TOPICS IN PROBABILITY, PARAMETRIC ESTIMATION AND STOCHASTIC CALCULUS

LEVI LOPES DE LIMA

ABSTRACT. We begin our journey by recalling the fundamentals of Probability Theory that underlie one of its most significant applications to real-world problems: Parametric Estimation. Throughout the text, we systematically develop this theme by presenting and discussing the main tools it encompasses—concentration inequalities, limit theorems, confidence intervals, maximum likelihood, least squares, and hypothesis testing—always with an eye toward both their theoretical underpinnings and practical relevance. While our approach follows the broad contours of conventional expositions, we depart from tradition by consistently exploring the geometric aspects of probability, particularly the invariance properties of normally distributed random vectors. This geometric perspective is taken further in an extended appendix, where we introduce the rudiments of Brownian motion and the corresponding stochastic calculus, culminating in Itô's celebrated change-of-variables formula. To highlight its scope and elegance, we present some of its most striking applications: the sharp Gaussian concentration inequality (a central example of the "concentration of measure phenomenon"), the Feynman–Kac formula (used to derive a path integral representation for the Laplacian heat kernel), and, as a concluding delicacy, the Black–Scholes strategy in Finance.

CONTENTS

1. Introduction

Probability Theory is a multifaceted intellectual enterprise, at once rigorous in its mathematical formulation and remarkably flexible in its applications. It has become indispensable across pure mathematics, applied sciences, and the ever-expanding universe of data-driven disciplines. Among its many usages in real-world problems, there is a particular body of knowledge that stands out: Parametric Estimation. Rooted in the classical contributions of K.F. Gauss and P.-S. Laplace—especially in their formulation of the method of least squares—and later transformed into a systematic research program by R.A. Fisher, J. Neyman, E.S. Pearson and others, this framework of ideas has grown into one of the central pillars of modern statistics. Its methods permeate fields as diverse as Biology, Medicine, and Evolutionary theory, Psychology, Sociology, and Economics, not to mention their longstanding role in Physics, Chemistry, and Engineering. More recently, suitable refinements of this classical theory have proven crucial in assessing the efficiency of statistical procedures

in modern Data Science and Machine Learning, as exemplified in the study of regression, classification, and sparsity-driven approaches within Statistical Learning [JWHT13].

The purpose of these notes is to provide a modest yet comprehensive introduction to this circle of ideas, written for those with adequate preparation in the necessary prerequisites—essentially Linear Algebra, Multivariate Calculus, and Measure Theory, with a touch of Fourier Analysis. Such background ensures the technical fluency required to follow the arguments and the mathematical maturity, at roughly the graduate level, to connect conceptual and computational aspects of the theory. The guiding principle is to move from the foundational elements of Probability Theory to a systematic development of estimation methods, always showing how abstract reasoning translates into applications to statistical problems.

With this in mind, the exposition opens in Sections 2 and 3 with the fundamental notions of Probability Theory, including the key concepts of independence and conditioning, thereby establishing the groundwork for the developments that follow. Section 4 builds on this foundation by introducing the essentials of the theory of normally distributed (Gaussian) random vectors, whose geometric properties reappear throughout the text as a unifying theme. From this point, Section 5 ventures into the domain of concentration inequalities, which provide non-asymptotic bounds for tail probabilities, with the discussion centering on inequalities derived from the elementary yet powerful Cramér–Chernoff method. In addition to showing how this approach leads to the Johnson–Lindenstrauss lemma and to phase-transition behavior in the Erdős–Rényi random graph model, we also discuss the Gaussian concentration inequality, emphasizing its connection to Poincaré's limit theorem and to the broader "concentration of measure phenomenon".

The exposition then transitions naturally to the asymptotic framework in Section 6, where the law of large numbers and the central limit theorem are derived within the Fourier-analytic setting of characteristic functions. These classical results form the theoretical foundation for constructing large-sample confidence intervals for the expectation of virtually any distribution of practical relevance and serve as a cornerstone for many developments in the subsequent sections. Section 7 turns to estimation proper, focusing on the role of statistical models (as formalized by Fisher) and the performance of estimators as measured by the mean squared error, alongside related notions such as consistency, bias-variance trade-off, and asymptotic normality. Building on this, Section 8 develops the method of maximum likelihood, analyzing its asymptotic behavior in the light of Fisher information and the celebrated Cramér–Rao lower bound, before establishing the optimal asymptotic normality of maximum likelihood estimators. Section 9 is devoted to the method of least squares. It begins with a careful presentation of the statistical model underlying ordinary least squares and proceeds to its inferential ramifications, including confidence intervals, hypothesis tests, and measures of goodness-of-fit. The discussion emphasizes both interpretability and predictive accuracy in the classical low-dimensional regime ($p \ll n$). The section then moves to the challenges posed by higher-dimensional settings, introducing regularization techniques and sparsity via the LASSO as a natural gateway to the methods that dominate modern Data Science; see, for instance, [JWHT13, Chapter 6]. Section 10 offers an introduction to the exponential family and the theory of generalized linear models, thereby unifying under a single framework many of the models encountered in practice (logistic, Poisson, gamma, among others), and extending some of the regression tools of Section 9 beyond the normal case. Section 11 discusses sufficiency, while Section 12 completes the standard estimation package with hypothesis testing, emphasizing likelihood ratio tests and illustrating them with canonical examples. Further perspectives appear in Section 13, which revisits "classical" estimation theory and highlights Fisher's enduring contributions, and in Section 14, which presents the Bayesian pathway as an alternative and increasingly influential approach.

A substantial Appendix, requiring only familiarity with the material up to Subsection 4.1, is devoted to Brownian motion and Itô's calculus. It begins with the construction of Brownian paths and culminates in Itô's celebrated change-of-variables formula. This framework allows for a complete proof of the sharp Gaussian concentration inequality, which provides the exact constant for the associated standard deviation, and demonstrates the strength of stochastic calculus through further applications such as the Feynman–Kac formula,

yielding a path-integral representation of the Laplacian heat kernel, and the Black–Scholes model in Finance, a landmark contribution that transformed financial mathematics and earned a Nobel Prize in 1997.

The presentation draws on both classical references and recent monographs, many of which are cited throughout[1]. At certain points, however, the exposition deliberately departs from the conventional treatment by emphasizing geometric perspectives in the theory. These include Fisher's elegant method for deriving the distributions of ubiquitous statistics such as the chi-square, Student's t, and correlation coefficients, the profound link between Scheffé-type simultaneous confidence bands for the mean response in the linear model and Weyl's tube formula, and the illuminating interplay between concentration inequalities and the framework of high-dimensional probability, a theme that has become pivotal in contemporary Data Science. A further example is provided by the well-known dichotomy between model interpretability and prediction accuracy in linear models, where the geometry migrates from projections in sample space (as in the realm of ordinary least squares) to constraints in parameter space through regularization methods (such as ridge regression and the LASSO), illustrating how classical approaches adapt to remain effective in modern contexts. Taken together, these excursions highlight the central geometric role of normally distributed random vectors across key aspects of the theory.

To complement the theoretical development, the text incorporates numerous illustrative applications and contextual remarks that accentuates real-world connections, helping to settle abstract concepts within the broader landscape of probability and statistics. In the same spirit of bridging abstract theory with concrete practice, the notes are accompanied by a series of computational labs made available at

`https://github.com/levilopesdelima/stat-inference-labs`

along with updates and corrections of this text[2]. These labs are not intended as a mere repository of R code snippets. Instead, they are designed to bridge the gap between abstract reasoning and practical implementation, as they combine concise theoretical recaps with step-by-step simulations, numerical experiments, and visualizations. By aligning formal arguments with computational exploration, they provide a structured way to consolidate conceptual understanding while also illustrating how probabilistic methods come to life in practice.

2. The fundamentals of Probability Theory

In this section, we present a concise overview of Probability Theory (or, more precisely, of those aspects most relevant to the applications that follow). Since this material is covered in detail in many standard references, and our primary goal here is merely to establish notation and recall essential facts, proofs are only outlined or omitted altogether.

2.1. **The probabilistic setup: random variables and their distributions.**

Let $\Omega \neq \varnothing$ be a set and consider $\mathcal{F}$ a collection of subsets of Ω.

Definition 2.1. We say that $\mathcal{F}$ is a *σ-algebra* if

- $\Omega \in \mathcal{F}$;
- $A \in \mathcal{F} \Rightarrow A^c \in \mathcal{F}$;

[1] I am also grateful for the many insightful conversations with colleagues, which have helped shape and refine not only these notes but also the companion computational labs. Special thanks go to C. Barroso and J. F. Montenegro (UFC), T. Alencar and J. Silva (UFCA), J. X. da Cruz Neto (UFPI) and J. Stoyanov.

[2] The topics currently covered in the labs include linear regression, the central limit theorem, the Johnson–Lindenstrauss lemma, maximum likelihood estimation and James–Stein estimators, among others.

- $\{A_i\}_{i=1}^{+\infty} \subset \mathcal{F} \Rightarrow \cup_{i=1}^{+\infty} A_i \in \mathcal{F}$.

Trivial examples of σ-algebras are $\mathcal{F} = \{\varnothing, \Omega\}$ and $\mathcal{F} = 2^{\Omega}$, the set of all subsets of Ω. More generally, if $\mathcal{U} = \{U_\lambda\}_{\lambda\in\Lambda}$ is any collection of subsets of Ω, we denote by $\mathcal{F}_{\mathcal{U}} = \mathcal{F}(U_\lambda)$ the σ-algebra generated by $\mathcal{U}$. By definition, this is the smallest σ-algebra contained all elements of $\mathcal{U}$. For example, if $\mathcal{O}^n$ is the set of open subsets in $\mathbb{R}^n$ then $\mathcal{B}^n := \mathcal{F}_{\mathcal{O}_n}$ is the σ-algebra of Borel subsets.

Definition 2.2. A *measure* on $(\Omega, \mathcal{F})$ is a real valued function P on $\mathcal{F}$ so that:

- $P(\varnothing) = 0$;
- $P(A) \geq 0$, for any $A \in \mathcal{F}$;
- if $\{A_i\}_{i=1}^{+\infty} \subset \mathcal{F}$ satisfies $A_i \cap A_j = \varnothing$, $i \neq j$, then

$$P(\cup_i A_i) = \sum_i P(A_i).$$

We say that a triple $(\Omega, \mathcal{F}, P)$ is a *measure space*. A classical example is $(\mathbb{R}^n, \mathcal{L}^n, \lambda^n)$, where $\mathcal{L}^n$ is the standard completion of $\mathcal{B}^n$ and λ^n is Lebesgue measure. If $P(\Omega) = 1$ then we say that $(\Omega, \mathcal{F}, P)$ is a *probability space*, a basic notion in Probability Theory. In this setting, and when no confuson arises, we will represent the corresponding Lebesgue spaces simply by $L^p(\Omega)$, $1 \leq p < +\infty$, with no further reference to the additional data defining the associated probability space. Also, each set $A \in \mathcal{F}$ is called an *event*.

Another key notion is that of *random vector*. If $(\Omega, \mathcal{F}, P)$ is a probability space, this is a function $X : (\Omega, \mathcal{F}) \to (\mathbb{R}^n, \mathcal{B}^n)$ which is measurable in the sense that $X^{-1}(B) \in \mathcal{F}$ for any $B \in \mathcal{B}^n$. When $n = 1$ we say that X is a *random variable*. If X is a random vector then we denote by $\mathcal{F}_X$ the σ-algebra generated by X, i.e. the σ-algebra generated by $\{X^{-1}(B); B \in \mathcal{B}^n\}$. Alternatively, $\mathcal{F}_X$ is the smallest σ-algebra of $\mathcal{F}$ with respect to which X remains measurable.

A central notion in Probability is that of *independence*[3]. Here we define it at several levels, shifting the emphasis from events to random variables.

Definition 2.3. In the setting above, we adopt the following notions of independence.

(1) A finite collection of events $A_1, \ldots, A_k \in \mathcal{F}$, $k \geq 2$, is said to be *independent* if

$$P(A_1 \cap \cdots \cap A_k) = P(A_1) \cdots P(A_k). \tag{2.1}$$

(2) Let $\{\mathcal{F}_\lambda\}_{\lambda\in\Lambda}$ be an arbitrary family of σ-subalgebras of $\mathcal{F}$. We say that $\{\mathcal{F}_\lambda\}_{\lambda\in\Lambda}$ is *independent* if for every finite subcollection $\{\mathcal{F}_{\lambda_\ell}\}_{\ell=1}^k$ and events $A_{\lambda_\ell} \in \mathcal{F}_{\lambda_\ell}$, the events $\{A_{\lambda_\ell}\}_{\ell=1}^k$ are independent.

(3) A family $\{X_\lambda\}_{\lambda\in\Lambda}$ of random variables defined on the same probability space is said to be *independent* if $\{\mathcal{F}_{X_\lambda}\}_{\lambda\in\Lambda}$ is independent. (Notation: for a pair of independent random variables, we write $X \perp Y$.)

Remark 2.4. A finite collection of events $A_1, \ldots, A_k \in \mathcal{F}$, as in (1) above, is said to be *mutually independent* if, for any $2 \leq l \leq k$ and any subcollection of indices $1 \leq i_1 < \cdots < i_l \leq k$, one has

$$P(A_{i_1} \cap \cdots \cap A_{i_l}) = P(A_{i_1}) \cdots P(A_{i_l}).$$

[3]As already noted in [Kol18, Secton I.5]: "Historically, the independence of experiments and random variables represents the very mathematical concept that has given the theory of probabilities its peculiar stamp".

This requirement is, of course, much stronger than (2.1), since it entails the verification of $2^k - k - 1$ conditions rather than a single one; see [Sto13, Section 3] for a thorough discussion of this well-known distinction. Instead of including this stronger notion in item (1) of Definition 2.3, we deliberately depart from tradition and adopt the weaker notion for at least two reasons. First, it fits naturally into the hierarchy of levels presented there, which progresses from events to random variables (cf. the comments immediately following Corollary 2.7). Second, the notion of independence of events is rarely (if ever!) used directly in the text, since the emphasis here is always on exploring independence at the level of random variables, which is properly defined in item (3) and later reinterpreted via the product rule for the corresponding joint distribution (Proposition 2.15). □

We now consider the *expectaction* (or *expected value*) of a random vector $X : \Omega \to \mathbb{R}^n$, which is given by

$$\mathbb{E}(X) := \int_\Omega X \, dP \in \mathbb{R}^n.$$

Usually we assume that this is finite (that is, X is integrable: $X \in L^1(\Omega)$). Another key notion is that of *covariance* of two random variables: if $X, Y : \Omega \to \mathbb{R}$ then this is given by

$$\begin{aligned} \operatorname{cov}(X,Y) &= \mathbb{E}\left((X - \mathbb{E}(X))(Y - \mathbb{E}(Y))\right) \\ &= \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y). \end{aligned}$$

Here, we require that $X, Y \in L^2(\Omega)$ as this implies that $XY \in L^1(\Omega)$ by Cauchy-Schwarz.

Definition 2.5. We say that X and Y are *uncorrelated* if $\operatorname{cov}(X,Y) = 0$.

That uncorrelatedness pertains to independence is a consequence of the next fundamental result.

Proposition 2.6. *If $X, Y : \Omega \to \mathbb{R}$ are independent random variables then $\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$.*

Proof. (sketch) By a simple approximation, it suffices to assume that X and Y are simple functions with $|X|, |Y| \leq M < +\infty$. Hence[4],

$$X = \sum_i a_i \mathbf{1}_{F_i}, \quad Y = \sum_j b_j \mathbf{1}_{G_j},$$

with $F_i = X^{-1}(a_i) \in \mathcal{F}_X$ and similarly for G_j. Hence,

$$XY = \sum_{ij} a_i b_j \mathbf{1}_{F_i \cap G_j},$$

so that

$$\mathbb{E}(XY) = \sum_{ij} a_i b_j P(F_i \cap G_j).$$

[4]Here and in the following, if $A \in \mathcal{F}$ we shall denote by $\mathbf{1}_A$ the corresponding *indicator function*

$$\mathbf{1}_A(x) = \begin{cases} 1 & \text{if } x \in A; \\ 0 & \text{otherwise} \end{cases}$$

Using that $\{\mathcal{F}_X, \mathcal{F}_Y\}$ is independent together with (2.1) we thus see that

$$\begin{aligned}\mathbb{E}(XY) &= \sum_{ij} a_i b_j P(F_i)P(G_j) \\ &= \sum_i a_i P(F_i) \cdot \sum_j b_j P(G_j) \\ &= \mathbb{E}(X)\mathbb{E}(Y),\end{aligned}$$

as desired. □

Corollary 2.7. *If X and Y are independent then they are uncorrelated.*

In a sense, the argument sketched in the proof of Proposition 2.6 respects the shift in levels adopted in Definition 2.3: we first work at level (1), that of events, which amounts to establishing the result for simple functions, and then use a suitable approximation (essentially the definition of integrals in Lebesgue theory) to move to level (3), where the concept of independence is formulated in terms of random variables. A straightforward proof of Proposition 2.6, which operates directly at the highest level, is provided in Remark 2.16 below. Note also that the converse of Corollary 2.7 does not hold in general. It does, however, hold in the important case where X and Y are the components of a normally distributed random vector; see Proposition 4.11.

Definition 2.8. If $X : \Omega \to \mathbb{R}^n$ is a random vector then its *distribution* (or *law*) is the probability measure $X_\sharp P$ on $\mathbb{R}^n$ given by

$$X_\sharp P(B) = P(X^{-1}(B)), \quad B \in \mathcal{B}^n.$$

We also represent $X_\sharp P$ by P_X and set $P(X \geq a) := P_X([a, +\infty))$, etc. Also, an element $\mathbf{x} \in \mathbb{R}^n$ in the image of a random vector $X : \Omega \to \mathbb{R}^n$ (or equivalently, in $\mathrm{supp}(X)$) is called a *realization* (or *observed value*) of X.

The moral is that *any* random variable X is doomed to mediate between the (rather abstract) probability measure P and its distribution P_X, a more tangible probability measure on $\mathbb{R}$. In this way, X links two complementary levels of description of randomness. While P is often difficult to visualize or describe, since it is defined on the sample space Ω, a purely mathematical construct whose internal structure is rarely made explicit, its distribution P_X lives on $\mathbb{R}$ and can therefore be analyzed through familiar descriptive tools such as densities, cumulative distribution functions, tail probabilities, and quantiles. Furthermore, because realizations of independent copies of X (the random samples in Remark 2.18) provide direct access to the features of P_X, it is this distribution that becomes the natural object of study in statistics, physics, and the applied sciences.

Definition 2.9. If $X : \Omega \to \mathbb{R}$ is a random variable then its *cumulative distribution function (cdf)* is the function $F_X : \mathbb{R} \to [0, 1]$ given by $F_X(x) = X_\sharp P((-\infty, x])$.

Notice that F_X completely determines $X_\sharp P = P_X$. Moreover,

$$F_X(x) = \int_{-\infty}^{x} dP_X, \quad x \in \mathbb{R}. \tag{2.2}$$

Definition 2.10. We say that randon variables $X : \Omega \to \mathbb{R}^n$ and $Y : \Omega' \to \mathbb{R}^n$ are *identically distributed (i.d.)* if $P_X = P_Y$.

Proposition 2.11. *We have*

$$\mathbb{E}(X) = \int_{\mathbb{R}^n} \mathbf{x}\, dP_X,$$

where $\mathbf{x} = (x_1, \cdots, x_n)$ *is the position vector. More generally, if* $f : \mathbb{R}^n \to \mathbb{R}^p$ *is measurable, so that* $f(X) = f \circ X : \Omega \to \mathbb{R}^p$ *is a random vector, then*

$$\mathbb{E}(f(X)) = \int_{\mathbb{R}^n} f(\mathbf{x})\, dP_X, \tag{2.3}$$

where $f(\mathbf{x}) = f \circ \mathbf{x}$.

The notion of distribution may be used to single out two important classes of random vectors.

Definition 2.12. Let $X : \Omega \to \mathbb{R}^n$ be a random vector. We say that

- X is *discrete* if its range $\mathrm{Ran}(X) := X(\Omega) \subset \mathbb{R}^n$ is countable:

$$\mathrm{Ran}(X) = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_k, \cdots\}.$$

  In this case, the map

$$\mathbf{x}_j \in \mathrm{Ran}(X) \mapsto p_j := P_X(\{\mathbf{x}_j\}) \in \mathbb{R}, \quad j = 1, 2, \cdots, \tag{2.4}$$

  is called the *mass distribution function (mdf)* and satisfies

$$\sum_j p_j = 1,$$

  with (2.3) meaning that

$$\mathbb{E}(f(X)) = \sum_j p_j f(\mathbf{x}_j). \tag{2.5}$$

  If needed, in (2.4) we may replace $\mathrm{Ran}(X)$ by $\mathrm{supp}(P_X)$, in which case each $p_j > 0$.
- X is *continuous* if P_X is absolutely continuous with respect to the Lebesgue measure $d\mathbf{x}$. In this case, the Radon-Nykodim derivative

$$\psi_X := \frac{dP_X}{d\mathbf{x}} : \mathbb{R}^n \to \mathbb{R}, \tag{2.6}$$

  is called the *probability density function (pdf)* of X, with (2.3) meaning that

$$\mathbb{E}(f(X)) = \int_{\mathbb{R}^n} f(\mathbf{x})\psi_X(\mathbf{x}) d\mathbf{x}. \tag{2.7}$$

In case X is real, we recall that the *support* of P_X is given by

$$\mathrm{supp}(P_X) = \{x \in \mathbb{R}; F_X(x+\varepsilon) - F_X(x-\varepsilon) > 0 \,\text{for all}\, \varepsilon > 0\}.$$

Also notice that, at least for a distribution whose support $\mathrm{supp}(P_X)$ is contained in some closed, bounded interval, the absolute continuity in the second item above means that the corresponding cdf F_X is absolutely continuous, or equivalently, the following assertions hold:

(1) F'_X exists a.s. and is integrable (both with respect to Lebesgue measure);
(2) there holds

$$F_X(b) - F_X(a) = \int_a^b F'_X(t)dt,$$

where $[a, b] \subset \text{supp}(P_X)$[5].

In particular,

$$\int_{-\infty}^{x} dP_X \overset{(2.2)}{=} F_X(x) = \int_{-\infty}^{x} F'_X(t)dt, \quad x \in \text{supp}(P_X), \tag{2.8}$$

so that from (2.6) we see that

$$\psi_X = F'_X \quad \text{a.s.} \tag{2.9}$$

Finally, note that from (2.7) with $f \equiv 1$ we get

$$\int_{\mathbb{R}^n} \psi_X(\mathbf{x})d\mathbf{x} = 1, \tag{2.10}$$

as expected. Clearly, for $n = 1$ this also follows from (2.8) and (2.9) with $x = +\infty$.

Convention 2.13. Given a random vector $X : \Omega \to \mathbb{R}^n$, we will, unless otherwise specified, assume that it is continuous in the sense introduced above, so that it possesses a probability density function ψ_X. To ensure that the standard tools of calculus (including the fundamental theorem) may be applied, we further assume that ψ_X is piecewise smooth with at most finitely many singularities. It should be stressed, however, that virtually every statement involving integration in the continuous case can, when properly interpreted, be reformulated for the discrete case, and conversely. For example, the right-hand side of (2.5) may be written as

$$\int_{\mathbb{R}^n} f(\mathbf{x})\, dP_X,$$

the abstract Lebesgue integral of f with respect to the discrete measure dP_X. Taking into account (2.6), this expression becomes formally indistinguishable from the "continuous" integral on the right-hand side of (2.7). This highlights the advantage of adopting Lebesgue integration from the outset in modern probability theory.

We now consider random vectors $X_j : (\Omega, \mathcal{F}) \to (\mathbb{R}^{p_j}, \mathcal{B}^{p_j})$, $j = 1, \cdots, n$ with distributions P_{X_j}. We may form the random vector

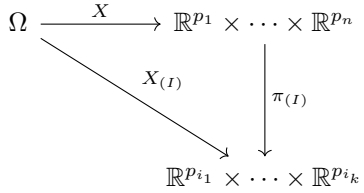
$$(X_1, \cdots, X_n) : (\Omega, \mathcal{F}) \to (\mathbb{R}^{p_1} \times \cdots \times \mathbb{R}^{p_n}, \mathcal{B}^{p_1} \otimes \cdots \otimes \mathcal{B}^{p_n})$$

given by $(X_1, \cdots, X_n)(\omega) = (X_1(\omega), \cdots, X_n(\omega))$, $\omega \in \Omega$, so that the *joint distribution* $P_{(X_1,\cdots,X_n)}$ on $\mathbb{R}^{p_1} \times \cdots \times \mathbb{R}^{p_n}$ is well defined. Moreover, each choice of k distinct indexes, say $I = \{i_1, \cdots, i_k\} \subset \{1, \cdots, n\}$, determines a *marginal distribution* $P_{(X_{i_1},\cdots,X_{i_k})}$ induced by

$$X_{(I)} = (X_{i_1}, \cdots, X_{i_k}) : (\Omega, \mathcal{F}) \to (\mathbb{R}^{p_{i_1}} \times \cdots \times \mathbb{R}^{p_{i_k}}, \mathcal{B}^{p_{i_1}} \otimes \cdots \otimes \mathcal{B}^{p_{i_k}}).$$

[5]For proofs of these claims we refer to [RF10, Sections 6.4, 6.5, 18.4 and 20.3].

In other words, we have the commutative diagram



where $\pi_{(I)}$ is the associated projection. From this, the next result follows immediately.

Proposition 2.14. *(Joint distribution and pdf of a marginal) With the notation above,*

$$P_{(X_{i_1},\cdots,X_{i_k})}(B) = P_{(X_1,\cdots,X_n)}(B \times \mathbb{R}^{p_{j_1}} \times \cdots \times \mathbb{R}^{p_{j_k}}),$$

where $B \in \mathcal{B}^{p_{i_1}} \otimes \cdots \otimes \mathcal{B}^{p_{i_k}}$ and $\{1,\cdots,n\} = \{i_1,\cdots,i_k\} \cup \{j_1,\cdots,j_{n-k}\}$, a disjoint union. In particular, in the continuos case,

$$\psi_{(X_{i_1},\cdots,X_{i_k})}(\mathbf{x}_{i_1},\cdots,\mathbf{x}_{i_k}) = \int_{\mathbb{R}^{p_{j_1}} \times \cdots \times \mathbb{R}^{p_{j_k}}} \psi_{(X_1,\cdots,X_n)}(\mathbf{x}_1,\cdots,\mathbf{x}_n) d\mathbf{x}_{j_1} \cdots d\mathbf{x}_{j_{n-k}}.$$

The next result provides a way of handling independence which is quite satisfactory from an operational viewpoint.

Proposition 2.15. *$\{X_j\}_{j=1}^n$ is independent if and only if*

$$P_{(X_1,\cdots,X_n)} = P_{X_1} \otimes \cdots \otimes P_{X_n},$$

the product measure. Equivalently, in terms of the corresponding pdfs,

$$\psi_{(X_1,\cdots,X_n)}(\mathbf{x}_1,\cdots,\mathbf{x}_n) = \psi_{X_1}(\mathbf{x}_1)\cdots\psi_{X_n}(\mathbf{x}_n), \quad (\mathbf{x}_1,\cdots,\mathbf{x}_n) \in \mathbb{R}^{p_1} \times \cdots \times \mathbb{R}^{p_n}.$$

If $\{X_j\}_{j=1}^n$ is independent, then, with the notation above, a straightforward combination of Propositions 2.15 and 2.14 shows that for any $2 \leq k \leq n$ and any collection of indices $\{i_1,\ldots,i_k\} \subset \{1,\ldots,n\}$, the subcollection $\{X_{i_l}\}_{l=1}^k$ is independent as well. Thus, independence at the level of the full collection $\{X_j\}_{j=1}^n$ automatically propagates to all finite subcollections. This observation shows that the notion of independence for random variables in Definition 2.3, item (3), is fully consistent with (and may be recovered from) the distribution-based formulation given in Proposition 2.15.

Remark 2.16. We may now give a direct proof of Proposition 2.6: since

$$\mathbb{E}(XY) = \iint_{\mathbb{R}^2} xy\, dP_{(X,Y)}(x,y),$$

it follows from Proposition 2.15 and Fubini that

$$\begin{aligned} \mathbb{E}(XY) &= \iint_{\mathbb{R}^2} xy\, dP_X(x) \otimes dP_Y(y) \\ &= \int_{\mathbb{R}} x\, dP_X(x) \int_{\mathbb{R}} y\, dP_Y(y) \\ &= \mathbb{E}(X)\mathbb{E}(Y), \end{aligned}$$

as desired. □

Remark 2.17. If $X : \Omega \to \mathbb{R}$ is a random variable and $Y = X^2$ we may compute ψ_Y in terms of ψ_X as follows. Since $F_Y \leq y$ if and only if $-\sqrt{y} \leq F_X \leq \sqrt{y}$, it follows that

$$F_{X^2}(y) = (F_X(\sqrt{y}) - F_X(-\sqrt{y}))\,\mathbf{1}_{[0,+\infty)]}(y),$$

so that from (2.9) we get

$$\psi_{X^2}(y) = F'_{X^2}(y) = \frac{1}{2\sqrt{y}}\left(\psi_X(\sqrt{y}) + \psi_X(\sqrt{-y})\right)\mathbf{1}_{(0,+\infty)]}(y).$$

A similar computation shows that

$$\psi_{\sqrt{X}}(x) = 2x\psi_X(x^2)\mathbf{1}_{[0,+\infty)]}(x) \tag{2.11}$$

if $X \geq 0$. Also, if Z and V are given with $V > 0$ and $Z \perp V$ then, in terms of the joint distribution $P_{(V,Z)}$,

$$F_{Z/V}(x) = P(Z \leq xV) = \iint_{\{z \leq xv\}} dP_{(V,Z)}(v,z).$$

By the independence and Proposition 2.15 we may write this as an iterated integral,

$$\begin{aligned} F_{Z/V}(x) &= \int_0^{+\infty}\left(\int_{\{z\leq xv\}} dP_Z(z)\right)dP_V(v) \\ &= \int_0^{+\infty} F_Z(xv)\psi_V(v)dv, \end{aligned}$$

and upon derivation we find that

$$\psi_{Z/V}(x) = \int_0^{+\infty} \psi_Z(xv)v\psi_V(v)dv. \tag{2.12}$$

Under the same conditions we also have that

$$F_{ZV}(x) = \int_0^{+\infty} F_Z(xv^{-1})\psi_V(v)dv, \tag{2.13}$$

and hence we obtain

$$\psi_{ZV}(x) = \int_0^{+\infty} \psi_Z(xv^{-1})v^{-1}\psi_V(v)dv,$$

again upon derivation □

Remark 2.18. (Drawing a random sample from a population) If $X_j : (\Omega_j, \mathcal{F}_j, P^{(j)}) \to (\mathbb{R}, \mathcal{B})$, $j = 1, \cdots, n$, are random variables, form the product probability space

$$(\Omega^\sharp, \mathcal{F}^\sharp, P^\sharp) = \otimes_j(\Omega_j, \mathcal{F}_j, P^{(j)}),$$

and define the random variables $Y_j : (\Omega^\sharp, \mathcal{F}^\sharp, P^\sharp) \to (\mathbb{R}, \mathcal{B})$, $Y_j = X_j \circ \pi_j$, where $\pi_j : \mathbb{R}^n \to \mathbb{R}$ is the canonical projection onto the j^{th} factor. Now, if $B_j \in \mathcal{B}$ we have

$$\begin{aligned} P^{(j)}_{X_j}(B_j) &= P^{(j)}(X_j^{-1}(B_j)) \\ &= P^\sharp(\Omega_1 \times \cdots \times \Omega_{j-1} \times X_j^{-1}(B_j) \times \Omega_{j+1} \times \cdots \times \Omega_n) \\ &= P^\sharp(\pi_j^{-1}(X_j^{-1}(B_j))) \\ &= P^\sharp(Y_j^{-1}(B_j)), \end{aligned}$$

so that $P^{(j)}_{X_j} = P^\sharp_{Y_j}$ for any j. On the other hand, if $P^\sharp_Y$ is the joint distribution of the random vector $Y = (Y_1, \cdots, Y_n) : (\Omega^\sharp, \mathcal{F}^\sharp, P^\sharp) \to (\mathbb{R}^n, \mathcal{B}^n)$,

$$\begin{aligned} P^\sharp_Y(B_1 \times \cdots \times B_n) &= P^\sharp(Y^{-1}(B_1 \times \cdots \times B_n)) \\ &= P^\sharp(X_1^{-1}(B_1) \times \cdots \times X_n^{-1}(B_n)) \\ &= \Pi_j P^{(j)}(X_j^{-1}(B_j)) \\ &= \Pi_j P^{(j)}_{X_j}(B_j) \\ &= \Pi_j P^\sharp_{Y_j}(B_j) \\ &= (\otimes_j P^\sharp_{Y_j})(B_1 \times \cdots \times B_n), \end{aligned}$$

so that $P^\sharp_Y = \otimes_j P^\sharp_{Y_j}$. Thus, by Proposition 2.15, $\{Y_j\}_{j=1}^n$ is independent. Note that if $\{X_j\}$ is identically distributed then $\{Y_j\}$ is identically distributed as well, so we conclude: *given any random variable X there exist $\{Y_j\}_{j=1}^n$ which is independent and identically distributed to X*. It turns out to be a bit more involved to extend this construction to *countably* many random variables, as this involves the consideration of a suitable *infinite* product of sample spaces [FG13, Section 9.6]. In any case, this stronger assertion, which is a rather special case of Kolmogorov's extension (Theorem A.3), may indeed be regarded as the *Fundamental Theorem of Statistical Inference*, since it provides the very mathematical foundation for the notion of an *infinite random sample* drawn from a given population, a terminology we shall use below without further notice. The existence of such a sequence $\{Y_j\}_{j\geq 1}$ of i.i.d. random variables is an indispensable assumption underlying virtually all limit results in Statistics—most notably, the *Law of Large Numbers* (Theorem 6.2) and the *Central Limit Theorem* (Theorem 6.5)—and, more broadly, the very formulation of a *statistical model* in the sense introduced by Fisher (Definition 7.2), whereby one postulates a parametric family of probability laws governing the distribution of these random samples. In this way, the abstract (possibly infinite) product construction above bridges the axiomatic framework of Probability with the empirical foundations of Statistical Inference. □

Definition 2.19. If $X_1, \cdots, X_n : \Omega \to \mathbb{R}$ are random variables (equivalently, $X = (X_1, \cdots, X_n) : \Omega \to \mathbb{R}^n$ is a random vector) we define its *covariance matrix* by

$$\mathrm{cov}(X, X)_{ij} = \mathrm{cov}(X_i, X_j),$$

where

$$\mathrm{cov}(X_i, X_j) = \mathbb{E}(X_i X_j) - \mathbb{E}(X_i)\mathbb{E}(X_j).$$

a symmetric matrix. In other words,

$$\mathrm{cov}(X, X) = \mathbb{E}(X \otimes X) - \mathbb{E}(X) \otimes \mathbb{E}(X),$$

where $\otimes$ denotes the Kronecker product of vectors and the expectation has been extended to random vectors and matrices in the obvious manner.

In case $n = 1$, $X = X_1$, this defines the *variance* of X:

$$\mathrm{var}(X) = \mathrm{cov}(X, X) = \mathbb{E}(X^2) - (\mathbb{E}(X))^2. \tag{2.14}$$

Since $\mathrm{var}(X) \geq 0$ we usually set $\sigma^2 := \mathrm{var}(X)$ and $\sigma := \sqrt{\mathrm{var}(X)}$, the *standard deviation*, which we also denote by $\mathrm{sd}(X)$. Note also that

$$\mathrm{var}\left(\sum_i X_i\right) = \sum_i \mathrm{var}(X_i) + \sum_{i \neq j} \mathrm{cov}(X_i, X_j).$$

Thus, if the X_i's are pairwise uncorrelated (in particular, if they are independent) then

$$\operatorname{var}\left(\sum_i X_i\right) = \sum_i \operatorname{var}(X_i). \tag{2.15}$$

To simplify the notation we sometimes denote $\operatorname{cov}(X, X)$ by either $\operatorname{cov}(X)$ or $\operatorname{var}(X)$ in case X is vector valued.

In order to properly compare distinct random variables it is sometimes convenient to pass to a suitable normalization. In most cases, this is accomplished as follows.

Definition 2.20. If $X : \Omega \to \mathbb{R}$ is a random variable with $\mathbb{E}(X) = \mu$ and $\operatorname{var}(X) = \sigma^2$ then its *standardization* is

$$Z = \frac{X - \mu}{\sigma}.$$

Note that $\mathbb{E}(Z) = 0$ and $\operatorname{var}(Z) = 1$, hence the terminology.

Im many applications it is useful to estimate from above the tail probabilities of a random variable whose expectation/variance is known. We now present a couple of elementary results in this direction, which can be regarded as examples of (quite conservative) concentration inequalities.

Proposition 2.21. *(Markov's inequality) If $X : \Omega \to \mathbb{R}$ is a non-negative random variable and $a > 0$ then*

$$P(X \geq a) \leq \frac{\mathbb{E}(X)}{a}. \tag{2.16}$$

Proof. Using (2.7) with $f(x) = x$ we compute

$$\begin{aligned} \mathbb{E}(X) &= \int_0^{+\infty} x\psi_X(x)dx \\ &\geq \int_a^{+\infty} x\psi_X(x)dx \\ &= a\int_a^{+\infty} \psi_X(x)dx \\ &= aP(X \geq a), \end{aligned}$$

as desired. □

Corollary 2.22. *(Chebyshev's inequality) Let $X : \Omega \to \mathbb{R}$ be a random variable with $0 < \sigma^2 := \operatorname{var}(X) < +\infty$. Then*

$$P(|X - \mathbb{E}(X| \geq a) \leq \frac{\sigma^2}{a^2}. \tag{2.17}$$

Equivalently,

$$P(|X - \mathbb{E}(X| \geq c\sigma) \leq c^{-2}, \quad c > 0. \tag{2.18}$$

Proof. Note that

$$P(|X - \mathbb{E}(X)| \geq a) = P(|X - \mathbb{E}(X)|^2 \geq a^2)$$

and use (2.16) with X replaced by $|X - \mathbb{E}(X)|^2$ and a replaced by a^2. □

We now discuss the various modes of convergence of random variables.

Definition 2.23. Let $\{X_j\}_{j=1}^{+\infty}$ a sequence of random variables and let X be another random variable (all defined on the same sample space $(\Omega, \mathcal{F}, P)$). We say that

- X_j converges to X *almost surely* (notation: $X_j \overset{a.s.}{\to} X$) if
$$P\left(\lim_{j\to+\infty} X_j = X\right) = 1.$$
- X_j converges to X *in probability* (notation: $X_j \overset{p}{\to} X$) if, for any $\varepsilon > 0$,
$$\lim_{j\to+\infty} P\left(|X_j - X| < \epsilon\right) = 1.$$
- X_j converges to X *in distribution* (notation: $X_j \overset{d}{\to} X$) if
$$\lim_{j\to+\infty} F_{X_j}(x) = F_X(x),$$
for any $x \in \mathbb{R}$ where F_X is continuous. Equivalently, $\mathbb{E}(\xi(X_j)) \to \mathbb{E}(\xi(X))$ for all $\xi : \mathbb{R} \to \mathbb{R}$ uniformly bounded and continuous.
- X_j converges to X *in the mean* (notation: $X_j \overset{m}{\to} X$) if
$$\lim_{j\to+\infty} \mathbb{E}(|X_j - X|^2) = 0.$$

Since, as memtioned in Definition 2.8, F_X as also known as the law of X, convergence in distribution is also referred to as convergence in law (notation: $X_j \overset{l}{\to} X$). Also, the equivalence between the two ways above of defining convergence in distribution is part of the Portmanteau theorem [VdV00, Lemma 2.2].

Proposition 2.24. *One has* $(\overset{a.s.}{\to}) \Rightarrow (\overset{p}{\to}) \Rightarrow (\overset{d}{\to})$. *Also,* $(\overset{m}{\to}) \Rightarrow (\overset{p}{\to})$ *and* $(\overset{d}{\to}) \Rightarrow (\overset{p}{\to})$ *if the limiting variable is constant.*

The following quite useful result is worth mentioning here[6].

Theorem 2.25. *(Slutsky) If* $X_j \overset{d}{\to} X$ *and* $Y_j \overset{p}{\to} c$, $c \in \mathbb{R}$, *then* $X_j + Y_j \overset{d}{\to} X + c$ *and* $X_jY_j \overset{d}{\to} cX$. *Also, if* $Y_j \neq 0$ *and* $c \neq 0$ *then* $X_j/Y_j \overset{d}{\to} X_n/c$. *Finally, these assertions hold true if* $(\overset{d}{\to})$ *gets replaced by* $(\overset{p}{\to})$ *everywhere.*

2.2. **The analytical setup: characteristic functions.** We now introduce an important notion which will allow us to make use of analytical techniques in the theory[7].

[6] We refer [Gut06, Chapter 5] for much more on the convergence properties of random variables.

[7] Here and in the following, we denote the inner product of vectors $\mathbf{a}, \mathbf{b} \in \mathbb{R}^n$ either by $\langle \mathbf{a}, \mathbf{b} \rangle$ or by $\mathbf{a}^\top \mathbf{b} = \mathbf{a}\mathbf{b}^\top$, where the superscript $\top$ indicates transpose of a vector (or of a matrix, more generally). Also, we set $\|\mathbf{a}\|^2 = \mathbf{a}^\top \mathbf{a}$ for the corresponding squared norm.

Definition 2.26. If $X : \Omega \to \mathbb{R}^n$ is a random vector then its *characteristic function* $\phi_X : \mathbb{R}^n \to \mathbb{C}$ is given by

$$\phi_X(\mathbf{u}) = \mathbb{E}(e^{\mathbf{i}\langle X,\mathbf{u}\rangle}) = \int_{\mathbb{R}^n} e^{\mathbf{i}\langle \mathbf{x},\mathbf{u}\rangle} dP_X(\mathbf{x}).$$

If X carries a pdf ψ_X then

$$\phi_X(\mathbf{u}) = \int_{\mathbb{R}^n} e^{\mathbf{i}\langle \mathbf{x},\mathbf{u}\rangle} \psi_X(\mathbf{x}) d\mathbf{x}.$$

Remark 2.27. It is immediate from Definition 2.23 that $X_j \xrightarrow{d} X$ implies $\phi_{X_j} \to \phi_X$ pointwise. □

Since ϕ_X is the (inverse) Fourier transform of P_X, we expect that it completely determines the corresponding cdf F_X. A proof of this general statement, at least in case X is real, may be found in [Gne18, Section 39], where an explicit formula for F_X in terms of ϕ_X is indicated; see also the discussion in [Luk70, Section 3.2]. We present here two instances where this expectation is confirmed (with explicit formulas).

Proposition 2.28. *The following hold:*

(1) *If X is $\mathbb{Z}$-valued and $p_k := P(X = k)$, $k \in \mathbb{Z}$, then*

$$p_k = \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{-\mathbf{i}ku} \phi_X(u) du, \quad k \in \mathbb{Z}. \tag{2.19}$$

(2) *If X is real and has a characteristic function $\phi_X : \mathbb{R} \to \mathbb{C}$ such that $|\phi_X|$ is integrable then its distribution is absolutely continuous with respect to Lebesgue measure with the corresponding pdf being continuous and given by*

$$\psi_X(x) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} e^{-\mathbf{i}xu} \phi_X(u) du, \quad x \in \mathbb{R}. \tag{2.20}$$

Proof. We only prove (2.19) here[8]. If $\operatorname{supp} P_X \subset \mathbb{Z}$ then it is immediate to check that ϕ_X is 2π-periodic. Also,

$$\phi_X(u) = \sum_{l \in \mathbb{Z}} e^{\mathbf{i}lu} p_l, \tag{2.21}$$

where the convergence is uniform. In particular, ϕ_X is continuous. Now integrate over $[-\pi, \pi]$ the product of this series by $e^{-\mathbf{i}ku}$ and use the well-known orthogonality relations for the basis $\{e^{\mathbf{i}mx}\}_{m=-\infty}^{+\infty}$ in order to obtain (2.19). □

Remark 2.29. The inversion formula (2.20) means that $\psi_X = \widehat{\phi_X}$, where the hat means Fourier transform. On the other hand, (2.21) provides the Fourier series expansion of ϕ_X with Fourier coefficients given by (2.19). □

We now describe a simple condition on a random variable ensuring that its characteristic function is sufficiently regular.

