \in D_F\}) = 1$, then X is called a *discrete* random variable (equivalently, X has a finite or countable range, with probability 1).
    – If $D_X = \emptyset$, then X is called a *continuous* random variable (the distribution function F_X is continuous).
    – It may very well happen that a random variable may be neither discrete nor continuous.

Probability Density and Probability Mass Functions

- The probability mass function (or frequency function) $f_X : \mathbb{R} \to [0, 1]$ of a discrete random variable X is defined as

$$f_X(x) = \mathbb{P}(X = x).$$

By its definition, a probability mass function satisfies
(i) $\mathbb{P}(X \in A) = \sum_{t\in A\cap\mathcal{X}} f_X(t)$, for $A \subseteq \mathbb{R}$ and $\mathcal{X} = \{x \in \mathbb{R} : f_X(x) > 0\}$.
(ii) $F_X(x) = \sum_{t\in(-\infty,x]\cap\mathcal{X}} f_X(t)$, for all $x \in \mathbb{R}$ and $\mathcal{X} = \{x \in \mathbb{R} : f_X(x) > 0\}$.
(iii) An immediate corollary is that $F_X(x)$ is piecewise constant with jumps at the points in $\mathcal{X} = \{x \in \mathbb{R} : f_X(x) > 0\}$.

- A continuous random variable X has probability density function $f_X : \mathbb{R} \to [0, +\infty)$ if

$$F_X(b) - F_X(a) = \int_a^b f_X(t)dt.$$

for all real numbers $a < b$. By its definition, a probability density satisfies

(i) $F_X(x) = \int_{-\infty}^{x} f_X(t)dx$

(ii) $f_X(x) = F_X'(x)$, whenever f_X is continuous at x.

(iii) Note that $f_X(x) \neq \mathbb{P}(X = x) = 0$. In fact, it can be $f(x) > 1$ for some x. It can even happen that f is unbounded.

Random Vectors and Joint Distributions

A random vector $\mathbf{X} = (X_1, \ldots, X_d)^\top$ is a finite collection of random variables, arranged as the coordinates of a vector. The point is that we may want to make probabilistic statements on the joint behaviour of all these random variables. In this case, we need to define their joint distribution, and respective joint density (or joint frequency).

- The joint distribution function of a random vector $\mathbf{X} = (X_1, \ldots, X_d)^\top$ is defined as:

$$F_{\mathbf{X}}(x_1, \ldots, x_d) = \mathbb{P}(X_1 \leq x_1, \ldots, X_d \leq x_d).$$

- Correspondingly, one defines the
  - joint frequency function, if the $\{X_i\}_{i=1}^d$ are all discrete,

$$f_{\mathbf{X}}(x_1, \ldots, x_d) = \mathbb{P}(X_1 = x_1, \ldots, X_d = x_d).$$

  - the joint density function, if there exists $f_{\mathbf{X}} : \mathbb{R}^d \to [0, +\infty)$ such that:

$$F_{\mathbf{X}}(x_1, \ldots, x_d) = \int_{-\infty}^{x_1} \ldots \int_{-\infty}^{x_d} f_{\mathbf{X}}(u_1, \ldots, u_d)du_1 \ldots du_d$$

In this case, when $f_{\mathbf{X}}$ is continuous at the point $\mathbf{x}$,

$$f_{\mathbf{X}}(x_1, \ldots, x_d) = \frac{\partial^d}{\partial x_1 \ldots \partial x_d} F_{\mathbf{X}}(x_1, \ldots, x_d)$$

Marginal Distributions

Given the joint distribution of the random vector $\mathbf{X} = (X_1, \ldots, X_d)^\top$, we can always isolate the distribution of a single coordinate, say X_i.

- In the discrete case, the marginal frequency function of X_i is given by $f_{X_i} : \mathbb{R} \to [0, +\infty)$:

$$f_{X_i}(x_i) = \mathbb{P}(X_i = x_i) = \sum_{x_1} \cdots \sum_{x_i-1} \sum_{x_i+1} \cdots \sum_{x_d} f_{\boldsymbol{X}}(x_1, \ldots, x_{i-1}, x_i, x_{i+1}, \ldots, x_d)$$

- In the continuous case, the marginal density function of X_i is given by $f_{X_i} : \mathbb{R} \to [0, +\infty)$:

$$f_{X_i}(x_i) = \int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} f_{\boldsymbol{X}}(y_1, \ldots, y_{i-1}, x_i, y_{i+1}, \ldots, y_d) dy_1 \ldots dy_{i-1} dy_{i+1} dy_d.$$

- More generally, we can define the joint frequency/density of a random vector formed by a subset of the coordinates of $\mathbf{X} = (X_1, \ldots, X_d)^\top$, say the first k (with $k < d$), $(X_1, \ldots, X_k)^\top$, via
  - Discrete case: $f_{X_1,\ldots,X_k}(x_1, \ldots, x_k) = \sum_{x_{k+1}} \cdots \sum_{x_d} f_{\boldsymbol{X}}(x_1, \ldots, x_k, x_{k+1}, \ldots, x_d)$.
  - Continuous case $f_{X_1,\ldots,X_k}(x_1, \ldots, x_k) = \int_{-\infty}^{+\infty} \ldots \int_{-\infty}^{+\infty} f_{\boldsymbol{X}}(x_1, \ldots, x_k, x_{k+1}, \ldots, x_d) dx_{k+1} \ldots dx_d$.
- In other words, in order to find a marginal density/frequency of a subset of random variables, we need to integrate/sum out the remaining variables from the overall joint density/frequency.
- It is important to note that the marginals do not uniquely determine the joint distribution.

Conditional Distributions

Similarly to the notion of conditional probability, we may wish to make probabilistic statements about the potential outcomes of one random variable, if we already know the outcome of another. For this we need the notion of conditional density and conditional frequency functions. If $(X_1, \ldots, X_d)$ is a continuous/discrete random vector, we define the conditional probability density/frequency function of $(X_1, \ldots, X_k)$ given $\{X_{k+1} = x_{k+1}, \ldots, X_d = x_d\}$ as

$$f_{X_1,\ldots,X_k|X_{k+1},\ldots,X_d}(x_1, \ldots, x_k | x_{k+1}, \ldots, x_d) = \frac{f_{X_1,\ldots,X_d}(x_1, \ldots, x_k, x_{k+1}, \ldots, x_d)}{f_{X_{k+1},\ldots,X_d}(x_{k+1}, \ldots, x_d)}$$

provided that $f_{X_{k+1},\ldots,X_d}(x_{k+1}, \ldots, x_d) > 0$. The corresponding distribution functions are:

- In the discrete case:

$$F_{X_1,\ldots,X_k|X_{k+1},\ldots,X_d}(x_1, \ldots, x_k | x_{k+1}, \ldots, x_d) = \sum_{u_1 \le x_1} \ldots \sum_{u_k \le x_k} f_{X_1,\ldots,X_k|X_{k+1},\ldots,X_d}(u_1, \ldots, u_k | x_{k+1}, \ldots, x_d).$$

- In the continuous case:

$$F_{X_1,\ldots,X_k|X_{k+1},\ldots,X_d}(x_1,\ldots,x_k|x_{k+1},\ldots,x_d) = \int_{-\infty}^{x_1}\ldots\int_{-\infty}^{x_k} f_{X_1,\ldots,X_k|X_{k+1},\ldots,X_d}(u_1,\ldots,u_k|x_{k+1},\ldots,x_d)du_1\ldots du_k.$$

Independent Random Variables

The random variables $X_1,\ldots,X_d$ are called independent if and only if, for all $x_1,\ldots,x_d \in \mathbb{R}$

$$F_{X_1,\ldots,X_d}(x_1,\ldots,x_d) = F_{X_1}(x_1)\times\ldots\times F_{X_d}(x_d).$$

Equivalently, $X_1,\ldots,X_d$ are independent if and only if, for all $x_1,\ldots,x_d \in \mathbb{R}$

$$f_{X_1,\ldots,X_d}(x_1,\ldots,x_d) = f_{X_1}(x_1)\times\ldots\times f_{X_d}(x_d).$$

Note that when random variables are independent, conditional distributions reduce to the corresponding marginal distributions. Intuitively, knowing the value of one of the random variables gives us no information about the distribution of the rest.

Expectation, Variance, Covariance

The expectation (or expected value) of a random variable X formalises the notion of the "average" value taken by that random variable (in a sense, the typical value, what we expect). It is defined as follows.

– For continuous variables:

$$\mathbb{E}[X] = \int_{-\infty}^{+\infty} x\, f_X(x)dx.$$

– For discrete variables:

$$\mathbb{E}[X] = \sum_{x\in\mathcal{X}} x\, f_X(x), \qquad \mathcal{X} = \{x\in\mathbb{R} : f_X(x) > 0\}.$$

The expectation satisfies the following properties:

- Linearity: $\mathbb{E}[X_1 + \alpha X_2] = \mathbb{E}[X_1] + \alpha\mathbb{E}[X_2]$.
- $\mathbb{E}[h(x)] = \sum_{x\in\mathcal{X}} h(x) f_X(x)$ (discrete case)
  or
  $\mathbb{E}[h(x)] = \int_{-\infty}^{+\infty} h(x) f(x)dx$ (continuous case).

The variance of a random variable X expresses how disperse the realisations of X are around its expectation.

$$\mathrm{Var}(X) = \mathbb{E}\left[(X - \mathbb{E}(X))^2\right] \qquad (\text{if } \mathbb{E}[X^2] < \infty).$$

Furthermore, the covariance of a random variable X_1 with another random variable X_2 expresses the degree of linear dependency between the two.

$$\mathrm{Cov}(X_1, X_2) = \mathbb{E}\left[(X_1 - \mathbb{E}(X_1))(X_2 - \mathbb{E}(X_2))\right] \qquad (\text{if } \mathbb{E}[X_i^2] < \infty).$$

The correlation between X_1 and X_2 is defined as

$$\mathrm{Corr}(X_1, X_2) = \frac{\mathrm{Cov}(X_1, X_2)}{\sqrt{\mathrm{Var}(X_1)\,\mathrm{Var}(X_2)}}.$$

It also expresses the degree of linear dependency. Its advantage is that it is invariant to changes of units of measurement, and moreover can be understood in absolute terms (it ranges in $[-1, 1]$), as a result of the correlation inequality (itself a consequence of the Cauchy–Schwarz inequality):

$$|\mathrm{Corr}(X_1, X_2)| \leq \sqrt{\mathrm{Var}(X_1)\,\mathrm{Var}(X_2)}.$$

Some useful formulae relating expectations, variance, and covariances are:

- $\mathrm{Var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 = \mathrm{Cov}(X, X)$
- $\mathrm{Var}(aX + b) = a^2\,\mathrm{Var}(X)$
- $\mathrm{Var}(\sum_i X_i) = \sum_i \mathrm{Var}(X_i) + \sum_{i \neq j} \mathrm{Cov}(X_i, X_j)$
- $\mathrm{Cov}(X_1, X_2) = \mathbb{E}[X_1 X_2] - \mathbb{E}[X_1]\mathbb{E}[X_2]$
- $\mathrm{Cov}(aX_1 + bX_2, Y) = a\mathrm{Cov}(X_1, Y) + b\mathrm{Cov}(X_2, Y)$
- if $\mathbb{E}[X_1^2] + \mathbb{E}[X_2^2] < \infty$, then the following are equivalent:
  (i) $\mathbb{E}[X_1 X_2] = \mathbb{E}[X_1]\mathbb{E}[X_2]$
  (ii) $\mathrm{Cov}(X_1, X_2) = 0$
  (iii) $\mathrm{Var}(X_1 \pm X_2) = \mathrm{Var}(X_1) + \mathrm{Var}(X_2)$
  Independence will imply these three last properties, but none of these properties imply independence.

A.2 Taylor's Formula and the Inverse Function Theorem

The following two classic analysis results will often be used. See Rudin [21] (Chaps. 5 and 9) for their proofs.[1]

[1] An elementary proof of the one-dimensional form of the inverse function theorem (which will be all that will be needed for this text as stated below) can also be found in Corwin and Szczarba [5], Chap. 9.

Theorem A.1 (Taylor's Formula with Lagrange Remainder) *Let $h(x) : \mathbb{R} \to \mathbb{R}$ be k-times continuously differentiable on the closed interval I with endpoints x and y, for some $k \geq 0$. If $f^{(k+1)}$ exists on the interior of I, then there exists $t \in (0, 1)$ such that*

$$h(x) = h(y) + h'(y)(x - y) + \frac{h''(y)}{2!}(x - y)^2 + \cdots + \frac{h^{(k)}(y)}{k!}(x - y)^k + \frac{h^{(k+1)}(\xi)}{(k + 1)!}(x - y)^{k+1}$$

for $\xi = tx + (1 - t)y$.

Theorem A.2 (Inverse Function Theorem) *Let $h(x) : \mathbb{R} \to \mathbb{R}$ be continuously differentiable, with a non-zero derivative at a point $x_o \in \mathbb{R}$. Then, there exists an $\varepsilon > 0$ such h^{-1} continuously differentiable on $(h(x_0) - \epsilon, h(x_0) + \epsilon)$, and in fact $(h^{-1})'(y) = [h'(h^{-1}(y))]^{-1}$ for $|y - h(x_0)| < \varepsilon$.*

A.3 Two Concentration Inequalities

Lemma A.3 (Markov's Inequality) *Let X be a non-negative random variable. Then, given any $\epsilon > 0$,*

$$\mathbb{P}[X \geq \epsilon] \leq \frac{\mathbb{E}[X]}{\epsilon}.$$

Proof Notice that $0 \leq \epsilon \mathbf{1}\{X \geq \epsilon\} \leq X$. Therefore, $\mathbb{E}[\epsilon \mathbf{1}\{X \geq \epsilon\}] \leq \mathbb{E}[X]$. But

$$\mathbb{E}[\epsilon \mathbf{1}\{X \geq \epsilon\}] = \epsilon \mathbb{E}[\mathbf{1}\{X \geq \epsilon\}] = \epsilon \left(1 \cdot \mathbb{P}[X \geq \epsilon] + 0 \cdot \mathbb{P}[X < \epsilon]\right) = \epsilon \mathbb{P}[X \geq \epsilon].$$

Combining our findings yields the result. □

Lemma A.4 (Chebyshev's Inequality) *Let X be a random variable with finite mean $\mathbb{E}[X] < \infty$. Then, given any $\epsilon > 0$,*

$$\mathbb{P}\Big[|X - \mathbb{E}[X]| \geq \epsilon\Big] \leq \frac{\mathrm{Var}[X]}{\epsilon^2}.$$

Proof Define $Y = (X - \mathbb{E}[X])^2$ and apply Markov's inequality to Y. □

A.4 Monotonicity and Covariance

Lemma A.5 (Covariance of X and $g(X)$) *Let X be a real random variable with $\mathbb{E}[X^2] < \infty$. Let $g : \mathbb{R} \to \mathbb{R}$ be a non-decreasing function such that $\mathbb{E}[g^2(X)] < \infty$. Then,*

$$\mathrm{Cov}[X, g(X)] \geq 0.$$

Proof By definition of covariance:

$$\begin{aligned}
\mathrm{Cov}[X, g(X)] &= \mathbb{E}\Big\{\Big(X - \mu\Big)\Big(g(X) - \mathbb{E}[g(X)]\Big)\Big\} \\
&= \mathbb{E}\Big\{\Big(X - \mu\Big)\Big(g(X) - g(\mu) + g(\mu) - \mathbb{E}[g(X)]\Big)\Big\} \\
&= \mathbb{E}\Big\{\Big(X - \mu\Big)\Big(g(X) - g(\mu)\Big)\Big\} + \underbrace{\mathbb{E}\Big\{\Big(X - \mu\Big)\Big(g(\mu) - \mathbb{E}[g(X)]\Big)\Big\}}_{=0}
\end{aligned}$$

Now g is non-decreasing so if $X \geq \mu$, then $g(X) \geq g(\mu)$. If $X \leq \mu$, on the other hand, then $g(X) \leq g(\mu)$ also. Therefore

$$(X - \mu)(g(X) - g(\mu)) \geq 0$$

and the result follows. □

A.5 Quantiles

Recall that, for a random variable X taking values in $\mathcal{X}$, we define its distribution function to be:

$$F_X : \mathbb{R} \to [0, 1],$$

$$F_X(x) = \mathbb{P}[X \leq x], \qquad x \in \mathbb{R}.$$

Simply put, the distribution function is the answer to the following question: given a real number $x \in \mathbb{R}$, what is the probability $\mathbb{P}[X \leq x]$ that X fall at or below x? We could also ask the opposite question:

$$\boxed{\text{Given a probability } \alpha \in (0, 1), \text{ is there a real number } x \text{ such that } \mathbb{P}[X \leq x] = \alpha?} \tag{A.1}$$

The motivates the definition of the so-called *quantile function*.