Proposition 2.30. *If a random variable X satisfies $\mathbb{E}(|X|^r) < +\infty$ for some $r \geq 1$ then $\phi_X \in C^r(\mathbb{R})$ and*

$$\phi_X^{(j)}(u) = \mathbf{i}^j \mathbb{E}(X^j e^{\mathbf{i}Xu}), \quad u \in \mathbb{R}, \quad j = 1, \ldots, r.$$

[8] A direct proof of the inversion formula (2.20) may be found in [FG13, Chapter 13]; see also [Luk70, Theorem 3.2.2].

In particular, as $u \to 0$,

$$\phi_X(u) = \sum_{j=0}^{r} \frac{\mathbf{i}^j}{j!}\mathbb{E}(X^j)u^j + o(|u|^r).$$

Proof. If $r = 1$ we have $\mathbb{E}(|X|) < +\infty$ and since

$$\frac{e^{\mathbf{i}X(u+h)} - e^{\mathbf{i}Xu}}{h} = \mathbf{i}Xe^{\mathbf{i}Xu} + o(h)$$

we may use dominated convergence to see that

$$\phi'_X(u) = \lim_{h\to 0} \mathbb{E}\left(\frac{e^{\mathbf{i}X(u+h)} - e^{\mathbf{i}Xu}}{h}\right) = \mathbf{i}\,\mathbb{E}\left(Xe^{\mathbf{i}Xu}\right),$$

which proves this case. The general assertion for $r \geq 2$ follows by induction taking into account that

$$\frac{(\mathbf{i}X)^j e^{\mathbf{i}X(u+h)} - (\mathbf{i}X)^j e^{\mathbf{i}Xu}}{h} = (\mathbf{i}X)^j e^{\mathbf{i}Xu} + o(h)$$

and that $\mathbb{E}(|X|^j) \leq \mathbb{E}(|X|^r)^{j/r}$ by Hölder inequality. □

Proposition 2.31. *A real random variable satisfies:*

1. *$\phi_{\alpha X}(u) = \phi_X(\alpha u)$, $\alpha \in \mathbb{R}$. In particular, $\phi_{-X} = \overline{\phi_X}$.*
2. *If $\mu = \mathbb{E}(X)$ is finite then $\phi_X(u) = 1 + u\mu\mathbf{i} + o(|u|)$ as $u \to 0$. Moreover, if $\mu = 0$ and $\sigma^2 = \mathbb{E}(X^2)$ is finite then*

$$\phi_X(u) = 1 - \frac{1}{2}\sigma^2u^2 + o(|u|^2)$$

Proof. (1) is obvious and (2) is an immediate consequence of Proposition 2.30 (after Taylor expanding ϕ_X around $u = 0$). □

We now examine how the characteristic functions of independent random variables contribute to the characteristic and density functions of their sum or difference.

Proposition 2.32. *The following properties hold for independent real random variables X and Y:*

1. *$\phi_{X+Y} = \phi_X\phi_Y$.*
2. *$\psi_{X+Y} = \psi_X \star \psi_Y$, where $\star$ means convolution.*
3. *moreover, if X and Y are identically distributed then $\phi_{X-Y} = |\phi_X|^2$.*

Proof. For (1) note that, in terms of the joint distribution $P_{(X,Y)}$,

$$\begin{aligned}
\phi_{X+Y}(u) &= \iint_{\mathbb{R}^2} e^{\mathbf{i}(x+y)u} dP_{(X,Y)}(x,y) \\
&\overset{(*)}{=} \iint_{\mathbb{R}^2} e^{\mathbf{i}xu}e^{\mathbf{i}yu} dP_X(x) \otimes dP_Y(y) \\
&= \phi_X(u)\phi_Y(u),
\end{aligned}$$

where we used Proposition 2.15 in $(*)$ and Fubini in the last step. Also, by Remark 2.29,

$$\begin{aligned}\psi_{X+Y} &= \widehat{\phi_{X+Y}} \\ &\overset{(3)}{=} \widehat{\phi_X \phi_Y} \\ &\overset{(**)}{=} \widehat{\phi_X} \star \widehat{\phi_X} \\ &= \psi_X \star \psi_Y,\end{aligned}$$

where we used a well-known property of the Fourier transform in $(**)$. Finally, (3) follows from (1) and Proposition 2.31 (1). □

Remark 2.33. The clear contrast between the two types of products that appear on the right-hand sides of items (1) and (2) above already indicates why it is often preferable to work with characteristic functions rather than with the pdfs themselves; see Remark 4.17 for an illustrative example.

Definition 2.34. A random variable X is *symmetric* (about 0) if X and $-X$ are identically distributed.

Proposition 2.35. *X is symmetric if and only if ϕ_X is $\mathbb{R}$-valued, in which case there holds*

$$\phi_X(u) = \mathbb{E}(\cos(Xu)). \tag{2.22}$$

Proof. Immediate from the previous results. □

Definition 2.36. We say that $\boldsymbol{\epsilon}$ is a *Rademacher variable* if $\operatorname{supp} P_{\boldsymbol{\epsilon}} = \{-1, 1\}$ with $P(\boldsymbol{\epsilon} = -1) = P(\boldsymbol{\epsilon} = 1) = 1/2$.

Proposition 2.37. *If $\{\boldsymbol{\epsilon}, X\}$ is independent with X symmetric then X and $\boldsymbol{\epsilon}X$ are identically distributed.*

Proof. The cdf of $\boldsymbol{\epsilon}X$ is

$$\begin{aligned}F_{\boldsymbol{\epsilon}X}(x) &= P(\boldsymbol{\epsilon}X \le x) \\ &= P(\{X \le x\} \cap \{\boldsymbol{\epsilon} = 1\}) + P(\{-X \le x\} \cap \{\boldsymbol{\epsilon} = -1\}),\end{aligned}$$

so independence gives

$$\begin{aligned}F_{\boldsymbol{\epsilon}X}(x) &= P(X \le x)\, P(\boldsymbol{\epsilon} = 1)) + P(-X \le x)\, P(\boldsymbol{\epsilon} = -1)) \\ &= \frac{1}{2}(F_X(x) + F_{-X}(x)) \\ &= F_X(x),\end{aligned}$$

where in the last step we used that $F_X = F_{-X}$. □

We now introduce another important notion which is closely related to characteristic functions.

Definition 2.38. The *moment generating function* (mgf) of a random vector $X : \Omega \to \mathbb{R}^n$ is given by

$$\varphi_X(\mathbf{u}) = \mathbb{E}(e^{\langle X, \mathbf{u}\rangle}), \quad \mathbf{u} \in \mathbb{R}^n, \tag{2.23}$$

whenever the right-hand side is finite.

Remark 2.39. We will always assume that φ_X is defined at least in a neighborhood $V \subset \mathbb{R}^n$ of the origin, which happens if ϕ_X is analytic there [Luk70, Section 7.2]. In this case we have $\varphi_X(\mathbf{u}) = \phi_X(-\mathbf{iu})$, $\mathbf{u} \in V$, a replacement we shall use in the sequel without further notice. Under these conditions, if $X \in \mathbb{R}$ then all the *moments* of X,

$$\alpha_k(X) := \int_{-\infty}^{+\infty} x^k dP_X(s), \quad k = 0, 1, 2, \ldots,$$

are finite with

$$\beta^{-1} := \limsup_k \left(\frac{\alpha_k(X)}{k!} \right)^{1/k} < +\infty,$$

so there holds

$$\varphi_X(u) = \sum_k \frac{\alpha_k(X)}{k!} u^k, \quad u \in (-\beta, \beta). \tag{2.24}$$

Thus, $\alpha_k(X) = \varphi_X^{(k)}(0)$, which justifies the mgf terminology. Note that the expectation and variance of X are given by

$$\mathbb{E}(X) = \varphi_X'(0), \quad \mathrm{var}(X) = \varphi_X''(0) - (\varphi_X'(0))^2, \tag{2.25}$$

with similar formulae holding for higher order centered moments. □

Example 2.40. (Binomial trials as the sum of independent Bernoulli trials) Set $\mathbb{N}^{(n)} := \{0, 1, \cdots, n\}$, $n \geq 1$, and consider a discrete random variable X whose probability distribution is supported on $\mathbb{N}^{(n)}$ and satisfies, for some $0 < p < 1$,

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad k \in \mathbb{N}^{(n)}. \tag{2.26}$$

We then say that $X \sim \mathsf{Bin}(p; n)$, the *binomial distribution* determined by the pair (p, n). Using that the characteristic function of X is

$$\begin{aligned} \phi_X(u) &= \sum_{k=0}^{n} e^{\mathbf{i}ku} \binom{n}{k} p^k (1-p)^{n-k} \\ &= \sum_{k=0}^{n} \binom{n}{k} (pe^{\mathbf{i}u})^k (1-p)^{n-k}, \end{aligned}$$

we obtain

$$\phi_X(u) = (1 - p + pe^{\mathbf{i}u})^n. \tag{2.27}$$

Together with Proposition 2.15, Proposition 2.32 (1) and Proposition 2.28 (1), this shows that, by possibly changing the underlying sample space, we may assume that $X = X_1 + \cdots + X_n$, where $\{X_j\}_{j=1}^n$ is independent and each $X_j \sim \mathsf{Bin}(p; 1) =: \mathsf{Ber}(p)$, the *Bernoulli distribution*[9], so that $\mathbb{E}(X_j) = p$ and $\mathrm{var}(X_j) = p(1-p)$. Finally,

$$\varphi_X(u) = (1 - p + pe^u)^n \tag{2.28}$$

follows immediately from (2.27). □

[9]This kind of distribution models any binary random experiment (such as coin toss, for instance) with $k = 1$ corresponding to success and $k = 0$ corresponding to failure.

Example 2.41. (Poisson trials). For each $n \geq 1$ consider a discrete random variable Y whose probability distribution is supported on $\mathbb{N}_0 = \{0\} \cup \mathbb{N}$ with

$$P(Y = k) = \frac{\lambda^k e^{-\lambda}}{k!}.$$

We represent this as $Y \sim \mathsf{Pois}(\lambda)$, the *Poisson distribution* with parameter λ. We compute:

$$\begin{aligned} \phi_Y(u) &= \sum_{k\geq 0} e^{\mathbf{i}ku} \frac{\lambda^k e^{-\lambda}}{k!} \\ &= e^{-\lambda} \sum_{k\geq 0} \frac{(\lambda e^{\mathbf{i}u})^k}{k!} \\ &= e^{-\lambda} e^{\lambda e^{\mathbf{i}u}}, \end{aligned}$$

which gives

$$\phi_Y(u) = e^{\lambda(e^{\mathbf{i}u}-1)}, \tag{2.29}$$

and hence

$$\varphi_Y(u) = e^{\lambda(e^{u}-1)}. \tag{2.30}$$

In particular, $\mathbb{E}(Y) = \mathrm{var}(Y) = \lambda$. Also, if $Y \sim \mathsf{Pois}(n)$, $n \in \mathbb{N}$, then it follows from (2.29) with $\lambda = n$ that we may decompose $Y = Y_1 + \cdots + Y_n$ with $\{Y_j\}_{j=1}^n$ independent and each $Y_j \sim \mathsf{Pois}(1)$, so that $\mathbb{E}(Y_j) = \mathrm{var}(Y_j) = 1$. □

Example 2.42. (The geometric distribution) Let Z be a discrete random variable whose probability distribution is supported on $\mathbb{N}$ with

$$P(Z = k) = (1-p)^{k-1}p, \quad k \geq 1,$$

which gives the probability that the first occurrence of success in a sequence of independent Bernoulli trials as in Example 2.40 requires exactly k steps. This is called the *geometric distribution* for an obvious reason:

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

Similarly,

$$\begin{aligned} \phi_Z(u) &= \sum_{k\geq 1} (1-p)^{k-1} p e^{\mathbf{i}u} \\ &= \frac{p}{1-p} \sum_{k\geq 1} \left((1-p)e^{\mathbf{i}u}\right)^k \\ &= \frac{pe^{\mathbf{i}u}}{1-(1-p)e^{\mathbf{i}u}}, \end{aligned}$$

so that

$$\varphi_Z(u) = \frac{p}{p-1+e^{-u}}, \quad u < -\ln(1-p).$$

From this and (2.25) we easily see that $\mathbb{E}(Z) = 1/p$ and $\mathrm{var}(Z) = (1-p)/p^2$. □

3. CONDITIONING

We now discuss the various ways of conditioning a given random variable.

3.1. **Conditional probability.** Let $X : \Omega \to \mathbb{R}^m$ and $Y : \Omega \to \mathbb{R}^p$ be random vectors with distributions P_X and P_Y, respectively, defined on a probability space $(\Omega, \mathcal{F}, P)$. We denote by $P_{(X,Y)}$ the joint distribution of $(X, Y) : \Omega \to \mathbb{R}^m \times \mathbb{R}^p$. As usual, we assume that all these distributions are absolutely continuous with respect to the Lebesgue measure and therefore admit probability density functions (pdfs). Given $B \in \mathcal{B}^p$, our aim is to define the *conditional probability* that $Y \in B$ given a realization $\mathbf{x} \in \mathbb{R}^m$ of X.

Definition 3.1. A *Markov kernel* is a map $\kappa : \mathbb{R}^m \times \mathcal{B}^p \to [0, 1]$ such that:

- for each $B \in \mathcal{B}^p$, the map $\mathbf{x} \mapsto \kappa(\mathbf{x}, B)$ is $\mathcal{B}^m$-measurable;
- for each $\mathbf{x} \in \mathbb{R}^m$, the map $B \mapsto \kappa(\mathbf{x}, B)$ is a probability measure on $(\mathbb{R}^p, \mathcal{B}^p)$.

Given a Markov kernel κ and a probability measure μ on $(\mathbb{R}^m, \mathcal{B}^m)$, the rule

$$(A, B) \longmapsto (\mu \star \kappa)(A, B) := \int_A \kappa(\mathbf{x}, B)\, d\mu(\mathbf{x}), \qquad (A, B) \in \mathcal{B}^m \times \mathcal{B}^p,$$

defines a probability measure on $(\mathbb{R}^m \times \mathbb{R}^p, \mathcal{B}^m \times \mathcal{B}^p)$. The following result, known as the *disintegration theorem*, asserts the existence of a unique Markov kernel that plays the role of a "conditional quotient" of $P_{(X,Y)}$ by P_X under the convolution operation $\star$.

Proposition 3.2. *There exists a unique Markov kernel $P_{Y|X}$ such that*

$$P_{(X,Y)} = P_X \star P_{Y|X}.$$

Equivalently,

$$P_{(X,Y)}(A, B) = \int_A P_{Y|X}(\mathbf{x}, B)\, dP_X(\mathbf{x}), \qquad (A, B) \in \mathcal{B}^m \times \mathcal{B}^p. \tag{3.1}$$

Proof. See [Kle13, Chapter 8]. □

The kernel $P_{Y|X}$ thus provides the precise object that realizes, in measure-theoretic terms, the intuitive idea of conditioning on a given value of X.

Definition 3.3. If $\mathbf{x} \in \mathbb{R}^m$ we define the *conditional probability* by

$$P_{Y|_{X=\mathbf{x}}} = P_{Y|X}(\mathbf{x}, \cdot),$$

which is a probability measure in $(\mathbb{R}^p, \mathcal{B}^p)$.

Thus,

$$P(Y \in B|_{X=\mathbf{x}}) := P_{Y|_{X=\mathbf{x}}}(B) = P_{Y|X}(\mathbf{x}, B), \quad B \in \mathcal{B}^p, \tag{3.2}$$

should be interpreted as the *conditional probability* that $Y \in B$ given that $X = \mathbf{x}$. It is immediate from (3.1) that the corresponding pdf's satisfy

$$\psi_{Y|_{X=\mathbf{x}}}(\mathbf{y}) = \frac{\psi_{(X,Y)}(\mathbf{x}, \mathbf{y})}{\psi_X(\mathbf{x})}, \quad \mathbf{y} \in \mathbb{R}^p, \tag{3.3}$$

whenever $\psi_X(\mathbf{x}) > 0$, so that the corresponding *conditional expectation function* and *conditional covariance function* are

$$\mathbb{E}(Y|_{X=\mathbf{x}}) = \int_{\mathbb{R}^p} \mathbf{y}\psi_{Y|_{X=\mathbf{x}}}(\mathbf{y})d\mathbf{y} \tag{3.4}$$

and

$$\operatorname{cov}(Y|_{X=\mathbf{x}}) = \int_{\mathbb{R}^p} (\mathbf{y} - \mathbb{E}(Y|_{X=\mathbf{x}})^2\psi_{Y|_{X=\mathbf{x}}}(\mathbf{y})d\mathbf{y}, \tag{3.5}$$

respectively.

Remark 3.4. Whenever possible, we may simply dispense with the existence theory sketched above and adopt (3.3) as the definition of the *conditional pdf* of Y given $X = \mathbf{x}$. □

We now turn to a few elementary yet useful consequences of the preceding theory. For simplicity, throughout the remainder of this subsection we assume that all random variables are real-valued, continuous, and possess probability density functions that are strictly positive on their domains. In particular, with a slight abuse of notation, (3.3) may be interpreted as the pdf of the *conditioned random variable* $Y|_{X=\mathbf{x}}$, thus allowing the natural extension of the concepts introduced so far for random variables to this broader setting.

Proposition 3.5. *The following hold:*
(1) *If $\{X, Y\}$ is independent then $\mathbb{E}(Y) = \mathbb{E}(Y|_{X=\mathbf{x}})$ for any $\mathbf{x}$;*
(2) *If $\{X, Y, Z\}$ is independent then $\{X|_{Z=\mathbf{z}}, Y|_{Z=\mathbf{z}}\}$ is independent for any $\mathbf{z}$.*

Proof. For (1) note that by (3.3) and Proposition 2.15,

$$\psi_{Y|_{X=\mathbf{x}}}(\mathbf{y}) = \frac{\psi_{(X,Y)}(\mathbf{x},\mathbf{y})}{\psi_X(\mathbf{x})} = \frac{\psi_X(\mathbf{x})\psi_Y(\mathbf{y})}{\psi_X(\mathbf{x})} = \psi_Y(\mathbf{y}). \tag{3.6}$$

As for (2), again by Proposition 2.15,

$$\begin{aligned} \psi_{(X|_{Z=\mathbf{z}},Y|_{Z=\mathbf{z}})}(\mathbf{x},\mathbf{y}) &= \frac{\psi_{(X,Y,Z)}(\mathbf{x},\mathbf{y},\mathbf{z})}{\psi_Z(\mathbf{z})} \\ &= \frac{\psi_X(\mathbf{x})\psi_Y(\mathbf{y})\psi_Z(\mathbf{z})}{\psi_Z(\mathbf{z})} \\ &\overset{(3.6)}{=} \psi_{X|_{Z=\mathbf{z}}}(\mathbf{x})\psi_{Y|_{Z=\mathbf{z}}}(\mathbf{y}), \end{aligned}$$

and the result follows from the "conditioned" version of Proposition 2.15. □

Remark 3.6. In general the converses to both items in Proposition 3.5 fail to hold true if the independence assumptions are removed. □

Finally, we present another consequence of (3.3) with notable applications to the so-called Bayesian approach to Statistical Inference; see Subsection 14.

Theorem 3.7. *(Bayes rule) If both ψ_X and ψ_Y are everywhere positive then*

$$\psi_{X|_{Y=\mathbf{y}}}(\mathbf{x}) = \frac{\psi_{Y|_{X=\mathbf{x}}}(\mathbf{y})\psi_X(\mathbf{x})}{\psi_Y(\mathbf{y})}, \quad (\mathbf{x},\mathbf{y}) \in \mathbb{R}^m \times \mathbb{R}^q,$$

with

$$\psi_Y(\mathbf{y}) = \int_{\mathbb{R}^m} \psi_{Y|_{X=\mathbf{x}}}(\mathbf{y})\psi_X(\mathbf{x})d\mathbf{x}.$$

Proof. It follows from (3.3) that

$$\psi_{(X,Y)}(\mathbf{x},\mathbf{y}) = \psi_{Y|_{X=\mathbf{x}}}(\mathbf{y})\psi_X(\mathbf{x}) = \psi_{X|_{Y=\mathbf{y}}}(\mathbf{x})\psi_Y(\mathbf{y}),$$

and the result follows. □

3.2. **Conditional expectation.** Here we discuss how to condition a random variable with respect to a σ-subalgebra and then relate this to the discussion in the previous subsection (via Proposition 3.14).

Proposition 3.8. *Let $(\Omega,\mathcal{F},P)$ be a probability space and let $\mathcal{G}\subset\mathcal{F}$ a σ-subalgebra. Given a random vector $X:\Omega\to\mathbb{R}^n$ there exists a unique random vector $Y:\Omega\to\mathbb{R}^n$ which is $\mathcal{G}$-measurable and satisfies*

$$\int_G Y dP = \int_G X dP, \quad G\in\mathcal{G}.$$

Proof. Define a measure Q on $\mathcal{P}$ by

$$Q(G) = \int_G X dP, \quad G\in\mathcal{G}.$$

Clearly, Q is absolutely continuous with respect to $P|_{\mathcal{G}}$. Now take $Y = dQ/dP|_{\mathcal{G}}$. □

Remark 3.9. Given $X\in L^2(\Omega,\mathcal{F},P)$ consider the closed subspace $L^2(\Omega,\mathcal{G},P|_{\mathcal{G}})\subset L^2(\Omega,\mathcal{F},P)$ and let $\pi: L^2(\Omega,\mathcal{F},P)\to L^2(\Omega,\mathcal{G},P|_{\mathcal{G}})$ be the standard orthogonal projection. It then follows that $Y=\pi X$. □

Definition 3.10. We call $Y=\mathbb{E}(X|\mathcal{G})$ the *conditional expectation* of X given $\mathcal{G}$. If $\mathcal{G}=\mathcal{F}_Z$ for some other Z then we set $\mathbb{E}(X|Z):=\mathbb{E}(X|\mathcal{F}_Z)$.

Note that $\mathbb{E}(X|\mathcal{G})$ is characterized by

$$\int_G \langle Z,\mathbb{E}(X|\mathcal{G})\rangle dP = \int_G \langle Z,X\rangle dP, \quad G\in\mathcal{G}, \tag{3.7}$$

for any $Z:\Omega\to\mathbb{R}^n$ $\mathcal{G}$-measurable.

Proposition 3.11. *Conditional expectation satisfies the following properties:*

1. $\mathbb{E}(aX+bX'|\mathcal{G}) = a\mathbb{E}(X|\mathcal{G})+b\mathbb{E}(X'|\mathcal{G})$;
2. $\mathbb{E}(\mathbb{E}(X|\mathcal{G})) = \mathbb{E}(X)$;
3. *if X is $\mathcal{G}$-measurable then* $\mathbb{E}(X|\mathcal{G}) = X$;
4. *if $X\perp\mathcal{G}$ then* $\mathbb{E}(X|\mathcal{G}) = \mathbb{E}(X)$;
5. *if $\mathcal{G}\subset\mathcal{H}$ then* $\mathbb{E}(X|\mathcal{G}) = \mathbb{E}(\mathbb{E}(X|\mathcal{H})|\mathcal{G})$;
6. *if $X\perp\mathcal{G}$ then* $\mathbb{E}(XY|\mathcal{G}) = \mathbb{E}(X)\mathbb{E}(Y|\mathcal{G})$. *In particular, if Y is $\mathcal{G}$-measurable then* $\mathbb{E}(XY|\mathcal{G}) = Y\mathbb{E}(X)$;
7. *if X is $\mathcal{G}$-measurable then* $\mathbb{E}(XY|\mathcal{G}) = XE(Y|\mathcal{G})$.

Proof. (1) and (2) are obvious. For (3), just think of $X : (\Omega, \mathcal{G}) \to \mathbb{R}^n$ as a random vector. For (4),

$$\begin{aligned}\int_G X dP &= \int_\Omega X \mathbf{1}_G dP \\ &\overset{(*)}{=} \int_\Omega X dP \int_\Omega \chi_G dP \\ &= \mathbb{E}(X)P(G) \\ &= \int_G \mathbb{E}(X) dP,\end{aligned}$$

where the assumption was used in $(*)$. The result then follows by uniqueness. For (5), note that $G \in \mathcal{G}$ implies $G \in \mathcal{H}$ and hence

$$\int_G \mathbb{E}(X|\mathcal{H}) dP = \int_G X dP = \int_G \mathbb{E}(X|\mathcal{G}) dP.$$

Also, (6) is the obvious generalization of (4): using that $X \perp Y \mathbf{1}_G$, $G \in \mathcal{G}$, we see that

$$\mathbb{E}(XY|\mathcal{G}) = \mathbb{E}(X|\mathcal{G})\mathbb{E}(Y|\mathcal{G}) = \mathbb{E}(X)\mathbb{E}(Y|\mathcal{G}).$$

Finally, if $G \in \mathcal{G}$,

$$\begin{aligned}\int_G X\mathbb{E}(Y|\mathcal{G}) dP &\overset{(3)}{=} \int_G \mathbb{E}(X|\mathcal{G})\mathbb{E}(Y|\mathcal{G}) dP \\ &\overset{(3.7)}{=} \int_G \mathbb{E}(X|\mathcal{G}) Y dP \\ &\overset{(3)}{=} \int_G XY dP,\end{aligned}$$

which proves (7). □

Example 3.12. (Birkhoff ergodic theorem) If $(\Omega, \mathcal{F}, P)$ is a probability space then $T : \Omega \to \Omega$ is *measure preserving* if T is $\mathcal{F}$-measurable and satisfies $P(T^{-1}(A)) = P(A)$ for any event $A \in \mathcal{F}$. Given a random variable $X : \Omega \to \mathbb{R}$ we then define, for $n \in \mathbb{N}$, $X_n^{(T)} : \Omega \to \mathbb{R}$ by

$$X_n^{(T)}(\omega) = \frac{1}{n} \sum_{j=0}^{n-1} X(T^j \omega).$$

A version of *Birkhoff's ergodic theorem* [Kre11, Section 1.2] says that there exists a random variable $X^{(T)}$ such that:

$$P\left(\omega \in \Omega; X_n^{(T)}(\omega) \to_{n\to+\infty} X^{(T)}(\omega)\right) = 1, \tag{3.8}$$

from which it follows that $X^{(T)} \circ T = X^{(T)}$ and $\mathbb{E}(X^{(T)}) = \mathbb{E}(X)$. In order to identify $X^{(T)}$ let us consider

$$\mathcal{G}_T = \left\{A \in \mathcal{F}; T^{-1}(A) = A\right\},$$

the σ-subalgebra of T-*invariant* events. If $G \in \mathcal{G}_T$ define $X_G := \mathbf{1}_G X$. Thus,

$$X_G(T^j\omega) = \mathbf{1}_G(T^j\omega)X(T^j\omega) = \mathbf{1}_G(\omega)X(T^j\omega),$$

so if we use (3.8) with X_G replacing X we see that $X_G^{(T)} = \mathbf{1}_G X^{(T)}$ and hence $\mathbb{E}(\mathbf{1}_G X^{(T)}) = \mathbb{E}(X_G) = \mathbb{E}(\mathbf{1}_G X)$, which means that $X^{(T)} = \mathbb{E}(X|\mathcal{G}_T)$. Also, if T is *ergodic* in the sense that

$$\mathcal{G}_T = \{A \in \mathcal{F}; P(A) = 0 \,\text{or}\, P(A) = 1\}$$

then it is immediate to check that $X \perp \mathcal{G}_T$ and it follows from Proposition 3.11 (4) that

$$P\left(\omega \in \Omega; X_n^{(T)}(\omega) \to_{n\to+\infty} \mathbb{E}(X)\right) = 1,$$

which is an improvement of (3.8). □

Example 3.13. (von Neumann ergodic theorem) The L^2 version of Example 3.12 goes as follows. For any measure preserving T as above let us consider

$$L^2\left(\Omega,\mathcal{G}_T,P|_{\mathcal{G}_T}\right)=\left\{\widetilde{X}\in L^2(\Omega,\mathcal{F},P);\widetilde{X}\circ T=\widetilde{X}\right\}.$$

Then *von Neumann's mean ergodic theorem* [Kre11, Section 1.1] assures that for any random variable $X\in L^2(\Omega,\mathcal{F},P)$ there exists a unique $X^{[T]}\in L^2\left(\Omega,\mathcal{G}_T,P|_{\mathcal{G}_T}\right)$ such that

$$\lim_{n\to+\infty}\mathbb{E}(|X_n^{(T)}-X^{[T]}|^2)=0. \tag{3.9}$$

Using that

$$\iota_T:L^2(\Omega,\mathcal{F},P)\to L^2(\Omega,\mathcal{F},P),\quad \iota_T(X)=X\circ T,$$

is an isometry we compute, for any $\widetilde{X}\in L^2(X,\mathcal{G}_T,P|_{\mathcal{G}_T})$,

$$\begin{aligned}\mathbb{E}(X^{[T]}\widetilde{X}) &= \lim_{n\to+\infty}\frac{1}{n}\sum_{j=0}^{n-1}\mathbb{E}\left((X\circ T^j)\widetilde{X}\right)\\ &= \lim_{n\to+\infty}\frac{1}{n}\sum_{j=0}^{n-1}\mathbb{E}\left((X\circ T^j)(\widetilde{X}\circ T^j)\right)\\ &= \lim_{n\to+\infty}\frac{1}{n}\sum_{j=0}^{n-1}\mathbb{E}\left(X\widetilde{X}\right)\\ &= \mathbb{E}(X\widetilde{X}),\end{aligned}$$

which means that $X^{[T]}$ is the L^2 projection of X over $L^2\left(\Omega,\mathcal{G}_T,P|_{\mathcal{G}_T}\right)$. It follows from Remark 3.9 that $\overline{X}^{[T]}=\mathbb{E}(X|\mathcal{G}_T)$, so that (3.9) is the "mean squared" version of (3.8). □

We now give an useful rewording of the random variable induced by the conditional expectation function (3.4) in terms of the notion of conditional expectation appearing in Definition 3.10.

Proposition 3.14. *If, as in (3.4), we set*

$$g(\mathbf{x})=\mathbb{E}(Y|_{X=\mathbf{x}}),\quad \mathbf{x}\in\mathbb{R}^m, \tag{3.10}$$

then

$$g(X)=\mathbb{E}(Y|X). \tag{3.11}$$

In particular,

$$\mathbb{E}(g(X))=\mathbb{E}(Y). \tag{3.12}$$

Proof. We need to check that

$$\int_C g(X)dP=\int_C Y\,dP,\quad C\in\mathcal{F}_X,$$

so we write $C=X^{-1}(A)$, $A\in\mathcal{B}^m$, in order to have

$$\mathbf{1}_C(\omega)=\mathbf{1}_A(X(\omega)),\quad \omega\in\Omega. \tag{3.13}$$

We first note that

$$\begin{aligned}\int_C g(X)dP &= \int_\Omega \mathbf{1}_C g(X)dP \\ &\overset{(3.13)}{=} \int_\Omega \mathbf{1}_A(X)g(X)dP \\ &= \int_{\mathbb{R}^m} \mathbf{1}_A(\mathbf{x})g(\mathbf{x})\psi_X(\mathbf{x})d\mathbf{x},\end{aligned}$$

and using both (3.10) and (3.4),

$$\int_C g(X)dP = \int_{\mathbb{R}^m} \mathbf{1}_A(\mathbf{x}) \left(\int_{\mathbb{R}^p} \mathbf{y}\psi_{Y|X=\mathbf{x}}(\mathbf{y})d\mathbf{y} \right) \psi_X(\mathbf{x})d\mathbf{x}.$$

From Fubini and (3.3) we thus get

$$\begin{aligned}\int_C g(X)dP &= \int\int_{\mathbb{R}^m\times\mathbb{R}^p} \mathbf{1}_A(\mathbf{x})\mathbf{y}\psi_{(X,Y)}(\mathbf{x},\mathbf{y})d\mathbf{x}d\mathbf{y} \\ &= \int\int_{\mathbb{R}^m\times\mathbb{R}^p} \mathbf{1}_A(\mathbf{x})\mathbf{y}dP_{(X,Y)} \\ &= \mathbb{E}(\mathbf{1}_A(X)Y) \\ &\overset{(3.13)}{=} \mathbb{E}(\mathbf{1}_C Y) \\ &= \int_C Y dP,\end{aligned}$$

which proves (3.11). Finally, (3.12) follows from Proposition 3.11, (2). □

Corollary 3.15. *(Laws of total expectation and variance) There hold*

$$\mathbb{E}(X) = \mathbb{E}(\mathbb{E}(X|Y)) \tag{3.14}$$

and

$$\operatorname{var}(X) = \mathbb{E}(\operatorname{var}(X|Y)) + \operatorname{var}(\mathbb{E}(X|Y)). \tag{3.15}$$

Note that (3.14) corresponds to item (2) in Proposition 3.11 with $\mathcal{G} = \mathcal{F}_Y$.

4. Normally distributed random variables and their friends

Here we single out and study some important families of random variables that are closely related to the normal distribution. A comprehensive treatment of this subject can be found in [Ton90].

4.1. Normally distributed random variables.

We begin with the family of random variables that is, without doubt, the most pervasive in Probability Theory and its applications.

Definition 4.1. We say that a random vector $X : \Omega \to \mathbb{R}^n$ is *normally distributed* (or a *Gaussian*, or simply that X is a *normal*) if its pdf $\psi_X : \mathbb{R}^n \to \mathbb{R}$ is given by

$$\psi_X(\mathbf{x}) = \frac{\sqrt{\det A}}{(2\pi)^{n/2}} e^{-\frac{1}{2}\langle A(\mathbf{x}-\boldsymbol{\mu}),\mathbf{x}-\boldsymbol{\mu}\rangle}, \tag{4.1}$$

where A is a positive definite, symmetric matrix and $\boldsymbol{\mu} \in \mathbb{R}^n$. We then write $X \sim \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$, or simply $X \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma} = A^{-1}$. If $n = 1$, in which case

$$\psi_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad x \in \mathbb{R}, \tag{4.2}$$

where $\mu \in \mathbb{R}$ and $\sigma^2 > 0$, we represent this as $X \sim \mathcal{N}(\mu, \sigma^2)$.

The next proposition shows that this is well defined.

Proposition 4.2. *One has* $\int_{\mathbb{R}^n} \psi_X(\mathbf{x}) d\mathbf{x} = 1$.

Proof. Take O an orthogonal matrix so that

$$OAO^{-1} = \Lambda = \operatorname{diag}(\lambda_1, \cdots, \lambda_n),$$

and define $\mathbf{y} = O(\mathbf{x} - \boldsymbol{\mu})$. It follows that

$$\psi_X(\mathbf{x}) = \frac{\sqrt{\det A}}{(2\pi)^{n/2}} e^{-\frac{1}{2}\langle \Lambda \mathbf{y}, \mathbf{y} \rangle},$$

so that

$$\int_{\mathbb{R}^n} \psi_X(\mathbf{x}) d\mathbf{x} = \frac{\sqrt{\det A}}{(2\pi)^{n/2}} \Pi_i \int_{\mathbb{R}} e^{-\frac{1}{2}\lambda_i y_i^2} dy_i.$$

Thus, if $z_i = \sqrt{\lambda_i/2}\, y_i$ then

$$\int_{\mathbb{R}} e^{-\frac{1}{2}\lambda_i x_i^2} dy_i = \sqrt{\frac{2}{\lambda_i}} \int_{\mathbb{R}} e^{-z_i^2} dz_i = \sqrt{\frac{2\pi}{\lambda_i}},$$

so that

$$\int_{\mathbb{R}^n} \psi_X(\mathbf{x}) d\mathbf{x} = \frac{\sqrt{\det A}}{(2\pi)^{n/2}} \frac{(2\pi)^{n/2}}{\Pi_i \sqrt{\lambda_i}} = 1,$$

where we used that $\det A = \Pi_i \lambda_i$ in the last step. □

In general, if $X : \Omega \to \mathbb{R}^n$ is a random vector we have defined its *expectation vector*

$$\boldsymbol{\mu}(X) = \mathbb{E}(X),$$

and its *covariance matrix*

$$\operatorname{cov}(X, X)_{ij} = \operatorname{cov}(X_i, X_j).$$

Sometimes we write this simply as $\operatorname{cov}(X)$. We now compute these invariants assuming that X is Gaussian random vector.

Proposition 4.3. *If* $X \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, $\boldsymbol{\Sigma} = A^{-1}$, *then* $\boldsymbol{\mu}(X) = \boldsymbol{\mu}$ *and* $\operatorname{cov}(X) = \boldsymbol{\Sigma}$.

Proof. In terms of the substitution above we have $\mathbf{x} = \boldsymbol{\mu} + Q\mathbf{y}$, $Q = O^{-1}$. Thus,

$$\begin{aligned} \mathbb{E}(X_i) &= \int_{\mathbb{R}^n} x_i \psi_X(\mathbf{x}) d\mathbf{x} \\ &= \frac{\sqrt{\det A}}{(2\pi)^{n/2}} \int_{\mathbb{R}^n} \left(\boldsymbol{\mu}_i + \sum_j Q_{ij} y_j \right) \Pi_k e^{-\frac{1}{2}\lambda_k y_k^2} dy_1 \cdots dy_n. \end{aligned}$$

But

$$\int_{\mathbb{R}} y_j e^{-\frac{1}{2}\lambda_j y_j^2} dy_j = 0, \tag{4.3}$$

so we get $\boldsymbol{\mu}(X)_i = \boldsymbol{\mu}_i$. Also,

$$\begin{aligned}\mathbb{E}(X_i X_j) &= \int_{\mathbb{R}^n} x_i x_j \psi_X(\mathbf{x}) d\mathbf{x} \\ &= \int_{\mathbb{R}^n} \left(\boldsymbol{\mu}_i + \sum_k Q_{ik} y_k\right)\left(\boldsymbol{\mu}_j + \sum_l Q_{jl} y_l\right) \psi_X(\mathbf{x}) d\mathbf{x},\end{aligned}$$

and using (4.3) again we get

$$\begin{aligned}\mathbb{E}(X_i X_j) &= \int_{\mathbb{R}^n} \boldsymbol{\mu}_i \boldsymbol{\mu}_j \psi_X(\mathbf{x}) d\mathbf{x} + \frac{\sqrt{\det A}}{(2\pi)^{n/2}} \sum_{kl} Q_{ik} Q_{jl} \int_{\mathbb{R}^n} y_k y_l \Pi_p (e^{-\frac{1}{2}\lambda_p y_p^2}) dy_p \\ &= \boldsymbol{\mu}_i \boldsymbol{\mu}_j + \frac{\sqrt{\det A}}{(2\pi)^{n/2}} \sum_k Q_{ik} Q_{jk} \int_{\mathbb{R}^n} y_k^2 \Pi_p e^{-\frac{1}{2}\lambda_p y_p^2} dy_p \\ &= \boldsymbol{\mu}_i \boldsymbol{\mu}_j + \frac{\sqrt{\det A}}{(2\pi)^{n/2}} \sum_k Q_{ik} Q_{jk} \int_{\mathbb{R}} y_k^2 e^{-\frac{1}{2}y_k^2} dy_k \times \Pi_{p\neq k} \int_{\mathbb{R}} e^{-\frac{1}{2}y_p^2} dy_p \\ &= \boldsymbol{\mu}_i \boldsymbol{\mu}_j + \frac{\sqrt{\det A}}{(2\pi)^{n/2}} (2\pi)^{\frac{n-1}{2}} \sum_k Q_{ik} Q_{jk} \int_{\mathbb{R}} y_k^2 e^{-\frac{1}{2}y_k^2} dy_k \times \Pi_{p\neq k} \frac{1}{\lambda_p^{1/2}}.\end{aligned}$$

But

$$\begin{aligned}\int_{\mathbb{R}} y_k^2 e^{-\frac{1}{2}\lambda_k y_k^2} dy_k &= -\frac{y_k}{\lambda_k} e^{-\frac{1}{2}\lambda_k y_k^2}|_{-\infty}^{+\infty} + \frac{1}{\lambda_k} \int_{\mathbb{R}^n} e^{-\frac{1}{2}\lambda_k y_k^2} dy_k \\ &= \frac{1}{\lambda_k} \frac{(2\pi)^{1/2}}{\lambda_k^{1/2}},\end{aligned}$$

so that

$$\mathbb{E}(X_i X_j) = \boldsymbol{\mu}_i \boldsymbol{\mu}_j + \sum_k \frac{Q_{ik} Q_{jk}}{\lambda_k} = \boldsymbol{\mu}_i \boldsymbol{\mu}_j + \boldsymbol{\Sigma}_{ij},$$

where we used that $\boldsymbol{\Sigma} = A^{-1} = Q\Lambda^{-1}Q^{-1}$. This completes the proof. □

We now compute the characteristic function of a normally distributed random vector.

Proposition 4.4. *If $X \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ then*

$$\phi_X(\mathbf{u}) = e^{\langle \boldsymbol{\mu}, \mathbf{u}\rangle \mathbf{i} - \frac{1}{2}\langle \boldsymbol{\Sigma}\mathbf{u}, \mathbf{u}\rangle}. \tag{4.4}$$

Proof. Recalling that $Q^{-1}AQ = \Lambda$ and $\mathbf{x} = \boldsymbol{\mu} + Q\mathbf{y}$, we have

$$\begin{aligned}\phi_X(\mathbf{u}) &= \int_{\mathbb{R}^n} e^{\langle \mathbf{x}, \mathbf{u}\rangle \mathbf{i}} \frac{\sqrt{\det A}}{(2\pi)^{n/2}} e^{-\frac{1}{2}\langle A(\mathbf{x}-\boldsymbol{\mu}), \mathbf{x}-\boldsymbol{\mu}\rangle} dx \\ &= \frac{\sqrt{\det A}}{(2\pi)^{n/2}} \int_{\mathbb{R}^n} e^{\langle \boldsymbol{\mu}, \mathbf{u}\rangle \mathbf{i} + \langle \mathbf{y}, \mathbf{v}\rangle \mathbf{i} - \frac{1}{2}\langle \Lambda \mathbf{y}, \mathbf{y}\rangle} d\mathbf{y},\end{aligned}$$

where $\mathbf{v} = Q^\top \mathbf{u}$. Now observe that if $(\ ,\ )$ is the sesquilinear product in $\mathbb{C}^n$ then

$$\begin{aligned} -\frac{1}{2}\left(\Lambda^{1/2}\mathbf{y} - \mathbf{i}\Lambda^{-1/2}\mathbf{v}, \Lambda^{1/2}\mathbf{y} - \mathbf{i}\Lambda^{-1/2}\mathbf{v}\right) &= -\frac{1}{2}\langle \Lambda \mathbf{y}, \mathbf{y}\rangle \\ &\quad + \frac{1}{2}\langle \Lambda^{-1}\mathbf{v}, \mathbf{v}\rangle + \mathbf{i}\langle \mathbf{v}, \mathbf{y}\rangle, \end{aligned}$$

which gives

$$\begin{aligned} \phi_X(\mathbf{u}) &= \frac{\sqrt{\det A}}{(2\pi)^{n/2}} e^{\langle \boldsymbol{\mu}, \mathbf{u}\rangle \mathbf{i} - \frac{1}{2}\langle \Lambda^{-1}\mathbf{v}, \mathbf{v}\rangle} \int_{\mathbb{R}^n} e^{-\frac{1}{2}\left(\Lambda^{1/2}\mathbf{y} - \mathbf{i}\Lambda^{-1/2}\mathbf{v}, \Lambda^{1/2}\mathbf{y} - \mathbf{i}\Lambda^{-1/2}\mathbf{v}\right)} d\mathbf{y} \\ &\overset{(*)}{=} \frac{\sqrt{\det A}}{(2\pi)^{n/2}} e^{\langle \boldsymbol{\mu}, \mathbf{u}\rangle \mathbf{i} - \frac{1}{2}\langle Q\Lambda^{-1}Q^{-1}\mathbf{u}, \mathbf{u}\rangle} \int_{\mathbb{R}^n} e^{-\frac{1}{2}\left(\Lambda^{1/2}\mathbf{y}, \Lambda^{1/2}\mathbf{y}\right)} d\mathbf{y} \\ &= e^{\langle \boldsymbol{\mu}, \mathbf{u}\rangle \mathbf{i} - \frac{1}{2}\langle \boldsymbol{\Sigma}\mathbf{u}, \mathbf{u}\rangle}, \end{aligned}$$

where in $(*)$ we changed the contour of integration (and used the appropriate multi-dimensional version of Cauchy's theorem in Complex Variables). □

Corollary 4.5. *If $X \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ then its mgf is*

$$\varphi_X(\mathbf{u}) = e^{\langle \boldsymbol{\mu}, \mathbf{u}\rangle + \frac{1}{2}\langle \boldsymbol{\Sigma}\mathbf{u}, \mathbf{u}\rangle}. \tag{4.5}$$

In particular, if $n = 1$ and $X \sim \mathcal{N}(\mu, \sigma^2)$ then

$$\varphi_X(u) = e^{u\mu + \frac{1}{2}\sigma^2 u^2}. \tag{4.6}$$

Corollary 4.6. *If X is a normally distributed random vector, $X \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then its distribution P_X is completely determined by its characteristic function.*

Proof. Note that

$$\int_{\mathbb{R}^n} |\phi_X(\mathbf{u})| d\mathbf{u} = \int_{\mathbb{R}^n} e^{-\frac{1}{2}\langle \boldsymbol{\Sigma}\mathbf{u}, \mathbf{u}\rangle} d\mathbf{u} < +\infty$$

and apply (the appropriate multi-variate version of) Proposition 2.28 (2). □

Corollary 4.7. *For a random vector $X \in \mathbb{R}^n$ the following hold:*

- *$X \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ if and only if $\langle \mathbf{u}, X\rangle \sim \mathcal{N}(\langle \mathbf{u}, \boldsymbol{\mu}\rangle, \langle \boldsymbol{\Sigma}\mathbf{u}, \mathbf{u}\rangle)$ for any $\mathbf{u} \in \mathbb{R}^n$.*
- *If $X \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ then*

$$CX \sim \mathcal{N}(C\boldsymbol{\mu}, C\boldsymbol{\Sigma}C^\top\}),$$

where C is an invertible $n \times n$ matrix.

Proof. The first assertion is an immediate consequence of the identity

$$\phi_X(u\mathbf{u}) = \phi_{\langle \mathbf{u}, X\rangle}(u), \quad u \in \mathbb{R}, \quad \mathbf{u} \in \mathbb{R}^n.$$

As for the second, use that $\langle \mathbf{u}, CX\rangle = \langle C^\top \mathbf{u}, X\rangle$. □

Next we explore further consequences of the theory above.

Proposition 4.8. *If $X : \Omega \to \mathbb{R}$ is normally distributed, $X \sim \mathcal{N}(\mu, \sigma^2)$, then*

(1) $\phi_X(u) = e^{\mu u \mathbf{i} - \frac{1}{2}\sigma^2 u^2}$;
(2) *if $(r, s) \in \mathbb{R}^2$, $r \neq 0$, then $rX + s \sim \mathcal{N}(r\mu + s, r^2\sigma^2)$.*
(3) *If $Y \sim \mathcal{N}(\overline{\mu}, \overline{\sigma}^2)$ and $Y \perp X$ then $X + Y \sim \mathcal{N}(\mu + \overline{\mu}, \sigma^2 + \overline{\sigma}^2)$. As a consequence, if $\{X_j\}_{j=1}^n$ is independent with $X_j \sim \mathcal{N}(\mu_j, \sigma_j^2)$ then*

$$\sum_j a_j X_j \sim \mathcal{N}\left(\sum_j a_j \mu_j, \sum_j a_j^2 \sigma_j^2\right), \quad a_j \in \mathbb{R}. \tag{4.7}$$

In particular, if $\mu_j = 0$ and $\sigma_j = \sigma$ then $X = (X_1, \cdots, X_n)$ satisfies

$$\langle X, \vec{a} \rangle \sim \mathcal{N}(0, \|\vec{a}\|^2 \sigma^2), \tag{4.8}$$

where $\vec{a} = (a_1, \cdots, a_n)$.

Proof. (1) is a special case of (4.4). As for (2), note that $\phi_s(u) = e^{su\mathbf{i}}$, $s \perp rX$ and use Proposition 2.32 (1) and Proposition 2.31 (1) to check that

$$\phi_{rX+s}(u) = e^{(r\mu+s)u\mathbf{i} - \frac{1}{2}r^2\sigma^2u^2},$$

and finally use Corollary 4.6. Clearly, (3) is proved with the same kind of argument. For instance,

$$\begin{aligned} \phi_{X+Y}(u) &= \phi_X(u)\phi_Y(u) \\ &= e^{\mathbf{i}u\mu - \frac{1}{2}u^2\sigma^2} e^{\mathbf{i}u\overline{\mu} - \frac{1}{2}u^2\overline{\sigma}^2} \\ &= e^{\mathbf{i}u(\mu+\overline{\mu}) - \frac{1}{2}u^2(\sigma^2+\overline{\sigma}^2)}, \end{aligned}$$

as desired. □

Example 4.9. (Moments of a normal) If $X \sim \mathcal{N}(0, \sigma^2)$ then (4.6) gives

$$\varphi_X(u) = e^{\frac{1}{2}\sigma^2u^2} = \sum_{k\geq 0} \frac{\sigma^{2k}}{2^k k!} u^{2k},$$

so if we compare with (2.24) we conclude that

$$\alpha_l(X) = \begin{cases} \frac{l!}{2^{l/2}(l/2)!}\sigma^{2k}, & l \text{ even} \\ 0, & l \text{ odd} \end{cases} \tag{4.9}$$

which provides explicit expressions for all the moments of X. □

Example 4.10. (The log-normal distribution) If $n = 1$ and $X \sim \mathcal{N}(\mu, \sigma^2)$ then (4.6) implies that $Y = e^X$ satisfies

$$\mathbb{E}(Y) = \mathbb{E}(e^X) = e^{\mu + \frac{1}{2}\sigma^2} \tag{4.10}$$

and

$$\mathbb{E}(Y^2) = \mathbb{E}(e^{2X}) = e^{2\mu + 2\sigma^2},$$

so that

$$\mathrm{var}(Y) = \mathbb{E}(Y^2) - \mathbb{E}(Y)^2 = (e^{\sigma^2} - 1)e^{2\mu+\sigma^2}. \tag{4.11}$$

Hence, we may summarize (4.10) and (4.11) by writing

$$Y = e^X \sim \mathcal{LN}(e^{\mu + \frac{1}{2}\sigma^2}, (e^{\sigma^2} - 1)e^{2\mu+\sigma^2}), \tag{4.12}$$

where $\mathcal{LN}$ stands for "log-normal" (which means that $X = \ln Y$ follows a normal). Alternatively, we may write

$$Y \sim \Lambda(\mu, \sigma^2), \tag{4.13}$$

if emphasis on the parameters of the underlying normal distribution is needed [AB69]. In this notation, it is immediate from Proposition 4.8-(2) that (4.13) implies

$$e^a Y \sim \Lambda(\mu + a, \sigma^2), \quad a \in \mathbb{R}. \tag{4.14}$$

Now, an (obvious) generalization of (4.12) is

$$Y^u \sim \mathcal{LN}(e^{\mu u + \frac{1}{2}\sigma^2 u^2}, (e^{\sigma^2 u^2} - 1)e^{2\mu u + \sigma^2 u^2}), \quad u \in \mathbb{R},$$

so that the corresponding *coefficient of variation*,

$$\mathrm{cv}(Y^u) := \frac{\mathrm{sd}(Y^u)}{\mathbb{E}(Y^u)}, \tag{4.15}$$

is given by

$$\mathrm{cv}(Y^u) = \sqrt{e^{\sigma^2 u^2} - 1}.$$

In particular, it does not depend on $\mu = \mathbb{E}(\ln Y)$ and satisfies the "scaling-plus-inversion invariance property"

$$\mathrm{cv}(\alpha Y^u) = \mathrm{cv}(Y^u) = \mathrm{cv}(Y^{-u}), \quad \alpha > 0. \tag{4.16}$$

For later reference, we also note that the pdf ψ_Y of $Y = e^X$ is

$$\psi_Y(x) = \frac{1}{\sigma x \sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{\ln x - \mu}{\sigma}\right)^2}, \tag{4.17}$$

so that the corresponding cdf is

$$F_Y(x) = \Phi\left(\frac{\ln x - \mu}{\sigma}\right), \tag{4.18}$$

where

$$\Phi(x) := \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{1}{2}y^2} dy \tag{4.19}$$

is the cdf of $\mathcal{N}(0, 1)$. □

We now turn to another elegant application of the formalism of characteristic functions to Gaussian random variables. Recall that if X and Y are independent random variables, then they are necessarily uncorrelated (Corollary 2.7). We shall now verify that the converse also holds in the case of the components of a normally distributed random vector.

Proposition 4.11. *If $X = (X_1, \cdots, X_k) : \Omega \to \mathbb{R}^k$ is a normally distributed random vector, say $X \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, with $\boldsymbol{\Sigma}$ diagonal then:*

- *$\{X_j\}_{j=1}^k$ is independent;*
- *$X_j \sim \mathcal{N}(\boldsymbol{\mu}_j, \sigma^2_{X_j})$, $\quad \sigma^2_{X_j} = \mathrm{var}(X_j)$.*

Proof. By assumption, $\mathbf{\Sigma} = \mathrm{diag}(\sigma^2_{X_1}, \cdots, \sigma^2_{X_j})$. If $\boldsymbol{\mu} = \mathbb{E}(X)$ we have from (4.4) that

$$\begin{aligned}\phi_X(\mathbf{u}) &= e^{\langle \boldsymbol{\mu}, \mathbf{u}\rangle \mathbf{i} - \frac{1}{2}\langle \mathbf{\Sigma}\mathbf{u}, \mathbf{u}\rangle} \\ &= e^{\left(\sum_j \boldsymbol{\mu}_j u_j\right)\mathbf{i} - \frac{1}{2}\sum_j \mathbf{\Sigma}_{jj} u_j^2} \\ &= \Pi_j e^{\boldsymbol{\mu}_j u_j \mathbf{i} - \frac{1}{2}\mathbf{\Sigma}_{jj} u_j^2} \\ &= \Pi_j \phi_{Y_j}(u_j),\end{aligned}$$

where $Y_j \sim \mathcal{N}(\boldsymbol{\mu}_j, \sigma^2_{X_j})$ by (4.6) and Proposition 2.28, (2). Using (the multi-variate version of) (2.20) we then have

$$\begin{aligned}\psi_X(\mathbf{x}) &= \frac{1}{(2\pi)^n}\int_{\mathbb{R}^n} e^{-\mathbf{i}\langle \mathbf{x}, \mathbf{u}\rangle}\phi_X(\mathbf{u})d\mathbf{u} \\ &= \Pi_j \frac{1}{2\pi}\int_{\mathbb{R}} e^{-\mathbf{i}x_j u_j}\phi_{Y_j}(u_j)du_j \\ &= \Pi_j \psi_{Y_j}(x_j),\end{aligned}$$

which not only proves that $\{X_j\}_{j=1}^k$ is independent (by Proposition 2.15) but also that $\psi_{X_j} = \psi_{Y_j}$ (by Proposition 2.14), which concludes the proof. □

Corollary 4.12. *The following assertions are equivalent:*

- $X = (X_1, \cdots, X_k) \sim \mathcal{N}(\vec{0}, \sigma^2 \mathrm{Id}_k)$.
- $\{X_j\}_{j=1}^k$ *is independent and* $X_j \sim \mathcal{N}(0, \sigma^2)$.

Corollary 4.13. *(Rotational invariance) Let $\{X_j\}_{j=1}^k$ be independent with $X_j \sim \mathcal{N}(0, \sigma^2)$ and consider $Y_l = \sum_{j=1}^k C_{lj}X_j$, where $C = \{C_{lj}\}$ is orthogonal. Then $\{Y_l\}_{l=1}^k$ is independent with $Y_l \sim \mathcal{N}(0, \sigma^2)$.*

Proof. Write $Y = CX$ with $X \sim \mathcal{N}(\vec{0}, \sigma^2 I)$. It follows that

$$\phi_Y(\mathbf{u}) = \phi_{CX}(\mathbf{u}) = \phi_X(C\mathbf{u}),$$

where we used Proposition 2.31 (1) and Proposition 2.32 (1) in the last step. It follows from Proposition 4.4 that

$$\begin{aligned}\phi_Y(\mathbf{u}) &= e^{-\frac{1}{2}\langle \sigma^2 C\mathbf{u}, C\mathbf{u}\rangle} \\ &= e^{-\frac{1}{2}\sigma^2\|\mathbf{u}\|^2},\end{aligned}$$

so that $Y \sim \mathcal{N}(\vec{0}, \sigma^2 I)$ as well (by Corollary 4.6) and the independence of $\{Y_j\}$ now follows from the proposition. □

Definition 4.14. If $X = (X_1, \cdots, X_k)$ satisfies any of the conditions in Corollary 4.12 with $\sigma = 1$ (so that $X \sim \mathcal{N}(\vec{0}, \mathrm{Id}_k)$) then we say that X is a *standard* normal random vector.

Remark 4.15. The projection property in (4.8) is an easy consequence of rotational invariance. Indeed, let $Y = OX$, where $X \sim \mathcal{N}(\vec{0}, \sigma^2 \mathrm{Id})$ and O is an orthogonal matrix whose first line is $\|\vec{a}\|^{-1}\vec{a}$. Then

$$\|\vec{a}\|^{-1}\langle X, \vec{a}\rangle = (OX)_1 = Y_1 \sim \mathcal{N}(0, \sigma^2),$$

where Corollary 4.13 has been used in the last step. It follows that $\langle X, \vec{a}\rangle \sim N(0, \|\vec{a}\|^2\sigma^2)$, as claimed. □

Remark 4.16. In Proposition 4.11 it is essential to assume that the normal random variables X_j, $j = 1, \cdots, n$, are *jointly* normally distributed in the sense that $X = (X_1, \cdots, X_n)$ is normally distributed. In fact, there exist random variables X_1 and X_2 with $\mathrm{cov}(X_1, X_2) = 0$, $X_1, X_2 \sim \mathcal{N}(0,1)$ but with $X = (X_1, X_2)$ not being normally distributed and hence with $\{X_1, X_2\}$ not being independent. The classical example is obtained by taking $X_1 \sim \mathcal{N}(0,1)$, $\boldsymbol{\epsilon}$ a Rademacher random variable as in Definition 2.36 which is independent from X_1 and $X_2 = \boldsymbol{\epsilon} X_1$. To check the claims above, we first compute

$$\begin{aligned} \mathrm{cov}(X_1, X_2) &= \mathbb{E}(X_1 X_2) - \mathbb{E}(X_1)\mathbb{E}(X_2) \\ &= \mathbb{E}(X_1 X_2) \\ &= \mathbb{E}(X_1^2 \boldsymbol{\epsilon}) \\ &= \mathbb{E}(X_1^2)\mathbb{E}(\boldsymbol{\epsilon}) \\ &= 0, \end{aligned}$$

where we used that $\mathbb{E}(\boldsymbol{\epsilon}) = 0$ in the last step. Also, the fact that $X_2 \sim \mathcal{N}(0,1)$ follows from Proposition 2.37. To check that X is not normally distributed just note that $X_1 + X_2$ vanishes on $\epsilon^{-1}(-1)$ and hence fails to follow a normal, so the claim follows from Corollary 4.7. Finally, if $\{X_1, X_2\}$ were independent then $\{|X_1|, |X_2|\}$ would be independent as well, which is a contradiction because $|X_1| = |X_2|$. □

Remark 4.17. (The effectiveness of the characteristic function) The simplest case $n = 2$ already illustrates the difficulty in trying to prove Proposition 4.11 by means of pdfs (thus directly relying on Proposition 2.15). Let us assume that

$$(X_1, X_2) \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \tag{4.20}$$

where, with self-explanatory notation,

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_{X_1} \\ \mu_{X_2} \end{pmatrix}$$

and

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_{X_1}^2 & \sigma_{X_1 X_2} \\ \sigma_{X_1 X_2} & \sigma_{X_1}^2 \end{pmatrix} = \begin{pmatrix} \sigma_{X_1}^2 & \rho\sigma_{X_1}\sigma_{X_2} \\ \rho\sigma_{X_1}\sigma_{X_2} & \sigma_{X_1}^2 \end{pmatrix},$$

where

$$\rho = \frac{\sigma_{X_1 X_2}}{\sigma_{X_1}\sigma_{X_2}}$$

is the *correlation coefficient*. Hence, one must check that $\rho = 0$ implies that $\{X_1, X_2\}$ is independent, with each marginal following the appropriate normal distribution. To proceed, note that

$$\det \boldsymbol{\Sigma} = (1 - \rho^2)\sigma_{X_1}^2 \sigma_{X_2}^2,$$

so that

$$\begin{aligned} \boldsymbol{\Sigma}^{-1} &= \frac{1}{(1-\rho^2)\sigma_{X_1}^2\sigma_{X_2}^2} \begin{pmatrix} \sigma_{X_2}^2 & -\rho\sigma_{X_1}\sigma_{X_2} \\ -\rho\sigma_{X_1}\sigma_{X_2} & \sigma_{X_1}^2 \end{pmatrix} \\ &= \frac{1}{1-\rho^2} \begin{pmatrix} 1/\sigma_{X_1}^2 & -\rho/\sigma_{X_1}\sigma_{X_2} \\ -\rho/\sigma_{X_1}\sigma_{X_2} & 1/\sigma_{X_2}^2 \end{pmatrix}, \end{aligned}$$

and leading this to (4.1), with $A = \boldsymbol{\Sigma}^{-1}$, we see that the joint density of (X_1, X_2) is

$$\begin{aligned} \psi_{(X_1,X_2)}(x_1, x_2) &= \frac{1}{2\pi\sigma_{X_1}\sigma_{X_2}\sqrt{1-\rho^2}} \times \\ &\quad \times e^{-\frac{1}{2(1-\rho^2)}\left(\frac{(x_1-\mu_{X_1})^2}{\sigma_{X_1}^2} - \frac{2\rho(x_1-\mu_{X_1})(x_2-\mu_{X_2})}{\sigma_{X_1}\sigma_{X_2}} + \frac{(x_2-\mu_{X_2})^2}{\sigma_{X_2}^2}\right)}. \end{aligned} \tag{4.21}$$

Thus, if $\rho = 0$ this decomposes as

$$\psi_{(X_1,X_2)}(x_1,x_2) = \frac{1}{\sqrt{2\pi}\sigma_{X_1}} e^{-\frac{(x_1-\mu_{X_1})^2}{2\sigma_{X_1}^2}} \times \frac{1}{\sqrt{2\pi}\sigma_{X_2}} e^{-\frac{(x_2-\mu_{X_2})^2}{2\sigma_{X_2}^2}},$$

from which the claim follows immediately. However, it is not clear how this argument, which involves explicitly inverting the covariance matrix $\mathbf{\Sigma}$, carries over as n gets indefinitely large. This should be compared with the general proof displayed above, which relies on the inversion formula (2.20) combined with the fact that $\mathbf{\Sigma}$ appears *linearly* in the exponent of (4.4). As yet another nice application of characteristic functions, let us note that, in general, if $X = (X_1, X_2)$ is given then the characteristic function of the marginal X_1 is

$$\phi_{X_1}(u_1) = \mathbb{E}(e^{\mathbf{i}u_1X_1}) = \mathbb{E}(e^{\mathbf{i}(u_1X_1+0X_2)},$$

that is,

$$\phi_{X_1}(u_1) = \phi_{(X_1,X_1)}(u_1,0),$$

which tells us how to calculate the characteristic function of a marginal in terms of the characteristic function of the joint distribution. In particular, when applied to a bi-variate normal as in (4.20), and *not* necessarily assuming that $\{X_1, X_2\}$ is independent, this clearly implies that the marginals are normally distributed in the expected way: $X_j \sim \mathcal{N}(\mu_{X_j}, \sigma_{X_j}^2)$. Needless to say, a similar result holds for the marginals of a multivariate, normally distributed random vector, with essentially the same proof. □

Remark 4.18. (Regression to the mean) If $X = (X_1, X_2)$ is a bi-variate normal as in (4.20) then we know from Remark 4.17 that $X_j \sim \mathcal{N}(\mu_{X_j}, \sigma_{X_j}^2)$, $j = 1, 2$. Using this, (4.21), (3.3) and a little algebra we get

$$\begin{aligned}\psi_{X_2|_{X_1=x_1)}}(x_2) &= \frac{1}{\sqrt{2\pi}\sqrt{1-\rho^2}\sigma_{X_2}} \times \\ &\quad \times e^{-\frac{1}{2(1-\rho^2)\sigma_{X_2}^2}\left(x_2-\mu_{X_2}-\rho\frac{\sigma_{X_2}}{\sigma_{X_1}}(x_1-\mu_{X_1})\right)^2},\end{aligned}$$

so that

$$X_2|_{X_1=x_1} \sim \mathcal{N}\left(\mu_{X_2} + \rho\frac{\sigma_{X_2}}{\sigma_{X_1}}(x_1-\mu_{X_1}), (1-\rho^2)\sigma_{X_2}^2\right)$$

or equivalently,

$$\frac{X_2|_{X_1=x_1} - \mu_{X_2}}{\sigma_{X_2}} \sim \mathcal{N}\left(\rho\frac{x_1-\mu_{X_1}}{\sigma_{X_1}}, 1-\rho^2\right).$$

In particular,

$$\frac{\mathbb{E}\left(X_2|_{X_1=x_1}\right) - \mu_{X_2}}{\sigma_{X_2}} = \rho\frac{x_1-\mu_{X_1}}{\sigma_{X_1}}, \tag{4.22}$$

which says that, on average, the proper standardization of $X_2|_{X_1=x_1}$ is proportional to the observed standardization of X_1 by a factor which is strictly less than 1 (in absolute value) unless X_1 and X_2 are perfectly correlated ($|\rho| = 1$). More specifically, let us suppose that the variables model random measurements of some hereditary trait (stature, for instance) that passes from parents (X_1) to offspring (X_2) and happens to be "stable" in the sense that both variables follow the same normal $\mathcal{N}(\mu, \sigma^2)$ (and of course are jointly normally distributed as well). We then obtain from (4.22) that

$$\mathbb{E}\left(X_2|_{X_1=x_1}\right) - x_1 = -(1-\rho)(x_1-\mu),$$

which means that, on average, $X_2|_{X_1=x_1}$ lies somewhere between x_1 and μ. This "regression to the mean", first (empirically) discovered by F. Galton, has played a fundamental role in the conceptual development of Multivariate Analysis [Sti90, Sti97, Gor16]. In order to relate this to the simple linear regression model as

discussed in Remark 9.17, note from (4.21) that the ellipses of "equal frequency" for the joint distribution are given by

$$(x_1-\mu)^2-2\rho(x_1-\mu)(x_2-\mu)+(x_2-\mu)^2=\text{const.},$$

so the contact points of the corresponding vertical tangent lines satisfy

$$x_2-\mu=\rho(x_1-\mu),$$

which, upon comparison with (9.49) and (9.50), identifies ρ to the slope of the associated regression line[10]. □

4.2. **Random variables related to the normal.** We now present a few distributions closely related to the normal.

Definition 4.19. A randon variable $Y:\Omega\to\mathbb{R}$ is $\mathsf{Gamma}(\alpha,\lambda)$-*distributed*, where $\alpha,\lambda>0$, if its pdf is

$$\Gamma_{\alpha,\lambda}(x)=\frac{\alpha^\lambda}{\Gamma(\lambda)}x^{\lambda-1}e^{-\alpha x}\mathbf{1}_{(0,+\infty)}(x), \tag{4.23}$$

where

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

is the Gamma function. We then say that α and λ are the *inverse scale* and *shape* parameters of X, respectively. In particular, Y is *chi-square distributed* with $k\geq 1$ degrees of freedom if its pdf is $\chi^2_k:=\Gamma_{1/2,k/2}$. Explicitly,

$$\chi^2_k(x)=\frac{1}{2^{k/2}\Gamma(k/2)}x^{\frac{k}{2}-1}e^{-x/2}\mathbf{1}_{(0,+\infty)}(x). \tag{4.24}$$

Proposition 4.20. *If $Y\sim\mathsf{Gamma}(\alpha,\lambda)$ then its mgf is $\varphi_Y(u)=(1-\alpha^{-1}u)^{-\lambda}$, $|u|<\alpha$. In particular, $\mathbb{E}(Y)=\lambda/\alpha$ and $\mathrm{var}(Y)=\lambda/\alpha^2$.*

Proof. We have

$$\begin{aligned}\varphi_Y(u) &= \frac{\alpha^\gamma}{\Gamma(\lambda)}\int_0^{+\infty}x^{\lambda-1}e^{-(\alpha-u)x}dx\\ &\overset{y=(\alpha-u)x}{=} \frac{\alpha^\gamma}{\Gamma(\lambda)}(\alpha-u)^{-\lambda}\int_0^{+\infty}y^{\lambda-1}e^{-y}dy\\ &= \alpha^\gamma(\alpha-u)^{-\lambda}.\end{aligned}$$

The last assertion follows from Remark 2.39. □

Corollary 4.21. *If $Y\sim\chi^2_k$ then $\varphi_Y(u)=(1-2u)^{-k/2}$, $|u|<1/2$. In particular, $\mathbb{E}(Y)=k$ and $\mathrm{var}(Y)=2k$.*

Corollary 4.22. *If $Y\sim\mathsf{Gamma}(\alpha,\lambda)$ then its characteristic function is given by $\phi_Y(u)=(1-\alpha^{-1}u\mathbf{i})^{-\lambda}$. In particular, if $Y\sim\chi^2_k$ then $\phi_Y(u)=(1-2u\mathbf{i})^{-k/2}$.*

[10]For assessments of the social and intellectual contexts of his time and the nasty ideology behind Galton's pursuit of this statistical result, we refer to [Cow72, Hil73, Mac81, Bul03].

Corollary 4.23. *If $a > 0$ and $Y \sim \mathsf{Gamma}(\alpha, \lambda)$ then $aY \sim \mathsf{Gamma}(\alpha/a, \lambda)$. In particular, if $Y \sim \chi^2_k$ then $aY \sim \mathsf{Gamma}(1/2a, k/2)$.*

Proof. Recall from Proposition 2.31 (1) that

$$\phi_{aY}(u) = \phi_Y(au) = (1 - \alpha^{-1} au\mathbf{i})^{-k/2}.$$

□

Note that this justifies the adopted terminology for α.

Corollary 4.24. *If $\{Y_j\}_{j=1}^k$ is independent and $Y_j \sim \mathsf{Gamma}(\alpha, \lambda_j)$ then*

$$\sum_j Y_j \sim \mathsf{Gamma}\left(\alpha, \sum_j \lambda_j\right).$$

In particular, if $Y_j \sim \chi^2_{k_j}$ then

$$\sum_j Y_j \sim \chi^2_{\sum_j k_j}.$$

Corollary 4.25. *If $\{Z_j\}_{j=1}^k$ is independent with $Z_j \sim \mathcal{N}(0,1)$ then*

$$Z := \sum_j Z_j^2 \sim \chi^2_k.$$

In particular, $\mathbb{E}(Z) = k$ and $\mathrm{var}(Z) = 2k$.

Proof. By Remark 2.17 and Proposition 4.8 (1),

$$\psi_{Z_j^2}(x) = \frac{1}{\sqrt{2\pi}} x^{-1/2} e^{-x/2} \mathbf{1}_{[0,+\infty)]}(x),$$

so that $Z_j^2 \sim \chi^2_1$ (recall that $\Gamma(1/2) = \sqrt{\pi}$). The result now follows from Corollary 4.24. □

Remark 4.26. (The geometric way to χ^2_k) Corollary 4.25 can be elegantly retrieved as an application of the "n-space computations" introduced by R. Fisher [Fis15, Fis21, Fis25]. Indeed, from Proposition 2.15 we know that the amount of probability density spanned by a standard normal vector $Z = (Z_1, \cdots, Z_k) \in \mathbb{R}^k$ in an infinitesimal region of volume $dz = dz_1 \cdots dz_k$ is

$$\begin{aligned} \frac{1}{(2\pi)^{k/2}} e^{-\|z\|^2/2} dz &= \frac{1}{(2\pi)^{k/2}} e^{-\|z\|^2/2} \|z\|^{k-1} d\|z\| d\theta \\ &= \frac{1}{2} \frac{1}{(2\pi)^{k/2}} e^{-\|z\|^2/2} (\|z\|^2)^{\frac{k}{2}-1} d\|z\|^2 d\theta, \end{aligned}$$

where $z = (\|z\|, \theta) \in (0, +\infty) \times \mathbb{S}^{k-1}$ is the polar decomposition of Z[11]. Again by Proposition 2.15, if we view this latter expression as the joint distribution of $(\|Z\|^2, \Theta)$ then $\{\|Z\|^2, \Theta\}$ is independent with $\Theta = X/\|X\|$

[11] In this and similar computations, as in Remark 7.31 and Example 7.39, we represent a realization of a random variable, say Z_j or Θ, by the corresponding lower-case symbol (in this case, z_j or θ).

being *uniformly* distributed over $\mathbb{S}^{n-1}$. Hence, by Proposition 2.14 the infinitesimal density of $\|Z\|^2$ is

$$\psi_{\|Z\|^2}(\|z\|^2)d\|z\|^2 = \frac{\omega_{k-1}}{2}\frac{1}{(2\pi)^{k/2}}e^{-\|z\|^2/2}(\|z\|^2)^{\frac{k}{2}-1}d\|z\|^2,$$

where ω_{k-1} is the volume of $\mathbb{S}^{k-1}$. Since

$$\omega_{k-1} = \frac{2\pi^{k/2}}{\Gamma(k/2)} \tag{4.25}$$

it suffices to set $x = \|z\|^2$ in order to recover (4.24). The same computation gives that $Y = (Y_1, \cdots, Y_k) \sim \mathcal{N}(\vec{0}, \sigma^2 \mathrm{Id}_k)$ implies $\|Y\|^2 \sim \mathsf{Gamma}(1/2\sigma^2, k/2)$; cf. Corollary 4.24. For $k = 3$ and $\sigma^2 = \kappa T/2$, where κ is the Boltzmann constant and T is the temperature, this gives

$$\psi_E(\epsilon)d\epsilon = \frac{2\sqrt{\epsilon}}{\sqrt{\pi}(\kappa T)^{3/2}}e^{-\frac{\epsilon}{\kappa T}}d\epsilon,$$

the *energy distribution* of a Maxwellian gas [Kit04, Section 13]. In particular, by Proposition 4.20, $\mathbb{E}(E) = 3\kappa T/2$, which confirms the *principle of equipartition of energy*. □

By Corollary 4.12 we may rephrase Corollary 4.25 as saying that $Z \sim \mathcal{N}(\vec{0}, \mathrm{Id}_k)$ implies $\|Z\|^2 \sim \chi^2_k$. It turns out that this is just a special case of a more general result which makes it clear the geometric meaning of the notion of degree of freedom for a chi-square distribution.

Proposition 4.27. *If $Z \sim \mathcal{N}(\vec{0}, \mathrm{Id}_k)$ and $W = \langle Y, QY\rangle$, where Q is a $n \times n$ symmetric and idempotent matrix with $\operatorname{rank} Q = r \leq k$, then $W \sim \chi^2_r$.*

Proof. Since $Q : \mathbb{R}^n \to \mathbb{R}^n$ defines a projection onto its range $\operatorname{Im} Q$, a linear subspace of dimension r, we may use the projection property in (4.8), with $\sigma = 1$ and $\vec{a}$ running over an orthonormal basis of $\operatorname{Im} Q$, to conclude that $QZ \sim \mathcal{N}(\vec{0}, \mathrm{Id}_r)$. Thus, $W = \langle Z, QZ\rangle = \|QZ\|^2 \sim \chi^2_r$ by Corollary 4.25. □

This kind of geometric argument has many useful applications, including the next one, whose proof we omit.

Proposition 4.28. *Let $Z \sim \mathcal{N}(\vec{0}, \mathrm{Id}_n)$, $c \in \mathbb{R}^n$ and A a symmetric $n \times n$ matrix. Then $\langle c, Z\rangle$ and $\langle Z, AZ\rangle$ are independent if and only if $Ac = \vec{0}$.*

We now discuss some more random variables related to the normal distribution.

Definition 4.29. A random variable X is $\mathfrak{t}$-*Student distributed* with $k \geq 1$ degrees of freedom if

$$\psi_X(x) = \mathfrak{t}_k(x) := \frac{\Gamma(\frac{k+1}{2})}{\sqrt{k\pi}\Gamma(k/2)}\left(1 + k^{-1}x^2\right)^{-(k+1)/2}. \tag{4.26}$$

Proposition 4.30. *If $Z \sim \mathcal{N}(0,1)$ and $W \sim \chi^2_k$ with $Z \perp W$ then $Z/\sqrt{W/k} \sim \mathfrak{t}_k$.*

Proof. Note that $Z/\sqrt{W/k} = \sqrt{k}Z/V$, where $\sqrt{k}Z \sim \mathcal{N}(0,k)$ and $V := \sqrt{W}$ so that $\psi_V(v) = 2v\chi_k^2(v^2)$ by (2.11). It follows from (2.12) that

$$\psi_{Z/\sqrt{W/k}}(x) = \frac{1}{\sqrt{k\pi}2^{(k-1)/2}\Gamma(k/2)} \int_0^{+\infty} e^{-\frac{1}{2}(1+k^{-1}x^2)v^2} v^k dv.$$

The substitution $w = \frac{1}{2}(1+k^{-1}x^2)v^2$ then finishes the job. □

Remark 4.31. Proposition 4.30 can also be derived using Fisher's geometric method, as illustrated in Remark 4.26. This line of reasoning was first presented in [Fis25] and is reproduced here in Remark 7.31, where the method is applied to obtain the pdf of Student's sampling distribution defined in (7.26) below. □

Definition 4.32. Given $k_1, k_2 \in \mathbb{N}$ we say that a random variable X is F_{k_1,k_2}*-distributed* if

$$\psi_X(x) = \mathsf{F}_{k_1,k_2}(x) := c_{k_1,k_2} x^{k_1/2-1} \left(1 + \frac{k_1}{k_2}x\right)^{-\frac{k_1+k_2}{2}} \mathbf{1}_{(0,+\infty)}(x),$$

where

$$c_{k_1,k_2} = \frac{\Gamma\left(\frac{k_1+k_2}{2}\right)}{\Gamma(k_1/2)\Gamma(k_2/2)} \left(\frac{k_1}{k_2}\right)^{k_1/2}.$$

Proposition 4.33. *If $W_1 \sim \chi^2_{k_1}$ and $W_2 \sim \chi^2_{k_2}$ with $W_1 \perp W_2$ then*

$$\frac{W_1/k_1}{W_2/k_2} \sim \mathsf{F}_{k_1,k_2}.$$

Proof. From (2.12) and (4.24) we find that

$$\psi_{W_1/W_2}(x) = \frac{x^{\frac{k_1}{2}-1}}{2^{\frac{k_1+k_2}{2}}\Gamma(k_1/2)\Gamma(k_2/2)} \int_0^{+\infty} v^{\frac{k_1+k_2}{2}-1} e^{-(1+x)v/2} dv,$$

so that the substitution $w = (1+x)v/2$ transforms this into

$$\psi_{W_1/W_2}(x) = \frac{\Gamma\left(\frac{k_1+k_2}{2}\right)}{\Gamma(k_1/2)\Gamma(k_2/2)} x^{\frac{k_1}{2}-1}(1+x)^{-\frac{k_1+k_2}{2}}. \tag{4.27}$$

The result now follows because

$$\psi_Y(x) = \frac{k_1}{k_2}\psi_{W_1/W_2}\left(\frac{k_1}{k_2}x\right).$$

□

Corollary 4.34. *If $Y \sim \mathsf{F}_{k_1,k_2}$ then $Y^{-1} \sim \mathsf{F}_{k_2,k_1}$. Also, if $T \sim \mathsf{t}_k$ then $T^2 \sim \mathsf{F}_{1,k}$.*

Definition 4.35. A random variable X is Beta-distributed with shape parameters $\alpha, \beta > 0$ if

$$\psi_X(x) = \mathsf{Beta}(\alpha,\beta)(x) := \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}\mathbf{1}_{(0,1)}(x), \quad x \in \mathbb{R}. \tag{4.28}$$

Proposition 4.36. *If $W_1 \sim \chi^2_{k_1}$ and $W_2 \sim \chi^2_{k_2}$ with $W_1 \perp W_2$ then*

$$\widehat{W} := \frac{W_1}{W_1 + W_2} \sim \mathsf{Beta}\left(\frac{k_1}{2}, \frac{k_2}{2}\right).$$

Proof. We have

$$\widehat{W} = \frac{W_1/W_2}{1 + W_1/W_2},$$

so that, for $x \in (0,1)$,

$$\begin{aligned} F_{\widehat{W}}(x) &= P\left(\widehat{W} \leq x\right) \\ &= P\left(\frac{W_1}{W_2} \leq \frac{x}{1-x}\right) \\ &= F_{W_1/W_2}\left(\frac{x}{1-x}\right). \end{aligned}$$

By taking derivative with respect to x,

$$\psi_{\widehat{W}}(x) = (1-x)^{-2}\psi_{W_1/W_2}\left(\frac{x}{1-x}\right),$$

and the result follows from (4.27). □

5. Concentration inequalities

We now elaborate on the idea that an exponential bound on the moment generating function (mgf) of a random variable naturally yields *non-asymptotic* estimates for its tail probabilities, usually referred to as *concentration inequalities*. As will be seen through concrete examples, the *Cramér–Chernoff method* offers a systematic approach to deriving such inequalities and has numerous applications in both pure and applied mathematics. For comprehensive treatments of concentration phenomena and their significance in modern Data Science, including their role in the analysis of high-dimensional regression models; see [Ver18, Wai19]. As introductory illustrations, we present below their connections with the Johnson–Lindenstrauss Lemma and with the Erdös–Rényi random graph model.

5.1. **Sub-exponential random variables and the Johnson-Lindenstrauss Lemma.** If $X \sim \mathcal{N}(\mu, \sigma^2)$ and $t > 0$ then the tail probability

$$P(|X - \mu| \geq t) = \frac{2}{\sqrt{2\pi}\sigma}\int_t^{+\infty} e^{-x^2/2\sigma^2}dx$$

may be easily estimated by observing that $x/t \geq 1$ implies

$$P(|X - \mu| \geq t) \leq \frac{2}{\sqrt{2\pi}\sigma}\int_t^{+\infty} \frac{x}{t}e^{-x^2/2\sigma^2}dx,$$

thus yielding the exponential tail bound

$$P(|X - \mu| \geq t) \leq \sqrt{\frac{2}{\pi}}\frac{\sigma}{t}e^{-t^2/2\sigma^2}, \tag{5.1}$$

which happens to blow up as $t \to 0$. We may remedy this by considering the function

$$\begin{aligned}\xi(t) &= P(X-\mu \geq t) - \frac{1}{2}e^{-t^2/2\sigma^2} \\ &= 1 - F_{X-\mu}(t) - \frac{1}{2}e^{-t^2/2\sigma^2}.\end{aligned}$$

Since

$$\xi'(t) = \left(\frac{t}{2\sigma^2} - \frac{1}{\sqrt{2\pi}\sigma}\right) e^{-t^2/2\sigma^2},$$

we see that x decreases in the interval $(0, \sqrt{2/\pi}\sigma)$ and increases in the interval $(\sqrt{2/\pi}\sigma), +\infty)$. Since $\xi(0) = 0$ and $\xi(t) \to 0$ as $t \to +\infty$ we conclude that

$$P(|X-\mu| \geq t) \leq e^{-t^2/2\sigma^2}, \tag{5.2}$$

which provides a sharp bound as $t \to 0$. As yet another way to obtain an exponential tail probability bound, observe that

$$\begin{aligned}P(|X-\mu| \geq t) &= 2P(X-\mu \geq t) \\ &\overset{u\geq 0}{=} 2P(e^{u(X-\mu)} \geq e^{ut}) \\ &\overset{(2.16)}{\leq} 2\frac{\mathbb{E}(e^{u(X-\mu)})}{e^{ut}} \quad \text{(Markov's inequality)} \\ &\overset{(4.6)}{=} 2e^{\frac{1}{2}\sigma^2u^2 - ut},\end{aligned}$$

and since the right-hand side reaches its minimal value at $u = t/\sigma^2$ we conclude that

$$P(|X-\mu| \geq t) \leq 2e^{-t^2/2\sigma^2}. \tag{5.3}$$

Although in general (5.1) and (5.2) give sharper bounds than (5.3), this latter argument seems to be more promising as it suggests that a suitable exponential control on the mgf of a random variable might yield corresponding bounds for its tail probabilities, a trick usually referred to as the *Cramér-Chernoff method*. Its flexibility is illustrated by considering the following class of random variables.

Definition 5.1. We say that X is *Sub-Gaussian* if

$$\mathbb{E}\left(e^{(X-\mu)u}\right) \leq e^{\frac{\sigma^2u^2}{2}}, \quad u \in \mathbb{R}, \tag{5.4}$$

which we represent as $X \in \mathsf{SubG}(\sigma)$.

Clearly, any $X \in \mathsf{SubG}(\sigma)$ satisfies (5.3). As a simple example distinct from a normal to which this applies, note that the Rademacher variable $\boldsymbol{\epsilon}$ in Definition 2.36 satisfies

$$\mathbb{E}_{\boldsymbol{\epsilon}}(e^{\boldsymbol{\epsilon}u}) = \cosh u \leq e^{\frac{u^2}{2}}, \quad u \in \mathbb{R}, \tag{5.5}$$

so that $\boldsymbol{\epsilon} \in \mathsf{SubG}(1)$. Also, if $\{X_j\}_{j=1}^N$ is independent with $X_j \in \mathsf{SubG}(\sigma_j)$ then

$$\sum_j X_j \in \mathsf{SubG}\left(\sqrt{\sum_j \sigma_j^2}\right). \tag{5.6}$$

Thus, if we apply this to $X^{(N)} = \epsilon_1 + \cdots + \epsilon_N$, where $\{\epsilon_j\}_{j=1}^N$ is a collection of independent Rademacher variables, we see that $X^{(N)} \in \mathsf{SubG}(\sqrt{N})$ and hence

$$P(|X^{(N)}| \geq t) \leq 2e^{-t^2/2N}, \tag{5.7}$$

or equivalently,

$$P\left(\frac{|X^{(N)}|}{N} \geq t\right) \leq 2e^{-Nt^2/2}. \tag{5.8}$$

The next result substantially enriches the class of sub-Gaussian random variables.

Proposition 5.2. *If X is bounded, say $a \leq X \leq b$, then $X \in \mathsf{SubG}(b-a)$.*

Proof. We may assume that $\mathbb{E}(X) = 0$. Let Y be an independent copy of X, so that $X - Y$ is symmetric by Proposition 2.32 (3). Hence, if $\boldsymbol{\epsilon}$ is a Rademacher variable independent from both X and Y then $\boldsymbol{\epsilon}(X-Y)$ and $X - Y$ are identically distributed by Proposition 2.37. It follows that

$$\begin{aligned}\mathbb{E}_{(X,Y)}(e^{(X-Y)u}) &= \mathbb{E}_{(X,Y)}\left(\mathbb{E}_{\boldsymbol{\epsilon}}\left(e^{(X-Y)u}\right)\right)\\ &= \mathbb{E}_{(X,Y)}\left(\mathbb{E}_{\boldsymbol{\epsilon}}\left(e^{\boldsymbol{\epsilon}(X-Y)u}\right)\right)\\ &\overset{(5.5)}{\leq} \mathbb{E}_{(X,Y)}\left(e^{\frac{(X-Y)^2u^2}{2}}\right),\end{aligned}$$

and since $|X - Y| \leq b - a$, we see that

$$\mathbb{E}_{(X,Y)}(e^{(X-Y)u}) \leq e^{\frac{(b-a)^2u^2}{2}}.$$

On the other hand,

$$\begin{aligned}\mathbb{E}_{(X,Y)}(e^{(X-Y)u}) &= \mathbb{E}_X\left(e^{Xu}\mathbb{E}_Y\left(e^{-Yu}\right)\right)\\ &\geq \mathbb{E}_X\left(e^{Xu}e^{-\mathbb{E}_Y(Y)u}\right)\\ &= \mathbb{E}_X\left(e^{Xu}\right),\end{aligned}$$

where we used Jensen inequality and $\mathbb{E}_Y(Y) = 0$. Putting all the pieces of this computation together we conclude that

$$\mathbb{E}_X\left(e^{Xu}\right) \leq e^{\frac{(b-a)^2u^2}{2}},$$

as desired. □

Remark 5.3. Under the conditions of Proposition 5.2, it is possible to show that $X \in \mathsf{SubG}((b-a)/2)$, which is known as the *Hoeffding lemma*. □

The Hoeffding-type concentration inequalities obtained from Proposition 5.2 (or Remark 5.3) completely ignore the dispersion of a random variable as measured by its standard deviation. For example, if $X \sim \mathsf{Bin}(p; n)$ and, using Example 2.40, we write it as a sum of independent Bernoulli variables so that (5.6) applies, we obtain

$$P\left(X - np > t\right) \leq e^{-\frac{t^2}{2n}}, \quad t > 0, \tag{5.9}$$

an estimate that does not depend on the sampling probability $p \in (0, 1)$ (see also Subsection 5.3). Moreover, the class of sub-Gaussian random variables does not capture certain concentration inequalities that frequently arise in applications. This motivates the study of other families of random variables for which meaningful concentration inequalities can be established.