Definition A.6 (Quantile Function and Quantiles)

Let X be a random variable and F_X be its distribution function. We define the quantile function of X to be the function

$$\begin{aligned} F_X^- &: (0,1) \to \mathbb{R} \\ F_X^-(\alpha) &= \inf\{t \in \mathbb{R} : F_X(t) \geq \alpha\}. \end{aligned}$$

Given an $\alpha \in (0,1)$, we call the real number

$$q_\alpha = F_X^-(\alpha)$$

the α-quantile of X (or, equivalently, of F_X).

Recall that F_X is always non-decreasing, by its definition. Hence, there are two possibilities:

(A) F_X is in fact *strictly increasing*.[2] Then F_X is also invertible, and we have

$$F_X^-(\alpha) = F_X^{-1}(\alpha), \qquad \forall \alpha \in (0,1).$$

In this case, our question (A.1) has a unique answer, and the interpretation is very simple.

(B) F_X is non-decreasing, but not strictly increasing.[3] Then there are two things that may happen:

(B1) There may be multiple real numbers x that satisfy $F_X(x) = \alpha$ (for example, take $\alpha = 1 - p$ and take X to be a Bern(p) random variable; then any $x \in (0,1)$ satisfies that $F_X(x) = 1 - p = \alpha$). In this case, $F_X^{-1}(\alpha)$ is a set, not a single real number,

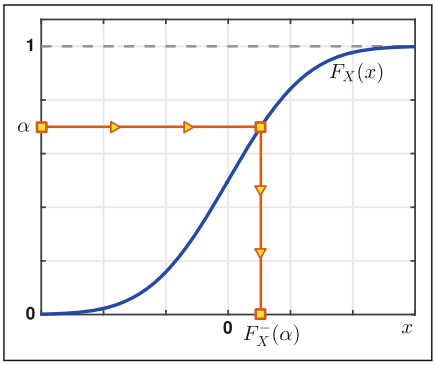
$$F_X^{-1}(\alpha) = \{x \in \mathbb{R} : F_X(x) = \alpha\}.$$

So, which of these numbers should we pick as the answer to our question (A.1)? The most mathematically appropriate choice turns out to be the infimum of this set.[4] Since F_X is right-continuous (being a probability distribution function) the infimum of this set equals $F_X^-(\alpha)$.

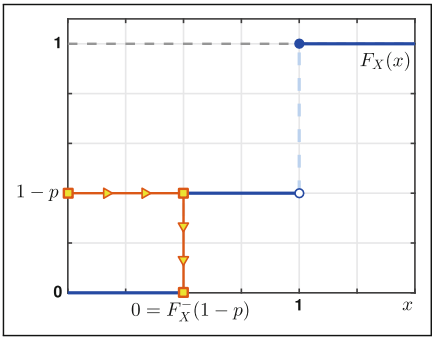
[2] This is the case if X is continuous with a density that satisfies $f_X(x) > 0 \; \forall x \in \mathbb{R}$.

[3] For regular models, this happens if X is discrete (so F_X is a step-function) or when X is continuous but there exists at least one open interval I such that $f_X(x) = 0, \forall x \in I$.

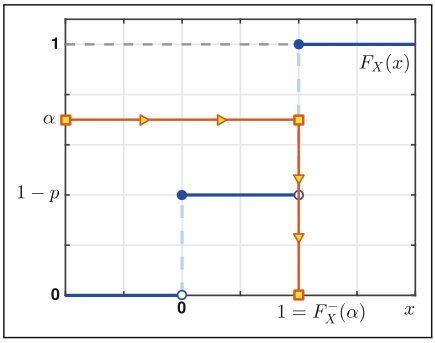
[4] This is due to the fact that, with this definition, we have $F(X) \geq \alpha \iff X \geq F^{-1}(\alpha)$, which is very useful when generating random variables with a prescribed distribution, see Exercise 11 (p. 22).



(a) Quantile in Scenario (A).



(b) Quantile in Scenario (B1).



(c) Quantile in Scenario (B2).

Fig. A.1 Evaluation of the quantile function for scenario (A), (B1) and (B2) above. Intuitively, in order to find q_α, we follow the *red arrows*. (**a**) Quantile in Scenario (A). (**b**) Quantile in Scenario (B1). (**c**) Quantile in Scenario (B2)

(B2) There may be no real number x such that $F_X(x) = \alpha$ (for example, take some $\alpha \in (1-p, 1)$ and take X to be a Bern(p) random variable). In this case, our question (A.1) has no answer. So, instead we have to settle for the first time that $F_X(x)$ "jumps" above α, which is again given by $F_X^-(x)$.

If all of this sounds complicated, Fig. A.1 gives an intuitive illustration that should clarify things.

Exercise 70 Let $X \sim \text{Exp}(\lambda)$ where $\lambda > 0$. Show that the $\alpha-$quantile of X is given by

$$q_\alpha = F_X^-(\alpha) = -\log(1-\alpha)/\lambda,$$

for $0 < \alpha < 1$.

Exercise 71 (Quantiles Determine Distributions) Let X and Y be random variables with respective distribution functions F_X and F_Y. Suppose that $F_X^-(\alpha) = F_Y^-(\alpha)$ for all $\alpha \in (0,1)$. Prove that $F_X = F_Y$.

A.6 Moment Generating Functions

The moment generating function (MGF) is a useful tool in probability theory that can often help us to prove independence of random variables or to determine their moments (hence the word moment generating).

Definition A.7 (Moment Generating Function)
Let X be a random variable taking values in $\mathbb{R}$. The MGF of X is defined as

$$M_X(t) : \mathbb{R} \to \mathbb{R} \cup \{\infty\}$$
$$M_X(t) = \mathbb{E}\Big[e^{tX}\Big], \qquad t \in \mathbb{R}.$$

Notice that $M_X(0) = 1$ always, so there exists at least one $t \in \mathbb{R}$ for which $M_X(t) < \infty$. When $M_X(t)$ is finite on an open neighbourhood of zero, then all the moments of X are defined, and can be determined by evaluating derivatives of M_X at zero.

Proposition A.8 (Moments via the MGF) *Let X be a random variable taking values in $\mathbb{R}$, and let I be an open interval such that $M_X(t) < \infty$ for all $t \in I$. It holds that*

1. *$\mathbb{E}[|X|^k e^{tX}] < \infty$ for all $k \in \mathbb{N}$ and all $t \in I$.*
2. *For all $t \in I$, the function M_X is k times differentiable, for all $k \in \mathbb{N}$ (hence infinitely differentiable on I).*
3. *For all $k \in \mathbb{N}$ and all $t \in I$, $\mathbb{E}[X^k e^{tX}] = \frac{d^k M_X}{dt^k}(t)$.*
4. *If $\{0\} \subset I$, then $\mathbb{E}[|X|^k] < \infty$ and $\mathbb{E}[X^k] = \frac{d^k M_X}{dt^k}(0)$, for all $k \in \mathbb{N}$.*

Proof We start with part 1. Fix $t_0 \in I$ and $k \in \mathbb{N}$. Since I is open, there exists a $\delta > 0$ such that $[t_0 - \delta, t_0 + \delta] \subset I$. Since the exponential function is increasing, we have

$$\begin{aligned} |X|^k e^{t_0 X} &= X^k e^{t_0 X} \mathbf{1}\{X \geq 0\} + (-X)^k e^{t_0 X} \mathbf{1}\{X < 0\} \\ &= e^{(t_0+\delta)X} u_{k,\delta}(X) \mathbf{1}\{X \geq 0\} + e^{(t_0-\delta)X} u_{k,\delta}(-X) \mathbf{1}\{X < 0\}, \end{aligned}$$

where $u_{k,\delta} : [0,\infty) \to [0,\infty)$ is given by

$$u_{k,\delta}(x) = x^k \exp(-\delta x), \qquad k \geq 0, \quad \delta > 0, \quad x \geq 0.$$

It's not hard to see that $C_{k,\delta} = \sup_{x\geq 0} u_{k,\delta}(x) < \infty$, since the exponential will decay faster than any polynomial. Specifically,

$$u'_{k,\delta}(x) = x^{k-1}e^{-\delta x}(k-\delta x)\begin{cases} >0 & x<\frac{k}{\delta} \\ <0 & x>\frac{k}{\delta}, \end{cases}$$

so that $u_{k,\delta}$ attains its maximum at $x = k/\delta$. We conclude that

$$\begin{aligned}\mathbb{E}|X|^k e^{t_0 X} &\leq C_{k,\delta}\mathbb{E}e^{(t_0+\delta)X}\mathbf{1}\{X\geq 0\} + C_{k,\delta}\mathbb{E}e^{(t_0-\delta)X}\mathbf{1}\{X<0\} \\ &\leq C_{k,\delta}M_X(t_0+\delta) + C_{k,\delta}M_X(t_0+\delta) < \infty.\end{aligned}$$

Since the choice of t_0 was arbitrary, we have proven part 1.

In order to prove parts 2 and 3, we proceed recursively. Both parts are trivially valid when $k = 0$. We will now show that if 2 and 3 are valid for $k-1$ (for all $t \in I$), then they must be valid for k, whenever $k \geq 1$.

Fix $t_0 \in I$. We need to show that

$$\lim_{t\to t_0}\frac{\mathbb{E}X^{k-1}e^{tX} - \mathbb{E}X^{k-1}e^{t_0X}}{t-t_0} = \lim_{t\to t_0}\frac{M_X^{(k-1)}(t) - M_X^{(k-1)}(t_0)}{t-t_0} = \mathbb{E}X^k e^{t_0X}. \tag{A.2}$$

Note that all the expectations in this equation are well defined (finite) as a result of part 1. Applying Taylor's formula (Theorem A.1, p. 159) to the function $h_x(t) = x^{k-1}e^{tx}$ (where x is seen as a fixed constant), we obtain

$$\frac{X^{k-1}e^{tX} - X^{k-1}e^{t_0X}}{t-t_0} = \frac{h_X(t)-h_X(t_0)}{t-t_0} = h'_X(\xi) = X^k e^{\xi X}, \quad |\xi - t_0| \leq |t-t_0|.$$

Note that since ξ depends on both t and X, it's in fact a random variable. Similarly,

$$\begin{aligned}\frac{X^{k-1}e^{tX} - X^{k-1}e^{t_0X}}{t-t_0} - X^k e^{t_0X} &= X^k e^{\xi X} - X^k e^{t_0X} \\ &= X^{k+1}e^{\xi' X}(\xi - t_0), \quad |\xi' - t_0| \leq |\xi - t_0|.\end{aligned}$$

We must thus show that the expectation on the right-hand side tends to zero as $t \to t_0$. Since $|\xi - t_0| \leq |t-t_0|$, it suffices to bound $\mathbb{E}X^{k+1}e^{\xi' X}$ uniformly in t. Let $\delta > 0$ be such that $[t_0 - 2\delta, t_0 + 2\delta] \subset I$. Suppose without loss of generality that $|t-t_0| < \delta$. It follows that $t_0 - \delta \leq \xi \leq t_0 + \delta$ and we can use the same approach as before to write:

$$\begin{aligned}|X|^{k+1}e^{\xi' X} &= X^{k+1}e^{\xi' X}1\{X\geq 0\} + (-X)^{k+1}e^{\xi' X}1\{X<0\} \\ &\leq X^{k+1}e^{(t_0+\delta)X}1\{X\geq 0\} + (-X)^{k+1}e^{(t_0-\delta)X}1\{X<0\} \\ &= e^{(t_0+2\delta)X}u_{k+1,\delta}(X)1\{X\geq 0\} + e^{(t_0-2\delta)X}u_{k+1,\delta}(-X)1\{X<0\}.\end{aligned}$$

It follows that

$$\mathbb{E}|X|^{k+1}e^{\xi' X} \leq C_{k+1,\delta}M_X(t_0+2\delta) + C_{k+1,\delta}M_X(t_0-2\delta) < \infty,$$

since $t_0 \pm 2\delta \in I$ and $C_{k+1,\delta} < \infty$. Hence

$$\mathbb{E}\left|\frac{X^{k-1}e^{tX} - X^{k-1}e^{t_0X}}{t-t_0} - X^k e^{t_0X}\right| \leq C_{k+1,\delta}[M_X(t_0+2\delta) + M_X(t_0-2\delta)]|t-t_0| \to 0, \quad t \to t_0.$$

Consequently, Eq. (A.2) holds true (since the term on the right-hand side of (A.2) is finite), which translates to

$$M_X^{(k)}(t_0) = \mathbb{E}X^k e^{t_0X} \qquad \forall\, t_0 \in I.$$

The recurrence thus holds true, which establishes 2 and 3. To complete the proof, observe that when $\{0\} \subset I$, part 4 follows directly from parts 1 and 3. □

A further important property of the MGF is that, provided that M_X exists on an open interval containing zero, it *uniquely* determines the distribution of X:

Proposition A.9 (Characterisation Property of the MGF) *Let X and Y be two random variables taking values in $\mathbb{R}$, and let F_X and F_Y be their respective distributions. Let $M_X, M_Y : \mathbb{R} \to \mathbb{R}$ be their MGFs. If there exists an open interval I containing zero, such that $M_X(t) < \infty$ and $M_Y(t) < \infty$ for all $t \in I$, then*

$$F_X = F_Y \iff M_X = M_Y.$$

We will not prove this result in its full generality, as this would require either notions related to Laplace transforms, or to characteristic functions (see, e.g., Billingsley [2], Sect. 30). We will only give a proof for the special case of non-negative random variables (following a saddlepoint argument of Dalang and Conus [8]). This suffices to cover the situations where we will use the theorem in this text.

Proof of Proposition A.9, Assuming $X, Y \geq 0$ We first consider the case of continuous random variables, and focus on the random variable $X \geq 0$. Since $X \geq 0$, it follows that $M_X(t) < \infty$ for all $t < 0$. Combining this fact with our assumption, means that there exists a $\delta > 0$ such that $M_X(t) < \infty$ for all $t < \delta$. By Proposition A.8, we now know that $\frac{d^k}{dt^k}M_X$ exists for all k and all $t < \delta$. Our strategy

will be to express F_X as a function of the derivatives of M_X. More specifically, define the function $G_X(t,x) : [0,\infty)^2 \to \mathbb{R}$ as

$$G_X(t,x) = \sum_{k=0}^{\lfloor tx \rfloor} \frac{t^k}{k!} \frac{d^k M_X}{dt^k}(-t),$$

where $\lfloor z \rfloor$ is the largest integer less than or equal to z. We will show that for any given $x \geq 0$,

$$\lim_{t\to\infty} G_X(t,x) = F_X(x).$$

Fix $x \geq 0$. Proposition A.8 shows that, for all $k \geq 0$,

$$\frac{d^k}{dt^k} M_X(t) = \mathbb{E}\left[X^k e^{tX}\right] = \int_0^\infty x^k e^{tx} f_X(x) dx,$$

where the last integral is over $[0,\infty)$ by non-negativity of X. Thus G may be re-expressed as

$$G_X(t,x) = \sum_{k=0}^{\lfloor tx \rfloor} \frac{t^k}{k!} \int_0^{+\infty} y^k e^{-ty} f_X(y) dy = \int_0^{+\infty} \left(\underbrace{\sum_{k=0}^{\lfloor tx \rfloor} \frac{t^k}{k!} y^k e^{-ty}}_{=\varphi_t(x,y)} \right) f_X(y) dy,$$

where $\varphi_t(x,y) = \mathbb{P}[W_{t,y} \leq tx]$ for $W_{t,x} \sim Poisson(ty)$. Consequently, when $y > x$, Chebyshev's inequality (Lemma A.4, p. 159) implies that

$$\begin{aligned} 0 \leq \varphi_t(x,y) &= \mathbb{P}[W_{t,y} \leq tx] = \mathbb{P}[W_{t,y} - ty \leq t(x-y)] \\ &\leq \mathbb{P}[|W_{t,y} - ty| \geq t(y-x)] \\ &\leq \frac{\mathrm{Var}[W_{t,y}]}{t^2(y-x)^2} = \frac{y}{t(y-x)^2}. \end{aligned}$$