Definition 5.4. A random variable Y with $\mathbb{E}(Y) = \mu$ is *sub-exponential* if there exist positive parameters (ν, β) such that

$$\mathbb{E}\left(e^{u(Y-\mu)}\right) \leq e^{\frac{\nu^2 u^2}{2}}, \quad |u| < \frac{1}{\beta}. \tag{5.10}$$

We represent this as $Y \in \mathsf{SubE}(\nu, \beta)$.

Remark 5.5. If $\mu = 0$ then $Y \in \mathsf{SubE}(\nu, \beta)$ implies $-Y \in \mathsf{SubE}(\nu, \beta)$. Also, if $\{Y_j\}_{j=1}^n$ is independent and $Y_j \in \mathsf{SubE}(\nu_j, \beta_j)$ with

$$\sum_j Y_j \in \mathsf{SubE}\left(\sqrt{\sum_j \nu_j^2}, \max_j \{\beta_j\}\right),$$

where we assume that $\mu_j = \mathbb{E}(Y_j) = 0$. □

Example 5.6. From Corollary 4.21 we have that $Y \sim \chi_k^2$ implies

$$\mathbb{E}\left(e^{u(Y-k)}\right) = e^{-ku}\left(1-2u\right)^{-k/2}, \quad |u| < \frac{1}{2}.$$

Note that $\mathbb{E}(e^{\frac{1}{2}(Y-k)}) = +\infty$ and hence Y is not sub-Gaussian. However, by Taylor expanding around $u = 0$ we find that

$$f(u) := \left(1-2u\right)^{-k/2} = 1 + ku + k\left(\frac{k}{2}+1\right)u^2 + \cdots$$

and

$$g(u) := e^{k(2u^2+u)} = 1 + ku + k\left(\frac{k}{2}+2\right)u^2 + \cdots,$$

from which we easily see that $g(u) \geq f(u)$ for $|u| < 1/4$ since both functions remain convex in this interval. In other words,

$$\mathbb{E}\left(e^{u(Y-k)}\right) \leq e^{2ku^2}, \quad |u| < \frac{1}{4},$$

and we conclude that $Y \in \mathsf{SubE}(2\sqrt{k}, 4)$. □

It turns out that sub-exponential random variables obey a concentration inequality that reveals a sharp threshold separating the sub-Gaussian regime from the purely sub-exponential one.

Proposition 5.7. *If $Y \in \mathsf{SubE}(\nu, \beta)$ with $\mathbb{E}(Y) = \mu$ then*

$$P(|Y-\mu| \geq t) \leq \begin{cases} 2e^{-\frac{t^2}{2\nu^2}}, & 0 \leq t < \frac{\nu^2}{\beta} \\ 2e^{-\frac{t}{2\beta}}, & t \geq \frac{\nu^2}{\beta} \end{cases}$$

Proof. Clearly, we may assume that $\mu = 0$, so Markov inequality (2.16) gives

$$P(Y \geq t) \leq e^{h_t(u)}, \quad 0 \leq u < \frac{1}{\beta},$$

with the graph of $h_t(u) = -ut + \nu^2 u^2/2$ being an upward pointing parabola passing through $(0,0)$ and with vertex located at $(t/\nu^2, -t^2/2\nu^2)$. Thus, in the first case, when $t/\nu^2 < 1/\beta$, the tail probability is bounded by

$$e^{h_t(t/\nu^2)} = e^{-\frac{t^2}{2\nu^2}},$$

whereas in the second case, when $t/\nu^2 \geq 1/\beta$, it is bounded by

$$e^{h_t(1/\beta)} \leq e^{-\frac{t}{\beta}+\frac{\nu^2}{2\beta^2}} \leq e^{-\frac{t}{2\beta}}.$$

By Remark 5.5, an identical estimate holds for $P(Y \leq -t)$, which concludes the proof. □

Corollary 5.8. *If $Y \sim \chi^2_k$ then*

$$P(|k^{-1}Y - 1| \geq t) \leq \begin{cases} 2e^{-\frac{kt^2}{8}}, & 0 < t < 1 \\ 2e^{-\frac{kt}{8}}, & t \geq 1 \end{cases} \tag{5.11}$$

Proof. From Example 5.6 and the proposition (with $\mu = k$ and $(\nu, \beta) = (2\sqrt{k}, 4)$) we know that

$$P(|Y - k| \geq \tau) \leq \begin{cases} 2e^{-\frac{\tau^2}{8k}}, & 0 < \tau < k \\ 2e^{-\frac{\tau}{8}}, & \tau \geq k \end{cases} \tag{5.12}$$

Since

$$P(|k^{-1}Y - 1| \geq k^{-1}\tau) = P(|Y - k| \geq \tau),$$

the substitution $\tau = kt$ finishes the proof. □

We now use the concentration inequality (5.11) to establish a celebrated result with a number of applications both in pure and applied mathematics.

Theorem 5.9. *(Johnson-Lindenstrauss). If $\mathcal{C} = \{x_1, \cdots, x_n\} \subset \mathbb{R}^p$ is a collection of n points then, given $\epsilon, \delta \in (0,1)$, there exist $m = O(\epsilon^{-2}\ln(n/\delta))$ and a map $F : \mathbb{R}^p \to \mathbb{R}^m$ such that*

$$1 - \epsilon \leq \frac{\|F(x_i) - F(x_j)\|^2}{\|x_i - x_j\|^2} \leq 1 + \epsilon, \tag{5.13}$$

for all $x_i \neq x_j$ in $\mathcal{C}$.

Proof. Use Remark 2.18 to construct a $m \times p$ random matrix $\mathbf{A}$ whose entries are independent and $\mathcal{N}(0,1)$-distributed random variables and define the linear map

$$F : \mathbb{R}^p \to \mathbb{R}^m, \quad F(x) = \frac{\mathbf{A}x}{\sqrt{m}}, \tag{5.14}$$

with m to be chosen later on. If $\mathbf{a}_i$ is the i^{th} line of $\mathbf{A}$ and $x \neq \vec{0}$ then $\langle \mathbf{a}_i, x/\|x\| \rangle \sim \mathcal{N}(0,1)$ by Proposition 4.8 (3), so Corollary 4.25 applies to ensure that

$$\frac{\|\mathbf{A}x\|^2}{\|x\|^2} = \sum_{i=1}^{m} \left\langle \mathbf{a}_i, \frac{x}{\|x\|} \right\rangle^2 \sim \chi^2_m.$$

Thus, by Corollary 5.8,

$$P\left(\left| m^{-1} \frac{\|\mathbf{A}x\|^2}{\|x\|^2} - 1 \right| \geq \epsilon \right) \leq 2e^{-m\epsilon^2/8}, \quad 0 < \epsilon < 1,$$

or equivalently,

$$P\left(\frac{\|F(x)\|^2}{\|x\|^2} \notin (1-\epsilon, 1+\epsilon) \right) \leq 2e^{-m\epsilon^2/8}, \quad 0 < \epsilon < 1.$$

From this we easily deduce that

$$P\left(\frac{\|F(x_i)-F(x_j)\|^2}{\|x_i-x_j\|^2} \notin (1-\epsilon,1+\epsilon) \text{ for some } x_i \neq x_j\right) \leq 2\begin{pmatrix} n \\ 2 \end{pmatrix} e^{-m\epsilon^2/8},$$

and the result follows if we impose that the right-hand side equals δ (which determines m as in the statement of the theorem), since this means that the probability that (5.13) holds true is $\geq 1-\delta > 0$. □

Remark 5.10. The argument above provides a clear illustration of the celebrated *probabilistic method*. In broad terms, this technique consists of reformulating the deterministic assertion to be proved as the occurrence of a random event. In the present case, the deterministic statement in (5.13) is recast via the introduction of the random matrix **A** in (5.14). Once it is shown that this event takes place with *positive* probability, the validity of the original statement follows immediately. For many further examples of this powerful method, particularly within Combinatorics, see [AS16]. □

Remark 5.11. In applications to Data Science [Ver18, BHK20], it is common for the number n of *samples* to be much smaller than the number p of *features*. The striking aspect of Theorem 5.9 is that, once a controlled distortion $\epsilon > 0$ is allowed in the "approximate projection" F, the set $\mathcal{C}$ acquires a sort of intrinsic dimension m[12]. Remarkably, this intrinsic dimension scales like $\ln n$ and is entirely independent of the ambient dimension p, which may even be infinite. □

Remark 5.12. If $Y \in \mathsf{SubE}(\nu,\beta)$ then Proposition 5.7 says that

$$P(|Y-\mathbb{E}(Y)| \geq t) \leq 2e^{-\min\left\{\frac{t^2}{2\nu^2},\frac{t}{2\beta}\right\}},$$

thus confirming that as $t \to +\infty$ this tail bound is much heavier than the one for $X \in \mathsf{SubG}(\sigma)$ in (5.3) since in this regime the upper bound here is $2e^{-t/2\beta}$. In a sense this reflects the fact that $X \in \mathsf{SubG}(\sigma)$ with $\mathbb{E}(X)=0$ implies $Y = X^2 \in \mathsf{SubE}(\nu,\beta)$ for some (ν,β) to be determined below. To check this claim, take $v \in (0,1)$, multiply both sides of the defining condition for X in (5.4) (with $\mu=0$) by $e^{-\sigma^2u^2/2v}$ and integrate to obtain

$$\begin{aligned}\int_{-\infty}^{+\infty} e^{\frac{\sigma^2u^2(v-1)}{2v}}du &\geq \int_{-\infty}^{+\infty}\mathbb{E}\left(e^{uX-\frac{\sigma^2u^2}{2v}}\right)du \\ &= \int_{-\infty}^{+\infty}\left(\int_{-\infty}^{+\infty} e^{ux-\frac{\sigma^2u^2}{2v}}dP_X(x)\right)du \\ &= \int_{-\infty}^{+\infty}\left(\int_{-\infty}^{+\infty} e^{ux-\frac{\sigma^2u^2}{2v}}du\right)dP_X(x).\end{aligned}$$

Using that

$$\int_{-\infty}^{+\infty} e^{-au^2+bu+c}du = e^{\frac{b^2+4ac}{4a}}\sqrt{\frac{\pi}{a}}, \quad a>0, \quad b,c\in\mathbb{R},$$

we may compute the Gaussian integrals above to conclude that

$$\mathbb{E}\left(e^{\frac{X^2v}{2\sigma^2}}\right) \leq \frac{1}{\sqrt{1-v}}, \quad 0 \leq v < 1,$$

or equivalently,

$$\mathbb{E}\left(e^{X^2u}\right) \leq \frac{1}{\sqrt{1-2\sigma^2u}}, \quad 0 \leq u < \frac{1}{2\sigma^2}.$$

[12]Concretely, there exists a subset $\mathcal{I} \subset \{1,\ldots,p\}$ of indices with $\sharp\mathcal{I} = p-m$ such that $y_j = 0$ for all $y \in F(\mathcal{C})$ and $j \in \mathcal{I}$. In other words, up to a global distortion measured by ϵ, m corresponds to the number of relevant (nonzero) features of the elements of $\mathcal{C}$, which explains the significance of the result for data reduction.

In particular, there holds

$$\mathbb{E}\left(e^{(X^2-\mathbb{E}(X^2))u}\right) < +\infty$$

for $u \in [0, \varepsilon)$, $\varepsilon > 0$. By Remark 2.39, $\mathbb{E}(|X|^4) < +\infty$ and we may Taylor expand up to second order:

$$\mathbb{E}\left(e^{(X^2-\mathbb{E}(X^2))u}\right) = 1 + \frac{\mathbb{E}((X^2 - \mathbb{E}(X^2))^2)}{2}u^2 + o(u^2).$$

Comparing this with

$$e^{\frac{\nu^2u^2}{2}} = 1 + \frac{\nu^2u^2}{2} + o(u^2),$$

we easily see that $Y = X^2 \in \mathsf{SubE}(\nu, \beta)$ if we take

$$\nu^2 > \mathbb{E}((X^2 - \mathbb{E}(X^2))^2)$$

and $\beta > 0$ large enough. Finally, note that if $\sigma = 1$ then we can take $\beta = 1$. □

Remark 5.13. If $\{Y_j\}_{j=1}^k$ is independent with $Y_j \in \mathsf{SubE}(\nu_j, \beta_j)$ and $\mathbb{E}(Y_j) = 0$ then we easily see that $Y = \sum_j Y_j \in \mathsf{SubE}(\nu, \beta)$, where

$$\nu^2 = \sum_j \nu_j^2, \quad \beta = \max_j \beta_j.$$

You may apply this to $Y_j = X_j^2$, $X_j \sim \mathcal{N}(0, 1)$. Using the well-known fact that, for a standard normal, $\mathbb{E}((X_j^2 - \mathbb{E}(X_j^2))^2) = 3$, and the computation in Example 5.12 above, we conclude that $\sum_j X_j^2 \in \mathsf{SubE}(\sqrt{2}k, \beta)$, for some $\beta > 0$, which provides essentially the same result as in Example 5.6. Clearly, this suffices for the applications culminating in Theorem 5.9 above. Notice that in this derivation no appeal to the explicit expression for the pdf of a chi-square distribution has been needed. □

5.2. **The Gaussian concentration inequality.** A suitable rewording of the inequalities in Corollary 5.8 provides valuable insights on the "high dimensional" behavior of standard normal random vectors, including the precise formulation of a remarkable dimension-free concentration inequality for Lipschitz functions of such vectors; see (5.18) below. Indeed, let $X = (X_1, \cdots, X_k)$ be such a vector, which means by definition that $\{X_j\}_{j=1}^k$ is independent with $X_j \sim \mathcal{N}(0, 1)$. Recalling that

$$|a - 1| \geq \delta \Longrightarrow |a^2 - 1| \geq \max\left\{\delta, \delta^2\right\}, \quad a \geq 0, \quad \delta > 0,$$

we have

$$\begin{aligned} P\left(\left|\frac{\|X\|}{\sqrt{k}} - 1\right| \geq \frac{\tau}{k}\right) &\leq P\left(\left|\frac{\|X\|^2}{k} - 1\right| \geq \max\left\{\frac{\tau}{k}, \frac{\tau^2}{k^2}\right\}\right) \\ &= P\left(\left|\|X\|^2 - k\right| \geq \max\left\{\tau, \frac{\tau^2}{k}\right\}\right) \\ &= P\left(\left|\|X\|^2 - k\right| \geq \left\{\begin{array}{ll} \tau, & \tau < k \\ \frac{\tau^2}{k} & \tau \geq k \end{array}\right.\right) \\ &\leq 2e^{-\frac{\tau^2}{8k}}, \quad \tau > 0, \end{aligned}$$

where we used (5.12) with $Y = \|X\|^2$ in the last step. By setting $t = \tau/k$ we then obtain

$$P\left(\left|\frac{\|X\|}{\sqrt{k}} - 1\right| \geq t\right) \leq 2e^{-\frac{kt^2}{8}}, \quad t > 0, \tag{5.15}$$

or equivalently,

$$P\left(\left|\|X\| - \sqrt{k}\right| \geq t\right) \leq 2e^{-\frac{t^2}{8}}, \quad t > 0, \tag{5.16}$$

These concentration inequalities say that, with a very high[13] probability, $X/\sqrt{k}$ remains at an *arbitrarily small* distance from the unit sphere $\mathbb{S}^{k-1} \subset \mathbb{R}^k$ or, equivalently, X remains at a *bounded* distance from the round sphere $\mathbb{S}^{k-1}_{\sqrt{k}} \subset \mathbb{R}^k$ of radius $\sqrt{k}$ as $k \to +\infty$. We note the striking similarity between (5.15) and (5.8): in both cases the relevant random variable, which turns out to be a function of a large collection of independent variables, becomes *almost constant* when properly re-scaled; for more on this perspective, which in a sense underlies the modern applications of the "concentration of measure phenomenon" to (high dimensional) probability, see [Tal96].

Remark 5.14. (Poincaré's limit theorem) From Remark 4.26 we know that a random vector $Z^{[k]}$ uniformly distributed over $\mathbb{S}^{k-1}_{\sqrt{k}}$ may be expressed as

$$Z^{[k]} = \sqrt{k}\Theta^{[k]}, \tag{5.17}$$

where $\Theta^{[k]} = X^{[k]}/\|X^{[k]}\|$, $X^{[k]} \sim \mathcal{N}(\vec{0}, \mathrm{Id}_k)$. Now, it follows from (5.15) that $\|X^{[k]}\|/\sqrt{k} \xrightarrow{p} 1$ as $k \to +\infty$[14]. On the other hand, if $x \in \mathbb{R} = \mathbb{R}^1 \subset \mathbb{R}^2 \subset \cdots \subset \mathbb{R}^k \subset \cdots$, $\|x\| = 1$, we know that $\langle X^{[k]}, x\rangle \sim \mathcal{N}(0,1)$ by Proposition 4.8 (3). Thus, from the identity

$$\langle Z^{[k]}, x\rangle = \frac{\sqrt{k}}{\|X^{[k]}\|}\langle X^{[k]}, x\rangle$$

and Theorem 2.25 we conclude that $\langle Z^{[k]}, x\rangle \xrightarrow{d} \mathcal{N}(0,1)$. Put in another way, as $k \to +\infty$ the marginals of $Z^{[k]}$ corresponding to a given set of $l \geq 1$ coordinates converge in distribution to a standard normal vector $Z^{[\infty]} \sim \mathcal{N}(\vec{0}, \mathrm{Id}_l)$, a statement usually referred to as "Poincaré's limit theorem" [HN64, McK73, DF87]. □

As yet another instance of an insight coming from the concentration inequalities above, if we compare the bounds in (5.15) and (5.16) with the normal bound in (5.3), we see that the corresponding fluctuations, as measured by the standard deviation, are $O(1/\sqrt{k})$ and $O(1)$, respectively. Noticing that

$$\frac{1}{\sqrt{k}} = \mathrm{Lip}\left(x \mapsto \frac{\|x\|}{\sqrt{k}}\right)$$

and

$$1 = \mathrm{Lip}\left(x \mapsto \|x\|\right),$$

where Lip denotes the Lispschitz constant of a function on $\mathbb{R}^k$ (with respect to the euclidean norm), we are thus led to suspect that the dimension-free inequality

$$P\left(|F(X) - \mathbb{E}(F(X))| > t\right) \leq 2e^{-\frac{Ct^2}{\mathrm{Lip}(F)^2}}, \quad t > 0, \tag{5.18}$$

should hold true for some universal constant $C > 0$ not depending on k, where $F : \mathbb{R}^k \to \mathbb{R}$ is assumed to be Lipschitz. Of course, the optimal possibility is $C = 1/2$, in which case (5.18) says that if $\mathrm{Lip}(F) = 1$ then $F(X)$ is at least as much concentrated around its mean as each X_j, regardless of the size k of the sample. For discussions on this *Gaussian Concentration Inequality* which explore the connection with several other mathematical topics, including the geometric notion of *isoperimetry* and the analytical concept of *hypercontractivity*, we refer to [Led01, Led06, LT13, BLM13]; see also Remark 5.15 below. An elegant approach to the sharpest version of (5.18), due to Maurey and Pisier [Pis06, Chapter 2] and relying heavily on Itô's Stochastic Calculus,

[13] That is, as close to 1 as we wish!

[14] This assertion also follows from the Law of Large Numbers (Theorem 6.2 below). Indeed, if $X \sim \mathcal{N}(\vec{0}, \mathrm{Id}_k)$ then

$$\|X\|^2 = \sum_{j=1}^{k} X_j^2,$$

where $X_j \sim \mathcal{N}(0,1)$ and hence $X_j^2 \sim \chi_1^2$ with $\mathbb{E}(X_j^2) = 1$ by Corollary 4.25. Thus, by making $X = X^{[k]}$, Theorem 6.2 applies to ensure that $\|X^{[k]}\|^2/k \xrightarrow{p} 1$, as desired.

is presented in Subsection A.4 below. We provide here a more pedestrian (but no less elegant!) argument, also available in [Pis06, Chapter 2], which delivers a constant slightly smaller than the optimal one ($1/2$ gets replaced by $2/\pi^2$).

Without loss of generality, we may assume that F is smooth (so that $\|\nabla F\| \leq \mathrm{Lip}(f)$ a.s.) and $\mathbb{E}(F(X)) = 0$. This latter assumption implies, via Jensen's inequality, that $\mathbb{E}(e^{vF(X)}) \geq e^{v\mathbb{E}(F(X))} = 1$ for any $v \in \mathbb{R}$, which gives

$$\mathbb{E}\left(e^{uF(X)}\right) \leq \mathbb{E}_{(X,X')}\left(e^{u\left(F(X)-F(X')\right)}\right), \quad u \geq 0, \tag{5.19}$$

where X' is an independent copy of X (so that $(X, X') \sim \mathcal{N}(\vec{0}, \mathrm{Id}_{2k})$). For each $\theta \in [0, \pi/2]$ define

$$X_\theta = \cos\theta\, X + \sin\theta\, X'$$

and

$$X'_\theta = \frac{dX_\theta}{d\theta} = -\sin\theta\, X + \cos\theta\, X'.$$

Since

$$\begin{pmatrix} X_\theta \\ X'_\theta \end{pmatrix} = \begin{pmatrix} \cos\theta\,\mathrm{Id}_k & \sin\theta\,\mathrm{Id}_k \\ -\sin\theta\,\mathrm{Id}_k & \cos\theta\,\mathrm{Id}_k \end{pmatrix} \begin{pmatrix} X \\ X' \end{pmatrix},$$

it follows from the rotational invariance in Corollary 4.13 that (X_θ, X'_θ) is identically distributed to (X, X') with X'_θ being an independent copy of X_θ as well. Now,

$$F(X) - F(X') = \int_0^{\pi/2} \langle (\nabla F)(X_\theta), X'_\theta \rangle d\theta,$$

so that Jensen's inequality gives

$$e^{u\left(F(X)-F(X')\right)} \leq \frac{2}{\pi} \int_0^{\pi/2} e^{\frac{\pi}{2} u \langle (\nabla F)(X_\theta), X'_\theta \rangle} d\theta,$$

and hence,

$$\begin{aligned} \mathbb{E}_{(X,X')}\left(e^{u\left(F(X)-F(X')\right)}\right) &\leq \frac{2}{\pi} \int_0^{\pi/2} \mathbb{E}_{(X,X')}\left(e^{\frac{\pi}{2} u \langle (\nabla F)(X_\theta), X'_\theta \rangle}\right) d\theta \\ &= \mathbb{E}_{(X,X')}\left(e^{\frac{\pi}{2} u \langle (\nabla F)(X), X' \rangle}\right) \\ &= \int_{\mathbb{R}^k} \left(\mathbb{E}_{X'}\left(e^{\frac{\pi}{2} u \langle (\nabla F)(x), X' \rangle}\right)\right) dP_X(x), \end{aligned}$$

where in the second step we used the consequences of the rotational invariance mentioned above to ensure that the expectation integrand does not depend on θ and in the last step we used the independence of $\{X, X'\}$. Now, for each x such that $\nabla F(x) \neq \vec{0}$ we known from Proposition 4.8 (3) that

$$\frac{\pi}{2} \langle (\nabla F)(x), X' \rangle \sim \mathcal{N}\left(0, \frac{\pi^2}{4} \|\nabla F(x)\|^2\right),$$

so that (4.6) gives

$$\mathbb{E}_{X'}\left(e^{\frac{\pi}{2} u \langle (\nabla F)(x), X' \rangle}\right) \leq e^{\frac{\pi^2}{8} \mathrm{L}(f)^2 u^2}.$$

Note that the right-hand side does not depend on x and the estimate remains true if $(\nabla F)(x) = \vec{0}$. Thus, if we put all the pieces of our calculation together we obtain

$$\mathbb{E}\left(e^{uF(X)}\right) \leq e^{\frac{\pi^2}{8} \mathrm{L}(f)^2 u^2},$$

which amounts to saying that $F(X) \in \mathsf{SubG}(\pi \mathrm{Lip}(f)/2)$, a quite good but not entirely satisfactory estimate due to the $\pi/2$ factor appearing in the sub-Gaussian parameter[15]. In any case, we may now appeal to the Cramér-Chernoff method introduced above: for each $t > 0$,

$$P\left(|F(X)| > t\right) \leq 2e^{\frac{\pi^2}{8}\mathrm{L}(f)^2 u^2 - ut}, \quad u \geq 0,$$

and minimizing the right-hand side we finally get

$$P\left(|F(X)| > t\right) \leq 2e^{-\frac{2t^2}{\pi^2 \mathrm{L}(f)^2}},$$

as desired.

Remark 5.15. (Poncaré's limit revisited) Let $\Pi_{k,l} : \mathbb{R}^k \to \mathbb{R}^l$ be the orthogonal projection associated to the natural embedding $\mathbb{R}^l \hookrightarrow \mathbb{R}^k$ and let P_k be the uniform probability measure on $\mathbb{S}^{k-1}_{\sqrt{k}}$. With this terminology, the "Poncaré's limit theorem" in Remark 5.14 says that the random vectors $\widetilde{\Pi}_{k,l} = \Pi_{k,l}|_{\mathbb{S}^{k-1}_{\sqrt{k}}} : (\mathbb{S}^{k-1}_{\sqrt{k}}, P_k) \to \mathbb{R}^l$ converge in distribution to $Z^{[\infty]} \sim \mathcal{N}(\vec{0}, \mathrm{Id}_l)$, which means that $\mathbb{E}(\xi(\widetilde{\Pi}_{k,l})) \to \mathbb{E}(\xi(Z^{[\infty]}))$ for any $\xi : \mathbb{R}^l \to \mathbb{R}$ uniformly bounded and continuous; cf. Definition 2.23. It turns out that with a bit more of effort it may be checked that this statement actually holds true with $\xi = \mathbf{1}_A$, the indicator function of an arbitrary Borel set $A \in \mathcal{B}^l$. Precisely,

$$\lim_{k \to +\infty} P_{\widetilde{\Pi}_{k,l}}(A) = \frac{1}{(2\pi)^{l/2}} \int_A e^{-\|y^2\|/2} dy. \tag{5.20}$$

To prove this claim, let us first observe that, with the notation of Remark 5.14,

$$\begin{aligned} P_{\widetilde{\Pi}_{k,l}}(A) &= P_k(\widetilde{\Pi}_{k,l}^{-1}(A)) \\ &= P_k\left(\Pi_{k,l}^{-1}(A) \cap \mathbb{S}^{k-1}_{\sqrt{k}}\right) \\ &= P\left(\Pi_{k,l}(Z^{[k]}) \in A\right), \end{aligned}$$

so if $R_m^2 := X_1^2 + \cdots + X_m^2$, $1 \leq m \leq k$, (5.17) gives

$$\begin{aligned} P_{\widetilde{\Pi}_{k,l}}(A) &= P\left(\frac{\sqrt{k}}{R_k}(X_1, \cdots, X_l) \in A\right) \\ &= P\left(\left(k\frac{R_l^2}{R_k^2}\right)^{1/2} \frac{1}{R_l}(X_1, \cdots, X_l) \in A\right). \end{aligned}$$

Now, by Remark 4.26,

$$\left\{R_l^2, R_k^2 - R_l^2, \frac{1}{R_l}(X_1, \cdots, X_l)\right\}$$

is independent and therefore

$$\left\{\frac{R_l^2}{R_k^2}, \frac{1}{R_l}(X_1, \cdots, X_l)\right\}$$

is independent as well. On the other hand, by Corollary 4.25 and Proposition 4.36,

$$\frac{R_l^2}{R_k^2} = \frac{R_l^2}{R_l^2 + (R_k^2 - R_l^2)} \sim \mathsf{Beta}\left(\frac{l}{2}, \frac{k-l}{2}\right),$$

[15]This should be compared to the sharp estimate in (A.23) obtained by means of the full machinery of the Stochastic Calculus.

so if we put together these facts we get, by Proposition 2.15,

$$\begin{aligned} P_{\widetilde{\Pi}_{k,l}}(A) &= \frac{\Gamma(k/2)}{\Gamma(l/2)\Gamma((k-l)/2)} \times \\ &\qquad \omega_{l-1}^{-1}\int_{\mathbb{S}_1^{l-1}}\int_0^1 \mathbf{1}_A(\sqrt{kx}\theta)x^{\frac{l}{2}-1}(1-x)^{\frac{k-l}{2}-1}d\mathbb{S}_1^{l-1}(\theta)dx, \end{aligned}$$

where $d\mathbb{S}_1^{l-1}(\theta)$ is the (unormalized) volume element of the unit sphere $\mathbb{S}_1^{l-1}$ (induced by the embedding $\mathbb{S}_1^{l-1}\hookrightarrow\mathbb{R}^l$) and $\omega_{l-1}=\mathrm{vol}_{l-1}(\mathbb{S}_1^{l-1})$. Thus, if we set $u=\sqrt{kx}$ and use (4.25) we find that

$$\begin{aligned} P_{\widetilde{\Pi}_{k,l}}(A) &= \frac{\Gamma(k/2)}{\Gamma((k-l)/2)}\frac{1}{\pi^{l/2}k^{l/2}} \times \\ &\qquad \int_{\mathbb{S}_1^{l-1}}\int_0^{\sqrt{k}} \mathbf{1}_A(u\theta)u^{l-1}\left(1-\frac{u^2}{k}\right)^{\frac{k-l}{2}-1} d\mathbb{S}_1^{l-1}(\theta)du. \end{aligned}$$

Since l is held fixed, the Stirling approximation in (6.1) below gives

$$\frac{\Gamma(k/2)}{\Gamma((k-l)/2)} \approx_{k\to+\infty} 2^{-l/2}k^{l/2}\left(\frac{k-l}{k}\right)^{-l/2},$$

so we end up with

$$\lim_{k\to+\infty} P_{\widetilde{\Pi}_{k,l}}(A) = \frac{1}{(2\pi)^{l/2}}\int_{\mathbb{S}_1^{l-1}}\int_0^{+\infty} \mathbf{1}_A(u\theta)u^{l-1}e^{-u^2/2}d\mathbb{S}_1^{l-1}(\theta)du,$$

which proves the claim because this double integral clearly equals the right-hand side of (5.20) under the substitution $y=u\theta$. The limit theorem in (5.20) is the key ingredient in explicitly solving the isoperimetric problem for the *Gaussian space* $(\mathbb{R}^l,\delta,(2\pi)^{-l/2}e^{-|y|^2/2}dy)$ by essentially viewing it as the limit of the corresponding problem for large, high-dimensional spheres $\mathbb{S}^{k-1}_{\sqrt{k}}$ as $k\to+\infty$ [Bor75, ST78], a celebrated result which by its turn may be used to establish a version of (5.18) with the optimal constant $C=1/2$, but this time with the mean replaced by the median [Led06, Chapter 2]. Needless to say, this is a prominent instance of the "concentration of measure phenomenon" extensively studied elsewhere [GKPS99, Led01, BLM13, Shi16]. □

5.3. **Chernoff-type bounds for binomial trials and the Erdös-Rényi model.** As illustrated in (5.9), the sub-Gaussian version of the Cramér-Chernoff method yields an estimate which fails to account for the dispersion of a binomial trial (as measured by its standard deviation). We may remedy this by directly applying the method to $X\sim\mathsf{Bin}(p;n)$ as in Example 2.40 in order to get, for $t>0$,

$$\begin{aligned} P\left(X\geq t\right) &= P\left(e^{Xu}\geq e^{tu}\right) \\ &= e^{-tu}\mathbb{E}\left(e^{Xu}\right) \quad \text{(Markov)} \\ &\overset{(2.28)}{=} e^{-tu}\left(1-p+pe^u\right)^n. \end{aligned}$$

Using that $1+x\leq e^x$, $x\geq 0$, we obtain

$$P\left(X\geq t\right)\leq e^{-tu+np(e^u-1)}, \tag{5.21}$$

and since the function on the exponent is minimized at $u=\ln(t/\lambda)$, where $\lambda=np=\mathbb{E}(X)$ is the expectation, we end up with the *Chernoff-type* inequality

$$P\left(X\geq t\right)\leq e^{-\lambda}\left(\frac{e\lambda}{t}\right)^t, \quad t>\lambda. \tag{5.22}$$

Regarding this analysis, the following comments are worth mentioning:

- From (5.22) we have

$$P(X \geq t) \leq C_1 e^{C_2 t - t \ln t}, \tag{5.23}$$

where $C_1 = e^{-\lambda}$ and $C_2 = 1 + \ln \lambda$, which for t large gives a tail behavior somehow interpolating between the sub-Gaussian and sub-exponential regimes.

- Past experience with the sub-exponential case in Proposition 5.7 suggests that we should be able to recover a sub-Gaussian tail for *small* deviations around the mean λ (which is the only critical point of the exponential function in the right-hand side of (5.23)). This is the case indeed: if we insert $t = (1+\varepsilon)\lambda$, $|\varepsilon| < 1$, in (5.22) we see that

$$\begin{aligned} P(X \geq (1+\varepsilon)\lambda) &\leq \left(e^{\varepsilon - (1+\varepsilon)\ln(1+\varepsilon)}\right)^{\lambda} \\ &= \left(e^{-\frac{\varepsilon^2}{2} + \frac{\varepsilon^3}{6} + o(|\varepsilon|^4)}\right)^{\lambda}, \end{aligned}$$

which easily leads to the estimate

$$P(|X - \lambda| \geq \varepsilon\lambda) \leq 2e^{-\frac{\varepsilon^2}{3}\lambda}, \quad 0 < \varepsilon < 1. \tag{5.24}$$

- From (5.21) and (2.30) with $\lambda = np$ we see that the Cramér-Chernoff method delivers the same estimate as in (5.22) had we started with $X \sim \mathsf{Pois}(\lambda)$, the Poisson variable with the same expectation as our original binomial variable; see Example 2.41. This suggests that, at least in the asymptotic regime $n \to +\infty$, the classes $\mathsf{Bin}(\lambda/n; n)$ and $\mathsf{Pois}(\lambda)$ are closely related, a claim substantiated by the Law of Rare Events (Theorem 6.13 below). This deep relationship between binomial and Poisson distributions finds many applications in the theory of random graphs, notably in connection with the Erdös-Rényi model studied in the sequel; see the proof of Proposition 5.18 below for a simple manifestation of this connection and [VDH24] for the general theory.

We now illustrate the of the Chernoff-type bounds for binomial trials developed so far in the art of precisely determining the exact threshold for the emergence of certain "phase transitions" in the most commonly studied class of random graphs.

For each $N \geq 2$ let us define $[N] = \{1, \cdots, N\}$, which we call a *set of vertices*. We represent by $[ij]$ the *unordered* pair derived from $\{i, j\} \in [N] \times [N]$ with $i \neq j$ (so that $[ij] = [ji]$). The union of all such objects is the *set of potential edges*, denoted E. Note that $\sharp(E) = N(N-1)/2$. We now inject a probabilistic ingredient in the construction of graphs starting with E. The most obvious possibility, which we adopt here, is simply to flip a (possibly biased) coin for each potential edge in order to decide whether it effectively occurs as a link between two vertices, with the provision that the flips should comprise independent events. In formal terms, for each $e \in E$ we consider a random variable $X_e \sim \mathsf{Ber}(p)$ so that $\operatorname{supp} P_{X_e} = \{0, 1\}$ with $P(X_e = 1) = 1 - P(X_e = 0) = p$. Now define $\Omega = \{0,1\}^{\binom{N}{2}}$, the cartesian product of $N(N-1)/2$ copies of $\{0, 1\}$, one for each $e \in E$, and $P = \otimes_{e \in E} P_{X_e}$, the product probability on Ω. Note that each element $\omega \in \Omega$ may be viewed as a function $\omega : \Omega \to \{0, 1\}$ and hence defines a graph whose *edge set* is

$$E_\omega = \{e \in E; \omega(e) = 1\}.$$

For this reason, the sample space $(\Omega, 2^\Omega, P)$ is called the *Erdös-Rényi model* for a random graph, usually denoted by $\mathbb{G}(N; p)$.

Proposition 5.16. *Let $\pi_e : \Omega \to \{0, 1\}$ be the canonical projection onto the factor corresponding to e. Then each π_e is identically distributed to X_e (in particular, $\pi_e \sim \mathsf{Ber}(p)$) with $\{\pi_e\}_{e \in E}$ being independent.*

Proof. This is a special case of the general procedure in Remark 2.18. □

By construction of $\mathbb{G}(N;p)$, any event (a subset of Ω) defines a specific collection of graphs. For instance, for each $e \in E$ we may consider

$$\Omega_e = \{\omega \in \Omega; \omega(e) = 1\},$$

the set of all graphs having e as a vertex. The next result confirms that a random graph in the Erdös-Rényi model is obtained by flipping a coin for each potential vertex with the flippings being independent moves.

Proposition 5.17. *$\{\Omega_e\}_{e\in E}$ is a set of independent events.*

Proof. Note that $\omega(e) = \pi_e(\omega)$ so that $\Omega_e = \pi_e^{-1}(1)$ and then apply Proposition 5.16. □

Since $\pi_e = \mathbf{1}_{\Omega_e}$ for each e, the total number of edges in a random graph ω is given by $\mathscr{E}_N(\omega)$, where

$$\mathscr{E}_N = \sum_{e\in E} \pi_e,$$

so that, from Proposition 5.16 and Example 2.40,

$$\mathscr{E}_N \sim \mathsf{Bin}\left(p; \binom{N}{2}\right) \Longrightarrow \mathbb{E}(\mathscr{E}_N) = \binom{N}{2} p.$$

It follows from (5.24) that

$$P\left(|\mathscr{E}_N - \mathbb{E}(\mathscr{E}_N)| < \varepsilon \mathbb{E}(\mathscr{E}_N)\right) \geq 1 - 2e^{-\frac{\varepsilon^2}{3}\mathbb{E}(\mathscr{E}_N)}, \quad 0 < \varepsilon < 1.$$

Thus, as $N \to +\infty$,

$$\frac{\mathscr{E}_N}{\mathbb{E}(\mathscr{E}_N)} \to 1 \quad \text{in probability}$$

so that the total number of edges *asymptotically* approaches its expected value. More generally, if we assume that $p = p_N$ (that is, the biased coin possibly changes with N) then the same conclusion holds as long as

$$\mathbb{E}(\mathscr{E}_N) = \binom{N}{2} p_N \to +\infty, \tag{5.25}$$

with the expectation now being computed with respect to $\mathsf{Bin}\left(p_N; \binom{N}{2}\right)$. At this point, a slightly more ambitious task would be to make sure that, with very high probability, a minimal amount of edges emerges in the regime determined by (5.25). That this is the case indeed follows from the next result, which actually shows that the *asymptotic* emergence of a fixed number of edges in the Erdös-Rényi model is explicitly determined by the limiting value of $\mathbb{E}(\mathscr{E}_N)$.

Proposition 5.18. *Under the conditions above, if $m \in \mathbb{N}$,*

$$\lim_{N\to+\infty} P\left(\mathscr{E}_N > m\right) = \begin{cases} 0 & \mathbb{E}(\mathscr{E}_N) \to 0 \\ 1 - e^{-\lambda}\sum_{k=0}^{m} \frac{\lambda^k}{k!} & \mathbb{E}(\mathscr{E}_N) \to \lambda \in \mathbb{R}_+ \\ 1 & \mathbb{E}(\mathscr{E}_N) \to +\infty \end{cases}$$

Proof. We only prove the convergence in the middle since the remaining items, at least formally, follow from this case. By the Law of Rare Events (Theorem 6.13 below) there exists a Poisson variable $Z \sim \mathsf{Pois}(\lambda)$ such

that $\mathscr{E}_N \xrightarrow{d} Z$ as $n \to +\infty$. Hence,

$$\begin{aligned}\lim_{N\to+\infty} P(\mathscr{E}_N > m) &= P(Z > m)\\ &= 1 - P(Z \leq m)\\ &= 1 - e^{-\lambda}\sum_{k=0}^{m}\frac{\lambda^k}{k!},\end{aligned}$$

as desired. □

We now turn to the incidence properties of $\mathbb{G}(N;p)$. For each vertex $i \in [N]$ consider the random variable

$$d_i = \sum_{j;j\neq i} \pi_{[ij]}.$$

Clearly, for each graph $\omega \in \Omega$, $d_i(\omega)$ measures the number of edges of ω having i as a vertex. We call d_i the *degree*.

Proposition 5.19. *For each i, $d_i \sim \mathsf{Bin}(p; N-1)$. In particular, $d := \mathbb{E}(d_i) = (N-1)p$.*

Proof. Immediate from Proposition 5.16 and Example 2.40. □

Recall that a random graph is *almost regular* if the degree of each vertex equals its expected value with very high probability. The next result identifies the threshold on the degree function beyond which almost regularity holds in the Erdös-Rényi model.

Proposition 5.20. *For any $\varepsilon, \delta \in (0,1)$ there exists $C = C_{\varepsilon,\delta} > 0$ such that $d \geq C \ln N$ implies*

$$P(|d_i - d| \leq \varepsilon d \,\text{for all}\, i) \geq 1 - \delta.$$

Proof. For each $i \in [N]$ we have from Proposition 5.19 and (5.24) that

$$P(|d_i - d| > \varepsilon d) \leq 2e^{-\frac{\varepsilon^2}{3}d},$$

so that

$$P(|d_i - d| > \varepsilon d \,\text{for some}\, i) \leq 2Ne^{-\frac{\varepsilon^2}{3}d},$$

and hence

$$P(|d_i - d| \leq \varepsilon d \,\text{for all}\, i) \geq 1 - 2Ne^{-\frac{\varepsilon^2}{3}d}.$$

Thus, we must find C such that

$$2Ne^{-\frac{\varepsilon^2}{3}C\ln N} \leq \delta,$$

or equivalently,

$$C \geq \frac{3}{\varepsilon^2}h_\delta(N), \quad h_\delta(N) = \frac{\ln(2/\delta) + \ln N}{\ln N}.$$

Now, as N varies $h_\delta(N)$ is uniformly bounded by $M_\delta = \ln(4/\delta)/\ln 2$, so it suffices to take $C \geq 3M_\delta/\varepsilon^2$. □

Note that the almost regularity in Proposition 5.20 implies a sort of homogeneous behavior of the random graph around each of its vertices[16]. In particular, the event that no vertex is isolated occurs with high probability. Now, it turns out that in the regime where $p_N \approx \ln N/N$ with $N \to +\infty$, this event is essentially equiprobable to the event defining connectedness of a random graph, which suggests that $\ln N/N$ should be a sharp threshold for the *asymptotic* occurrence of this topological property. Indeed, arguing along these lines it may be shown that if $p_N = c_N \ln N/N$ and

$$K := \lim_{N\to+\infty} (c_N - 1)\ln N = \lim_{N\to+\infty} (Np_N - \ln N)$$

exists as an extended real number then

$$\lim_{N\to+\infty} P\left(\{\omega \in \mathbb{G}(N;p_N) : \omega \,\text{is connected}\}\right) = e^{-e^{-K}}.$$

In particular,

$$\lim_{N\to+\infty} P\left(\{\omega \in \mathbb{G}(N;p_N) : \omega \,\text{is connected}\}\right) = \begin{cases} 0 & c_N \to c < 1 \\ 1 & c_N \to c > 1 \end{cases}$$

For full discussions on this and similar "phase transition" phenomena exhibiting a sharp threshold in the Erdös-Rényi model, see [JLR11, FK16, VDH24].

6. THE FUNDAMENTAL LIMIT THEOREMS

We now present two asymptotic results that play a central role in the theory. It should be emphasized, however, that in contrast with the concentration estimates of Section 5, whose strength lies precisely in their *non-asymptotic* nature, the applicability of limit theorems becomes reliable only in the asymptotic regime, that is, when the number of random variables in the random sample grows without bound; see Remark 6.7. The proofs outlined below rely on a particular instance of a profound convergence theorem due to Lévy [Wil91, Theorem 18.1]. The version adopted here provides the appropriate converse to Remark 2.27 and may be treated by means of Fourier Analysis.