Similarly, in the case $y < x$, we have

$$\begin{aligned} 0 \leq 1 - \varphi_t(x,y) &= \mathbb{P}[W_{t,y} > tx] = \mathbb{P}[W_{t,y} - ty > t(x-y)] \\ &\leq \mathbb{P}[|W_{t,y} - ty| > t(x-y)] \\ &\leq \frac{\mathrm{Var}[W_{t,y}]}{t^2(x-y)^2} = \frac{y}{t(x-y)^2}. \end{aligned}$$

Now let $\epsilon > 0$. Choose $h > 0$ sufficiently small so that $F_X(x+h) - F_X(x) < \epsilon/3$ and $F_X(x) - F_X(x-h) < \epsilon/3$ (such a choice is ensured by continuity of F_X). Then choose $t > 0$ sufficiently large so that $t > 6x/\epsilon h^2$. We have

$$\begin{aligned}
|G_X(t,x) - F_X(x)| &= \left| \int_0^{+\infty} \varphi_t(x,y) f_X(y) dy - \int_0^x f_X(y) dy \right| \\
&= \left| \int_0^{x-h} (\varphi_t(x,y) - 1) f_X(y) dy + \int_{x-h}^{x} (\varphi_t(x,y) - 1) f_X(y) dy \right. \\
&\qquad \left. + \int_x^{x+h} \varphi_t(x,y) f_X(y) dy + \int_{x+h}^{\infty} \varphi_t(x,y) f_X(y) dy \right| \\
&\le \int_0^{x-h} |\varphi_t(x,y) - 1| f_X(y) dy + \int_{x-h}^{x} |\varphi_t(x,y) - 1| f_X(y) dy \\
&\qquad + \int_x^{x+h} |\varphi_t(x,y)| f_X(y) dy + \int_{x+h}^{\infty} |\varphi_t(x,y)| f_X(y) dy.
\end{aligned}$$

Let us consider the terms on the right-hand side one at a time, and bound them suitably (note that if $x = 0$, we only need to consider the last two integrals). We have

$$\begin{aligned}
\int_0^{x-h} |\varphi_t(x,y) - 1| f_X(y) dy &\le \frac{1}{t} \int_0^{x-h} \frac{y}{(x-y)^2} f_X(y) dy \\
&\le \frac{x-h}{th^2} \int_0^{x-h} f_X(y) dy \le \frac{x-h}{th^2},
\end{aligned}$$

by our earlier calculation. Similarly,

$$\int_{x+h}^{\infty} |\varphi_t(x,y)| f_X(y) dy \le \frac{x+h}{th^2}.$$

Furthermore, $|\varphi_t(x,y) - 1| \le 1$ and $|\varphi_t(x,y)| \le 1$ for all $x, y \ge 0$, so that

$$\int_{x-h}^{x} |\varphi_t(x,y) - 1| f_X(y) dy \le \int_{x-h}^{x} f_X(y) dy = F_X(x) - F_X(x-h)$$

and

$$\int_x^{x+h} |\varphi_t(x,y)| f_X(y) dy \le \int_x^{x+h} f_X(y) dy = F_X(x+h) - F_X(x).$$

In summary, we have shown that for all $t > \frac{6x}{\epsilon h^2}$,

$$\begin{aligned}|G_X(t,x) - F_X(x)| &\leq \frac{x-h}{th^2} + [F_X(x) - F_X(x-h)] \\ &\quad + [F_X(x+h) - F_X(x)] + \frac{x+h}{th^2} \\ &= [F_X(x) - F_X(x-h)] + [F_X(x+h) - F_X(x)] + \frac{2x}{th^2} \\ &= \frac{\epsilon}{3} + \frac{\epsilon}{3} + \frac{\epsilon}{3} = \epsilon.\end{aligned}$$

In other words, we have shown that $|G_X(t,x) - F_X(x)| < \epsilon$ for any $\epsilon > 0$ and t sufficiently large, which proves that $\lim_{t\to\infty} G_X(t,x) = F_X(x)$. The exact same arguments show that $\lim_{t\to\infty} G_Y(t,x) = F_Y(x)$, where $G_Y(t,x)$ is defined in analogous fashion as $G_X(t,x)$. But $G_X = G_Y$ since $M_X = M_Y$, which proves that $F_X = F_Y$ and completes the proof in the case when the random variables X, Y are continuous. For the discrete case, we follow the exact same argument, replacing integrals by sums, and proving that $\lim_{t\to\infty} G_X(t,x) = F_X(x)$ for all continuity points x of $F_X(x)$. For discontinuity points, we then simply use the right-continuity of F_X. The proof is now complete. □

The next lemma is useful when trying to establish the distribution of a sum or independent random variables.

Lemma A.10 (Sums and MGFs) *Let X and Y be two independent random variables taking values in $\mathbb{R}$, and let $Z = X + Y$. If $M_X(t) < \infty$ and $M_Y(t) < \infty$ for all t in an open interval I, then $M_Z(t) < \infty$ for all $t \in I$ and*

$$M_Z(t) = M_X(t)M_Y(t).$$

Proof By independence, we may write

$$\begin{aligned}\infty > M_X(t)M_Y(t) &= \mathbb{E}[e^{tX}]\mathbb{E}[e^{tY}] = \mathbb{E}[e^{tX}e^{tY}] \\ &= \mathbb{E}[\exp\{t(X+Y)\}] = M_Z(t), \qquad t \in I.\end{aligned}$$

□

A.7 Continuous Mapping and Slutsky's Theorem

In order to prove these two results, we will first need a couple of results regarding distribution functions and their convergence.

Lemma A.11 *Let F be a cumulative distribution function. Then F has at most countably many discontinuities.*

Proof Let D_F be the set of discontinuity points of F. Given any $x \in D_F$, we have

$$\lim_{\epsilon \downarrow 0} F(x-\epsilon) < \lim_{\epsilon \downarrow 0} F(x+\epsilon)$$

since F is non-decreasing. It follows that there exists a rational number $q(x)$ such that

$$\lim_{\epsilon \downarrow 0} F(x-\epsilon) < q(x) < \lim_{\epsilon \downarrow 0} F(x+\epsilon), \qquad \forall x \in D_F.$$

Furthermore, whenever $x_1 < x_2$ (so that we may write $x_2 = x_1 + \delta$, for some $\delta > 0$), the fact that F is non-decreasing implies that

$$q(x_1) < \lim_{\epsilon \downarrow 0} F(x_1+\epsilon) \leq F(x_1+\delta/2) = F(x_2-\delta/2) \leq \lim_{\epsilon \downarrow 0} F(x-\epsilon) < q(x_2).$$

Summarising, we have constructed an injection $q : D_F \to \mathbb{Q}$, and thus D_F must be countable. □

Lemma A.12 *Given a sequence of random variables $X, X_1, X_2, \ldots$, the following two statements are equivalent:*

1. $X_n \xrightarrow{d} X$.
2. *For all closed subsets $C \subseteq \mathbb{R}$, one has*

$$\limsup_{n \to \infty} \mathbb{P}(X_n \in C) \leq P(X \in C).$$

Proof Assume first that (2) holds true, so that for $C_1 = (-\infty, a]$ and $C_2 = [a, \infty)$, we have

$$\mathbb{P}(X < a) = 1 - \mathbb{P}(X \geq a) \leq 1 - \limsup_{n \to \infty} \mathbb{P}(X_n \geq a) = \liminf_{n \to \infty} \mathbb{P}(X_n < a)$$

$$\leq \liminf_{n \to \infty} \mathbb{P}(X_n \leq a) \leq \limsup_{n \to \infty} \mathbb{P}(X_n \leq a) \leq \mathbb{P}(X \leq a).$$

If a is a continuity point of the distribution function of X, it must be that $\mathbb{P}(X < a) = \mathbb{P}(X \le a)$ and so $\mathbb{P}(X_n \le a) \to \mathbb{P}(X \le a)$. This establishes that $X_n \xrightarrow{d} X$.

To prove the converse, assume initially that $C = [a, b]$, where $-\infty < a \le b < \infty$. There exist sequences $0 \le \epsilon_k \searrow 0, 0 \le \delta_k \searrow 0$ such that $F(x) = \mathbb{P}(X \le x)$ is continuous at the points $a - \delta_k$ and $b + \epsilon_k$ for all k (Lemma A.11). Consequently,

$$\limsup_{n\to\infty} \mathbb{P}(X_n \in C) \le \limsup_{n\to\infty} \mathbb{P}(a - \delta_k < X_n \le b + \epsilon_k) = \limsup_{n\to\infty} \mathbb{P}(X_n \le b + \epsilon_k)$$

$$-\mathbb{P}(X_n \le a - \delta_k) = \mathbb{P}(X \le b + \epsilon_k) - \mathbb{P}(X \le a - \delta_k) = \mathbb{P}(a - \delta_k < X \le b + \epsilon_k).$$

Letting $k \to \infty$, continuity from above of probability measures yields

$$\begin{aligned}\limsup_{n\to\infty} \mathbb{P}(X_n \in C) &\le \lim_{k\to\infty} \mathbb{P}(a - \delta_k < X \le b + \epsilon_k) \\ &= \mathbb{P}\left(\bigcap_{k=1}^{\infty} \{a - \delta_k < X \le b + \epsilon_k\}\right) = \mathbb{P}(X \in C).\end{aligned}$$

If $a = -\infty$ or $b = \infty$, the statement can be shown to be true by a similar argument. Thus (2) is true when C is an interval.

If $C = \cup C_k$ is the countable union of (potentially infinitely many) closed disjoint intervals, the subadditivity of limit superior yields

$$\begin{aligned}\limsup_{n\to\infty} \mathbb{P}(X_n \in C) &= \limsup_{n\to\infty} \sum_{k=1}^{\infty} \mathbb{P}(X_n \in C_k) \le \sum_{k=1}^{\infty} \limsup_{n\to\infty} \mathbb{P}(X_n \in C_k) \\ &\le \sum_{k=1}^{\infty} \mathbb{P}(X \in C_k) = \mathbb{P}(X \in C).\end{aligned}$$

Suppose now that $C = \cap C_k$, where each C_k is a disjoint union of countably many closed intervals, and $C_{k+1} \subseteq C_k$ for all k. Following the same course as in the first part of the proof,

$$\limsup_{n\to\infty} \mathbb{P}(X_n \in C) \le \limsup_{n\to\infty} \mathbb{P}(X_n \in C_k) \le \mathbb{P}(X \in C_k) \to \mathbb{P}(X \in C), \quad k \to \infty.$$

To complete the proof, thus, it suffices to show that any closed set $C \subseteq \mathbb{R}$ can be written in this form.

For every k, divide $\mathbb{R}$ into closed intervals of length 2^{-k}, that is $I_j^{(k)} = 2^{-k}[j, j + 1]$. Let C_k be the union of those intervals $\{I_j^{(k)}\}$ that have a non-empty intersection with C:

$$C_k = \bigcup_{j\in\mathbb{Z}: I_j^{(k)} \cap C \neq \emptyset} I_j^{(k)}.$$

It is clear that C_k is the countable union of countably many closed intervals, and that $C_k \supseteq C$. If $x \notin C$, there exists an interval I such that $C \cap I = \emptyset$ that contains x. For k such that $2^{-k} < m(I)/2$ it follows that $x \notin C_k$. We may thus conclude that $C = \cap C_k$. The fact that C_k is closed follows by a similar reasoning, but we can argue differently: let $x_n \in C_k$ be a sequence converging to x. There must exist an M such that the sequence is contained in $C_k \cap [-M, M]$. This last set is closed, as it is the union of finitely many closed intervals. Hence $x \in C_k \cap [-M, M]$ and so C_k is closed.

It remains to show that $C_{k+1} \subseteq C_k$. Let $x \in C_{k+1}$. There exists $j \in \mathbb{Z}$ such that $x \in I_j^{(k+1)} \subseteq C_{k+1}$. Or, $I_j^{(k+1)} \subset I_{\lfloor j/2 \rfloor}^{(k)}$, and thus this last set has a non-empty intersection with C. It follows that $x \in I_{\lfloor j/2 \rfloor}^{(k)} \subseteq C_k$, and the proof is complete. □

Proof of the Continuous Mapping Theorem (Theorem 2.25, p. 57) By Lemma A.12, it suffices to prove that $X_n \xrightarrow{d} X$ implies $\limsup_{n\to\infty} \mathbb{P}[g(X_n) \leq y] \leq \mathbb{P}[g(X) \in C]$ for all closed $C \subseteq \mathbb{R}$. To this aim, let $C \subseteq \mathbb{R}$ be an arbitrary closed set, let

$$A = \{x \in \mathbb{R} : g(x) \in C\}$$

be the inverse image of C via g, and let $\overline{A}$ denote the closure of A. If D_g is the set of discontinuities of g, we may write

$$\overline{A} = \left\{\overline{A} \cap D_g\right\} \cup \left\{\overline{A} \cap D_g^c\right\} \subseteq D_g \cup \left\{\overline{A} \cap D_g^c\right\}.$$

Now if $x \in \overline{A} \cap D_g^c$, then there exists a sequence $\{x_k\} \subset A$ such that $\lim_{k\to\infty} x_k = x$ (by definition of the closure, $\overline{A}$). Furthermore, it holds that $g(x) = \lim_{k\to\infty} g(x_k) \in C$, because $x \in D_g^c$ also. Consequently $x \in A$, and we have proven that $\overline{A} \cap D_g^c \subseteq A$.

Summarising, we have

$$\overline{A} \subseteq A_y \cup D_g. \tag{A.3}$$

We now exploit this inclusion in order to write

$$\mathbb{P}[g(X_n) \in C] = \mathbb{P}[X_n \in A] \leq \mathbb{P}[X_n \in \overline{A}].$$

But,

$$\begin{aligned}\limsup_{n\to\infty} \mathbb{P}[X_n \in \overline{A}] &\leq \mathbb{P}[X \in \overline{A}] \quad [\text{using } X_n \xrightarrow{d} X, \text{combined with Lemma A.12}] \\ &\leq \mathbb{P}[X \in A \cup D_g] \qquad [\text{by (A.3)}] \\ &\leq \mathbb{P}[X \in A] + \underbrace{\mathbb{P}[X \in D_g]}_{=0} \\ &= \mathbb{P}[g(X) \in C].\end{aligned}$$

It follows that $\limsup_{n\to\infty} \mathbb{P}[g(X_n) \in C] \leq \mathbb{P}[g(X) \in C]$ and our proof is complete.