Theorem 6.1. *(Lévy's convergence) Let $\{Z_j\}_{j=1}^{\infty}$ be a sequence of random variables such that ϕ_{Z_j} converges pointwise to ϕ_Z, where Z is another random variable. Then $Z_j \to Z$ in distribution.*

Proof. By a simple approximation we may assume that ξ in Definition 2.23 is an arbitrary Schwartz function. Hence,

$$\begin{aligned} \mathbb{E}(\xi(Z_j)) &= \int_{-\infty}^{+\infty} \xi(z_j) dP_{Z_j}(z_j) \\ &= \int_{-\infty}^{+\infty} \left(\int_{-\infty}^{+\infty} \widehat{\xi}(u) e^{\mathbf{i} z_j u} du \right) dP_{Z_j}(z_j), \end{aligned}$$

[16]Incidentally, this homogeneity confirms that the Erdös-Rényi random graph fails to reliably model real-world complex networks, where a sizable amount of variability of the incidence pattern of the vertices is observed.

where we used Fourier inversion in order to recover ξ from its Fourier transform $\widehat{\xi}$, which is Schwarz as well and hence uniformly bounded. Using Fubini and dominated convergence we get

$$\begin{aligned}
\mathbb{E}(\xi(Z_j)) &= \int_{-\infty}^{+\infty} \widehat{\xi}(u) \left(\int_{-\infty}^{+\infty} e^{\mathbf{i}z_j u} dP_{Z_j}(z_j) \right) du \\
&= \int_{-\infty}^{+\infty} \widehat{\xi}(u) \phi_{Z_j}(u) du \\
&\stackrel{j\to+\infty}{\longrightarrow} \int_{-\infty}^{+\infty} \widehat{\xi}(u) \phi_{Z}(u) du \\
&\vdots \quad \text{(the same computation as above in reverse order)} \\
&= \mathbb{E}(\xi(Z)),
\end{aligned}$$

and the result follows. □

We may now present the first fundamental limit theorem.

Theorem 6.2. *(Law of large numbers, LLN) If $\{X_j\}_{j\geq 1}$ is a sequence of i.i.d. (that is, independent and identically distributed) real random variables with $\mathbb{E}(X_j) = \mu$ then the sequence of random variables*

$$\overline{X}_n := \frac{1}{n}(X_1 + \cdots + X_n)$$

converges in probability to μ as $n \to +\infty$.

Proof. By Proposition 2.24, it suffices to prove that $\overline{X}_n \to \mu$ in distribution. By Propositions 2.31 and 2.32, if $|u|/n$ is small,

$$\phi_{\overline{X}_n}(u) = \Pi_{j=1}^n \phi_{X_j}(u/n) = \left[1 + \mu \frac{u}{n} \mathbf{i} + o\left(\frac{|u|}{n} \right) \right]^n, \quad n \to +\infty,$$

so that, for *any* $u \in \mathbb{R}$,

$$\lim_{n\to+\infty} \phi_{\overline{X}_n}(u) = e^{u\mu\mathbf{i}} = \phi_\mu(u),$$

where ϕ_μ is the characteristic function of the random variable identically constant to μ. The result now follows from Theorem 6.1. □

Remark 6.3. We append two complements to this result:

(1) If we further assume that $\mathbb{E}(|X_j|^2) < +\infty$ then it also follows from the argument based on (7.33) below, which relies on Chebyshev's inequality and hence provides a quite effective (i.e. *non*-asymptotic) estimate;
(2) For obvious reasons, Theorem 6.2 is usually referred as the *weak* LLN. With some more effort we may show that the convergence holds in a rather strong sense: $\overline{X}_n \stackrel{a.s}{\to} \mu$. This latter result is usually known as *Kolmogorov's LLN* (in [Kre11, Section 1.4] it is shown how it follows from Birkhoff's ergodic theorem discussed in Example 3.12). □

Remark 6.4. The limiting behavior of the Student's t-distribution $\mathfrak{t}_k$ in Definition 4.29 as the number of degrees of freedom k grows indefinitely may be determined if one makes use of the Stirling asymptotics for the gamma function:

$$\Gamma(k) \approx_{k\to+\infty} \sqrt{2\pi} k^{k-\frac{1}{2}} e^{-k}; \tag{6.1}$$

see Remark 6.8 below for a probabilistic proof of this result. Using this, a little computation starting with (4.26) then shows that

$$\lim_{k\to+\infty} \mathfrak{t}_k(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}, \quad x \in \mathbb{R},$$

so that

$$\mathfrak{t}_k \xrightarrow{d} \mathcal{N}(0,1) \tag{6.2}$$

by Scheffé's lemma [Sch47]. We point out that this may also be justified with a simple application of Theorem 6.2. Indeed, by Remark 2.18 we may pick an independent sample $Z, X_1, \cdots, X_k \sim \mathcal{N}(0,1)$, so that $X_j^2 \sim \chi_1^2$ by Proposition 4.25. By LLN, as $k \to +\infty$ we have that $W_k := \sum_{j=1}^k X_j^2$ satisfies

$$\sqrt{\frac{W_k}{k}} \to \sqrt{\mathbb{E}(\chi_1^2)} = 1$$

in probability. Hence, $Z/\sqrt{W_k/k} \to \mathcal{N}(0,1)$ in distribution so that (6.2) may be verified using that $Z/\sqrt{W_k/k}$ is $\mathfrak{t}_k$-distributed by Proposition 4.30. As another instance of this kind of argument, let us check that if $X \sim \mathsf{F}_{k_1,k_2}$ then

$$k_1 X \xrightarrow{d} \chi_{k_1}^2 \quad \text{as } k_2 \to +\infty.$$

Indeed, from Proposition 4.33 we may write

$$X = \frac{W_1/k_1}{W_2/k_2},$$

where $W_1 \perp W_2$ and $W_j \sim \chi_{k_j}^2$, $j = 1, 2$. Thus,

$$k_1 X = \frac{W_1}{W_2/k_2}$$

and since $W_2/k_2 \xrightarrow{p} 1$ the claim follows. □

The next result provides an accurate asymptotic description of the distribution of a rescaled version of $\overline{X}_n$ and highlights the pervasive role of the normal distribution in Probability Theory. The underlying rationale can be sketched as follows. From Theorem 6.2, one suspects the existence of a (possibly monotone) function $\nu : \mathbb{N} \to \mathbb{R}$ with $\nu(n) \to +\infty$ as $n \to \infty$, such that, informally,

$$\overline{X}_n \approx \mu + O(\nu(n)^{-1}).$$

In this case,

$$\nu(n)\left(\overline{X}_n - \mu\right) \approx O(1),$$

which suggests that the rescaled deviation might converge to a finite distribution. To identify ν, assume temporarily that the sample is normally distributed, $X_j \sim \mathcal{N}(\mu, \sigma^2)$. Proposition 4.8 then shows that $\sqrt{n}(\overline{X}_n - \mu) \sim \mathcal{N}(0, \sigma^2)$, indicating that $\nu(n) = \sqrt{n}$. The remarkable feature of the forthcoming result is that this relation, which holds exactly for normal samples, in fact extends asymptotically to arbitrary distributions (with finite variance). In line with Theorem 6.2, one obtains convergence in distribution to a normal law, regardless of the original distribution of the sample.

Theorem 6.5. *(Central Limit Theorem, CLT) Let $\{X_j\}_{j\geq 1}$ be a sequence of i.i.d. real random variables with $\mathbb{E}(X_j) = \mu$ and $\operatorname{var}(X_j) = \sigma^2 > 0$. Then the sequence formed by the standardization of the sample mean,*

$$Z_n := \frac{\sum_{j=1}^n X_j - n\mu}{\sqrt{n}\sigma} = \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}}, \tag{6.3}$$

converges in distribution to a random variable whose pdf is the standard normal distribution $\mathcal{N}(0,1)$. Equivalently, $\sqrt{n}(\overline{X}_n - \mu) \xrightarrow{d} \mathcal{N}(0,\sigma^2)$.

Proof. Define $Y_j = (X_j - \mu)/\sigma$ so that $\mathbb{E}(Y_j) = 0$ and $\mathrm{cov}(Y_j) = 1$. Since $Z_n = \sum_j Y_j/\sqrt{n}$, by Propositions 2.31 and 2.32 we get, for $|u|/\sqrt{n}$ small,

$$\phi_{Z_n}(u) = \Pi_{j=1}^n \phi_{Y_j}\left(\frac{u}{\sqrt{n}}\right) = \left[1 - \frac{u^2}{2n} + o\left(\frac{|u|^2}{n}\right)\right]^n,$$

so that

$$\lim_{n\to+\infty} \phi_{Z_n}(u) = e^{-\frac{1}{2}u^2}, \quad u \in \mathbb{R}. \tag{6.4}$$

By Corollary 4.6 and Proposition 4.8 (1), the right-hand side is the characteristic function of a random variable $Z \sim \mathcal{N}(0,1)$, so we may apply Theorem 6.1 to conclude the proof. □

Remark 6.6. An enlightening discussion of several proofs of Theorem 6.5, including the one above, may be found in [Tao12, Chapter 2]. □

Remark 6.7. A common misconception in the practical use of Theorem 6.5 is to assume that the sample mean $\overline{X}_n$ itself converges in distribution to a normal law. An even more misleading belief, often repeated in applications, is that there exists a fixed threshold for the sample size (frequently taken as $n = 30$) beyond which $\overline{X}_n$ becomes *exactly* normally distributed. Neither of these statements is supported by the theorem. On the contrary, both contradict Theorem 6.2, which establishes that $\overline{X}_n$ converges (almost surely, by Remark 6.3 (2)) to the constant population mean μ. What Theorem 6.5 does guarantee is the *approximation* of $\overline{X}_n$ in distribution by $\mathcal{N}(\mu, \sigma^2/n)$ as $n \to \infty$. Thus, for any $-\infty \le a < b \le +\infty$ we may, for practical purposes, write

$$P(a \le \overline{X}_n \le b) \approx_{n\to+\infty} \frac{\sqrt{n}}{\sqrt{2\pi}\sigma} \int_a^b e^{-\frac{n(x-\mu)^2}{2\sigma^2}} dx, \tag{6.5}$$

where $\approx_{n\to+\infty}$ indicates that the equality holds only asymptotically, in the regime of very large samples. In this notation, the CLT may be expressed as

$$\overline{X}_n \approx_{n\to+\infty} \mathcal{N}(\mu, \sigma^2/n), \tag{6.6}$$

or, equivalently,

$$X^{(n)} := X_1 + \cdots + X_n \approx_{n\to+\infty} \mathcal{N}(n\mu, n\sigma^2). \tag{6.7}$$

It is important, especially in applications, to note that the rate of convergence in these approximations can be made explicit. For instance, under the assumptions of Theorem 6.5, with $\mathbb{E}(X_j) = 0$, $\mathrm{var}(X_j) = \sigma^2$ and with the additional requirement that $\rho := \mathbb{E}(|X - \mu|^3) < +\infty$, the classical Berry–Esseen theorem guarantees that

$$\sup_{x\in\mathbb{R}} \left| F_{\sqrt{n}\overline{X}_n/\sigma}(x) - \Phi(x) \right| \le \frac{C\rho}{\sqrt{n}\sigma^3}, \quad C > 0.$$

Hence, apart from the universal constant C and the dependence on sample size through $1/\sqrt{n}$, the convergence rate is governed by the *shape factor* $\varepsilon := \rho/\sigma^3$, which measures the skewness of the parent distribution. As for the convergence behavior under sample size, although in particular cases the convergence may be faster than $O(n^{-1/2})$, there are distributions for which this worst-case bound is sharp, even when higher-order moments are finite. A simple example is provided by the Rademacher variable discussed in Remark 4.16 [Saz81, Chapter 1]. □

Remark 6.8. We insist that the proof presented above *does* cover the case in which the initial i.i.d. sequence $\{X_j\}$ is *discrete*. In fact, this is how the CLT first appeared, incarnated in the famous De Moivre-Laplace formulas (6.9)-(6.10) below [Fis11]. Let $\{X_j\}_{j=1}^n$ be independent with $X_j \sim \mathsf{Ber}(p)$, the Bernoulli distribution. From Example 2.40, we know that $X^{(n)} = X_1 + \cdots + X_n \sim \mathsf{Bin}(p; n)$, the binomial distribution. Since $\mathbb{E}(X_j) = p$ and $\mathrm{var}(X_j) = p(1-p)$, CLT applies[17] to give

$$Z_n = \sqrt{n}\frac{n^{-1}X^{(n)} - p}{\sqrt{p(1-p)}} \xrightarrow{d} \mathcal{N}(0,1), \quad n \to +\infty, \tag{6.8}$$

or equivalently, if we combine (6.7) and (2.26),

$$\sum_{a \leq k \leq b} \left(\begin{array}{c} n \\ k \end{array} \right) p^k (1-p)^{n-k} \approx_{n\to+\infty} \frac{1}{\sqrt{2\pi np(1-p)}} \int_a^b e^{-\frac{(x-np)^2}{2np(1-p)}} dx, \quad a < b. \tag{6.9}$$

It is not hard to check that this is the same as having

$$\left(\begin{array}{c} n \\ k \end{array} \right) p^k (1-p)^{n-k} \approx_{n\to+\infty} \frac{1}{\sqrt{2\pi np(1-p)}} e^{-\frac{(k-np)^2}{2np(1-p)}} \tag{6.10}$$

uniformly in k satisfying

$$k = np + \sqrt{np(1-p)}O(1), \tag{6.11}$$

which may be proved by using Stirling's formula in (6.13) below and the fact that (6.11) implies that $k/n \to p$ as $n \to +\infty$; see [CA06, Section 7.3]. As yet another application of CLT in the discrete setting, let us assume that $\{Y_j\}_{j=1}^n$ is independent with $Y_j \sim \mathsf{Pois}(1)$, the Poison distribution as in Example 2.41. Thus, $Y^{(n)} = Y_1 + \cdots + Y_n \sim \mathsf{Pois}(n)$, and CLT applies to yield

$$\sqrt{n}(n^{-1}Y^{(n)} - 1) \xrightarrow{d} \mathcal{N}(0,1), \quad n \to +\infty, \tag{6.12}$$

that is,

$$\sum_{a \leq k \leq b} \frac{n^k e^{-n}}{k!} \approx_{n\to+\infty} \frac{1}{\sqrt{2\pi n}} \int_a^b e^{-\frac{(x-n)^2}{2n}} dx,$$

which is the same as having

$$\frac{n^k e^{-n}}{k!} \approx_{n\to+\infty} \frac{1}{\sqrt{2\pi n}} e^{-\frac{(k-n)^2}{2n}}$$

uniformly in k such that

$$k = n + \sqrt{n}O(1).$$

Taking $k = n$ gives

$$n! \approx_{n\to+\infty} \sqrt{2\pi} n^{n+\frac{1}{2}} e^{-n}, \tag{6.13}$$

which is *Stirling's asymptotic formula*. If we take into account that $\Gamma(n) = (n-1)!$, this clearly implies (6.1). □

Remark 6.9. We may directly justify (6.8), the CLT for a Bernoulli population, as follows. Propositions 2.31 and 2.32 applied to (6.8) give

$$\phi_{Z_n}(u) = \phi_{X^{(n)}}(u) e^{-\mathbf{i}\frac{np}{\sqrt{npq}}u},$$

so that (2.27) leads to

$$\begin{aligned} \phi_{Z_n}(u) &= \left(q + pe^{\mathbf{i}\frac{u}{\sqrt{npq}}}\right)^n e^{-\mathbf{i}\frac{np}{\sqrt{npq}}u} \\ &= \left(qe^{-\mathbf{i}\sqrt{\frac{p}{nq}}u} + pe^{\mathbf{i}\sqrt{\frac{q}{np}}u}\right)^n. \end{aligned}$$

[17] See Remark 6.9 below for a direct justification of this step along the lines of the proof of Theorem 6.5.

Expanding the exponential terms in parentheses and performing some cancellations we find that

$$\phi_{Z_n}(u) = \left(1 - \frac{u^2}{2n} + o\left(\frac{u^2}{n}\right)\right)^n \overset{n\to+\infty}{\longrightarrow} e^{-u^2/2}, \tag{6.14}$$

which reproduces (6.4) in this case. We may also obtain a proof of (6.12), the CLT for a Poisson population, along the same lines. Indeed, this time the left-hand side of (6.12) is

$$Z_n = \frac{Y^{(n)}}{\sqrt{n}} - \sqrt{n},$$

so that

$$\begin{aligned}
\phi_{Z_n}(u) &= \phi_{Y^{(n)}}\left(\frac{u}{\sqrt{n}}\right) e^{-\mathbf{i}\sqrt{n}u} \\
&\overset{(2.29)}{=} e^{n\left(e^{\mathbf{i}\frac{u}{\sqrt{n}}}-1\right)} e^{-\mathbf{i}\sqrt{n}u} \\
&= \left(e^{e^{\mathbf{i}\frac{u}{\sqrt{n}}}-1-\mathbf{i}\frac{u}{\sqrt{n}}}\right)^n \\
&= \left(e^{-u^2/2n+o(u^2/n)}\right)^n \\
&= \left(1 - \frac{u^2}{2n} + o\left(\frac{u^2}{n}\right)\right)^n,
\end{aligned}$$

so we may proceed as in (6.14), as desired. □

Example 6.10. If $\{X_j\}$ is i.i.d. with a common cdf F then its *empirical distribution function* is the random variable

$$\mathbb{F}_n(x) = \frac{1}{n}\sum_{j=1}^n \mathbf{1}_{\{X_j \le x\}}, \quad n \in \mathbb{N}, \quad x \in \mathbb{R}.$$

Since, for each fixed x, $\mathbf{1}_{\{X_j\le x\}} \sim \mathsf{Ber}(F(x))$, CLT applies:

$$\sqrt{n}(\mathbb{F}_n(x) - F(x)) \xrightarrow{d} \mathcal{N}(0, F(x)(1-F(x))).$$

In particular, $\mathbb{F}_n(x) \xrightarrow{p} F(x)$ and in fact a.s. convergence takes place; cf. Remark 6.3. With a bit more of work we can prove that the convergence is actually uniform in x: $\|\mathbb{F}_n - F\|_\infty \to 0$ a.s, an assertion known as the *Glivenko-Cantelli theorem* [VdV00]. □

The following immediate consequence of CLT, which uses the notation of Example 4.10, is also worth mentioning here.

Theorem 6.11. *(Multiplicative CLT) If $\{Y_j\}_{j\ge1}$ is a i.i.d. sequence of* positive *random variables satisfying $\mathbb{E}(\ln Y_j) = \mu$ and $\mathrm{var}(\ln Y_j) = \sigma^2$ then*

$$\sqrt[n]{\Pi_{j=1}^n Y_j} \approx_{n\to+\infty} \Lambda(\mu, \sigma^2/n) = \mathcal{LN}(e^{\mu+\frac{\sigma^2}{2n}}, (e^{\frac{\sigma^2}{n}}-1)e^{2\mu+\frac{\sigma^2}{n}}).$$

Example 6.12. We say that a sequence of positive random variables $\{X_j\}_{j=0}^{+\infty}$ satisfy *Gibrat's law of proportionate effect* if there exist random variables $\{Y_j\}_{j=1}^{+\infty}$ such that $Y_j \perp X_{j-1}$ and the corresponding cdfs satisfy

$$F_{X_j}(z) = \int_0^{+\infty} F_{Y_j}(xu^{-1}) dF_{X_{j-1}}(u), \quad j \ge 1.$$

It then follows from (2.13) that $X_j = Y_j X_{j-1}$ and hence

$$X_0^{-1} X_n = \Pi_{j=1}^n Y_j, \quad n \geq 1.$$

Thus, if $\{Y_j\}$ is as in Theorem 6.11 we see that

$$X_0^{-1} X_n \approx_{n \to +\infty} \Lambda(n\mu, n\sigma^2) = \mathcal{LN}(e^{n\mu + n\frac{\sigma^2}{2}}, (e^{n\sigma^2} - 1)e^{2n\mu + n\sigma^2}).$$

Variations of this simple argument go a long way toward explaining the occurrence of lognormal distributions in a large class of natural and social phenomena [AB69]. □

As a final illustration of the usefulness of Theorem 6.1, we now present a result describing the limiting distribution of a sequence of binominal distributions $\mathsf{Bin}(p_n; n)$, with np_n approaching a positive constant, as a Poisson distribution.

Theorem 6.13. *(Law of Rare Events) If $X_n \sim \mathsf{Bin}(p_n; n)$ and $np_n \to \lambda > 0$ as $n \to +\infty$ then there exists $Z \sim \mathsf{Pois}(\lambda)$ such that*

$$X_n \xrightarrow{d} Z.$$

Proof. We compute for any $u \in \mathbb{R}$:

$$\begin{aligned}
\phi_{X_n}(u) &\overset{(2.27)}{=} \left(1 - p_n + p_n e^{\mathbf{i}u}\right)^n \\
&= \left(1 - \frac{\lambda}{n} + \frac{\lambda}{n} e^{\mathbf{i}u} + o(n^{-1})\right)^n \\
&= \left(1 + \frac{\lambda}{n}\left(e^{\mathbf{i}u} - 1\right) + o(n^{-1})\right)^n \\
&\to e^{\lambda\left(e^{\mathbf{i}u} - 1\right)} \\
&\overset{(2.29)}{=} \phi_Z(u).
\end{aligned}$$

Now apply Theorem 6.1. □

7. Estimation

Here we shall use the theory developed so far to provide an introduction to Estimation Theory, an important topic in Statistics with countless applications.

7.1. **Parametric estimation and the mean squared error.**

With the preliminary "large sample" results of Section 6 at hand, we now turn our attention to a *non-asymptotic* problem that appears very often in real world applications, where we only have access to *finitely* many measurements.

Definition 7.1. A *random sample* is a finite family $\{X_j\}_{j=1}^n$ of i.i.d. random variables.

We usually represent a random sample by

$$X_1, \cdots, X_n \sim \psi,$$

or simply by $X_j \sim \psi$, where ψ is the common pdf. Also, in the following we set $\mathbb{E}(X_j) = \mu$ and $\mathrm{var}(X_j) = \sigma^2$, $j = 1, \cdots, n$.

Definition 7.2. A (*parametric*) *statistical model* is a random sample

$$X_1, \cdots, X_n \sim \psi_\theta,$$

where the associated pdf is allowed to depend on the unknown parameter θ running in a given subset $\Theta \subset \mathbb{R}^q$.

Remark 7.3. Implicit in this definition is the existence of an underlying family of probability spaces, say $(\Omega, \mathcal{F}, \{\mathcal{P}_\theta\}_{\theta\in\Theta})$, so that $\{X_j\}$ is i.i.d. with respect to each element in this family. Also, by Proposition 2.15 the joint pdf of $(X_1, \cdots, X_n) : \Omega \to \mathbb{R}^n$ is

$$\mathbf{x} = (x_1, \cdots, x_n) \mapsto \psi_\theta(\mathbf{x}) := \Pi_{j=1}^n \psi(x_j; \theta), \tag{7.1}$$

where $\psi(x_j; \theta) = \psi_\theta(x_j)$.. □

Remark 7.4. Sometimes it is convenient to enlarge the scope of Definition 7.2 above in order to include samples $X_j \sim \psi_\theta$ for which the "identically distributed" assumption no long holds, so that only independence is retained. In the following, whenever we make use of this extended version of a statistical model, we will make explicit reference to this remark. □

Given the statistical model $X_j \sim \psi_\theta$ as above, we will always assume that it is *identifiable* in the sense that the map $\theta \mapsto \psi_\theta$ is injective. In any case, the corresponding *point estimator problem* consists of finding an *estimator*

$$\widehat{\theta} = h(X_1, \cdots, X_n) \tag{7.2}$$

for some *statistic*[18] $h : \mathbb{R}^n \to \Theta \subset \mathbb{R}^q$, which is supposed to yield an "efficient" guess of the true (and unknown) parameter $\theta \in \Theta$. The evaluation $\widehat{\theta}(\mathbf{x})$ of an estimator at a realization $\mathbf{x} \in \mathbb{R}^n$ of a given random sample $X = (X_1, \cdots, X_n)$ is called an *estimate*.

The statistical analysis of point estimators naturally divides into two complementary tasks:

- One must first evaluate the performance of a given estimator against competing alternatives in order to identify the most suitable choice for the problem under consideration. Typically, the candidates are restricted to a predetermined family of estimators, though there is no guarantee that the "best" option within this family will remain optimal in a broader sense. An illustration of this point appears in the performance analysis, under mean squared error, of the variance estimators $\widehat{\sigma}_c^2$, $c > 0$, in Subsection 7.2.
- Once an estimator has been selected, it is important to recognize that the true parameter value will almost never coincide exactly with its point estimate. Accordingly, one needs methods to quantify the variability of the random estimate around the true value, rather than reporting the point estimate alone. This leads naturally to the notion of confidence intervals, discussed in Subsection 7.3.

We thus start here by considering the first issue.

Definition 7.5. The *bias* of an estimator $\widehat{\theta}$ is given by

$$\text{bias}(\widehat{\theta}) = \mathbb{E}(\widehat{\theta} - \theta).$$

[18]To be precise, a statistic is any measurable function $h = h(X_1, \ldots, X_n)$. An estimator for θ, denoted by $\widehat{\theta}$, is a statistic that does not depend on the unknown parameter θ.

An estimator $\widehat{\theta}$ is said to be *unbiased* if $\mathrm{bias}(\widehat{\theta}) = 0$ (equivalently, $\mathbb{E}(\widehat{\theta}) = \theta$ for any θ). Also, the *mean squared error* (mse) of $\widehat{\theta}$ is

$$\mathrm{mse}\,(\widehat{\theta}) = \mathbb{E}(\|\widehat{\theta} - \theta\|^2). \tag{7.3}$$

Remark 7.6. Strictly speaking, the dependence of the invariants above on θ should be emphasized. For instance, the unbiasedness condition actually means that $\mathbb{E}_{\mathcal{P}_\theta}(\widehat{\theta}) = \theta$, where

$$\mathbb{E}_{\mathcal{P}_\theta}(X) = \int_{\mathbb{R}^n} \mathbf{x}\, dP_\theta(\mathbf{x}) \overset{(7.1)}{=} \int_{\mathbb{R}^n} \mathbf{x}\psi_\theta(\mathbf{x})d\mathbf{x},$$

where $P_\theta = X_\sharp \mathcal{P}_\theta$ is the distribution of X coming from $\mathcal{P}_\theta$ (see Remark 7.3). However, in order to keep the notation light, we usually refrain from doing so. Notice also that our notation ignores the dependence of $\widehat{\theta}$ on the size n of the random sample. Whenever emphasizing this is needed, we write $\widehat{\theta} = \widehat{\theta}_n$. □

Proposition 7.7. *(bias-variance trade-off) There holds*

$$\mathrm{mse}\,(\widehat{\theta}) = \mathrm{tr}\,\mathrm{cov}(\widehat{\theta}) + \|\mathrm{bias}(\widehat{\theta})\|^2. \tag{7.4}$$

Proof. If $\theta \in \mathbb{R}$ expand

$$\mathrm{mse}\,(\widehat{\theta}) = \mathbb{E}((\widehat{\theta} - \mathbb{E}(\widehat{\theta}) + \mathbb{E}(\widehat{\theta}) - \theta)^2)$$

and check that the crossed terms cancel, thus yielding (7.5) below. The vector case then follows because

$$\begin{aligned} \mathrm{mse}\,(\widehat{\theta}) &= \sum_j \mathbb{E}((\widehat{\theta}_j - \theta_j)^2) \\ &\overset{(7.5)}{=} \sum_j \mathrm{var}(\widehat{\theta}_j) + \sum_j |\mathrm{bias}(\widehat{\theta}_j)|^2 \\ &= \mathrm{tr}\,\mathrm{cov}(\widehat{\theta}) + \|\mathrm{bias}(\widehat{\theta})\|^2. \end{aligned}$$

□

Convention 7.8. Unless otherwise explicitly stated, we always assume in the sequel that $\theta \in \Theta \subset \mathbb{R}$, the *uni-dimensional* case, so that (7.4) reduces to

$$\mathrm{mse}\,(\widehat{\theta}) = \mathrm{var}(\widehat{\theta}) + |\mathrm{bias}(\widehat{\theta})|^2. \tag{7.5}$$

Here we adopt the viewpoint that the measure of the "performance" of an estimator is encoded in the "smallness" of the corresponding mse. In particular, a bound of the type $\mathrm{mse}(\widehat{\theta}) \leq Cn^{-\alpha}$, $\alpha > 0$, immediately provides an $O(n^{-\alpha/2})$ convergence rate estimate (in the mean) on how $\widehat{\theta}$ approaches θ as $n \to +\infty$. Another kind of convergence of estimators appears in the next definition.

Definition 7.9. We say that an estimator $\widehat{\theta}_n = \widehat{\theta}$ as above is *consistent* if $\widehat{\theta}_n \to \theta$ in probability (with respect to θ).

Proposition 7.10. *If $\widehat{\theta}_n$ is consistent with a uniformly bounded variance then* $\mathrm{bias}(\widehat{\theta}_n) \to 0$ *as* $n \to +\infty$ *(thus,* $\widehat{\theta}_n$ *is* asymptotically unbiased*).*

Proof. By Proposition 2.24, $\widehat{\theta}_n \to \theta$ in distribution so that $\mathbb{E}(\widehat{\theta}_n) \to \mathbb{E}(\theta) = \theta$ and hence $\mathbb{E}(\widehat{\theta}_n)$ is uniformly bounded (for each θ). Combining this with the bound on the variance we see that $\mathbb{E}(|\widehat{\theta}_n|^2) \leq M_\theta$ for some $M_\theta > 0$. Now, for any $\varepsilon > 0$ we have

$$\begin{aligned} |\mathbb{E}(\widehat{\theta}_n - \theta)| &\leq |\mathbb{E}((\widehat{\theta}_n - \theta)\mathbf{1}_{|\widehat{\theta}_n - \theta| < \varepsilon})| + |\mathbb{E}((\widehat{\theta}_n - \theta)\mathbf{1}_{|\widehat{\theta}_n - \theta| \geq \varepsilon})| \\ &< \varepsilon + \mathbb{E}(|\widehat{\theta}_n|\mathbf{1}_{|\widehat{\theta}_n - \theta| \geq \varepsilon}) + \mathbb{E}(|\theta|\mathbf{1}_{|\widehat{\theta}_n - \theta| \geq \varepsilon}) \\ &\leq \varepsilon + \sqrt{\mathbb{E}(|\widehat{\theta}_n|^2)}\sqrt{\mathbb{E}(\mathbf{1}_{|\widehat{\theta}_n - \theta| \geq \varepsilon})} + |\theta|\mathbb{E}(\mathbf{1}_{|\widehat{\theta}_n - \theta| \geq \varepsilon}), \end{aligned}$$

where we used Cauchy-Schwarz in the last step. Thus,

$$|\mathrm{bias}(\widehat{\theta}_n)| < \varepsilon + \sqrt{M_\theta}\sqrt{P(|\widehat{\theta}_n - \theta| \geq \varepsilon)} + |\theta| P(|\widehat{\theta}_n - \theta| \geq \varepsilon)$$

and since $P(|\widehat{\theta}_n - \theta| \geq \varepsilon) \to 0$ the result follows. □

Proposition 7.11. *If* $\mathrm{mse}(\widehat{\theta}_n) \to 0$ *as* $n \to +\infty$ *then* $\widehat{\theta}_n$ *is consistent.*

Proof. By Chebychev's inequality (2.18), for any $\varepsilon > 0$,

$$P_\theta\left(|\widehat{\theta}_n - \theta - \mathrm{bias}(\widehat{\theta}_n)| \geq \varepsilon\right) \leq \frac{\mathrm{var}(\widehat{\theta}_n - \theta)}{\varepsilon^2} \to 0,$$

which means that $\widehat{\theta}_n - \mathrm{bias}(\widehat{\theta}_n) \to \theta$ in probability (with respect to θ). Since $\mathrm{bias}(\widehat{\theta}_n) \to 0$ as well, Theorem 2.25 applies to ensure that $\widehat{\theta}_n \to \theta$ in probability. □

Definition 7.12. An estimator $\widehat{\theta}_n$ as above is *asymptotically normal* with *asymptotic variance* $\sigma_\theta^2 > 0$, $\theta \in \Theta$, if there exists $Z_\theta \sim \mathcal{N}(0, \sigma_\theta^2)$ such that $\sqrt{n}(\widehat{\theta}_n - \theta) \to Z_\theta$ in distribution (with respect to θ).

Proposition 7.13. *If $\widehat{\theta}_n$ is asymptotically normal then it is consistent.*

Proof. If $Z_{\theta,n} = Z_\theta/\sqrt{n} \sim \mathcal{N}(0, \sigma_\theta^2/n)$ then

$$\widehat{\theta}_n - \theta - Z_{\theta,n} = \frac{1}{\sqrt{n}}\left(\sqrt{n}(\widehat{\theta}_n - \theta - Z_{\theta,n})\right) \xrightarrow{p} 0.$$

But Chebychev's inequality gives, for any $\varepsilon > 0$,

$$P(|Z_{\theta,n}| \geq \varepsilon) \leq \frac{\sigma_\theta^2}{n\varepsilon^2} \to 0,$$

that is, $Z_{\theta,n} \xrightarrow{p} 0$ and the result follows by Theorem 2.25. □

Remark 7.14. The true nature of the asymptotic variance σ_θ^2 has not been explored in the previous argument, which makes sense because consistency concerns position rather than dispersion. It is worth noting, however, that asymptotic normality immediately entails that $\mathrm{var}(\widehat{\theta}_n)$ remains uniformly bounded (for each θ). Consequently, Proposition 7.10 ensures that $\widehat{\theta}_n$ is asymptotically unbiased. This conclusion may also be drawn directly from the definition, since $\sqrt{n}\,\mathbb{E}(\widehat{\theta}_n - \theta) \to \mathbb{E}(Z_\theta) = 0$. □

We include here an useful consequence of asymptotic normality.

Proposition 7.15. *(the delta method) If $\widehat{\theta}_n$ is asymptotically normal (as in Definition 7.12) and $g : \Theta \subset \mathbb{R} \to \mathbb{R}$ is a C^1 function whose derivative vanishes nowhere then $\sqrt{n}(g(\widehat{\theta}_n) - g(\theta)) \to Z_{\theta,g}$ in distribution, where $Z_{\theta,g} \sim \mathcal{N}(0, |g'(\theta)|^2\sigma_\theta^2)$.*

Proof. By Taylor,

$$\sqrt{n}\left(g(\widehat{\theta}_n) - g(\theta)\right) = g'(\widetilde{\theta}_n)\sqrt{n}\left(\widehat{\theta}_n - \theta\right),$$

for some $\widetilde{\theta}_n$ between $\widehat{\theta}_n$ and θ. Since $\widehat{\theta}_n \xrightarrow{p} \theta$ by Proposition 7.13, it is not hard to check that $g'(\widetilde{\theta}_n) \xrightarrow{p} g'(\theta)$, so the result follows from Theorem 2.25. □

Example 7.16. For *any* random sample $\{X_j\}$ as above it is immediate to check that the *sample mean*

$$\overline{X}_n := \frac{1}{n}\,(X_1 + \cdots + X_n) \tag{7.6}$$

is an unbiased estimator for the expected value of the underlying distribution. In other words, $\mathbb{E}(\overline{X}_n) = \mu$, where $\mu = \mathbb{E}(X_j)$ is the common expectation. Also, if $\sigma^2 = \mathrm{var}(X_j)$ is the common variance of the sample (the population variance) then it follows from (2.15) that

$$\mathrm{var}(\overline{X}_n) = \frac{1}{n^2}\sum_{j=1}^{n}\mathrm{var}(X_j) = \frac{1}{n^2}\sum_{j=1}^{n}\sigma^2,$$

that is,

$$\mathrm{var}(\overline{X}_n) = \frac{\sigma^2}{n} \tag{7.7}$$

and hence $\mathrm{mse}(\overline{X}_n) = \sigma^2/n$. This is the reason why we call

$$Z_n = \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}}$$

the *standardization* of the sample mean; compare with (6.3). Note that $\mathbb{E}(Z_n) = 0$ and $\mathrm{var}(Z_n) = 1$. Finally, note that $\overline{X}_n$ is consistent (as an estimator for μ) either by LLN or by Proposition 7.11 and that for any g as in Proposition 7.15, CLT applies to ensure that $g(\overline{X}_n)$, as an estimator of $g(\mu)$, satisfies

$$\sqrt{n}\left(g(\overline{X}_n) - g(\mu)\right) \xrightarrow{d} \mathcal{N}(0, |g'(\mu)|^2\sigma^2),$$

hence being asymptotically normal as well. □

Example 7.17. (weighted estimators for the population mean) If $w = (w_1, \cdots, w_n)$ is a *weight vector* (which means that $\sum_j w_j = 1$), then we may consider the corresponding *weighted estimator* for μ given by

$$\overline{X}_n^w = \sum_{j=1}^{n} w_j X_j,$$

which includes (7.6) as a rather special case. One easily verifies that $\mathrm{bias}(\overline{X}_n^w) = 0$ and $\mathrm{mse}(\overline{X}_n^w) = \mathrm{var}(\overline{X}_n^w) = \sigma^2 \sum_j w_j^2$, so the estimator with the least mse in this class is obtained by minimizing $w \mapsto |w|^2$ under the constraint $\sum_j w_j = 1$, which gives $w = (1/n, \cdots, 1/n)$, corresponding to the sample mean $\overline{X}_n$. □

Example 7.18. (Monte Carlo estimator) Let $X : \Omega \to \mathbb{R}^m$ be a random vector with a pdf ψ whose support is contained in the unit cube $[0,1]^m$ and let $f : [0,1]^m \to \mathbb{R}$ be such that $f\psi$ is Lebesgue integrable. If $X_j \sim \psi$ is a i.i.d. sample it is immediate from Example 7.16 that the *Monte Carlo estimator*

$$\widehat{\mu}_{(n)}^f := \frac{1}{n}\sum_{j=1}^n f(X_j),$$

is an unbiased estimator for the unknown parameter

$$\mu^f := \mathbb{E}(f(X_j)) = \int_{[0,1]^m} f(x)\psi(x)dx \tag{7.8}$$

which is consistent because

$$\lim_{n\to+\infty} \widehat{\mu}_{(n)}^f = \mu^f \tag{7.9}$$

in probability by LLN. As usual, this also follows from Proposition 7.11, given that

$$\mathrm{mse}(\widehat{\mu}_{(n)}^f) = \mathrm{var}(\widehat{\mu}_{(n)}^f) = \frac{\sigma_f^2}{n}, \tag{7.10}$$

where σ_f^2 is the (common) variance of $f(X_j)$. Notice that this conveys a $O(n^{-1/2})$ convergence rate for (7.9) which can be made explicit if we apply Chebyshev's inequality (2.18) with $X = \widehat{\mu}_{(n)}^f$, $\sigma = \sigma_f/\sqrt{n}$ and $c = 1/\sqrt{\delta}$, $\delta > 0$, so that

$$P\left(\left|\widehat{\mu}_{(n)}^f - \mu^f\right| \le \frac{1}{\sqrt{\delta}}\frac{\sigma_f}{\sqrt{n}}\right) \ge 1 - \delta.$$

Thus, one needs at least

$$n \approx \frac{1}{\delta}\frac{\sigma_f^2}{\varepsilon^2} \tag{7.11}$$

samples in order to obtain a dispersion of at most ε of the estimator around the expected value μ^f with probability at least $1-\delta$. This may be substantially improved if we appeal to CLT (Theorem 6.5) to obtain

$$\lim_{n\to+\infty} P\left(\left|\widehat{\mu}_{(n)}^f - \mu^f\right| \le \eta\frac{\sigma_f}{\sqrt{n}}\right) = \frac{1}{\sqrt{2\pi}}\int_{-\eta}^{\eta} e^{-x^2/2}dx \overset{(5.2)}{\approx} 1 - e^{-\eta^2/2}, \quad \eta \to 0,$$

which allows us to replace the previous estimate by

$$n \approx 2\ln(1/\delta)\frac{\sigma_f^2}{\varepsilon^2}. \tag{7.12}$$

One should be aware, however, that whereas the estimate (7.11) holds non-asymptotically, as its comes from Chebyshev's inequality, the estimate (7.12) becomes reliable only in the asymptotic regime ($n \to +\infty$); see Remarks 6.3 and 6.7. Regardless of the shape of their dependence on δ, the estimates above share the nice property of not depending on m, so that the "dimensionality curse" is not present here. Of course this is one of the reasons why Monte Carlo methods, based on (7.9), are quite versatile in approximating multiple integrals as those in the right-hand side of (7.8)[19]. Besides its slow $O(n^{-1/2})$ convergence rate, an obvious drawback of this method is its explicit dependence on the standard deviation σ_f, which is at least as hard to compute as μ^f itself. The simplest choices avoiding this latter problem (by explicitly bounding the variance) corresponds

[19]This method should be compared with the usual numerical approach which requires evaluation of $f\psi$ on a ε-net and hence has a complexity that grows like ε^{-m}.

to taking $\{X_j\}$ *uniformly* distributed in $[0,1]^m$ (so that $\psi = \mathbf{1}_{[0,1]^m}$), and $f = \mathbf{1}_B$, the indicator of a Borel set $B \subset [0,1]^m$, so that $\mu^f = \mathrm{vol}_m(B)$, the m-volume of M, and $\sigma_f^2 = \mathrm{vol}_m(B)(1 - \mathrm{vol}_m(B)) \leq 1/4$. Thus, at least the volumes of (well-behaved) Borel subsets can be efficiently calculated if we are able to provide low cost simulations of independent, uniformly distributed random variables on $[0,1]^m$ [RC10]. □

7.2. **Computing the mean squared error of $\widehat{\sigma}_c^2$.** We further illustrate the concepts introduced in Definition 7.5 by determining the "best" estimator for the variance σ^2 in the family

$$\widehat{\sigma}_c^2 = h_c(X_1, \cdots, X_n), \quad c > 0, \tag{7.13}$$

where

$$h_c(X_1, \cdots, X_n) = c\sum_{j=1}^{n}(X_j - \overline{X}_n)^2$$

and

$$\overline{X}_n = \frac{1}{n}(X_1 + \cdots + X_n)$$

is the sample mean; see Example 7.16. Thus, $\theta = \sigma^2 > 0$ and $\Theta = \mathbb{R}_+$. This involves minimizing the corresponding mean squared error $\mathrm{mse}\,(\widehat{\sigma}_c^2)$ viewed as a function of c.

Proposition 7.19. *One has*

$$\mathrm{bias}(\widehat{\sigma}_c^2) = (c(n-1) - 1)\,\sigma^2. \tag{7.14}$$

Proof. Using that

$$X_j - \overline{X}_n = \frac{n-1}{n}X_j - \frac{1}{n}\sum_{k\neq j} X_k \tag{7.15}$$

we first note that $\mathbb{E}(X_j - \overline{X}_n) = 0$ and hence

$$\mathbb{E}(\widehat{\sigma}_c^2) = c\sum_{i=1}^{n}\mathbb{E}((X_j - \overline{X}_n)^2) = c\sum_{j=1}^{n}\mathrm{var}(X_j - \overline{X}_n).$$

From the independence assumption and (2.15) we get, again using (7.15),

$$\mathrm{var}(X_j - \overline{X}_n) = \frac{(n-1)^2}{n^2}\mathrm{var}(X_j) + \frac{1}{n^2}\sum_{k\neq j}\mathrm{var}(X_k) = \frac{n-1}{n}\sigma^2,$$

so that

$$\mathbb{E}(\widehat{\sigma}_c^2) = cn\frac{n-1}{n}\sigma^2 = c(n-1)\sigma^2, \tag{7.16}$$

and the result follows. □

Corollary 7.20. *$\widehat{\sigma}_c^2$ is unbiased only if $c = (n-1)^{-1}$.*

Definition 7.21. The *sample variance* of the sample $\{X_j\}$ as above is

$$S_n^2 := \widehat{\sigma}^2_{(n-1)^{-1}} = \frac{1}{n-1}\sum_{j=1}^{n}(X_j - \overline{X}_n)^2.$$

Also, $S_n = \sqrt{S_n^2}$ is the *sample standard deviation*.

Recall that in general an unbiased estimator $\widehat{\theta}$ satisfies $\mathbb{E}(\widehat{\theta}) = \theta$, so that the target parameter θ is the expected value of the corresponding sample distribution. Intuitively, the unbiasedness property says that on average the estimator hits the right target. This is the main reason why unbiased estimators are often used in applications and we provide below two classical results (Theorems 9.7 and 8.18) ensuring that unbiased estimators minimize their variance (and hence their mse) within certain classes of competing unbiased estimators. However, we point out that the family $\widehat{\sigma}_c^2$ above may be used to illustrate that in general the best variance estimator might not be unbiased (that is, the function $c \mapsto \mathrm{mse}(\widehat{\sigma}_c^2)$ is minimized for some $c \neq (n-1)^{-1}$), which turns out to be a manifestation of the variance-bias trade-off in (7.5). For this we need to compute $\mathrm{var}(\widehat{\sigma}_c^2)$, which we do by assuming in the rest of the calculation that each X_j is normally distributed: $X_j \sim \mathcal{N}(\mu, \sigma^2)$. Let us set

$$U^2 = \sum_{j=1}^{n} \left(\frac{X_j - \mu}{\sigma} \right)^2, \quad V^2 = n \left(\frac{\overline{X}_n - \mu}{\sigma} \right)^2.$$

Proposition 7.22. *If $\{X_j\}$ is independent with $X_j \sim \mathcal{N}(\mu, \sigma^2)$ then $U^2 \sim \chi_n^2$ and $V^2 \sim \chi_1^2$.*

Proof. Note that $\sigma^{-1}(X_j - \mu) \sim \mathcal{N}(0,1)$ by Proposition 4.8, which can also be used to check that $V^2 = W^2$, where

$$W = \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1).$$

The results then follow from Proposition 4.25. □

Proposition 7.23. *If $\{X_j\}$ is independent and $X_j \sim \mathcal{N}(\mu, \sigma^2)$ then $\overline{X}_n$ and*

$$\mathcal{S}^2 := \sum_j (X_j - \overline{X}_n)^2$$

are independent. In particular, $\overline{X}_n \perp \widehat{\sigma}_c^2$, $c > 0$.

Proof. We may assume that $\mu = 0$. Note that

$$\mathcal{S}^2 = \sum_j X_j^2 - Y_1^2,$$

where

$$Y_1 = \sqrt{n}\overline{X}_n = \sum_j \frac{X_j}{\sqrt{n}} \tag{7.17}$$

is normal. By Gramm-Schmidt there exists an orthogonal $n \times n$ matrix, say O, whose first line is the vector $(1/\sqrt{n}, \ldots, 1/\sqrt{n})$ i.e. $O_{1j} = 1/\sqrt{n}$. If $Y = OX$ then Y_1 is indeed given by (7.17) so that

$$\mathcal{S}^2 = \|X\|^2 - Y_1^2 = \|Y\|^2 - Y_1^2 = \sum_{l=2}^{n} Y_l^2.$$

From Corollary 4.13, $Y' = (Y_2, \cdots, Y_n)$ is normally distributed and $\{Y_l\}_{l=2}^n$ is independent as well so its covariance matrix is diagonal. Moreover, using again the independence of $\{X_j\}$, we compute for $l \geq 2$ that

$$\begin{aligned} \operatorname{cov}(Y_1, Y_l) &= \operatorname{cov}\left(\sum_j O_{1j}X_j, \sum_k O_{lk}X_k\right) \\ &= \sigma^2 \sum_j O_{1j}O_{lj} \\ &= 0, \end{aligned}$$

so that $\{Y_j\}_{j=1}^n$ is independent by Proposition 4.11. In particular, $\mathcal{S}^2 = \|Y'\|^2 \perp Y_1/\sqrt{n} = \overline{X}_n$, as desired. □

Proposition 7.24. *If*

$$\Sigma^2 := \sigma^{-2} \sum_j (X_j - \overline{X}_n)^2 \tag{7.18}$$

then

$$\Sigma^2 \sim \chi^2_{n-1}. \tag{7.19}$$

In particular,

$$\operatorname{var}(\Sigma^2) = 2(n-1). \tag{7.20}$$

Proof. Upon multiplication by σ^{-2}, the elementary algebraic identity

$$\sum_{j=1}^n (X_j - \mu)^2 = \sum_{j=1}^n \left(X_j - \overline{X}_n\right)^2 + n\left(\overline{X}_n - \mu\right)^2 \tag{7.21}$$

becomes

$$U^2 = \Sigma^2 + V^2. \tag{7.22}$$

Since $\{\Sigma^2, V^2\}$ is independent (by Proposition 7.23) we have $\phi_{U^2} = \phi_{\Sigma^2}\phi_{V^2}$ so that Corollary 4.22 applies to give

$$\phi_{\Sigma^2}(u) = (1 - 2u\mathbf{i})^{-(n-1)/2},$$

which yields (7.19). Finally, (7.20) follows from Corollary 4.21. □

Remark 7.25. The parameter μ plays no essential role in the validity of the identity (7.21), the only relevant point being that $\overline{X}_n$ is the arithmetic mean of $\{X_j\}_{j=1}^n$. Thus, (7.21) remains true if μ gets replaced by any real number:

$$\sum_{j=1}^n (\mathfrak{y}_j - c)^2 = \sum_{j=1}^n (\mathfrak{y}_j - \overline{\mathfrak{y}}_n)^2 + n(\overline{\mathfrak{y}}_n - c)^2, \quad c \in \mathbb{R}, \tag{7.23}$$

where

$$\overline{\mathfrak{y}}_n = \frac{\mathfrak{y}_1 + \cdots + \mathfrak{y}_n}{n}.$$

In this more general form, this important identity resurfaces at many points below (see Remark 7.31, and Examples 7.39 and 7.40). □

We may record the result of our computation as follows.

Proposition 7.26. *If $\{X_j\}$ is independent with $X_j \sim \mathcal{N}(\mu, \sigma^2)$ then*

$$\operatorname{var}(\widehat{\sigma}_c^2) = 2(n-1)c^2\sigma^4. \tag{7.24}$$

As a consequence,

$$\operatorname{mse}(\widehat{\sigma}_c^2) = \left((n-1)(n+1)c^2 - 2(n-1)c + 1\right)\sigma^4. \tag{7.25}$$

Proof. Combine (7.5), (7.14) and (7.20). □

Corollary 7.27. *Under the conditions above,* $\operatorname{mse}(\widehat{\sigma}_c^2)$ *is minimized for* $c = (n+1)^{-1}$.

Remark 7.28. Since

$$\operatorname{bias}\widehat{\sigma}^2_{(n+1)^{-1}} = -2(n+1)^{-1}\sigma^2,$$

which only vanishes in the asymptotic limit $n \to +\infty$, as already advertised Corollary 7.27 illustrates that an unbiased estimator may fail to be the most efficient one (if the "performance" is measured by mse); a quite similar phenomenon, involving the so-called James-Stein estimator for the mean of certain normal random vectors, appears in Example 8.22. □

Remark 7.29. (Studentized mean) If $Z \sim \mathcal{N}(0,1)$ and $W \sim \chi^2_k$ then Proposition 4.30 says that

$$\frac{Z}{\sqrt{W/k}} \sim \mathrm{t}_k,$$

the Student's t-distribution with $k \geq 1$ degrees of freedom [Stu08b, Fis25]. In the setting of Proposition 7.26 (that is, under sample normality) we may apply this to $Z = \sqrt{n}(\overline{X}_n - \mu)/\sigma$ (which is $\mathcal{N}(0,1)$ by Proposition 4.8) and $W = \widehat{\sigma}^2_{\sigma^{-2}}$ (just use (7.19)) to conclude that

$$T_{n-1} := \frac{\overline{X}_n - \mu}{S_n/\sqrt{n}} \sim \mathrm{t}_{n-1}, \tag{7.26}$$

where S_n is the sample standard deviation (Definition 7.21). Notice that $Z \perp W$ here by Proposition 7.23[20]. We say that T_{n-1} is the *studentized mean* of the normal sample $\{X_j\}$. It plays a key role in finding "small sample" estimates for the population mean of a normally distributed sample with no prior knowledge of the population variance; see Subsection 7.3 below. □

Remark 7.30. If the random sample $\{X_j\}_{j=1}^n$ is not necessarily normal then a somewhat tedious computation gives

$$\operatorname{var}(\widehat{\sigma}_c^2) = \frac{(n-1)^2}{n}c^2\sigma^4\left(\kappa(X_j) - \frac{n-3}{n-1}\right), \tag{7.27}$$

where

$$\kappa(X) = \frac{\mathbb{E}((X - \mathbb{E}(X))^4)}{\operatorname{var}(X)^2}$$

is the *kurtosis* of X, which is finite if we require that $\mathbb{E}(|X|^4) < +\infty$; see [ONe14] for this and many other moment computations. In the normal case we have $\kappa(X_j) = 3$ and (7.27) reduces to (7.24). Of course, this general computation suffices if we are merely interested in the conclusions of Proposition 7.26, but we stress that the elegant argument above based on sample normality has the added bonus of yielding an explicit expression for the sampling distribution of the studentized mean T_{n-1} considered in Remark 7.29. □

[20]This independence between the sample mean $\overline{X}_n$ and the sample standard deviation S_n, which is crucial in precisely determining the shape of the sampling distribution of T_{n-1}, turns out to be a characteristic feature of normal samples; see [Luk42] for a proof which is a clever application of Proposition 2.32 (3).

Remark 7.31. (The geometric way to Student) The calculation leading to Proposition 7.24, in particular the independence between the sample mean $\overline{X}_n$ and the sample variance S_n^2 in Proposition 7.23 which plays a central role in accessing Student's distribution in (7.26) above, may be retrieved by means of the "n-space computations" due to R. Fisher already mentioned in Remarks 4.26 and 4.31[21]. Indeed, the probability density spanned by the independent normal random vector $X = (X_1, \cdots, X_n)$, $X_j \sim \mathcal{N}(\mu, \sigma^2)$, in an infinitesimal region of volume $dx = dx_1 \cdots dx_n$ is

$$\frac{1}{(2\pi)^{n/2}\sigma^n} e^{-\frac{\sum_j (x_j-\mu)^2}{2\sigma^2}} dx = \frac{1}{(2\pi)^{n/2}\sigma^n} e^{-\frac{n(\overline{x}_n-\mu)^2}{2\sigma^2}} e^{-\frac{(n-1)s_n^2}{2\sigma^2}} dx, \tag{7.28}$$

where Remark 7.25 has been used. Following [Fis25] we now observe that $\overline{x}_n$ is proportional to the height of x with respect to the hyperplane H^{n-1} defined by $\Sigma_j x_j = 0$, whereas s_n is proportional to the distance of x to the line l^1 given by $x_1 = \cdots = x_n$, with $\mathscr{X}_n := (\overline{x}_n, \cdots, \overline{x}_n) \in l^1$ realizing this distance. Since H^{n-1} and l^1 are perpendicular to each other, we may use the corresponding "cylindrical" coordinate system to check that dx is proportional to $s_n^{n-2} d\overline{x}_n ds_n d\theta$, where $d\theta$ is the volume element of the unit sphere $\mathbb{S}^{n-2} \subset H^{n-1}$ with center located at $H^{n-1} \cap l^1$, the origin of H^{n-1}. Leading this to (7.28), integrating with respect to θ and using Proposition 2.14 shows that the infinitesimal joint probability density of the random vector $(\overline{X}_n, S_n^2)$ is

$$\psi_{(\overline{X}_n, S_n^2)}(\overline{x}_n, s_n^2) d\overline{x}_n ds_n^2 \approx e^{-\frac{n(\overline{x}_n-\mu)^2}{2\sigma^2}} d\overline{x}_n \times e^{-\frac{(n-1)s_n^2}{2\sigma^2}} (s_n^2)^{\frac{n-3}{2}} ds_n^2, \tag{7.29}$$

where $\approx$ here means that we are neglecting certain normalizing constants which will take care of themselves. Incidentally, this geometric argument makes it obvious the connection to the previous computational proof of Proposition 7.23: the orthogonal map $Y = OX$ used there carries H^{n-1} onto the coordinate hyperplane $Y_1 = 0$, which has the net effect of reducing the size of the sample data by one, thus allowing for an induction argument based on Corollary 4.13. Moreover, it has a number of consequences which we now describe.

- Clearly, (7.29) implies that $\{\overline{X}_n, S_n^2\}$ is independent.
- Also, it follows from (7.29) that

$$\psi_{S_n^2}(s_n^2) ds_n^2 \approx e^{-\frac{(n-1)s_n^2}{2\sigma^2}} (s_n^2)^{\frac{n-3}{2}} ds_n^2,$$

  so if we combine this with (7.18) we see that

$$\psi_{\Sigma^2}(s_\sigma^2) ds_\sigma^2 \approx e^{-s_\sigma^2/2} (s_\sigma^2)^{\frac{n-3}{2}} ds_\sigma^2, \quad s_\sigma^2 = \frac{(n-1)s_n^2}{\sigma^2}, \tag{7.30}$$

  from which we easily deduce (7.19); compare with the computation in Remark 4.26.
- Finally, the geometric argument also provides another way of explicitly computing the probability density of the studentized mean in (7.26). Indeed, (7.29) implies that

$$\psi_{\overline{X}_n}(\overline{x}_n) d\overline{x}_n \approx e^{-\frac{n(\overline{x}_n-\mu)^2}{2\sigma^2}} d\overline{x}_n \approx (s_n^2)^{1/2} e^{-\frac{s_n^2 t_{n-1}^2}{2\sigma^2}} dt_{n-1}, \tag{7.31}$$

  where for s_n^2 fixed we set

$$t_{n-1} = \frac{\overline{x}_n - \mu}{s_n/\sqrt{n}},$$

  so that

$$d\overline{x}_n = \frac{s_n}{\sqrt{n}} dt_{n-1}$$

  by the previously established independence of $\{\overline{X}_n, S_n\}$. Replacing (7.31) back into (7.29) yields an explicit expression for the joint density $\psi_{(T_{n-1}, S_n^2)}(t_{n-1}, s_n^2) dt_{n-1} ds_n^2$, so that integration with respect to

[21] Recall that we always represent a realization of a random variable, say $\overline{X}_n$, by the corresponding lower-case symbol, in this case $\overline{x}_n$.

s_n^2 gives

$$\begin{aligned}\psi_{T_{n-1}}(t_{n-1})dt_{n-1} &\approx \left(\int_0^{+\infty}(s_n^2)^{\frac{n-2}{2}}e^{-\frac{s_n^2(n-1+t_{n-1}^2)}{2\sigma^2}}ds_n^2\right)dt_{n-1}\\ &\approx \left(n-1+t_{n-1}^2\right)^{-n/2}dt_{n-1}\\ &\approx \mathfrak{t}_{n-1}(t_{n-1})dt_{n-1},\end{aligned}$$

as desired. □

7.3. **Confidence intervals.** If $\widehat{\theta}$ is an *unbiased* estimator for the parameter θ whose standard deviation $\sigma_{\widehat{\theta}}$ is known then Chebyshev's inequality (2.18) gives

$$P(|\widehat{\theta}-\theta|\leq c\sigma_{\widehat{\theta}})\geq 1-c^{-2},\quad c>1,$$

which translates into a "confidence interval" estimate for the unknown parameter:

$$\theta\in[\widehat{\theta}\mp c\sigma_{\widehat{\theta}}]\text{ with prob. at least }1-c^{-2}, \tag{7.32}$$

where here and in the following we denote the interval $[a-b,a+b]$ simply by $[a\mp b]$ whenever convenient. In particular, this applies to $\widehat{\theta}=\overline{X}_n$, the mean sample estimator, which is an unbiased estimator for the population mean μ (by Example 7.16). Since $\sigma_{\overline{X}_n}=\sigma/\sqrt{n}$, if we take $c^{-2}\approx\delta$ and $c\sigma/\sqrt{n}\approx\epsilon$, where δ and ϵ are arbitrarily small positive real numbers, then we see that

$$n>\frac{\sigma^2}{\epsilon^2\delta}\Longrightarrow\mu\in[\overline{X}_n\mp\epsilon]\text{ with prob. }\approx 1-\delta. \tag{7.33}$$

Notice that this retrieves the "convergence in probability" version of LLN (under the additional assumption that σ is finite); see Remarks 6.3 and 6.7. We may also turn this into a confidence interval estimate as in (7.32):

$$\mu\in\left[\overline{X}_n\mp\frac{1}{\sqrt{\delta}}\frac{\sigma}{\sqrt{n}}\right]\text{ with prob. }\approx 1-\delta, \tag{7.34}$$

If we further require that the sample is normally distributed ($X_j\sim\mathcal{N}(\mu,\sigma^2)$) then we can employ

$$Z_n:=\frac{\overline{X}_n-\mu}{\sigma/\sqrt{n}}\sim\mathcal{N}(0,1), \tag{7.35}$$

where we used Proposition 4.8, to find a "small sample" confidence interval for the unknown expected value:

$$\mu\in\left[\overline{X}_n\mp z_{1-\delta/2}\frac{\sigma}{\sqrt{n}}\right]\text{ with prob. }\approx 1-\delta, \tag{7.36}$$

where, for a given $0<\beta<1$, the *normal quantile* $z_\beta>0$ is determined by $P(Z\leq z_\beta)=\beta$, where $Z\sim\mathcal{N}(0,1)$[22]. Notice that if we (more realistically!) relax the normality assumption then (7.36) becomes a "large sample" estimate since (7.35) holds asymptotically as $n\to+\infty$ due to CLT. Upon comparison with (7.36) we see that this amounts to replacing $1/\sqrt{\delta}$ by $z_{1-\delta/2}$ in the estimate for the dispersion around the sample mean[23]. In any case, the estimates (7.34) and (7.36) remain ineffective as long as σ is unknown, in which case (and coming back to a normal sample $X_j\sim\mathcal{N}(\mu,\sigma^2)$), Remark 7.29 suggests replacing (7.35) by (7.26) so as to obtain the "small sample" estimate

$$\mu\in\left[\overline{X}_n\mp\mathfrak{t}_{n-1,1-\delta/2}\frac{S_n}{\sqrt{n}}\right]\text{ with prob. }\approx 1-\delta, \tag{7.37}$$

[22]In order to unify the notation for quantiles, we will always impose that their subscripts denote cumulative probabilities. For instance, if $\delta=0.05$ then $z_{1-\delta/2}\approx 1.96$, which shows that roughly two standard deviations around the normal mean suffice to ensure the customary 95% confidence statement.

[23]Thus, if $\delta=0.05$ we are replacing $1/\sqrt{0.05}\approx 4.47$ by 1.96, which shrinks the dispersion by a factor of $4.47/1.96\approx 2.28$ while still retaining the same confidence level. But recall that this reduction only becomes reliable in the asymptotic regime (Remarks 6.3 and 6.7).

where $P(T_{n-1} \geq \mathfrak{t}_{n-1,\beta}) = 1-\beta$ is the "tail" probability associated to the t-distribution $\mathfrak{t}_{n-1}$ which defines the corresponding quantile $\mathfrak{t}_{n-1,\beta}$. The obvious advantage of (7.37) over (7.36) is that no previous knowledge of σ is used. Finally, note that in Remark 8.27 below it is shown that $\widehat{\sigma}^2_{n^{-1}} \xrightarrow{p} \sigma^2$ under this normality assumption. Since

$$\widehat{\sigma}^2_{(n-1)^{-1}} = \frac{n}{n-1}\widehat{\sigma}^2_{n^{-1}},$$

we see from Theorem 2.25 that

$$S_n = \sqrt{\widehat{\sigma}^2_{(n-1)^{-1}}} \xrightarrow{p} \sigma. \tag{7.38}$$

It then follows from

$$T_{n-1} = \frac{\sigma}{S_n} Z_n$$

and Theorem 2.25 that (7.36) and (7.37) provide essentially the same information in this asymptotic regime (in the sense that $T_{n-1} - Z_n \xrightarrow{p} 0$). We stress, however, the usefulness of (7.37) when dealing with small samples, which attests in favor of Student's fundamental contribution coming from Remark 7.29.

Remark 7.32. The convergence in (7.38) holds more generally (that is, with no normality assumption) if we assume that the random sample satisfies $\mathbb{E}(|X_j|^4) < +\infty$. Indeed, we already know from Corollary 7.20 that $\mathrm{bias}(S_n^2) = 0$. Also, from (7.27) with $c = (n-1)^{-1}$ we see that $\mathrm{var}(S_n^2) \to 0$. Thus, $\mathrm{mse}(S_n^2) \to 0$ and Proposition 7.11 applies to ensure that $S_n^2 \xrightarrow{p} \sigma^2$, from which (7.38) follows. □

Remark 7.33. It is important to have in mind that if we evaluate the sample mean $\overline{X}_n$ through a measurement so as to obtain a numerical value, say μ_n, then the corresponding realization of (7.36), namely,

$$\mu \in \left[\mu_n \mp z_{1-\delta/2}\frac{\sigma}{\sqrt{n}}\right] \text{ with prob. } \approx 1-\delta, \tag{7.39}$$

is completely devoid of sense. Indeed, since any trace of randomness has been removed from the interval in (7.39) (it has now become deterministic!) then either μ definitely belongs to it or not, with probability 0 or 1. Thus, the proper way to interpret (7.36) is to regard the corresponding interval as stochastic in nature and to adopt the "frequentist" perspective according to which the "relative frequency" that

$$\mu \in \left[\mu_n - z_{1-\delta/2}\frac{\sigma}{\sqrt{n}}, \mu_n + z_{1-\delta/2}\frac{\sigma}{\sqrt{n}}\right]$$

approaches $1-\delta$ as the number of successive realizations μ_n of $\overline{X}_n$ becomes larger and larger. Needless to say, similar remarks hold for (7.37) under realizations of $\overline{X}_n$ and S_n. □

Remark 7.34. Strictly speaking, Z_n and T_{n-1} do not qualify as estimators as they are statistics which depend on the underlying parameters. Instead, they are referred to as *pivotal quantities*, a terminology incorporating the appreciated property that their distributions do not depend on these parameters. □

Example 7.35. (Sampling from a Bernoulli population) If $X_j \sim \mathsf{Ber}(p)$ as in Remark 6.8 then the computation leading to (6.8) also gives

$$\sqrt{n}\frac{\overline{X}_n - p}{\sqrt{p(1-p)}} \xrightarrow{d} \mathcal{N}(0,1), \tag{7.40}$$

where $\overline{X}_n = X^{(n)}/n$ is the corresponding sample mean. This translates into the "large sample" estimate

$$p \in \left[\overline{X}_n \mp z_{1-\delta/2}\frac{\sqrt{p(1-p)}}{\sqrt{n}}\right] \text{ with prob. } \approx 1-\delta,$$

which displays the usual drawback, namely, the size of the confidence interval depends on p, the parameter we want to estimate. The conservative way of remedying this is to implement the rather crude estimate $p(1-p) \leq$

$1/4$ to eliminate the dependence on p. Another, certainly more effective, route consists of combining Theorem 2.25 and LLN to replace (7.40) by

$$\sqrt{n}\frac{\overline{X}_n - p}{\sqrt{\overline{X}_n(1-\overline{X}_n)}} \xrightarrow{d} \mathcal{N}(0,1),$$

which gives

$$p \in \left[\overline{X}_n \mp z_{1-\delta/2}\frac{\sqrt{\overline{X}_n(1-\overline{X}_n)}}{\sqrt{n}}\right] \text{ with prob. } \approx 1-\delta.$$

Since $\sqrt{\overline{X}_n(1-\overline{X}_n)} \leq \frac{1}{2}$, an increase in sample size by a factor of $\gamma > 0$ yields only a $\gamma^{-1/2}$-order reduction in the dispersion around $\overline{X}_n$, the center of the confidence interval. This simple observation has many applications, as Bernoulli trials provide a convenient model for a wide range of binary random experiments such as coin toss, two-candidate election polls, and male–female birth ratios, among others. □

Example 7.36. (The difference of means of two normal samples) Let $\{X_j\}_{j=1}^m$ and $\{Y_k\}_{k=1}^n$ be normally distributed random samples, say with $X_j \sim \mathcal{N}(\mu_X, \sigma_X^2)$ and $Y_k \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$, which we assume to be independent to one another. In general, we also assume that the true parameter $\theta = (\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2)$ is unknown. In order to estimate the difference of means $\mu := \mu_X - \mu_Y$ we first note that

$$\mathbb{E}\left(\overline{X}_m - \overline{Y}_n\right) = \mu, \quad \text{var}\left(\overline{X}_m - \overline{Y}_n\right) = \frac{\sigma_X^2}{m} + \frac{\sigma_Y^2}{n},$$

and hence

$$Z_{X,Y} := \frac{\overline{D}_{mn} - \mu}{\sqrt{\frac{\sigma_X^2}{m} + \frac{\sigma_Y^2}{n}}} \sim \mathcal{N}(0,1),$$

where $\overline{D}_{mn} = \overline{X}_m - \overline{Y}_n$ is an unbiased estimator for μ. Thus, if both σ_X and σ_Y are known we get the "small sample" confidence interval

$$\mu \in \left[\overline{D}_{mn} \mp z_{1-\delta/2}\sqrt{\frac{\sigma_X^2}{m} + \frac{\sigma_Y^2}{n}}\right] \text{ with prob. } \approx 1-\delta.$$

In the general case we may proceed as follows. If

$$S_X^2 := \frac{1}{m-1}\sum_j \left(X_j - \overline{X}_m\right)^2, \quad S_Y^2 := \frac{1}{n-1}\sum_k \left(Y_k - \overline{Y}_n\right)^2,$$

are the unbiased estimators for σ_X^2 and σ_Y^2 respectively (by Corollary 7.20) then Corollary 7.24 implies that

$$\Sigma_X^2 := (m-1)\frac{S_X^2}{\sigma_X^2} \sim \chi_{m-1}^2 \text{ and } \Sigma_Y^2 := (n-1)\frac{S_Y^2}{\sigma_Y^2} \sim \chi_{n-1}^2 \tag{7.41}$$

are independent and hence, by Corollary 4.24,

$$W_{X,Y} := \Sigma_X^2 + \Sigma_Y^2 \sim \chi_{m+n-2}^2.$$

If we set

$$S_{X,Y}^2(\eta) = \frac{(m-1)S_X^2 + (n-1)\eta S_Y^2}{m+n-2}, \quad \eta := \frac{\sigma_X^2}{\sigma_Y^2},$$

then it follows from Proposition 4.30 that

$$\begin{aligned} T_{X,Y} &:= \frac{Z_{X,Y}}{\sqrt{W_{X,Y}/(m+n-2)}} \\ &= \frac{\overline{D}_{mn}-\mu}{\sqrt{c_{mn}(\eta)S^2_{X,Y}(\eta)}}, \quad c_{mn}(\eta)=\frac{1}{m}+\frac{1}{n\eta}, \end{aligned}$$

satisfies

$$T_{X,Y} \sim \mathfrak{t}_{m+n-2}, \tag{7.42}$$

thus being a pivotal quantity with respect to the parameter η; see Remark 7.34. Hence, if the population variances, though unknown, are such that their ratio η is known, then there holds

$$\mu \in \left[\overline{D}_{mn} \mp t_{m+n-2,1-\delta/2}\sqrt{c_{mn}(\eta)S^2_{XY}(\eta)}\right] \text{ with prob. } \approx 1-\delta, \tag{7.43}$$

a "small sample" confidence interval for μ quite similar in spirit to (7.37), which handles the case of a single normal sample. In particular, if the population variances are assumed to be equal then (7.43) holds with $\eta = 1$, in which case one has

$$T_{X,Y} = \frac{\overline{D}_{mn}-\mu}{\sqrt{c_{mn}(1)S^2_{XY}}}, \tag{7.44}$$

where

$$S^2_{XY} = \frac{(m-1)S^2_X + (m-1)S^2_Y}{m+n-2},$$

is the *pooled variance*, a weighted sum of the sample variances. Otherwise, (7.43) has no practical usefulness as it provides a dispersion around $\overline{D}_{mn}$ depending on the unknown parameter η. Proceeding as before, we are thus led to consider the *Behrens-Fisher-Welch statistics*

$$\mathcal{Z}_{X,Y} = \frac{\overline{D}_{mn}-\mu}{\sqrt{\frac{S^2_X}{m}+\frac{S^2_Y}{n}}},$$

the obvious counterpart of $Z_{X,Y}$. Unfortunately, no simple, closed expression for the pdf of $\mathcal{Z}_{X,Y}$ does seem to exist. To appreciate the difficulties involved, note that $\mathcal{Z}_{X,Y} = Z_{X,Y}/\sqrt{\mathcal{W}_{X,Y}}$, where if

$$\beta_X = \frac{1}{m-1}\left(1+\frac{m}{n}\eta^{-1}\right)^{-1}, \quad \beta_Y = \frac{1}{n-1}\left(1+\frac{n}{m}\eta\right)^{-1} \tag{7.45}$$

then

$$\mathcal{W}_{X,Y} = \beta_X\Sigma^2_X + \beta_Y\Sigma^2_Y,$$

so that Corollary 4.23 and (7.41) may be used to ensure that

$$\mathcal{W}_{X,Y} \sim \mathsf{Gamma}\left(\frac{1}{2\beta_X}, \frac{m-1}{2}\right) + \mathsf{Gamma}\left(\frac{1}{2\beta_Y}, \frac{n-1}{2}\right), \tag{7.46}$$

where $(m-1)\beta_X + (n-1)\beta_Y = 1$ by (7.45). Since $Z_{X,Y} \perp \mathcal{W}_{X,Y}$ and $Z_{X,Y} \sim \mathcal{N}(0,1)$, it follows from Remark 2.17 that computing $\psi_{\mathcal{Z}_{X,Y}}$ essentially reduces to figuring out $\psi_{\mathcal{W}_{X,Y}}$, the pdf of a sum of independent Gamma-distributed random variables whose inverse scale parameters are distinct except when

$$\eta = \frac{m(m-1)}{n(n-1)}.$$

Thus, in most cases this sum is not Gamma-distributed; cf. Corollary 4.24. In fact, it is known that the exact expression for $\psi_{\mathcal{W}_{X,Y}}$ and, more generally, for the pdf of a linear combination of chi-square distributions, requires an infinite series expansion involving certain transcendental functions [RP61, Mos85, HCP22]. This, in

turn, makes it necessary to rely on suitable approximations for $\psi_{\mathcal{Z}_{X,Y}}$, a situation that has prompted extensive research into the computational accuracy and efficiency of such methods[24]. To complicate matters further, as is apparent from (7.45) and (7.46), $\psi_{\mathcal{Z}_{X,Y}}$ explicitly depends on the "nuisance" parameter η, implying that $\mathcal{Z}_{X,Y}$ cannot serve as a pivotal quantity; see Remark 7.34. Thus, despite significant progress in its practical treatment [KC98], the broader challenge of obtaining the most efficient estimate of μ when η is unknown, commonly known as the *Behrens–Fisher problem* [Wel96, Section 3.8], remains largely unresolved. □

Remark 7.37. (F-test for the equality of variances) The reliability of the assumption on the equality of the population variances which led to (7.44) may be statistically justified (or not!) by running an F-test; for more on this see Section 12 below. We start by noticing that under the corresponding *null hypothesis*

$$H_0: \quad \sigma_X^2 = \sigma_Y^2,$$

(7.41) and Proposition 4.33 ensure that the appropriate test statistics

$$U := \frac{S_X^2}{S_Y^2} = \frac{\Sigma_X^2/(m-1)}{\Sigma_Y^2/(n-1)} \sim \mathsf{F}_{m-1,n-1}. \tag{7.47}$$

Now, a very rough analysis of the departure from H_0 goes as follows. If this hypothesis is not satisfied (so that $\sigma_X^2 = \eta\sigma_Y^2$, $\eta \neq 1$) then (7.47) becomes

$$U = \frac{\frac{\eta}{m-1}\Sigma_X^2}{\frac{1}{n-1}\Sigma_Y^2},$$

and we see from Corollary 4.23 and Proposition 4.20 that:

- The denominator satisfies
$$\frac{1}{n-1}\Sigma_Y^2 \sim \mathsf{Gamma}\left(\frac{n-1}{2}, \frac{n-1}{2}\right),$$
so its distribution remains the same regardless of the validity of H_0;
- On the other hand, the numerator satisfies
$$\frac{\eta}{m-1}\Sigma_X^2 \sim \mathsf{Gamma}\left(\frac{m-1}{2\eta}, \frac{m-1}{2}\right),$$
a Gamma distribution whose scale factor is proportional to
$$\eta = \mathbb{E}\left(\frac{\eta}{m-1}\Sigma_X^2\right),$$
the parameter quantifying the departure from H_0.

Thus, at least on average, very small or very large realizations for U (substantially departing from $u = 1$) provide statistical evidence for *rejecting* H_0. Precisely, if we fix $0 < \beta < 1$ and consider the corresponding F-*quantile* $\mathsf{f}_{k_1,k_2,\beta}$ determined by

$$F_{\mathsf{F}_{k_1,k_2}}(\mathsf{f}_{k_1,k_2,\beta}) = \beta, \tag{7.48}$$

where $F_{\mathsf{F}_{k_1,k_2}}$ is the cdf of F_{k_1,k_2}, then H_0 should be rejected "at significance level α" if the realization u of the statistics in (7.47) satisfies

$$u \in \left(0, \mathsf{f}_{m-1,n-1,\alpha/2}\right] \bigcup \left[\mathsf{f}_{m-1,n-1,1-\alpha/2}, +\infty\right). \tag{7.49}$$

A more convincing justification for this rather informal argument may be found in Section 12 below. □

[24] See [Bau13] for a critical review of recent developments on this topic.

Remark 7.38. (Reciprocity of the F-quantiles) Regarding the F-quantiles defined in (7.48), let us take $X \sim \mathsf{F}_{m,n}$ so that $X^{-1} \sim \mathsf{F}_{n,m}$ by Corollary 4.34. For any $\alpha > 0$ we then have

$$\frac{\alpha}{2} = P\left(X \leq \mathsf{f}_{m,n,\alpha/2}\right) = P\left(X^{-1} \geq \frac{1}{\mathsf{f}_{m,n,\alpha/2}}\right),$$

so that

$$P\left(X^{-1} \leq \frac{1}{\mathsf{f}_{m,n,\alpha/2}}\right) = 1 - \frac{\alpha}{2} = P\left(X^{-1} \leq \mathsf{f}_{n,m,1-\alpha/2}\right),$$

from which the identity

$$\mathsf{f}_{m,n,\alpha/2}\mathsf{f}_{n,m,1-\alpha/2} = 1$$

follows. □

Example 7.39. (The sampling distribution of the correlation coefficient) Let us retain the notation of Example 7.36, but this times with $m = n$ and assuming that the *independent* random sample

$$(X, Y) := \{(X_1, Y_1), \cdots, (X_m, Y_m)\}$$

has been drawn from a bi-variate normal population whose marginals are not necessarily independent. Thus,

$$(X_j, Y_j) \sim \mathcal{N}\left(\begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix}, \begin{pmatrix} \sigma_X^2 & \sigma_{XY} \\ \sigma_{XY} & \sigma_Y^2 \end{pmatrix}\right), \quad j = 1, \cdots, n, \tag{7.50}$$

where $\sigma_{XY} = \mathrm{cov}(X_j, Y_j)$ is the *population covariance* [25], so by (4.21) the joint distribution of (X, Y) is

$$\begin{aligned} \psi_{(X,Y)}(x,y)dxdy &= \frac{1}{(2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2})^m} \times \\ &\times e^{-\frac{1}{2(1-\rho^2)}\sum_j\left(\frac{(x_j-\mu_X)^2}{\sigma_X^2} - \frac{2\rho(x_j-\mu_X)(y_j-\mu_Y)}{\sigma_X\sigma_Y} + \frac{(y_j-\mu_Y)^2}{\sigma_Y^2}\right)}dxdy, \end{aligned} \tag{7.51}$$

where

$$\rho = \frac{\sigma_{XY}}{\sigma_X\sigma_Y}$$

is the *correlation coefficient*, a population parameter whose estimation is a central theme in Multivariate Statistical Analysis [And03][26]. It turns out that Fisher's geometric approach in Remark 7.31 can be successfully employed to this end [Fis15]. Indeed, using Remark 7.25 we have the identities

$$\sum_j (x_j - \mu_X)^2 = m\left((\overline{x}_m - \mu_X)^2 + \widehat{\sigma}^2_{m^{-1}}(x)\right)$$

and

$$\sum_j (y_j - \mu_y)^2 = m\left((\overline{y}_m - \mu_X)^2 + \widehat{\sigma}^2_{m^{-1}}(y)\right),$$

where $\widehat{\sigma}^2_{m^{-1}}(x)$ and $\widehat{\sigma}^2_{m^{-1}}(y)$ are the realizations of the variance estimators appearing in (7.13) with $c = m^{-1}$. Also, we will need their polarized version

$$\sum_j (x_j - \mu_X)(y_j - \mu_Y) = m\left((\overline{x}_m - \mu_X)(\overline{y}_m - \mu_Y) + \widehat{\sigma}^2_{m^{-1}}(x,y)\right),$$

where

$$\widehat{\sigma}^2_{m^{-1}}(X,Y) := \frac{1}{m}\sum_j \left(X_j - \overline{X}_m\right)\left(Y_j - \overline{Y}_m\right).$$

[25] By Proposition 4.11, $\{X_j, Y_j\}$ is independent if and only if $\sigma_{XY} = 0$.

[26] As usual, we assume that $|\rho| < 1$, thus avoiding the degenerate cases $\rho = \pm 1$.

Leading these identities to (7.51) we get

$$\begin{aligned}\psi_{(X,Y)}dxdy &= \frac{1}{(2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2})^m}\times \\ &\times e^{-\frac{m}{2(1-\rho^2)}\left(\frac{(\overline{x}_m-\mu_X)^2}{\sigma_X^2}-\frac{2\rho(\overline{x}_m-\mu_X)(\overline{y}_m-\mu_Y)}{\sigma_X\sigma_Y}+\frac{(\overline{y}_m-\mu_Y)^2}{\sigma_Y^2}\right)}\times \\ &e^{-\frac{m}{2(1-\rho^2)}\left(\frac{\widehat{\sigma}^2_{m^{-1}}(x)}{\sigma_X^2}-\frac{2\rho\widehat{\rho}\widehat{\sigma}_{m^{-1}}(x)\widehat{\sigma}_{m^{-1}}(y)}{\sigma_X\sigma_Y}+\frac{\widehat{\sigma}^2_{m^{-1}}(y)}{\sigma_Y^2}\right)}dxdy,\end{aligned}$$

where

$$\widehat{\rho}=\frac{\widehat{\sigma}^2_{m^{-1}}(X,Y)}{\widehat{\sigma}_{m^{-1}}(X)\widehat{\sigma}_{m^{-1}}(Y)}$$

is the *sample correlation coefficient*, the natural estimator for ρ; see Example 8.6 for a justification of this latter claim. If we could view realizations of the samples X and Y as *independent* elements of $\mathbb{R}^m_X$ and $\mathbb{R}^m_Y$, respectively, then the geometric reasoning in Remark 7.31 would ensure that

$$dxdy\approx\widehat{\sigma}^{m-2}_{m^{-1}}(x)d\widehat{\sigma}_{m^{-1}}(x)d\overline{x}_m d\theta_X\times\widehat{\sigma}^{m-2}_{m^{-1}}(y)d\widehat{\sigma}_{m^{-1}}(y)d\overline{y}_m d\theta_Y, \tag{7.52}$$

from which we would compute $\psi_{(\widehat{\sigma}_{m^{-1}}(X),\widehat{\sigma}_{m^{-1}}(Y),\widehat{\rho})}$ after integrating $\psi_{(X,Y)}dxdy$ above against $d\overline{x}_m d\theta_X d\overline{y}_m d\theta_Y$. However, and this is the key point here, x and y are *not* allowed to vary freely as they are constrained to move in such a way that $x\in\mathbb{S}^{m-2}_{\sqrt{m}\widehat{\sigma}_{m^{-1}}(x)}(\mathscr{X}_m)$ and $y\in\mathbb{S}^{m-2}_{\sqrt{m}\widehat{\sigma}_{m^{-1}}(y)}(\mathscr{Y}_m)$ with

$$\widehat{\rho}=\cos\theta,\quad \theta=\measuredangle(\overline{\mathscr{X}_m x},\overline{\mathscr{Y}_m y}).$$

Thus,

$$\begin{aligned}\psi_{(\widehat{\sigma}_{m^{-1}}(X),\widehat{\sigma}_{m^{-1}}(Y),\widehat{\rho})}dv &\approx e^{-\frac{m}{2(1-\rho^2)}\left(\frac{\widehat{\sigma}^2_{m^{-1}}(s)}{\sigma_X^2}-\frac{2\rho\widehat{\rho}\widehat{\sigma}_{m^{-1}}(x)\widehat{\sigma}_{m^{-1}}(y)}{\sigma_X\sigma_Y}+\frac{\widehat{\sigma}^2_{m^{-1}}(y)}{\sigma_Y^2}\right)}\times \\ &\times\widehat{\sigma}^{m-2}_{m^{-1}}(x)\widehat{\sigma}^{m-2}_{m^{-1}}(y)f(\widehat{\rho})dv,\end{aligned}$$

where $dv=d\widehat{\sigma}_{m^{-1}}(x)d\widehat{\sigma}_{m^{-1}}(y)d\widehat{\rho}$ and the extra factor $f(\widehat{\rho})$ comes from the constraint referred to above. Incidentally, this already shows that $\{\overline{X}_m,\overline{Y}_m\}$ is independent from $\{\widehat{\sigma}^2_{m^{-1}}(X),\widehat{\sigma}^2_{m^{-1}}(Y),\widehat{\sigma}_{m^{-1}}(XY)\}$, as in the uni-variate case; cf. Proposition 7.23. Now, in order to determine $f(\widehat{\rho})$ note that if x is fixed then, corresponding to an infinitesimal displacement $d\theta$, the segment $\overline{\mathscr{Y}_m y}$ describes an infinitesimal spherical slab in $\mathbb{S}^{m-2}_{\sqrt{m}\widehat{\sigma}_{m^{-1}}(y)}(\mathscr{Y}_m)$ with radius

$$\sqrt{m}\widehat{\sigma}_{m^{-1}}(y)\sin\theta=\sqrt{m}\widehat{\sigma}_{m^{-1}}(y)\sqrt{1-\widehat{\rho}^2}$$

and height

$$\sqrt{m}\widehat{\sigma}_{m^{-1}}(y)|d\theta|=\sqrt{m}\widehat{\sigma}_{m^{-1}}(y)\frac{d\widehat{\rho}}{\sqrt{1-\widehat{\rho}^2}},$$

thus tracing a volume proportionate to

$$(\sqrt{m}\widehat{\sigma}_{m^{-1}}(y)\sin\theta)^{m-3}\sqrt{m}\widehat{\sigma}_{m^{-1}}(y)|d\theta|=\widehat{\sigma}^{m-2}_{m^{-1}}(Y)\underbrace{(1-\widehat{\rho}^2)^{\frac{m-4}{2}}}_{=f(\widehat{\rho})}d\widehat{\rho},$$

which finally gives

$$\begin{aligned}\psi_{(\widehat{\sigma}_{m^{-1}}(X),\widehat{\sigma}_{m^{-1}}(Y),\widehat{\rho})}dv &\approx e^{-\frac{m}{2(1-\rho^2)}\left(\frac{\widehat{\sigma}^2_{m^{-1}}(x)}{\sigma_X^2}-\frac{2\rho\widehat{\rho}\widehat{\sigma}_{m^{-1}}(x)\widehat{\sigma}_{m^{-1}}(y)}{\sigma_X\sigma_Y}+\frac{\widehat{\sigma}^2_{m^{-1}}(y)}{\sigma_Y^2}\right)}\times \\ &\times\widehat{\sigma}^{m-2}_{m^{-1}}(x)\widehat{\sigma}^{m-2}_{m^{-1}}(y)d\widehat{\sigma}_{m^{-1}}(x)d\widehat{\sigma}_{m^{-1}}(y)\times \\ &\times(1-\widehat{\rho}^2)^{\frac{m-4}{2}}d\widehat{\rho}.\end{aligned}$$

As usual, explicit, albeit quite complicated, expressions for the desired pdf $\psi_{\widehat{\rho}}$, which may even be chosen so as to only involve elementary functions, are obtained by integrating this against the area element $d\widehat{\sigma}_{m^{-1}}(x)d\widehat{\sigma}_{m^{-1}}(y)$, with a further integration against $d\widehat{\rho}$ being needed to restore the normalizing constant. Of course, the computational difficulty here comes from the mixed term in the exponential which prevents $\{\widehat{\sigma}^2_{m^{-1}}(X), \widehat{\sigma}^2_{m^{-1}}(Y), \widehat{\rho}\}$ from being independent (except when $\rho = 0$). In any case, the resulting expressions are found to depend on the parameters of the underlying normal bi-variate population only through ρ (and not on any other combination of the entries of the variance matrix in (7.50)), and in fact they all reduce to

$$\psi_{\widehat{\rho}}(r) \approx (1-r^2)^{\frac{m-4}{2}} \mathbf{1}_{(-1,1)}(r), \quad r \in \mathbb{R},$$

when $\rho = 0$, which suffices to efficiently testing the mutual independence of $\{X_j, Y_j\}$ for any value of m along the lines of the general theory developed in Section 12. Otherwise, one has to appeal to asymptotic methods in order to construct "large sample" confidence intervals for ρ; see Example 8.30 below. We refer to the original sources [Stu08a, Fis15], as well as to [Ken46, Chapter 14] and [And03, Chapter 4], for discussions of the basic properties of $\psi_{\widehat{\rho}}$ and their applications. □

Example 7.40. (One way ANOVA) Fix $p \in \mathbb{N}$, $p \geq 3$, a finite sequence $\{n_j\}_{j=1}^p \subset \mathbb{N}$ and for each j consider a random sample $\{X_{jk}\}_{k=1}^{n_j}$ with $X_{jk} \sim \mathcal{N}(\mu_j, \sigma^2)$ such that all these $n := \sum_j n_j$ samples X_{ij} form an independent set. In other words, we are dealing here with p independent normal random samples with varied sizes and expectations but sharing the *same* variance, with the parameters $\{\mu_1, \cdots, \mu_p, \sigma^2\}$ being regarded as unknown. Within each sample we have the decomposition coming from Remark 7.25,

$$\sum_{k=1}^{n_j} \left(X_{jk} - \overline{X}_{\bullet\bullet}\right)^2 = \sum_{k=1}^{n_j} \left(X_{jk} - \overline{X}_{j\bullet}\right)^2 + n_j \left(\overline{X}_{j\bullet} - \overline{X}_{\bullet\bullet}\right)^2, \tag{7.53}$$

where

$$\overline{X}_{j\bullet} = \frac{1}{n_j} \sum_{k=1}^{n_j} X_{jk}$$

and

$$\overline{X}_{\bullet\bullet} = \frac{1}{n} \sum_{j=1}^{p} \sum_{k=1}^{n_j} X_{ij} = \frac{1}{n} \sum_{j=1}^{p} n_j \overline{X}_{j\bullet}.$$

Note that

$$\mathbb{E}(\overline{X}^2_{j\bullet}) = \mu_j^2 + \frac{\sigma^2}{n_j}, \tag{7.54}$$

and

$$\mathbb{E}(\overline{X}^2_{\bullet\bullet}) = \frac{1}{n^2} \left(\sum_{j=1}^{p} n_j \mu_j \right)^2 + \frac{\sigma^2}{n}. \tag{7.55}$$

Now, the same argument leading to the proof of Proposition 7.24 implies that

$$\sigma^{-2} \sum_{k=1}^{n_j} \left(X_{jk} - \overline{X}_{j\bullet}\right)^2 \sim \chi^2_{n_j - 1} \tag{7.56}$$

is independent of $\overline{X}_{j\bullet}$ and hence of $\left(\overline{X}_{j\bullet} - \overline{X}_{\bullet\bullet}\right)^2$. Thus, if we set

$$S^2_{\text{Total}} = \sum_{j=1}^{p} \sum_{k=1}^{n_j} \left(X_{jk} - \overline{X}_{\bullet\bullet}\right)^2,$$

the *total* sum of squares,

$$S^2_{\text{Within}} = \sum_{j=1}^{p} \sum_{k=1}^{n_j} \left(X_{jk} - \overline{X}_{j\bullet}\right)^2,$$

the sum of squares *within* the samples, and

$$S^2_{\text{Between}} = \sum_{j=1}^{p} n_j \left(\overline{X}_{j\bullet} - \overline{X}_{\bullet\bullet}\right)^2,$$

the sum of squares *between* the samples, then

$$S^2_{\text{Total}} = S^2_{\text{Within}} + S^2_{\text{Between}}, \tag{7.57}$$

with

$$\sigma^{-2} S^2_{\text{Within}} \sim \chi^2_{n-p}$$

being independent of S^2_{Between}. On the other hand, again by the argument leading to (7.56), but this time under the *null hypothesis*

$$H_0 : \quad \mu_1 = \cdots = \mu_p, \tag{7.58}$$

we have

$$\sigma^{-2} S^2_{\text{Total}} \sim \chi^2_{n-1}$$

and hence

$$\sigma^{-2} S^2_{\text{Between}} \sim \chi^2_{p-1},$$

which gives

$$\mathbb{E}(S^2_{\text{Between}}) = (p-1)\sigma^2 \tag{7.59}$$

by Corollary 4.21. It then follows from Proposition 4.33 that

$$V := \frac{S^2_{\text{Between}}/(p-1)}{S^2_{\text{Within}}/(n-p)} \sim \mathsf{F}_{p-1,n-p} \text{ under } H_0. \tag{7.60}$$

In order to proceed we now observe that:

- The decomposition (7.57), which is an easy consequence of the fundamental algebraic identity in Remark 7.25, plays a central role in our analysis as it displays S^2_{Total}, the total sum of squares, as resulting from the contribution of two terms of rather distinct types: S^2_{Within} collects together the variations *within* the various samples whereas S^2_{Between} measures the variation *between* the samples;
- In consonance with the previous item, the computation leading to the statistics in (7.60) shows that the distribution of its numerator is conditioned to the validity of H_0 whereas the distribution of its denominator remains the same regardless of the validity of this hypothesis;
- If H_0 is not necessarily satisfied then starting from the fact that

  $$S^2_{\text{Between}} = \sum_{j=1}^{p} n_j \overline{X}^2_{j\bullet} - n\overline{X}^2_{\bullet\bullet},$$

  we easily deduce by means of (7.54) and (7.55) that

  $$\mathbb{E}(S^2_{\text{Between}}) = (p-1)\sigma^2 + \sum_{j=1}^{p} n_j(\mu_j - \overline{\mu})^2, \quad \overline{\mu} = \frac{1}{n}\sum_{j=1}^{p} n_j \mu_j,$$

  which assumes its minimal value, given by (7.59), exactly when H_0 holds true.