□

Proof of Slutsky's Theorem (Theorem 2.26, p. 57) For the first part, assume that $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{p} c$. We may assume without loss of generality that $c = 0$. Let x be a continuity point of F_X. We have

$$\begin{aligned}\mathbb{P}[X_n + Y_n \leq x] &= \mathbb{P}[X_n + Y_n \leq x, |Y_n| \leq \epsilon] + \mathbb{P}[X_n + Y_n \leq x, |Y_n| > \epsilon] \\ &\leq \mathbb{P}[X_n \leq x + \epsilon] + \mathbb{P}[|Y_n| > \epsilon]\end{aligned}$$

because $\{X_n + Y_n \leq x \,\&\, |Y_n| \leq \epsilon\}$ implies that $\{X_n \leq x + \epsilon\}$. Similarly, we may obtain the inequality

$$\mathbb{P}[X_n \leq x - \epsilon] \leq \mathbb{P}[X_n + Y_n \leq x] + \mathbb{P}[|Y_n| > \epsilon].$$

Rearranging and collecting terms yields:

$$\mathbb{P}[X_n \leq x - \epsilon] - \mathbb{P}[|Y_n| > \epsilon] \leq \mathbb{P}[X_n + Y_n \leq x] \leq \mathbb{P}[X_n \leq x + \epsilon] + \mathbb{P}[|Y_n| > \epsilon]$$

$$\lim_{n\to\infty} \mathbb{P}[X_n \leq x - \epsilon] - 0 \leq \lim_{n\to\infty} \mathbb{P}[X_n + Y_n \leq x] \leq \lim_{n\to\infty} \mathbb{P}[X_n \leq x + \epsilon] + 0$$

By Lemma A.11, we may find a sequence $0 < \epsilon_k \downarrow 0$ such that $x + \epsilon_k$ is a continuity point, for all k. Replacing ϵ by ϵ_k gives

$$F_X(x - \epsilon_k) \leq \lim_{n\to\infty} \mathbb{P}[X_n + Y_n \leq x] \leq F_X(x + \epsilon_k).$$

Since x is a continuity point of F_X, letting $k \to \infty$ establishes $X_n + Y_n \xrightarrow{d} X$.

To prove the second part, let $Z_n = Y_n - c$, so that $Z_n \xrightarrow{p} 0$. Thus, if we can show $X_n Z_n \xrightarrow{d} 0$, then the conclusion follows by first part of the theorem, which is already proven. Let $\epsilon > 0$ and $M_k \uparrow \infty$ be positive sequence such that ϵM_k is a continuity point of $F_{|X|}$ for all k (this choice is feasible by Lemma A.11). Note

also that $|X_n| \xrightarrow{d} |X|$ by the continuous mapping theorem (Theorem 2.25, 57). Combining these ingredients yields:

$$\begin{aligned}\mathbb{P}[|X_n Z_n| > \epsilon] &\leq \mathbb{P}[|X_n Z_n| > \epsilon, |Z_n| \leq 1/M_k] + \mathbb{P}[|Z_n| \geq 1/M_k] \\ &\leq \mathbb{P}[|X_n| > \epsilon M_k] + \mathbb{P}[|Z_n| \geq 1/M_k] \\ &\leq 1 - \mathbb{P}[|X_n| \leq \epsilon M_k] + \mathbb{P}[|Z_n| \geq 1/M_k] \\ \Longrightarrow \lim_{n\to\infty} \mathbb{P}[|X_n Z_n| > \epsilon] &\leq \mathbb{P}[|X| > \epsilon M_k].\end{aligned}$$

The right-hand side can be made arbitrarily small by choosing k sufficiently large. Thus $Z_n X_n \xrightarrow{p} 0$. Since $X_n Y_n = Z_n X_n + c X_n$, we use the first part of the theorem (already proven) to conclude that $X_n Y_n \xrightarrow{p} 0$. □

A.8 On the Proof of the Central Limit Theorem

The standard proof of the central limit theorem makes use of the *characteristic function*, and thus involves notions from complex analysis, and more specifically the Lévy continuity theorem (see, e.g., Billingsley [2], Sect. 29). Since the latter result is beyond the scope of this text, we will provide an elementary proof here due to Lindeberg [17] (as presented in Dalang [7]), that is based on stronger assumptions, namely existence of a third absolute moment.[5,6]

We first need three intermediate results. In what follows, $C_b^3(\mathbb{R})$ denotes the set of all thrice continuously differentiable bounded functions $\mathbb{R} \to \mathbb{R}$, that are bounded, and whose first three derivatives are also bounded.

Lemma A.13 *Let Z be a continuous random variable, and $\{Z_n\}_{n\geq 1}$ a sequence of random variables such that*

$$\mathbb{E}[g(Z_n)] \xrightarrow{n\to\infty} \mathbb{E}[g(Z)]$$

for all $g \in C_b^3(\mathbb{R})$. Then

$$F_{Z_n}(x) \xrightarrow{n\to\infty} F_Z(x), \qquad \forall\, x \in \mathbb{R}.$$

[5] As a matter of fact, even this weaker version of the theorem would suffice for the asymptotic results presented in this text: these require the sufficient statistic of an exponential family to satisfy the central limit theorem (as Corollary 2.24, p. 56), and the latter statistic will have finite moments of all orders (see Eq. (2.11), p. 51, in the proof of Proposition 2.11).

[6] The same method of proof can be "upgraded" to work under only second moment assumptions, assuming knowledge of measure theory, in particular the monotone convergence theorem (Dalang [7]).

Proof Let $x \in \mathbb{R}$ and $k \geq 1$ be given. Note that we may always choose a function $g_k \in C_b^3(\mathbb{R})$ that satisfies the envelope relation

$$\mathbf{1}\{z \in (-\infty, x]\} \leq g_k(z) \leq \mathbf{1}\{z \in (-\infty, x + 1/k]\}. \tag{A.4}$$

Then, for all $n \geq 1$,

$$F_{Z_n}(x) = \mathbb{P}[Z_n \leq x] = \mathbb{E}[\mathbf{1}\{z \in (-\infty, x]\}] \leq \mathbb{E}[g_k(Z_n)],$$

and hence by our assumption we have

$$\begin{aligned} \limsup_{n\to\infty} F_{Z_n}(x) &\leq \lim_{n\to\infty} \mathbb{E}[g_k(Z_n)] = \mathbb{E}[g_k(Z)] \\ &\leq \mathbb{E}[\mathbf{1}\{z \in (-\infty, x + 1/k\}] = F_Z(x + 1/k). \end{aligned}$$

The same type of argument shows that $\liminf_{n\to\infty} F_{Z_n}(x) \geq F_Z(x - 1/k)$. Since the choice of k was arbitrary, and since F_Z is everywhere continuous, we have that $F_{Z_n}(x) \overset{n\to\infty}{\longrightarrow} F_Z(x)$, completing the proof. □

Lemma A.14 *Let $g \in C_b^3(\mathbb{R})$, and let $\sup_{x\in\mathbb{R}} |g'''(x)| = C < \infty$. Let (Y, Z) be independent random variables such that $\mathbb{E}[Y] = \mathbb{E}[Z]$, and $\mathbb{E}[Y^2] = \mathbb{E}[Z^2]$. If X is independent of Y and Z, we have*

$$\left|\mathbb{E}[g(X + Y) - g(X + Z)]\right| \leq \frac{C}{6}\left(\mathbb{E}|Y|^3 + \mathbb{E}|Z|^3\right).$$

Proof Taylor's theorem (Theorem A.1, p. 159) yields that

$$g(x + y) = g(x) + yg'(x) + \frac{1}{2}y^2 g''(x) + \frac{1}{6}y^3 g'''(u),$$

where u lies between x and $x + y$. It follows now by independence that

$$\mathbb{E}[g(X + Y)] = \mathbb{E}[g(X)] + \mathbb{E}[Y]\mathbb{E}[g'(X)] + \frac{1}{2}\mathbb{E}[Y^2]\mathbb{E}[g''(X)] + \frac{1}{6}\mathbb{E}[Y^3 g'''(U)]$$

$$\mathbb{E}[g(X + Z)] = \mathbb{E}[g(X)] + \mathbb{E}[Z]\mathbb{E}[g'(X)] + \frac{1}{2}\mathbb{E}[Z^2]\mathbb{E}[g''(X)] + \frac{1}{6}\mathbb{E}[Z^3 g'''(V)]$$

for a random variable U that lies between X and $X + Y$ almost surely, and a random variable V that lies between X and $X + Z$ almost surely. Consequently,

our assumptions yield that

$$\begin{aligned}\Big|\mathbb{E}[g(X+Y)-g(X+Z)]\Big| &= \Big|\frac{1}{6}\mathbb{E}[Y^3 g'''(U)] - \frac{1}{6}\mathbb{E}[Z^3 g'''(V)]\Big| \\ &\le \frac{1}{6}\mathbb{E}\Big|Y^3 g'''(U)\Big| + \frac{1}{6}\mathbb{E}\Big|Z^3 g'''(V)\Big| \\ &\le \frac{C}{6}\left(\mathbb{E}|Y|^3 + \mathbb{E}|Z|^3\right).\end{aligned}$$

□

Lemma A.15 *Let $\{\tilde{Y}_n\}_{n\ge 1}$ be a sequence of iid random variables such that $\mathbb{E}|\tilde{Y}_1|^3 < \infty$, $\mathbb{E}[\tilde{Y}_1^2] = 1$, and $\mathbb{E}[\tilde{Y}_1] = 0$. If $g \in C_b^3(\mathbb{R})$, then it holds that*

$$\mathbb{E}\left[g\left(\frac{\sum_{i=1}^n \tilde{Y}_i}{\sqrt{n}}\right)\right] \overset{n\to\infty}{\longrightarrow} \mathbb{E}\Big[g(\tilde{Z})\Big],$$

where $\tilde{Z} \sim N(0,1)$.