Thus, at least on average, a sufficiently large value of V provides statistical evidence for *rejecting* H_0. Precisely, if we fix $0 < \alpha < 1$ then H_0 should be rejected "at significance level α" if the realization v of V in (7.60) satisfies

$$v \in [\mathsf{f}_{p-1,n-p,1-\alpha}, +\infty)\,. \tag{7.61}$$

Here, we use the notation for F-quantiles introduced in (7.48). Again, we refer to Section 12 for a more theoretically inclined justification of this procedure, in particular for the proper understanding of why the "rejection subsets" in the right-hand side of the F-tests in (7.49) and (7.61) differ in their "connectedness". □

Example 7.41. (Estimating the variance of a normal population) If $X_j \sim \mathcal{N}(\mu, \sigma^2)$ we now estimate σ^2, the population variance. We first assume that μ is known and consider the estimator

$$\mathscr{S}_n^2 = \sum_{j=1}^{n} (X_j - \mu)^2,$$

so that

$$\frac{\mathscr{S}_n^2}{\sigma^2} = \sum_{j=1}^{n} \left(\frac{X_j - \mu}{\sigma}\right)^2 \sim \chi_n^2$$

by Corollary 4.25. If, as usual, we define the χ_k^2-*quantile* associated to $0 < \beta < 1$ by

$$F_{\chi_k^2}(\chi_{k,\beta}^2) = \beta, \tag{7.62}$$

then

$$\begin{aligned} P\left(\frac{\mathscr{S}_n^2}{\chi_{n,1-\delta/2}^2} \leq \sigma^2 \leq \frac{\mathscr{S}_n^2}{\chi_{n,\delta/2}^2}\right) &= P\left(\chi_{n,\delta/2}^2 \leq \frac{\mathscr{S}_n^2}{\sigma^2} \leq \chi_{n,1-\delta/2}^2\right) \\ &= 1 - \delta, \end{aligned}$$

so we obtain the confidence interval

$$\sigma^2 \in \left[\frac{\mathscr{S}_n^2}{\chi_{n,1-\delta/2}^2}, \frac{\mathscr{S}_n^2}{\chi_{n,\delta/2}^2}\right] \quad \text{with prob. } 1 - \delta.$$

It is instructive to determine how the center c_n and the length l_n of this (clearly asymmetric) interval behaves as $n \to +\infty$. For this we note that:

- From LLN (Theorem 6.2), $\mathscr{S}_n^2/n \xrightarrow{p} \sigma^2$ so that $\mathscr{S}_n^2 \approx n\sigma^2$ as $n \to +\infty$; see Remark 8.27 for the details.
- Since $Y_j \sim \chi_1^2$ implies $Y^{(n)} = Y_1 + \cdots + Y_n \sim \chi_n^2$ then CLT (Theorem 6.5) and Corollary 4.21 give

$$\frac{Y^{(n)} - n}{\sqrt{2n}} \xrightarrow{d} \mathcal{N}(0,1),$$

so that

$$\chi_{n,\beta}^2 \approx n + \sqrt{2n} z_\beta + O(1), \quad 0 < \beta < 1,$$

which yields

$$\chi_{n,1-\delta/2}^2 \approx n + \sqrt{2n} z + O(1) \;\text{ and }\; \chi_{n,\delta/2}^2 \approx n - \sqrt{2n} z + O(1), \quad z = z_{1-\delta/2}.$$

From these facts we easily deduce that the endpoints of the interval satisfy

$$\frac{\mathscr{S}_n^2}{\chi_{n,1-\delta/2}^2} = \sigma^2 \left(1 - \sqrt{2} z n^{-1/2} + O(n^{-1})\right)$$

and

$$\frac{\mathscr{S}_n^2}{\chi_{n,\delta/2}^2} = \sigma^2 \left(1 + \sqrt{2} z n^{-1/2} + O(n^{-1})\right),$$

so that

$$c_n \approx \sigma^2 + O(n^{-1}) \text{ and } l_n \approx \sigma^2 \left(2\sqrt{2}zn^{-1/2} + O(n^{-1})\right).$$

As expected, the interval is asymptotically centered at σ^2 and has length decaying at the rate $O(n^{-1/2})$. If μ is unknown it is natural to use instead the estimator

$$\widetilde{\mathscr{S}}_n^2 = \sum_{j=1}^{n}(X_j - \overline{X}_n)^2 = (n-1)S_n^2,$$

which corresponds to replacing μ by $\overline{X}_n$ in the definition of $\mathscr{S}_n^2$. Since $\widetilde{\mathscr{S}}_n^2/\sigma^2 \sim \chi^2_{n-1}$ by Proposition 7.24, we end up with the confidence interval

$$\sigma^2 \in \left[\frac{\widetilde{\mathscr{S}}_n^2}{\chi^2_{n-1,1-\delta/2}}, \frac{\widetilde{\mathscr{S}}_n^2}{\chi^2_{n-1,\delta/2}}\right] \text{ with prob. } 1-\delta,$$

whose center and length satisfy

$$\widetilde{c}_n \approx \sigma^2 + O((n-1)^{-1}) \text{ and } \widetilde{l}_n \approx \sigma^2 \left(2\sqrt{2}z(n-1)^{-1/2} + O((n-1)^{-1})\right) = O((n-1)^{-1/2}).$$

Although $\widetilde{l}_n > l_n$ for each n, there holds $\widetilde{l}_n/l_n \to 1$ as $n \to +\infty$, so the intervals are essentially identical for large samples. We note, however, that both intervals are quite conservative for n small. □

8. MAXIMUM LIKELIHOOD

We now present a remarkable class of estimators, introduced by R. Fisher, which displays, under suitable regularity assumptions, many desirable asymptotic properties, including asymptotic normality (Theorem 8.24).

8.1. **Maximum likehood estimators.** We start with an *independent* family $\{X_j\}_{j=1}^n$ of random variables with $X_j \sim \psi_j(x_j;\theta) > 0$, $\theta \in \Theta$.

Definition 8.1. The *likelihood function* of the random vector $X = (X_1, \cdots, X_n) : \Omega \to \mathbb{R}^n$ is

$$L(\mathbf{x};\theta) = \Pi_{j=1}^n \psi_j(x_j;\theta), \tag{8.1}$$

where $\mathbf{x} = (x_1, \cdots, x_n) \in \mathbb{R}^n$ is viewed as a realization of X. In particular, if $\{X_j\}$ is i.i.d. ($X_j \sim \psi(x_j;\theta)$ then

$$L(\mathbf{x};\theta) = \Pi_{j=1}^n \psi_\theta(x_j), \quad \psi_\theta(x_j) = \psi(x_j;\theta). \tag{8.2}$$

We begin with a motivation showing that, in the i.i.d. setting (which covers most cases considered here), maximum likelihood estimators arise as approximate solutions to a natural variational problem.

Definition 8.2. If $\theta_0 \in \Theta$ is the unknown parameter to estimate, we define the *Kullback-Leibler divergence* (centered at θ_0) by

$$\theta \in \Theta \mapsto D_{\theta_0}^{KL}(\theta) := \int_{\mathbb{R}} \psi_{\theta_0}(x) \ln\left(\frac{\psi_{\theta_0}(x)}{\psi_\theta(x)}\right) dx. \tag{8.3}$$

Using Jensen's inequality and assuming, as throughout, that our statistic model $X_j \sim \psi_\theta$ is identifiable[27], we readily see that $D_{\theta_0}^{KL}(\theta) \geq 0$ for any θ, with the equality holding only if $\theta = \theta_0$. This observation naturally leads us to seek an estimator θ^* satisfying

$$\theta^* = \mathrm{argmin}_{\theta\in\Theta} D_{\theta_0}^{KL}(\theta).$$

Since

$$D_{\theta_0}^{KL}(\theta) = \mathrm{const}_{\theta_0} - \mathbb{E}_{\theta_0}(\ln(\psi_\theta)), \tag{8.4}$$

this minimization problem is equivalent to

$$\theta^* = \mathrm{argmax}_{\theta\in\Theta} \mathbb{E}_{\theta_0}(\ln(\psi_\theta)).$$

By LLN we may approximate, for n sufficiently large,

$$\mathbb{E}_{\theta_0}(\ln(\psi_\theta)) \approx \frac{1}{n}\sum_{j=1}^{n} \ln\psi_\theta(X_j) = \mathbb{E}_{\widetilde{\theta}_X}\left(\psi_\theta(X_j)\right), \tag{8.5}$$

where

$$\widetilde{\theta}_X = \frac{1}{n}\sum_{j=1}^{n} \delta_{X_j}$$

is the *empirical distribution* associated with a sample X drawn from the unknown distribution ψ_{θ_0}; cf. Example 6.10. Combining (8.4) and (8.5), we obtain

$$\frac{1}{n} l(X;\theta) \approx \mathrm{const}_{\theta_0} - D_{\widetilde{\theta}_X}^{KL}(\theta), \tag{8.6}$$

where

$$l(\mathbf{x};\theta) = \ln L(\mathbf{x};\theta) \tag{8.7}$$

is the *log-likelihood function*, so that the choice

$$\theta^*(X) \approx \frac{1}{n}\mathrm{argmax}_{\theta\in\Theta} l(X;\theta) = \mathrm{argmax}_{\theta\in\Theta} L(X;\theta) \tag{8.8}$$

emerges as a natural candidate for estimating the unknown parameter θ_0 based on the observed value $\mathbf{x}$. In this sense, the resulting estimator approximately minimizes the Kullback–Leibler divergence from the model to the empirical distribution. Taking into account that the empirical distribution $\widetilde{\theta}_X$ is consistent for θ_0, we see that, at least in the i.i.d. case, the ML estimator asymptotically minimizes the Kullback-Leibler "distance" to θ_0 in (8.3).

This heuristic argument motivates the following fundamental construction due to R. A. Fisher, which remains the most widely used method for deriving estimators.

Definition 8.3. (Maximum Likelihood Estimation, MLE) Under the conditions above, a maximum likelihood (ML) estimator $\widehat{\theta} = \widehat{\theta}(X)$ is defined by

$$\widehat{\theta}(\mathbf{x}) = \mathrm{argmax}_{\theta\in\Theta} L(\mathbf{x};\theta) = \mathrm{argmax}_{\theta\in\Theta} l(\mathbf{x};\theta)., \tag{8.9}$$

where $\mathbf{x}$ is an observed value of X.

As usual, we assume that the log-likelihood l is strictly concave in θ so the solution to (8.9) is unique (whenever it exists).

[27]Recall that identifiability means that the map $\theta \in \Theta \mapsto \psi_\theta$ is injective.

Remark 8.4. Regarding the heuristic discussion above, we note that the interpretation of maximum likelihood estimation as the minimization of Kullback–Leibler divergence goes back to S. Kullback and R. Leibler [KL51] and was later made explicit by H. Akaike [Aka73]. Systematic treatments of this circle of ideas from a decision-theoretic perspective, leading to the consideration of a host of information criteria in model selection, may be found in [SIK86, KK08, BA13]; see also Remark 8.34, where we discuss the Akaike Information Criterion (AIC). □

Example 8.5. (MLE from a normal population) If $X_j \sim \mathcal{N}(\mu, \sigma^2)$ then

$$L(\mathbf{x};\theta) = (2\pi\theta_2)^{-n/2} e^{-\frac{1}{2\theta_2}\sum_{j=1}^n (x_j-\theta_1)^2}, \tag{8.10}$$

where $\theta = (\theta_1, \theta_2) = (\mu, \sigma^2) \in \Theta = \mathbb{R} \times \mathbb{R}_+$, so that

$$l(\mathbf{x};\theta) = \ln L(\mathbf{x};\theta) = -\frac{n}{2}\ln(2\pi\theta_2) - \frac{1}{2\theta_2}\sum_j (x_j-\theta_1)^2. \tag{8.11}$$

The usual first derivative test shows that the ML estimator $\widehat{\theta}_1$ and $\widehat{\theta}_1$ corresponding to (8.11) should satisfy

$$0 = \frac{\partial l}{\partial \theta_1}(\widehat{\theta}) = \frac{1}{\theta_2}\sum_j (X_j - \widehat{\theta}_1), \quad 0 = \frac{\partial l}{\partial \theta_2}(\widehat{\theta}) = -\frac{n}{2\theta_2} + \frac{1}{2\theta_2^2}\sum_j (X_j - \widehat{\theta}_1)^2,$$

which gives

$$\widehat{\theta}_1 = \overline{X}_n = \frac{1}{n}\sum_j X_j, \quad \widehat{\theta}_2 = \widehat{\sigma}^2_{n^{-1}} = \frac{1}{n}\Sigma_j (X_j - \overline{X}_n)^2.$$

We thus see that the ML estimator $\widehat{\theta}_2 = \widehat{\sigma}^2_{n^{-1}}$ for the variance coming from (8.11) not only fails to be unbiased but also satisfies

$$\mathrm{mse}(\widehat{\sigma}^2_{(n-1)^{-1}}) > \mathrm{mse}(\widehat{\sigma}^2_{n^{-1}}) > \mathrm{mse}(\widehat{\sigma}^2_{(n+1)^{-1}}),$$

so its performance, as measured by mse, lies somewhere between those of the variance estimators considered so far. We point out that the asymptotic performance of a (sufficiently regular and consistent) ML estimator is examined in Theorem 8.24 below. In particular, asymptotically normality is established there, which confirms that $\widehat{\sigma}^2_{n^{-1}}$ stands out as the most efficient estimator from this viewpoint. □

Example 8.6. (MLE from a jointly normal population) Using the notation of Example 7.39, we see that the right-hand side of (7.51) allows us to write down the likelihood function of the jointly normal random sample $\{X, Y\}$ as

$$\begin{aligned} L(\mathbf{x},\mathbf{y};\theta) &= \frac{1}{(2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2})^m} \times \\ &\quad \times e^{-\frac{1}{2(1-\rho^2)}\sum_j \left(\frac{A(x_j)}{\sigma_X^2} - \frac{2\rho B(x_j,y_j)}{\sigma_X\sigma_Y} + \frac{C(y_j)}{\sigma_Y^2}\right)}, \end{aligned}$$

where $\theta = (\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \rho)$ and

$$A(x_j) = (x_j-\mu_X)^2, \quad B(x_j,y_j) = (x_j-\mu_X)(y_j-\mu_Y), \quad C(x_j) = (y_j-\mu_Y)^2, \tag{8.12}$$

so that

$$\begin{aligned} l(\mathbf{x},\mathbf{y};\theta) &= -m\ln(2\pi\sigma_X\sigma_Y) - \frac{m}{2}\ln(1-\rho^2) - \\ &\quad -\frac{1}{2(1-\rho^2)}\sum_j \left(\frac{A(x_j)}{\sigma_X^2} - \frac{2\rho B(x_j,y_j)}{\sigma_X\sigma_Y} + \frac{C(y_j)}{\sigma_Y^2}\right), \end{aligned} \tag{8.13}$$

Starting from this, a straightforward analysis involving the first derivative test $\nabla_\theta l = 0$ confirms that $\widehat{\theta} := (\overline{X}_m, \overline{Y}_m, \widehat{\sigma}^2_{m^{-1}}(X), \widehat{\sigma}^2_{m^{-1}}(Y), \widehat{\rho})$ is the ML estimator of θ; see [Ken46, Section 14.11] or [And03, Corollary 3.2.2] for the details. □

Example 8.7. (MLE from an exponential population) If $X_j \sim \mathrm{Exp}(\lambda)$, the exponential distribution with parameter $\lambda > 0$, where

$$\mathrm{Exp}(\lambda)(x) = \lambda e^{-\lambda x}\mathbf{1}_{[0,+\infty)},$$

then

$$L(\mathbf{x};\lambda) = \lambda^n e^{-\lambda \sum_j x_j} \Longrightarrow l(\mathbf{x};\lambda) = n\ln\lambda - \lambda\sum_j x_j,$$

so that

$$\widehat{\lambda} = \frac{n}{\sum_j X_j},$$

which matches the fact that $\mathbb{E}(X_j) = 1/\lambda$. □

Example 8.8. If $X_j \sim \mathsf{Ber}(p)$, the Bernoulli distribution in Remark 6.8, then for $x \in \{0,1\}$ we have $P(X_j = x) = p^x(1-p)^{1-x}$, where $p \in (0,1)$ is the unknown parameter. If we assume as always that $\{X_j\}$ is independent the associated likelihood function is

$$L(\mathbf{x};p) = p^{\sum_j x_j}(1-p)^{n-\sum_j x_j},$$

and hence,

$$l(\mathbf{x};p) = \left(\sum_j x_j\right)\ln p + \left(n - \sum_j x_j\right)\ln(1-p). \tag{8.14}$$

The usual first derivative test for a minimum at $p = \widehat{p}$ is

$$0 = \frac{\partial l}{\partial p}(\widehat{p}) = \frac{\sum_j x_j}{\widehat{p}} - \frac{n - \sum_j x_j}{1-\widehat{p}},$$

so that

$$\widehat{p} = \frac{1}{n}\sum_j X_j,$$

the sample mean. Also, if $X_j \sim \mathsf{Pois}(\rho)$, the Poisson distribution with parameter $\rho > 0$, so that $P(X_j = x) = \rho^x e^{-\rho}/x!$, $x \in \{0,1,2,\cdots,\}$, then a simple computation shows that

$$\widehat{\rho} = \frac{1}{n}\sum_j X_j,$$

which confirms that the corresponding ML estimator is also the sample mean. □

Remark 8.9. The analysis in the examples above should be complemented with the usual second derivative test to check in each case that the ML estimator attains the (unique) global maximum of the corresponding likelihood function. □

8.2. **Fisher information and Cramér-Rao lower bound.** We now present a universal lower bound for the covariance matrix in each class of estimators with a prescribed expectation (in particular, for unbiased estimators) in terms of an invariant (Fisher information) depending on the likelihood function of the given statistical model. Instead of restricting ourselves to statistical models, let us assume for the moment only that $X_j : \Omega \to \mathbb{R}$ are independent random variables, $i = 1,\cdots,n$, giving rise to a random vector $X = (X_1,\cdots,X_n)$[28]. Let $L = L(\mathbf{x};\theta) > 0$ be the corresponding likelihood function and $l(\mathbf{x};\theta) = \ln L(\mathbf{x};\theta)$ the log-likelihood function, where $\mathbf{x} = (x_1,\cdots,x_n)$ and $\theta \in \Theta \subset \mathbb{R}^q$, the space of parameters. Clearly, the notions of statistics and estimators can be easily adapted to this broader setting. In the sequel we assume that l is regular enough so that all the differential/integral manipulations hold true.

[28]In other words, we consider here a statistical model in the extended sense of Remark 7.4.

Definition 8.10. Under the conditions above, we define the *score vector* (of the given sample X) as

$$s(X;\theta) = \nabla_\theta l(X;\theta).$$

Also, the corresponding *Fisher information matrix* is

$$\mathscr{F}^X(\theta) = \operatorname{cov}(s(X;\theta)). \tag{8.15}$$

We henceforth assume that the symmetric matrix $\mathscr{F} = \mathscr{F}^X$ in (8.10) is positive definite, so the inverse matrix $\mathscr{F}^{-1}$ exists. Since the random effects present in the sample have been averaged out after taking covariance of the score, $\mathscr{F} = \mathscr{F}(\theta)$ is an invariant of the given model, in particular not being attached to any potential estimator.

Remark 8.11. In order to illustrate the importance of requiring that $\{X_j\}$ is independent, let us assume that we are in the unidimensional case, $\Theta \subset \mathbb{R}$, so we call the scalar $\mathscr{F}$ simply the *Fisher information*. One has

$$s(X;\theta) = \sum_j \frac{\frac{d}{d\theta}\psi_j(X_j;\theta)}{\psi_j(X_j;\theta)},$$

a sum of *independent* random variables, so that by (2.15),

$$\operatorname{var}(s(X;\theta)) = \sum_j \operatorname{var}\left(\frac{\frac{d}{d\theta}\psi_j(X_j;\theta)}{\psi_j(X_j;\theta)}\right),$$

which means that

$$\mathscr{F}^X = \sum_j \mathscr{F}^{X_j}.$$

Thus, independence leads to a simple additive formula describing how the Fisher information of the whole sample decomposes as a sum of contributions coming from its parts. If we additionally require that $X_j \sim \psi_\theta$ is i.d.d., which is the only case treated in all examples below, then this becomes

$$\mathscr{F}_{(n)} = n\mathscr{F}_{(1)},$$

with $\mathscr{F}_{(n)} = \mathscr{F}^X$ refering as before to the whole sample whereas $\mathscr{F}_{(1)} = \mathscr{F}^{X_j}$ refers to *any* single observation. □

Example 8.12. If $X_j \sim \mathsf{Ber}(p)$ then (8.14) gives

$$s(X;p) = \frac{\partial}{\partial p} l(X;p) = \frac{\sum_j X_j}{p} - \frac{n - \sum_j X_j}{1-p} = \frac{n}{p(1-p)}\overline{X} - \frac{n}{1-p},$$

and since $\operatorname{cov}(\overline{X}) = \operatorname{cov}(X_j)/n = p(1-p)/n$, we conclude that

$$\mathscr{F}_{(n)}(p) = \frac{n}{p(1-p)}. \tag{8.16}$$

Thus, the Fisher information increases with the sample size according to a rate which is inversely proportional to the "fluctuation" (as measured by the population variance). In a sense, this simple example justifies the qualification of "information" for this concept; for more on this point see Remark 8.25 below. □

Example 8.13. If $X_j \sim \mathcal{N}(\theta_1, \theta_2)$ is drawn from a normal population, where $\theta_1 = \mu$ and $\theta_2 = \sigma^2$, then (8.11) leads to

$$s(X;\theta) = \nabla_\theta l(X;\theta) = \begin{pmatrix} \frac{1}{\theta_2}\sum_j (X_j - \theta_1) \\ -\frac{n}{2\theta_2} + \frac{1}{2\theta_2^2}\sum_j (X_j - \theta_1)^2 \end{pmatrix}, \quad \theta = \begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix}. \tag{8.17}$$

Since $X_j - \theta_1 \sim \mathcal{N}(0, \theta_2)$, independence implies that

$$\operatorname{var}\left(\frac{1}{\theta_2}\sum_j (X_j - \theta_1)\right) = \frac{n}{\theta_2}. \tag{8.18}$$

On the other hand,

$$-\frac{n}{2\theta_2} + \frac{1}{2\theta_2^2}\sum_j (X_j - \theta_1)^2 = -\frac{n}{2\theta_2} + \frac{1}{2\theta_2}\sum_j \left(\frac{X_j - \theta_1}{\sqrt{\theta_2}}\right)^2,$$

and since $(X_j - \theta_1)/\sqrt{\theta_2} \sim \mathcal{N}(0,1)$, we see that

$$\sum_j \left(\frac{X_j - \theta_1}{\sqrt{\theta_2}}\right)^2 \sim \chi^2_n$$

by Corollary 4.25, so that Corollary 4.21 applies to give

$$\operatorname{var}\left(\sum_j \left(\frac{X_j - \theta_1}{\sqrt{\theta_2}}\right)^2\right) = 2n, \tag{8.19}$$

and hence,

$$\operatorname{var}\left(-\frac{n}{2\theta_2} + \frac{1}{2\theta_2^2}\sum_j (X_j - \theta_1)^2\right) = \frac{n}{2\theta_2^2}. \tag{8.20}$$

Note that (8.18) and (8.20) provide the diagonal terms of the corresponding Fisher information matrix. In order to compute the off-diagonal terms we observe that

$$\mathbb{E}\left(\frac{1}{\theta_2}\sum_j (X_j - \theta_1)\right) = 0$$

and

$$\begin{aligned}\mathbb{E}\left(-\frac{n}{2\theta_2} + \frac{1}{2\theta_2^2}\sum_j (X_j - \theta_1)^2\right) &= -\frac{n}{2\theta_2} + \frac{1}{2\theta_2}\mathbb{E}\left(\sum_j \left(\frac{X_j - \theta_1}{\sqrt{\theta_2}}\right)^2\right)\\ &= -\frac{n}{2\theta_2} + \frac{n}{2\theta_2}\\ &= 0,\end{aligned}$$

where we used (8.19) and Corollary 4.21 in the next to the last step; this should be compared with the general result in Corollary 8.15 below. It follows that

$$\operatorname{cov}\left(\frac{1}{\theta_2}\sum_j (X_j - \theta_1), -\frac{n}{2\theta_2} + \frac{1}{2\theta_2^2}\sum_j (X_j - \theta_1)^2\right) = \frac{1}{2\theta_2^3}\sum_j \mathbb{E}\left((X_j - \theta_1)^3\right),$$

which clearly vanishes. We thus conclude that

$$\mathscr{F}_{(n)}(\theta) = \begin{pmatrix} n/\theta_2 & 0 \\ 0 & n/2\theta_2^2 \end{pmatrix}. \tag{8.21}$$

In particular, if we consider θ_2 as known,

$$\mathscr{F}_{(n)}(\theta_1) = \frac{n}{\theta_2}. \tag{8.22}$$

We will see in Remark 8.17 below a much simpler route to retrieve (8.21). □

In both (8.22) and (8.16), we observe that the Fisher information $\mathscr{F} = \mathscr{F}_{(n)}$ equals the reciprocal of the variance of the corresponding estimator, in both cases the (unbiased) sample mean. This reflects the remarkable general fact that the information matrix provides a universal lower bound for the covariance of any class of estimators with a prescribed bias, including in particular the unbiased ones; see Theorem 8.18 below. The next result is the first step toward establishing this fundamental connection.

Proposition 8.14. *For any (sufficiently regular) vector $t = t(\mathbf{x};\theta)$ there holds*

$$\mathbb{E}(s \otimes t) = \nabla_\theta \mathbb{E}(t) - \mathbb{E}(\nabla_\theta t).$$

Proof. We compute:

$$\begin{aligned}
\mathbb{E}(s \otimes t) &= \int_{\mathbb{R}^n} L(\mathbf{x};\theta)^{-1} \nabla_\theta L(\mathbf{x};\theta) \otimes t(\mathbf{x};\theta) L(\mathbf{x};\theta) d\mathbf{x} \\
&= \int_{\mathbb{R}^n} \nabla_\theta (L(\mathbf{x};\theta) t(\mathbf{x};\theta)) d\mathbf{x} - \int_{\mathbb{R}^n} L(\mathbf{x};\theta) \nabla_\theta t(\mathbf{x};\theta) d\mathbf{x} \\
&= \nabla_\theta \int_{\mathbb{R}^n} L(\mathbf{x};\theta) t(\mathbf{x};\theta) d\mathbf{x} - \int_{\mathbb{R}^n} L(\mathbf{x};\theta) \nabla_\theta t(\mathbf{x};\theta) d\mathbf{x},
\end{aligned}$$

as desired. □

Corollary 8.15. *There holds*

(8.23) $$\mathbb{E}(s) = \vec{0}.$$

In particular,

(8.24) $$\mathscr{F} = \mathbb{E}(s \otimes s) = -\mathbb{E}(\nabla^2_{\theta\theta} l).$$

Proof. Take $t = (1, \cdots, 1)$. □

Corollary 8.16. *If $t = t(\mathbf{x})$ then $\mathbb{E}(s \otimes t) = \nabla_\theta \mathbb{E}(t)$. In particular, if $t = \widehat{\theta}$ is an estimator with $g(\theta) := \mathbb{E}(\widehat{\theta})$ then $\mathbb{E}(s \otimes \widehat{\theta}) = \nabla_\theta g$.*

Proof. The first assertion is immediate and the second one follows from the fact that $\widehat{\theta}$, as an estimator, does *not* depend on θ. □

Remark 8.17. As a checking we may use (8.24) to recalculate the Fisher information matrix of a normal sample $X_j \sim \mathcal{N}(\theta_1, \theta_2)$ as in Example 8.13. From (8.17) we have

$$\nabla_{\theta\theta} l(X;\theta) = \begin{pmatrix} -\frac{n}{\theta_1} & -\frac{1}{\theta_2^2}\sum_j (X_j - \theta_1) \\ -\frac{1}{\theta_2^2}\sum_j (X_j - \theta_1) & \frac{n}{2\theta_2^2} - \frac{1}{\theta_2^3}\sum_j (X_j - \theta_1)^2 \end{pmatrix},$$

so that

$$\mathscr{F}_{(n)}(\theta) = \begin{pmatrix} \frac{n}{\theta_1} & 0 \\ 0 & -\frac{n}{2\theta_2^2} + \frac{1}{\theta_2^3}\sum_j \mathbb{E}\left((X_j - \theta_1)^2\right) \end{pmatrix},$$

and since $\mathbb{E}((X_j-\theta_1)^2)=\mathrm{var}(X_j)=\theta_2$, we recover (8.21). Note that this computation is much simpler because it is based on computing expectations, at the cost of taking one more derivative of log-likelihood function but with no need to compute covariances, and hence bypasses any appeal to the connection between sums of squares of normals and chi-squares. □

Let $\mathcal{E}$ be the set of all estimators (for θ). Given $g:\Theta\to\mathbb{R}^q$ define

$$\mathcal{E}_g=\left\{\widehat{\theta}\in\mathcal{E};\mathbb{E}(\widehat{\theta})=g(\theta)\right\}.$$

Equivalently, $\mathcal{E}_g$ is the set of all *unbiased* estimators for $g(\theta)$. Note that each $\widehat{\theta}\in\mathcal{E}_g$ satisfies

$$\mathrm{bias}(\widehat{\theta})=g(\theta)-\theta \tag{8.25}$$

and hence

$$\mathrm{mse}(\widehat{\theta})=\|g(\theta)-\theta)\|^2+\mathrm{tr}\,\mathrm{cov}(\widehat{\theta}).$$

The next result provides a uniform lower bound for the covariance (and hence for the mse) of estimators in each class $\mathcal{E}_g$ (provided it is not empty).

Theorem 8.18. *(Cramér-Rao) There holds*

$$\mathrm{cov}(\widehat{\theta})\geq\nabla_\theta g\mathscr{F}(\theta)^{-1}\nabla_\theta g^\top \tag{8.26}$$

for any $\widehat{\theta}\in\mathcal{E}_g$. In particular,

$$\mathrm{cov}(\widehat{\theta})\geq\mathscr{F}(\theta)^{-1} \tag{8.27}$$

if $\widehat{\theta}$ is unbiased ($g(\theta)=\theta$).

Proof. We first consider the uni-dimensional case $\Theta\subset\mathbb{R}$. From Corollaries 8.16 and 8.15 we have

$$\begin{aligned}g'(\theta) &= \int_{\mathbb{R}^n}s(\mathbf{x};\theta)\widehat{\theta}(\mathbf{x})L(\mathbf{x};\theta)dx\\ &= \int_{\mathbb{R}^n}s(\mathbf{x};\theta)\left(\widehat{\theta}(\mathbf{x})-g(\theta)\right)L(\mathbf{x};\theta)d\mathbf{x},\end{aligned}$$

so that Cauchy-Schwartz inequality gives

$$|g'(\theta)|^2\leq\mathrm{var}(s(X;\theta))\mathrm{var}(\widehat{\theta}(X)),$$

as desired. The proof of the multi-dimensional case is quite similar and makes use of a well-known algebraic inequality: for any random vectors $Z,W\in\mathbb{R}^q$ with $\mathrm{cov}(W)>0$ there holds

$$\mathrm{cov}(Z)\geq\mathrm{cov}(Z,W)\mathrm{cov}(W)^{-1}\mathrm{cov}(W,Z).$$

Taking $Z=\widehat{\theta}$ and $W=s$ we get

$$\begin{aligned}\mathrm{cov}(\widehat{\theta}) &\geq \mathrm{cov}(\widehat{\theta},s)\mathrm{cov}(s)^{-1}\mathrm{cov}(s,\widehat{\theta})\\ &= \mathbb{E}(\widehat{\theta}\otimes s)\mathbb{E}(s\otimes s)^{-1}\mathbb{E}(s\otimes\widehat{\theta})\\ &= \nabla_\theta g\mathscr{F}(\theta)^{-1}\nabla_\theta g^\top,\end{aligned}$$

as desired. □

Corollary 8.19. *The best estimator in $\mathcal{E}_g$ (if it exists) is the one whose covariance matrix attains the lower bound in (8.26). In particular, an unbiased estimator whose covariance matrix attains the lower bound in (8.27) has the best performance (as measured by the mse).*

Example 8.20. (The sample mean as the best unbiased estimator of the expected value of a Bernoulli or Poisson population) It follows from (8.16) that, for a Bernoulli population,

$$\mathscr{F}_{(n)}(p) = \frac{n}{p(1-p)} = \frac{1}{\mathrm{var}(\widehat{p})},$$

so Corollary 8.19 applies and the sample mean $\widehat{p}$ is the best unbiased estimator for the expected value. A similar reasoning, based on the explicit computation of the corresponding Fisher information, confirms that the sample mean is the best unbiased estimator for the expected value of a Poisson random sample as in Example 8.8.

Example 8.21. (The sample mean as the best estimator of the expected value of a normal population) As observed in [HL51], the Cramér–Rao inequality in Theorem 8.18 may be used to prove that the sample mean $\widehat{\theta}_1 = \overline{X}$ is the *best* estimator for the mean $\theta_1 = \mu$ of a normal population, in the sense that it attains the smallest possible mse among *all* such estimators.[29]; see also [LC06, Example 5.2.8] for an alternative proof of this fact. Here we adopt the notation of Example 8.13, so that our sample satisfies $X_j \sim \mathcal{N}(\theta_1, \theta_2)$, $j = 1, \ldots, n$. Let $\widehat{\theta}_\bullet$ be an estimator of θ_1 such that $\mathrm{mse}(\widehat{\theta}_\bullet) \leq \mathrm{mse}(\widehat{\theta}_1) = \theta_2/n$. Setting $b(\theta_1) = \mathrm{bias}_{\theta_1}(\widehat{\theta}_\bullet)$ and noting that $\mathcal{F}(\theta_1) = n/\theta_2$ by (8.21), we obtain from (8.26) and (8.25) that

$$b(\theta_1)^2 + \frac{\theta_2}{n}\left(1 + b'(\theta_1)\right)^2 \leq \frac{\theta_2}{n}. \tag{8.28}$$

It follows that b is uniformly bounded, $|b| \leq \sqrt{\theta_2/n}$, and that $b' \leq 0$ everywhere. If there existed $\epsilon > 0$ such that $b' \leq -\epsilon$ for $\theta_1 \to \pm\infty$, then b would not remain bounded, which is impossible. Hence $b'(\pm\infty) = 0$, and from (8.28) we deduce that $b(\pm\infty) \to 0$. Since b is nonincreasing, this forces $b \equiv 0$. Therefore $\widehat{\theta}_\bullet$ is unbiased and satisfies $\mathrm{mse}(\widehat{\theta}_\bullet) = \mathrm{mse}(\widehat{\theta}_1)$, as claimed. As also noted in [HL51], this argument can be adapted to the setting of Bernoulli and Poisson populations in Example 8.20, thereby eliminating the need for an unbiasedness assumption on the competing estimators. □

Example 8.22. (The James-Stein estimator [JS61]) Starting with a single sample $X \sim \mathcal{N}(\boldsymbol{\mu}, \sigma^2 \mathrm{Id}_p)$, where σ^2 is known, the log-likehood function

$$l(\mathbf{x}; \boldsymbol{\mu}) = -\frac{p}{2}\ln 2\pi\sigma^2 - \frac{1}{2\sigma^2}\|\mathbf{x} - \boldsymbol{\mu}\|^2 \tag{8.29}$$

tells us that the MLE for $\boldsymbol{\mu}$ is $\widehat{\boldsymbol{\mu}} = X$. Since $\sigma^{-2}\|X - \boldsymbol{\mu}\|^2 \sim \chi^2_p$ we know that

$$\mathrm{mse}(\widehat{\boldsymbol{\mu}}) = \mathbb{E}(\|X - \boldsymbol{\mu}\|^2) = p\sigma^2, \tag{8.30}$$

with only the variance contributing (since $\widehat{\boldsymbol{\mu}}$ is unbiased). Now let us compare $\widehat{\boldsymbol{\mu}}$ with the *James-Stein estimator*

$$\widehat{\boldsymbol{\mu}}_{JS} = \left(1 - \frac{(p-2)\sigma^2}{\|\widehat{\boldsymbol{\mu}}\|^2}\right)\widehat{\boldsymbol{\mu}} = X - (p-2)\,\sigma^2\frac{X}{\|X\|^2}, \tag{8.31}$$

[29] In this case we say that $\widehat{\theta}_1$ is *admissible*, meaning that there exists no other estimator $\widehat{\theta}_\bullet$ such that $\mathrm{mse}(\widehat{\theta}_\bullet) < \mathrm{mse}(\widehat{\theta}_1)$.

whose mean squared error is

$$\begin{aligned}\mathrm{mse}(\widehat{\boldsymbol{\mu}}_{JS}) &= \mathbb{E}\left(\|\widehat{\boldsymbol{\mu}}_{JS}-\boldsymbol{\mu}\|^2\right)\\ &= \mathbb{E}\left(\left\|\widehat{\boldsymbol{\mu}}-\boldsymbol{\mu}-(p-2)\sigma^2\frac{\widehat{\boldsymbol{\mu}}}{\|\widehat{\boldsymbol{\mu}}\|^2}\right\|^2\right)\\ &= \mathrm{mse}(\widehat{\boldsymbol{\mu}})-2(p-2)\sigma^2\mathbb{E}\left(\frac{\langle X, X-\boldsymbol{\mu}\rangle}{\|X\|^2}\right)+(p-2)^2\sigma^4\mathbb{E}\left(\|X\|^{-2}\right)\\ &= p\sigma^2-2(p-2)\sigma^2\mathbb{E}\left(\frac{\langle X, X-\boldsymbol{\mu}\rangle}{\|X\|^2}\right)+(p-2)^2\sigma^4\mathbb{E}\left(\|X\|^{-2}\right),\end{aligned}$$

where we used (8.30) in the last step. In order to handle the mixed term in the right-hand side we first note from (8.29) that the score vector is

$$s(\mathbf{x};\boldsymbol{\mu})=\nabla_{\boldsymbol{\mu}} l(\mathbf{x};\boldsymbol{\mu})=\sigma^{-2}\left(\mathbf{x}-\boldsymbol{\mu}\right),$$

so if we make $t=t(X)\in\mathbb{R}$ in Corollary 8.16 we obtain

$$\mathbb{E}\left(\frac{\partial}{\partial x_j}t(X)\right)=\frac{\partial}{\partial \mu_j}\mathbb{E}\left(t(X)\right)=\sigma^{-2}\mathbb{E}\left(t(X)(X_j-\boldsymbol{\mu}_j)\right),\quad j=1,\cdots,p, \tag{8.32}$$

a result usually known as *Stein's equation*[30]. By taking $t(X)=X_j/\|X\|^2$ and summing over j we realize that

$$\mathbb{E}\left(\frac{\langle X, X-\boldsymbol{\mu}\rangle}{\|X\|^2}\right)=(p-2)\sigma^2\mathbb{E}\left(\|X\|^{-2}\right), \tag{8.33}$$

which gives

$$\mathrm{mse}(\widehat{\boldsymbol{\mu}}_{JS})=p\sigma^2-(p-2)^2\sigma^4\mathbb{E}\left(\|X\|^{-2}\right).$$

On the other hand, it follows from (2.7) that

$$\mathbb{E}\left(\|X\|^{-2}\right)=\frac{1}{(2\pi)^{p/2}}\int_{\mathbb{R}^p}\|\mathbf{x}\|^{-2}e^{-\frac{1}{2}\|\mathbf{x}-\boldsymbol{\mu}\|^2}d\mathbf{x},$$

an integral which becomes finite if $\int r^{p-3}dr$ converges near $r=0$. Thus, we conclude that $\mathrm{mse}_{\boldsymbol{\mu}}(\widehat{\boldsymbol{\mu}}_{JS})<p\sigma^2$ if $p\geq 3$ for *any* $\boldsymbol{\mu}$, which confirms that in those cases the unbiased MLE estimator $\widehat{\boldsymbol{\mu}}$ fails to be the most efficient one (if the "performance" is measured by the mean squared error); cf. Remark 7.28. Regarding this remarkable estimator, we add the following comments.