Proof Let $g \in C_b^3(\mathbb{R})$, and $n \ge 1$. Let $\{\tilde{Z}_i\}_{i=1}^n \overset{\text{iid}}{\sim} N(0,1)$ (independent of the $\{\tilde{Y}_i\}$) and define

$$Y_i = \tilde{Y}_i/\sqrt{n} \qquad \& \qquad Z_i = \tilde{Z}_i/\sqrt{n}.$$

Since $\{\tilde{Z}_i\}_{i=1}^n \overset{\text{iid}}{\sim} N(0,1/n)$, it follows that $\sum_{i=1}^n Z_i \sim N(0,1)$ (by Corollary 1.35, p. 25). It thus suffices to show that

$$\Big|\mathbb{E}[g(Y_1+\cdots Y_n)] - \mathbb{E}[g(Z_1+\cdots Z_n)\Big| \le \frac{C}{6}\frac{\mathbb{E}[|\tilde{Y}_1|^3] + \mathbb{E}[|\tilde{Z}_1|^3]}{\sqrt{n}} \tag{A.5}$$

for $C = \sup_{x\in\mathbb{R}} |g'''(x)| < \infty$. Define

$$\begin{aligned}U_i &= Y_1 + \cdots + Y_{i-1} + Y_i + Z_{i+1} + \cdots + Z_n \\ V_i &= Y_1 + \cdots + Y_{i-1} + 0 + Z_{i+1} + \cdots + Z_n\end{aligned}$$

and observe that these satisfy

$$U_i = V_i + Y_i \qquad \& \qquad U_{i-1} = V_i + Z_i$$

so that we may re-write the left-hand side of Eq. (A.5) as

$$\begin{aligned}\mathbb{E}[g(U_n)] - \mathbb{E}[g(U_0)] &= \sum_{i=1}^{n} \left(\mathbb{E}[g(U_i)] - \mathbb{E}[g(U_{i-1})]\right) \\ &= \sum_{i=1}^{n} \left(\mathbb{E}[g(V_i + Y_i)] - \mathbb{E}[g(V_i + Z_i)]\right).\end{aligned}$$

We now use Lemma A.14 to bound the last expression by

$$\sum_{i=1}^{n} \frac{C}{6}\left(\mathbb{E}[|Y_i|^3] - \mathbb{E}[|Z_i|^3]\right) = n\frac{C}{6}n^{-3/2}\left(\mathbb{E}[|\tilde{Y}_1|^3] + \mathbb{E}[|\tilde{Z}_1|^3]\right)$$

thus establishing the validity of inequality A.5, and completing the proof. □

Theorem A.16 (Third Moment Central Limit Theorem) *Let $Y_1, \ldots, Y_n$ be iid random variables such that $\mathbb{E}[Y_i] = \mu < \infty$, $\mathrm{Var}[Y_i] = \sigma^2$, and $\mathbb{E}|Y_i|^3 < \infty$. Let $\overline{Y}_n = \frac{1}{n}\sum_{i=1}^{n} Y_i$. Then,*

$$\sqrt{n}(\overline{Y}_n - \mu) \xrightarrow{d} N(0, \sigma^2).$$

Proof The random variables $\tilde{Y}_i = \frac{Y_i - \mu}{\sigma}$ satisfy the conditions of Lemma A.15. Thus, if we define

$$Z_n := \frac{\tilde{Y}_1 + \ldots + \tilde{Y}_n}{\sqrt{n}} = \frac{\sqrt{n}(\overline{Y}_n - \mu)}{\sigma},$$

we must have

$$\mathbb{E}\left[g\left(Z_n\right)\right] \xrightarrow{n\to\infty} \mathbb{E}\Big[g(Z)\Big], \qquad \forall\, g \in C_b^3(\mathbb{R}),$$

for $Z \sim N(0, 1)$. Lemma A.13 now implies that $F_{Z_n}(x) \overset{n\to\infty}{\to} F_Z(x)$ for all $x \in \mathbb{R}$, and so $\sigma Z_n = \sqrt{n}(\overline{Y}_n - \mu) \xrightarrow{d} N(0, \sigma^2)$. □

Bibliography

1. Bickel, P. J., & Doksum, K. A. (2001). *Mathematical statistics: Basic ideas and selected topics*. Upper Saddle River: Prentice Hall.
2. Billingsley, P. (1986). *Probability and measure*. New York: Wiley.
3. Blitzstein, J. K., & Hwang, J. (2015). *Introduction to probability*. London: Chapman & Hall/CRC.
4. Casella, G., & Berger, R. L. (2002). *Statistical inference*. Pacific Grove: Duxbury Press.
5. Corwin, L. J., & Szczarba, R. H. (1982). *Multivariable calculus*. New York: Marcel Dekker.
6. Cox, D. R., & Hinkley, D. V. (1979). *Theoretical statistics*. London: Chapman & Hall/CRC.
7. Dalang, R. C. (2006). Une démonstration élémentaire du théorème central limite. *Elemente der Mathematik, 61*(2), 65–73.
8. Dalang, R. C., & Conus, D. (2008). *Introduction à la théorie des probabilités*. Lausanne: Presses Polytechniques et Universitaires Romandes.
9. Davison, A. C. (2003). *Statistical models*. Cambridge: Cambridge University Press.
10. Durrett, R. (1996). *Probability: Theory and examples*. Pacific Grove:: Duxbury Press.
11. Grimmett, G., & Welsh, D. (2014). *Probability: An introduction*. Oxford: Oxford University Press.
12. Hogg, R. V., & Craig, A. T. (1970). *Introduction to mathematical statistics*. New York: Macmillan.
13. Hogg, R. V., & Tanis, E. A. (2000). *Probability and statistical inference*. Upper Saddle River: Prentice Hall.
14. Knight, K. (2000). *Mathematical statistics*. Boca Raton: Chapman & Hall/CRC.
15. Lehmann, E. L., & Casella, G. (2003). *Theory of point estimation*. New York: Springer.
16. Lehmann, E. L., & Romano, J. P. (2008). *Testing statistical hypotheses*. New York: Springer.
17. Lindeberg, J. (1922). Eine neue Herleitung des Exponentialgesetzes in der Wahrscheinlichkeitsrechnung. *Mathematische Zeitschrift, 15*, 211–225.
18. Pitman, J. (1993). *Probability*. New York: Springer.
19. Rice, J. A. (2006). *Mathematical statistics and data analysis*. Belmont: Duxbury Press.
20. Ross, S. M. (2010). *A first course in probability*. Upper Saddle River: Prentice Hall.
21. Rudin, W. (1976). *Principles of mathematical analysis*. New York: McGraw-Hill.
22. Schervish, M. J. (2010). *Theory of statistics*. New York: Springer.
23. Shao, J. (2008). *Mathematical statistics*. New York: Springer.
24. Silvey, S. D. (2003). *Statistical inference*. London: Chapman & Hall/CRC.
25. Wasserman, L. (2004). *All of statistics: A concise course in statistical inference*. New York: Springer.
26. Young, G. A., & Smith, R. L. (2005). *Essentials of statistical inference*. Cambridge: Cambridge University Press.

V.M. Panaretos, *Statistics for Mathematicians*, Compact Textbooks in Mathematics,
DOI 10.1007/978-3-319-28341-8