- Since $\widehat{\boldsymbol{\mu}}=X$ is simply the sample mean as we have just a single observation at our disposal, this is in sharp contrast with the result in Example 8.21, which says that the sample mean is the best estimator for $\boldsymbol{\mu}$ is $p=1$. To reinforce this analogy, let us take a random sample $X_j\sim\mathcal{N}(\boldsymbol{\mu},\sigma^2\mathrm{Id}_p)$, $j=1,\cdots,n$, so that n observations of the underlying multivariate normal population are available. Now, the log-likelihood function is

  $$l(\mathbf{x};\boldsymbol{\mu})=-\frac{np}{2}\ln 2\pi\sigma^2-\frac{1}{2\sigma^2}\sum_j\|\mathbf{x}_j-\boldsymbol{\mu}\|^2,\quad \mathbf{x}=(\mathbf{x}_1,\cdots,\mathbf{x}_n)\in\mathbb{R}^{np},$$

  so the MLE for $\boldsymbol{\mu}$ is $\widehat{\boldsymbol{\mu}}^{(n)}=\overline{X}$, where

  $$\overline{X}=\frac{1}{n}\sum_j X_j\in\mathbb{R}^p.$$

[30]Remarkably enough, the validity of (8.32) for all t varying in a suitable class of test functions completely characterizes $\sigma^{-1}(X-\boldsymbol{\mu})$ as a standard normal random vector, which turns out to be the starting point of Stein's approach to the Berry-Esseen theorem discussed in Remark 6.7 [Che21].

Since $\sigma^{-2}\|X_j-\boldsymbol{\mu}\|^2 \sim \chi^2_p$ for each j and $\{X_j-\boldsymbol{\mu}\}_{j=1}^n$ is independent, we have

$$\begin{aligned}\mathrm{mse}(\widehat{\boldsymbol{\mu}}^{(n)}) &= \mathbb{E}(\|\overline{X}-\boldsymbol{\mu}\|^2)\\ &= \frac{1}{n^2}\mathbb{E}\left(\left\|\sum_j (X_j-\boldsymbol{\mu})\right\|^2\right)\\ &= \frac{1}{n^2}\sum_j \mathbb{E}\left(\|X_j-\boldsymbol{\mu}\|^2\right)\\ &= \frac{p}{n}\sigma^2,\end{aligned}$$

where again only the covariance contributes (since $\widehat{\boldsymbol{\mu}}^{(n)}$ is unbiased). It turns out that essentially the same argument as above confirms that $\widehat{\boldsymbol{\mu}}^{(n)}$ fails to be admissible if $p \geq 3$, as the corresponding James-Stein estimator

$$\widehat{\boldsymbol{\mu}}^{(n)}_{JS} = \left(1-\frac{(p-2)\sigma^2/n}{\|\widehat{\boldsymbol{\mu}}^{(n)}\|^2}\right)\widehat{\boldsymbol{\mu}}^{(n)} = \overline{X} - (p-2)\,\frac{\sigma^2}{n}\,\frac{\overline{X}}{\|\overline{X}\|^2}$$

satisfies $\mathrm{mse}(\widehat{\boldsymbol{\mu}}^{(n)}_{JS}) < p\sigma^2/n$.

- In case σ^2 is unknown, and restricting ourselves to the case $n=1$ for simplicity, let us replace (8.31) by

$$\widehat{\boldsymbol{\mu}}_{JS_l} = \left(1-\frac{(p-2)c_l\mathfrak{s}}{\|\widehat{\boldsymbol{\mu}}\|^2}\right)\widehat{\boldsymbol{\mu}} = X - (p-2)\,c_l\mathfrak{s}\frac{X}{\|X\|^2},$$

where $p \geq 3$, c_l is a positive constant (depending on a positive integer l given in advance) to be determined below and $\mathfrak{s}$ is the appropriate estimator of σ^2 in the sense that $\sigma^{-2}\mathfrak{s} \sim \chi^2_l$ and $\{\mathfrak{s}, X\}$ is independent. Setting $\boldsymbol{\mu}_\bullet = \sigma^{-1}\boldsymbol{\mu}$, $\widehat{\boldsymbol{\mu}}_\bullet = \sigma^{-1}\widehat{\boldsymbol{\mu}}$, $X_\bullet = \sigma^{-1}X$ and $\mathfrak{s}_\bullet = \sigma^{-2}\mathfrak{s}$, we compute

$$\begin{aligned}\mathrm{mse}(\widehat{\boldsymbol{\mu}}_{JS_l}) &= \mathbb{E}\left(\|\widehat{\boldsymbol{\mu}}_{JS_l}-\boldsymbol{\mu}\|^2\right)\\ &= \sigma^2\mathbb{E}\left(\left\|\widehat{\boldsymbol{\mu}}_\bullet - \boldsymbol{\mu}_\bullet - (p-2)c_l\mathfrak{s}_\bullet\frac{\widehat{\boldsymbol{\mu}}_\bullet}{\|\widehat{\boldsymbol{\mu}}_\bullet\|^2}\right\|^2\right)\\ &= \sigma^2\left(\mathrm{mse}(\widehat{\boldsymbol{\mu}}_\bullet) - 2(p-2)c_l l\mathbb{E}\left(\frac{\langle X_\bullet, X_\bullet-\boldsymbol{\mu}_\bullet\rangle}{\|X_\bullet\|^2}\right) + (p-2)^2c_l^2 l(l+2)\mathbb{E}\left(\|X_\bullet\|^{-2}\right)\right),\end{aligned}$$

where we used the independence and that $\mathbb{E}(\mathfrak{s}_\bullet) = l$ and $\mathbb{E}(\mathfrak{s}_\bullet^2) = l(l+2)$ in the last step. Combining this with the obvious counterpart of (8.33) we end up with

$$\mathrm{mse}(\widehat{\boldsymbol{\mu}}_{JS_l}) = \sigma^2\left(p - (p-2)^2 l\left[2c_l - c_l^2(l+2)\right]\mathbb{E}(\|X_\bullet\|^{-2})\right),$$

from which we see that the best choice is $c_l = 1/(l+2)$, in which case

$$\widehat{\boldsymbol{\mu}}_{JS_l} = \sigma^2\left(k - \frac{p-2}{l+2}\frac{\mathfrak{s}}{\|\widehat{\boldsymbol{\mu}}\|^2}\right)\widehat{\boldsymbol{\mu}}$$

certainly satisfies $\mathrm{mse}(\widehat{\boldsymbol{\mu}}_{JS_l}) < p\sigma^2$.

- As the formulas above make clear (see, for instance, (8.31)), the James–Stein estimator shrinks the sample mean toward the origin. This adjustment introduces a small amount of bias by pulling the estimate away from its observed value, yet when the number p of components in the underlying normal mean vector is sufficiently large, the reduction in variance more than compensates for this bias. The outcome is an estimator with a smaller total error, as measured by the mean squared error[31]. This principle of shrinkage represented more than a technical refinement; it marked a genuine paradigm

[31] A similar phenomenon appears in Corollary 7.27, where $\widehat{\sigma}^2_{(n+1)^{-1}}$ may be interpreted as a shrinkage of both the unbiased estimator $\widehat{\sigma}^2_{(n-1)^{-1}}$ and the maximum likelihood estimator $\widehat{\sigma}^2_{n^{-1}}$.

shift in Statistics. It laid the foundation for regularization techniques such as Ridge regression and the Lasso, which have since become essential tools in Data Science and Machine Learning, particularly in high-dimensional contexts; see Subsection 9.3 for further discussion in the setting of linear regression.

- Instead of shrinking toward the origin, it is often convenient to choose some $\boldsymbol{\nu} \in \mathbb{R}^k$ and replace (8.31) by

$$\widehat{\boldsymbol{\mu}}_{JS_{\boldsymbol{\nu}}} = \left(1 - \frac{(p-2)\sigma^2}{\|\widehat{\boldsymbol{\mu}} - \boldsymbol{\nu}\|^2}\right)(\widehat{\boldsymbol{\mu}} - \boldsymbol{\nu}) + \boldsymbol{\nu},$$

thereby performing shrinkage toward $\boldsymbol{\nu}$. The resulting estimator always satisfies $\mathrm{mse}(\widehat{\boldsymbol{\mu}}_{JS_{\boldsymbol{\nu}}}) < \sigma^2 k$. Although the optimal choice of $\boldsymbol{\nu}$ is not generally known, this formulation introduces a useful degree of flexibility. A particularly natural, data-driven option is the *grand mean vector* $\overline{X}_{\mathrm{gm}}\mathbf{1}$, where $\overline{X}_{\mathrm{gm}}$ denotes the arithmetic mean of the components of the observed sample mean X. This choice was used in the classical analysis of the baseball batting averages data set by [EM77]. □

Remark 8.23. By rewriting (8.27) as

$$\mathrm{cov}(\widehat{\theta})\mathscr{F}(\theta) \geq \mathrm{Id}_n,$$

it is patent the resemblance of the Cramér-Rao lower bound to the uncertainty principle in Quantum Mechanics. □

8.3. **Asymptotic normality of ML estimators.** We now check that under suitable regularity assumptions (which are too complicated to reproduce here) the ML estimator achieves the Cramér-Rao lower bound as the sample size n grows indefinitely, which follows from the fact that consistent ML estimators are asymptotically normal (in the sense of Definition 7.12), with their asymptotic covariance σ_θ^2 determined by the (inverse of the) Fisher information matrix. As usual we consider an infinite family $X_j \sim \psi_\theta$ of i.i.d. random variables, so that for each n the log-likelihood of $X^{[n]} = (X_1, \cdots, X_n)$ is given by

$$l^{(n)}(\mathbf{x}; \theta) = \sum_{j=1}^{n} \ln \psi_\theta(x_j), \tag{8.34}$$

where $\theta \in \Theta$ is the true (but unknown) parameter. For simplicity, let us assume that $\Theta \subset \mathbb{R}$ (the unidimensional case) so that $\mathscr{F}(\theta) > 0$ is the Fisher information. For each n let $\widehat{\theta}_n$ be the corresponding (and unique!) ML estimator so that

$$\frac{d}{d\theta} l^{(n)}(\mathbf{x}; \widehat{\theta}_n) = 0. \tag{8.35}$$

Theorem 8.24. *(Asymptotic normality) Under the conditions above, if $\widehat{\theta}_n$ is consistent (in the sense of Definition 7.9) then*

$$\sqrt{n}(\widehat{\theta}_n - \theta) \to \mathcal{N}(0, \mathscr{F}_{(1)}(\theta)^{-1}) \tag{8.36}$$

in distribution (with respect to θ), where $\mathscr{F}_{(1)}$ is the Fisher information of a single observation (say, X_1). As a consequence,

$$\widehat{\theta}_n \approx_{n\to+\infty} \mathcal{N}(\theta, \mathscr{F}_{(n)}(\theta)^{-1}), \tag{8.37}$$

where $\mathscr{F}_{(n)} = n\mathscr{F}_{(1)}$ is the Fisher information of the whole sample $X^{[n]}$.

Remark 8.25. (Asymptotic efficiency) It follows from the asymptotic normality established in Theorem 8.24 that, as $n \to +\infty$,

$$n\,\mathrm{var}(\widehat{\theta}_n) \;\longrightarrow\; \mathscr{F}_{(1)}(\theta)^{-1}.$$

Since the limiting value coincides with the Cramér–Rao lower bound (8.27) for the variance of any sufficiently regular (though not necessarily asymptotically normal) unbiased estimator of θ based on a single observation, this property is usually referred to as *asymptotic efficiency*. Equivalently, the fluctuations of $\widehat{\theta}_n$ around θ, as measured by its standard deviation, decay at the expected rate $n^{-1/2}$, with proportionality constant given by the reciprocal of the square root of the Fisher information for a single observation. More generally, if $g : \Theta \subset \mathbb{R} \to \mathbb{R}$ is a C^1 function with nowhere-vanishing derivative, then (8.36), together with the delta method (Proposition 7.15), yields

$$\sqrt{n}\big(g(\widehat{\theta}_n) - g(\theta)\big) \xrightarrow{d} \mathcal{N}\big(0,\ |g'(\theta)|^2\, \mathscr{F}_{(1)}(\theta)^{-1}\big)\,.$$

This shows that the asymptotic variance of the transformed estimator $g(\widehat{\theta}_n)$ attains the Cramér–Rao lower bound (8.27) within the class $\mathcal{E}_g$. Hence, $g(\widehat{\theta}_n)$ is asymptotically efficient for estimating $g(\theta)$. □

Remark 8.26. (Large sample confidence intervals for θ via the "consistency trick") From (8.37) we know that

$$P(a \le \widehat{\theta}_n \le b) \approx_{n\to+\infty} \sqrt{\frac{\mathcal{F}_{(n)}(\theta)}{2\pi}} \int_a^b e^{-\frac{\mathcal{F}_{(n)}(\theta)(x-\theta)^2}{2}}\, dx,$$

but we can rely on Theorem 2.25 to replace θ by $\widehat{\theta}_n$ in the right-hand side because $\widehat{\theta}_n \to \theta$ in probability (consistency), so as to obtain

$$P(a \le \widehat{\theta}_n \le b) \approx_{n\to+\infty} \sqrt{\frac{\mathcal{F}_{(n)}(\widehat{\theta}_n)}{2\pi}} \int_a^b e^{-\frac{\mathcal{F}_{(n)}(\widehat{\theta}_n)(x-\widehat{\theta}_n)^2}{2}}\, dx.$$

The key point here is that the right-hand side depends solely on sample data, and only through the estimator $\widehat{\theta}$. In the language of confidence intervals of Subsection 7.3, this translates into

$$\theta \in \left[\widehat{\theta}_n \mp \frac{z_{1-\delta/2}}{\sqrt{\mathcal{F}_{(n)}(\widehat{\theta}_n)}}\right] \text{ with prob. } \approx 1-\delta, \tag{8.38}$$

the "large sample" estimate for θ. □

Remark 8.27. The consistency requirement in Theorem 8.24 may be often justified under suitable regularity assumptions on the underlying pdf's, which in particular apply to the ML estimator $\widehat{\sigma}^2_{n^{-1}}$ in Example 8.5 [NM94, Theorem 2.5]. We may also directly retrieve the consistency of $\widehat{\sigma}^2_{n^{-1}}$ as follows. First note from (7.23) that

$$\widehat{\sigma}^2_{n^{-1}} = \frac{1}{n}\sum_{j=1}^n \sigma^2 \left(\frac{X_j-\mu}{\sigma}\right)^2 - (\overline{X}_n - \mu)^2. \tag{8.39}$$

Also, recalling that X_j is drawn from a normal population, $\sigma^{-1}(X_j-\mu) \sim \mathcal{N}(0,1)$ implies that $\sigma^{-2}(X_j-\mu)^2 \sim \chi^2_1$ by Corollary 4.25 and hence

$$\mathbb{E}\left(\sigma^2\left(\frac{X_j-\mu}{\sigma}\right)^2\right) = \sigma^2$$

by Corollary 4.21. Thus, LLN (Theorem 6.2) applies to ensure that

$$\frac{1}{n}\sum_{j=1}^n \sigma^2 \left(\frac{X_j-\overline{X}_n}{\sigma}\right)^2 \xrightarrow{p} \sigma^2.$$

On the other hand, it also follows from LLN that $(\overline{X}_n-\mu)^2 \xrightarrow{p} 0$. Thus, $\widehat{\sigma}^2_{n^{-1}} \xrightarrow{p} \sigma^2$ by (8.39). Another approach to this same conclusion follows by taking $c = n^{-1}$ in (7.25) to check that $\mathrm{mse}(\widehat{\sigma}^2_{n^{-1}}) \to 0$ as $n \to +\infty$, so that consistency follows by Proposition 7.11. In fact, any of these methods may be adapted to check that the ML

estimators in Examples 8.7 and 8.8 above are consistent as well. We also note that, in general, Proposition 7.13 applies to ensure that consistency is a necessary condition for asymptotic normality. □

Proof. (of Theorem 8.24) Set $\widetilde{l}^{(n)} = n^{-1}l^{(n)}$ and note that by (8.35) and the Mean Value Theorem,

$$0 = \frac{d}{d\theta}\widetilde{l}^{(n)}(\widehat{\theta}_n) = \frac{d}{d\theta}\widetilde{l}^{(n)}(\theta) + \frac{d^2}{d\theta^2}\widetilde{l}^{(n)}(\theta_n^\bullet)(\widehat{\theta}_n - \theta), \tag{8.40}$$

for some $\theta_n^\bullet$ lying between $\widehat{\theta}_n$ and θ. A computation shows that for any θ we have

$$\frac{d^2}{d\theta^2}\widetilde{l}^{(n)}(\theta) = \frac{1}{n}\sum_{j=1}^{n}\left(\frac{d^2}{d\theta^2}\ln\psi_\theta(X_j)\right) \to \mathbb{E}\left(\frac{d^2}{d\theta^2}\ln\psi_\theta(X_1)\right),$$

where the convergence is in probability by LLN. Since $\widehat{\theta}_n \to \theta$ in probability, we conclude that

$$\frac{d^2}{d\theta^2}\widetilde{l}^{(n)}(\theta_n^\bullet) \to \mathbb{E}\left(\frac{d^2}{d\theta^2}\ln\psi_\theta(X_1)\right) = -\mathscr{F}_{(1)}(\theta), \tag{8.41}$$

where the convergence is in probability and we used (8.24) in the last step. On the other hand,

$$\sqrt{n}\frac{d}{d\theta}\widetilde{l}^{(n)}(\theta) = \sqrt{n}\left(\frac{1}{n}\sum_{j=1}^{n}\frac{d}{d\theta}\ln\psi_\theta(X_j)\right),$$

which may be rewritten as

$$\sqrt{n}\frac{d}{d\theta}\widetilde{l}^{(n)}(\theta) = \sqrt{n}\left(\frac{1}{n}\sum_{j=1}^{n}\frac{d}{d\theta}\ln\psi_\theta(X_j) - \mathbb{E}\left(\frac{d}{d\theta}\ln\psi_\theta(X_1)\right)\right),$$

as the term within the expectation is a score (Corollary 8.15). Thus we may apply CLT to see that

$$\sqrt{n}\frac{d}{d\theta}\widetilde{l}^{(n)}(\theta) \to \mathcal{N}\left(0, \operatorname{cov}\left(\frac{d}{d\theta}\ln\psi_\theta(X_1)\right)\right) = \mathcal{N}(0, \mathscr{F}_{(1)}(\theta)), \tag{8.42}$$

where the convergence is in distribution and the last step follows from the definition of $\mathscr{F}_{(1)}$. Since (8.40) leads to

$$\sqrt{n}(\widehat{\theta}_n - \theta) = -\frac{\sqrt{n}\frac{d}{d\theta}\widetilde{l}^{(n)}(\theta)}{\frac{d^2}{d\theta^2}\widetilde{l}^{(n)}(\theta_n^\bullet)},$$

we may use Theorem 2.25, (8.41) and (8.42) to complete the proof. □

Remark 8.28. Although Theorem 8.24 is proved under i.i.d. sampling for simplicity, the same asymptotic normality holds when the sample $\{X_1, \ldots, X_n\}$ consists of independent, though not identically distributed, random variables satisfying the usual Lindeberg–Feller and regularity conditions [Ame85, Chapter 4]. In this case, (8.36) gets replaced by

$$\sqrt{n}\,(\widehat{\theta}_n - \theta) \xrightarrow{d} \mathcal{N}\left(0, \overline{\mathscr{F}}(\theta)^{-1}\right),$$

where

$$\overline{\mathscr{F}}(\theta) = \lim_{n\to+\infty}\frac{1}{n}\sum_{j=1}^{n}\mathscr{F}^{X_j}(\theta),$$

so that (8.37) becomes

$$\overline{\theta}_n \approx \mathcal{N}\left(\theta, \overline{\mathscr{F}}_{(n)}(\theta)^{-1}\right), \quad \overline{\mathscr{F}}_{(n)}(\theta) = n\overline{\mathscr{F}}(\theta).$$

This more general setting covers, in particular, the regression models considered in Chapter 9 (see Example 9.21, for instance) and in the discussion of generalized linear models in Section 10, where the responses are independent but their conditional distributions depend on covariates. □

We now discuss the implications of this theory for some of the statistical models discussed earlier.

Example 8.29. We start with

$$\widehat{\theta} = (\widehat{\theta}_1, \widehat{\theta}_2) = (\overline{X}_n, \widehat{\sigma}^2_{n^{-1}}),$$

the ML estimator for the bi-dimensional parameter $(\theta_1, \theta_2) = (\mu, \sigma^2)$ coming from a normal population as in Example 8.5 above. We first look only at $\widehat{\theta}_2$, which amounts to declaring that μ is known. As observed in Remark 8.27, this estimator is consistent and hence asymptotically normal by Theorem 8.24. In order to determine the associated limiting normal distribution by means of Theorem 8.24 we need to recall the corresponding Fisher information. From (8.21) with $n = 1$,

$$\mathscr{F}_{(1)}(\theta_2) = \mathscr{F}_{(1)}(\theta)_{22} = \frac{1}{2\theta_2^2}, \Longrightarrow \mathscr{F}_{(n)}(\theta_2) = \frac{n}{2\theta_2^2}, \tag{8.43}$$

so (8.36) and (8.37) apply to give

$$\sqrt{n}(\widehat{\sigma}^2_{n^{-1}} - \sigma^2) \xrightarrow{d} \mathcal{N}(0, 2\sigma^4) \tag{8.44}$$

and

$$\widehat{\sigma}^2_{n^{-1}} \approx_{n\to+\infty} \mathcal{N}(\sigma^2, 2\sigma^4/n).$$

Thus, in view of (8.43), (8.38) translates into

$$\sigma^2 \in \left[\left(1 - \sqrt{\frac{2}{n}} z_{1-\delta/2}\right)\widehat{\sigma}^2_{n^{-1}}, \left(1 + \sqrt{\frac{2}{n}} z_{1-\delta/2}\right)\widehat{\sigma}^2_{n^{-1}}\right] \text{ with prob. at least } 1-\delta. \tag{8.45}$$

We next consider the bi-dimensional case $\theta = (\theta_1, \theta_2)$. Again by (8.21),

$$\mathscr{F}_{(1)}(\theta_1, \theta_2) = \begin{pmatrix} \sigma^2 & 0 \\ 0 & 2\sigma^4 \end{pmatrix},$$

which gives

$$\sqrt{n}\left(\begin{pmatrix} \widehat{\theta}_1 \\ \widehat{\theta}_2 \end{pmatrix} - \begin{pmatrix} \mu \\ \sigma^2 \end{pmatrix}\right) \xrightarrow{d} \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2 & 0 \\ 0 & 2\sigma^4 \end{pmatrix}\right), \tag{8.46}$$

or equivalently,

$$\begin{pmatrix} \widehat{\theta}_1 \\ \widehat{\theta}_2 \end{pmatrix} \approx_{n\to+\infty} \mathcal{N}\left(\begin{pmatrix} \mu \\ \sigma^2 \end{pmatrix}, \begin{pmatrix} \sigma^2/n & 0 \\ 0 & 2\sigma^4/n \end{pmatrix}\right). \tag{8.47}$$

A key point here is that the asymptotic covariance matrix in (8.46) only depends on $\theta_2 = \sigma^2$, which allows us to proceed as in Remark 8.26: consistency permits to replace σ^2 by $\widehat{\theta}_2$ in the covariance matrix of (8.47) to obtain

$$\begin{pmatrix} \widehat{\theta}_1 \\ \widehat{\theta}_2 \end{pmatrix} \approx_{n\to+\infty} \mathcal{N}\left(\begin{pmatrix} \mu \\ \sigma^2 \end{pmatrix}, \begin{pmatrix} \widehat{\theta}_2/n & 0 \\ 0 & 2\widehat{\theta}_2^2/n \end{pmatrix}\right), \tag{8.48}$$

an asymptotic estimate in which the "fluctuation" around the center $(\widehat{\theta}_1, \widehat{\theta}_2)$ of the "confidence region" where the unknown parameter (μ, σ^2) is supposed to be (with probability at least $1-\delta$) depends on sample data, and only through the estimator $\widehat{\theta}_2 = \widehat{\sigma}^2_{n^{-1}}$. □

Example 8.30. Let us refer to the notation and terminology of Examples 7.39 and 8.6, with the (simplifying and justifiable) assumption that $\mu_X = \mu_Y = 0$, so that $\theta = (\sigma_X^2, \sigma_Y^2, \rho) \in \mathbb{R}^+ \times \mathbb{R}^+ \times (-1, 1)$. In order to determine the asymptotic behaviour of the ML estimator $\widehat{\theta}_m = (\widehat{\sigma}^2_{m^{-1}}(X), \widehat{\sigma}^2_{m^{-1}}(Y), \widehat{\rho}_m)$ for the unknown population parameter θ as $m \to +\infty$ (for a given jointly normal sample $\{X_j, Y_j\}$) we start with (8.13) and, after a somewhat tedious computation, we end up with a complicated expression for the 3×3 matrix $\nabla_{\theta\theta} l(X, Y; \theta)$ whose entries

depend *linearly* on the symbols in (8.12) evaluated on the sample, with the corresponding coefficients being algebraic on the components of θ. Using that

$$\mathbb{E}(A(X_j)) = \sigma_X^2, \quad \mathbb{E}(B(X_j, Y_j)) = \sigma_{XY} = \rho\sigma_X\sigma_Y, \quad \mathbb{E}(C(Y_j)) = \sigma_Y^2,$$

and (8.24) we conclude that the corresponding Fisher information matrix is

$$\mathscr{F} = \frac{1}{1-\rho^2}\begin{pmatrix} \frac{2-\rho^2}{4\sigma_X^4} & -\frac{\rho^2}{4\sigma_X^2\sigma_Y^2} & -\frac{\rho}{2\sigma_X^2} \\ -\frac{\rho^2}{4\sigma_X^2\sigma_Y^2} & \frac{2-\rho^2}{4\sigma_Y^4} & -\frac{\rho}{2\sigma_Y^2} \\ -\frac{\rho}{2\sigma_X^2} & -\frac{\rho}{2\sigma_Y^2} & \frac{1+\rho^2}{1-\rho^2} \end{pmatrix},$$

so that

$$\mathscr{F}^{-1} = \begin{pmatrix} 2\sigma_X^4 & 2\rho^2\sigma_X^2\sigma_Y^2 & \rho(1-\rho^2)\sigma_X^2 \\ 2\rho^2\sigma_X^2\sigma_Y^2 & 2\sigma_X^4 & \rho(1-\rho^2)\sigma_Y^2 \\ \rho(1-\rho^2)\sigma_X^2 & \rho(1-\rho^2)\sigma_Y^2 & (1-\rho^2)^2 \end{pmatrix}.$$

From this and Theorem 8.24 we thus derive not only that

$$\sqrt{m}\left(\widehat{\sigma}^2_{m^{-1}}(X) - \sigma_X^2\right) \xrightarrow{d} \mathcal{N}(0, 2\sigma_X^2), \quad \sqrt{m}\left(\widehat{\sigma}^2_{m^{-1}}(Y) - \sigma_Y^2\right) \xrightarrow{d} \mathcal{N}(0, 2\sigma_Y^2),$$

which are fully compatible with (8.44), but also that

$$\sqrt{m}\,(\widehat{\rho}_m - \rho) \xrightarrow{d} \mathcal{N}(0, (1-\rho^2)^2), \tag{8.49}$$

which identifies the asymptotic variance of the sample correlation coefficient $\widehat{\rho}_m$ as being $(1-\rho^2)^2$; cf. Definition 7.12. Moreover, Remark 8.25 guarantees that this asymptotic invariance equals the Crámer-Rao lower bound (8.27) for the variances of all *unbiased* estimators for ρ. We remark, however, that $\widehat{\rho}_m$ itself is *not* unbiased even though there holds

$$F(\widehat{\rho}_m) - \widehat{\rho}_m \xrightarrow{p} 0$$

for any *unbiased* estimator of the form $F(\widehat{\rho}_m)$, where F is assumed to be odd [OP58]. In particular,

$$\sqrt{m}\,(F(\widehat{\rho}_m) - \rho) \xrightarrow{d} \mathcal{N}(0, (1-\rho^2)^2).$$

As in Example 8.29 above, we may combine (8.49) with the consistency of $\widehat{\rho}_m$ to construct a large sample confidence interval for the unknown correlation coefficient ρ, namely,

$$\rho \in \left[\widehat{\rho}_m \mp z_{1-\delta/2}\frac{1-\widehat{\rho}_m^2}{\sqrt{m}}\right] \text{ with prob. } \approx 1-\delta$$

whose "fluctuation" around its center $\widehat{\rho}_m$ depends only on sample data, and through the estimator $\widehat{\rho}_m$. An alternate route is to apply the delta method (Proposition 7.15) with the *Fisher z-transformation*

$$z = g(\rho) = \frac{1}{2}\ln\left(\frac{1+\rho}{1-\rho}\right)$$

to (8.49) so as to get

$$\sqrt{m}\,(\widehat{z}_m - z) \xrightarrow{d} \mathcal{N}(0, 1),$$

which allows us to obtain large sample estimates for z in terms of the familiar normal quantiles $z_{1-\delta/2}$ and then transform them back to corresponding estimates for $\rho = \tanh z$; see [Ken46, Section 14.18] and [And03, Subsection 4.2.3] □

Example 8.31. (The coefficient of variation of a normal population) Let $g : \Theta \to \mathbb{R}$ be any smooth function satisfying $\nabla g \neq \vec{0}$ everywhere, where $\Theta = \mathbb{R} \times \mathbb{R}^+$ is the parameter space of a normal population as above; cf.

Example 8.5. Recall that $\theta = (\theta_1, \theta_2) = (\mu, \sigma^2)$ in this case. It then follows from (8.46) and the multi-dimensional version of the delta method in Proposition 7.15 that

$$\sqrt{n}\left(g\begin{pmatrix}\widehat{\theta}_1\\ \widehat{\theta}_2\end{pmatrix} - g\begin{pmatrix}\mu\\ \sigma^2\end{pmatrix}\right) \xrightarrow{d} \mathcal{N}\left(0, \nabla g(\theta)^\top \begin{pmatrix}\sigma^2 & 0\\ 0 & 2\sigma^4\end{pmatrix} \nabla g(\theta)\right). \tag{8.50}$$

Assuming that $\mu \neq 0$, we may apply this to $g(\theta) = \sqrt{\theta_2}/\theta_1 = \sigma/\mu$, the *coefficient of variation*; cf. (4.15). Since

$$\nabla g\left(\mu, \sigma^2\right)^\top = \left(-\frac{\sigma}{\mu^2}, \frac{1}{2\mu\sigma}\right)$$

we obtain that the corresponding estimator, $\widehat{\sigma}_{n^{-1}}/\overline{X}_n$, is asymptotically normal,

$$\sqrt{n}\left(\frac{\widehat{\sigma}_{n^{-1}}}{\overline{X}_n} - \frac{\sigma}{\mu}\right) \xrightarrow{d} \mathcal{N}\left(0, \frac{\sigma^2}{\mu^2}\left(\frac{1}{2} + \frac{\sigma^2}{\mu^2}\right)\right),$$

with the asymptotic variance depending on σ/μ itself. As usual, we may use consistency to get

$$\frac{\sigma}{\mu} \in \left[\frac{\widehat{\sigma}_{n^{-1}}}{\overline{X}_n} \mp z_{1-\delta/2}\frac{\widehat{\sigma}_{n^{-1}}}{\sqrt{n}\overline{X}_n}\left(\frac{1}{2} + \frac{\widehat{\sigma}^2_{n^{-1}}}{\overline{X}^2_n}\right)^{1/2}\right] \text{ with prob. } \approx 1-\delta,$$

a large sample confidence interval estimate for the coefficient of variation. □

Example 8.32. (MLE for a Gamma population) It follows from (4.23) that the log-likelihood function of a Gamma distribution $\mathsf{Gamma}(\alpha, \lambda)$ is

$$l(\mathbf{x}; \theta) = n\left(\lambda \ln \alpha - \ln \Gamma(\lambda) + (\lambda - 1)\overline{\ln x} - \alpha\overline{x}\right), \quad \theta = (\alpha, \lambda), \tag{8.51}$$

where $\overline{\ln x}$ is the arithmetic mean of $\{\ln x_1, \cdots, \ln x_n\}$; recall that $x_j > 0$ for each j. Hence, the score vector is

$$s(\mathbf{x}; \theta) = n\begin{pmatrix}\frac{\lambda}{\alpha} - \overline{x}\\ \ln \alpha - \psi(\lambda) + \overline{\ln x},\end{pmatrix}, \quad \psi(\lambda) = \frac{d}{d\lambda}\ln\Gamma(\lambda), \tag{8.52}$$

so the ML estimator $\widehat{\theta} = (\widehat{\alpha}, \widehat{\gamma})$ satisfies

$$\begin{cases} \frac{\widehat{\lambda}}{\widehat{\alpha}} &= \overline{x}\\ \psi(\widehat{\lambda}) - \ln\widehat{\alpha} &= \overline{\ln x}\end{cases} \tag{8.53}$$

Note that trying to find a solution for this system in closed form is out of question so possible strategies here are:

- to use our favorite optimization package to find
$$\widehat{\theta} = \mathrm{argmax}_\theta l(\mathbf{x}; \theta)$$
starting from (8.51);
- to use the first equation in (8.53) to eliminate $\widehat{\alpha}$ in the second equation, *numerically* solve for $\widehat{\lambda}$ in the resulting equation, namely,
$$\psi(\widehat{\lambda}) - \ln\widehat{\lambda} = \overline{\ln x} - \ln\overline{x}, \tag{8.54}$$
and then replacing the result back in the first equation in order to get $\widehat{\alpha}$[32].

[32]That a unique solution $\widehat{\lambda}$ to (8.54) exists for any given x is a consequence of the facts that i) $\overline{\ln x} < \ln\overline{x}$ if each $x_j > 0$; ii) the function $\lambda \mapsto \Psi(\lambda) = \psi(\lambda) - \ln\lambda$ is monotone continuous and satisfies

$$\lim_{\lambda\to 0}\Psi(\lambda) = -\infty \text{ and } \lim_{\lambda\to-\infty}\Psi(\lambda) = 0.$$

In any case, with the ML estimator so determined, we may proceed to compute the associated Fisher information matrix by means of (8.52) and (8.24):

$$\mathscr{F}_{(n)}(\theta) = n \begin{pmatrix} \lambda/\alpha^2 & -1/\alpha \\ -1/\alpha & \psi_1(\lambda) \end{pmatrix} = n\mathscr{F}_{(1)}(\theta), \quad \psi_1 = d\psi/d\lambda.$$

It is not hard to check that $\det \mathscr{F}_{(1)}(\theta) = (\lambda\psi_1(\lambda) - 1)/\alpha^2 > 0$, so Theorem 8.24 gives asymptotic normality for $\widehat{\theta}_n$:

$$\sqrt{n}\left(\begin{pmatrix} \widehat{\alpha}_n \\ \widehat{\lambda}_n \end{pmatrix} - \begin{pmatrix} \alpha \\ \lambda \end{pmatrix}\right) \xrightarrow{d} \mathcal{N}\left(\vec{0}, \mathscr{F}_{(1)}^{-1}\right) = \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \frac{1}{\lambda\psi_1(\lambda) - 1}\begin{pmatrix} \alpha^2\psi_1(\lambda) & \alpha \\ \alpha & \lambda \end{pmatrix}\right).$$

As usual, may combine this with consistency in order to obtain

$$\begin{pmatrix} \widehat{\alpha}_n \\ \widehat{\lambda}_n \end{pmatrix} \approx_{n\to+\infty} \mathcal{N}\left(\begin{pmatrix} \alpha \\ \lambda \end{pmatrix}, \mathscr{F}_{(n)}^{-1}\right) = \mathcal{N}\left(\begin{pmatrix} \alpha \\ \lambda \end{pmatrix}, \frac{1}{n(\widehat{\lambda}_n\psi_1(\widehat{\lambda}_n) - 1)}\begin{pmatrix} \widehat{\alpha}_n^2\psi_1(\widehat{\lambda}_n) & \widehat{\alpha}_n \\ \widehat{\alpha}_n & \widehat{\lambda}_n \end{pmatrix}\right), \tag{8.55}$$

which may be used to construct not only large sample confidence intervals for α and λ (separately) but also large sample confidence regions for the whole vector parameter θ; see Remark 8.33 below for this latter kind of construction.

Remark 8.33. (Confidence region for the unknown parameter θ via asymptotic normality of the ML estimator) Starting with the (possibly multivariate) version of (8.37), where $\theta \in \mathbb{R}^p$, $p \geq 1$, and $\mathscr{F}_{(n)}(\theta)$ is a $p\times p$ symmetric, positive definite matrix, consistency of the ML estimator $\widehat{\theta}_n$ leads to the asymptotic normality relation

$$\widehat{\theta}_n \approx_{n\to+\infty} \mathcal{N}(\theta, \mathscr{F}_{(n)}(\widehat{\theta}_n)^{-1}),$$

from which (8.55) is a rather special case (with $p = 2$). In order to extract from this a confidence region for the unknown vector parameter θ, let us write $\mathscr{F}_{(n)}(\widehat{\theta}_n) = A^\top A$ so that Corollary 4.7 gives

$$A(\widehat{\theta}_n - \theta) \sim \mathcal{N}(0, \mathrm{Id}_p)$$

and hence

$$(\widehat{\theta}_n - \theta)^\top \mathscr{F}_{(n)}(\widehat{\theta}_n)(\widehat{\theta}_n - \theta) = \|A(\widehat{\theta}_n - \theta)\|^2 \sim \chi_p^2.$$

In other words, the quadratic form in the left-hand side is a pivotal quantity to which the standard method may be applied: if we recall the definition of the χ^2-quantile in (7.62) then

$$P\left((\widehat{\theta}_n - \theta)^\top \mathscr{F}_{(n)}(\widehat{\theta}_n)(\widehat{\theta}_n - \theta) \leq \chi_{p,1-\alpha}^2\right) \approx 1 - \alpha.$$

Since $\mathscr{F}_{(n)}(\widehat{\theta}_n)$ is positive definite, the random confidence region where θ is supposed to lie (within the given confidence level) is ellipsoidal in nature, with its size, shape, and orientation being completely determined by the totality of the elements of $\mathscr{F}_{(n)}(\widehat{\theta}_n)$. Moreover, since its construction takes into account the possible correlations among the various components of $\widehat{\theta}_n$, as encoded in the off-diagonal elements of the asymptotic covariance matrix $\mathscr{F}_{(n)}(\widehat{\theta}_n)^{-1}$, in such cases it certainly encloses a much tighter volume than the p-cube which is the product of the separate confidence intervals for the entries of θ.

Remark 8.34. (Akaike Information Criterion) The information-theoretic motivation behind maximum likelihood estimation discussed in Remark 8.4 leads to an efficient way of measuring the goodness of fit of the model $\psi_{\widehat{\theta}}$, where $\widehat{\theta} = \widehat{\theta}(X)$ denotes the ML estimator associated with the statistical model

$$X_1, \ldots, X_n \sim \psi_\theta, \qquad \theta \in \Theta \subset \mathbb{R}^p.$$

More precisely, we wish to assess how effective $\psi(Z;\widehat{\theta})$ is at predicting $\psi(Z;\theta_0)$, where θ_0 is the unknown true parameter and $Z \sim \psi_{\theta_0}$ is independent of the sample X. The heuristic discussion leading to Definition 8.1, which emphasizes the minimization of the Kullback-Leibler divergence, suggests that the *expected log-likelihood*,

$$m(\theta) := n\,\mathbb{E}_Z(\ln \psi(Z;\theta)) = c_n - nD_{\theta_0}^{KL}(\theta),$$

is the quantity to be maximized. This naturally leads to considering $m(\widehat{\theta})$ as a measure of goodness of fit. Since this quantity still depends on the sample X, we pass to the *mean expected log-likelihood* $\mathbb{E}_X(m(\widehat{\theta}))$, which should be viewed as an unknown quantity to be estimated by the purely data-driven *maximized log-likelihood* $l(\widehat{\theta})$, where $l(\theta) := l(X;\theta)$ denotes the log-likelihood evaluated at the sample. It turns out, however, that this estimation carries a bias. A neat geometric interpretation for this bias is available (see Remark 8.35), but for present purposes it is convenient to make it explicit by writing

$$\begin{aligned}\text{bias} &:= \mathbb{E}_X\Big(l(\widehat{\theta})\Big) - \mathbb{E}_X\Big(m(\widehat{\theta})\Big) \\ &= \Big(\mathbb{E}_X\Big(l(\widehat{\theta})\Big) - m(\theta_0)\Big) - \Big(\mathbb{E}_X\Big(m(\widehat{\theta})\Big) - m(\theta_0)\Big),\end{aligned}$$

which shows that it arises from the fact that both $m(\widehat{\theta})$ and $l(\widehat{\theta})$ are biased as estimators of $m(\theta_0)$. Although the exact biases are generally out of reach for fixed n, they may be determined in the asymptotic regime with relatively little effort. To this end, we first expand $m(\theta)$ around the true value $\theta = \theta_0$, obtaining

$$\begin{aligned}m(\theta) &\approx m(\theta_0) + n\,\mathbb{E}_Z(\nabla_\theta \ln\psi(Z;\theta_0))^\top(\theta-\theta_0) \\ &\quad + \frac{1}{2}n(\theta-\theta_0)^\top\mathbb{E}_Z(\nabla_{\theta\theta}\ln\psi(Z;\theta_0))(\theta-\theta_0) \\ &= m(\theta_0) - \frac{1}{2}(\theta-\theta_0)^\top\mathscr{F}_{(n)}(\theta_0)(\theta-\theta_0),\end{aligned}$$

where we used Corollary 8.15 in the last step. Since $\widehat{\theta} \xrightarrow{p} \theta_0$ by consistency, we may take $\theta = \widehat{\theta}$ for n sufficiently large. Arguing as in Remark 8.33, we then have

$$(\widehat{\theta}-\theta_0)^\top\mathscr{F}_{(n)}(\theta_0)(\widehat{\theta}-\theta_0) \xrightarrow{d} \chi^2_p,$$

so that, upon taking expectations,

$$\mathbb{E}_X\Big(m(\widehat{\theta})\Big) \approx m(\theta_0) - \frac{p}{2}, \tag{8.56}$$

which indicates a bias loss of $p/2$ when using $m(\widehat{\theta})$ to estimate $m(\theta_0)$. On the other hand, expanding $l(\theta) = l(X;\theta)$ around $\theta = \widehat{\theta}$ yields

$$\begin{aligned}l(\theta) &\approx l(\widehat{\theta}) + \nabla_\theta l(\widehat{\theta})^\top(\theta-\widehat{\theta}) \\ &\quad + \frac{1}{2}(\theta-\widehat{\theta})^\top\nabla_{\theta\theta}l(\widehat{\theta})(\theta-\widehat{\theta}) \\ &= l(\widehat{\theta}) + \frac{1}{2}(\theta-\widehat{\theta})^\top\nabla_{\theta\theta}l(\widehat{\theta})(\theta-\widehat{\theta}),\end{aligned}$$

since the first-order term vanishes by the defining property of $\widehat{\theta}$ as a maximizer of l. By LLN,

$$\frac{1}{n}\nabla_{\theta\theta}l(\theta_0) = \frac{1}{n}\sum_{j=1}^n \nabla_{\theta\theta}\ln\psi(X_j;\theta_0) \xrightarrow{p} -\mathscr{F}_{(1)}(\theta_0),$$

and consistency therefore implies

$$\nabla_{\theta\theta}l(\widehat{\theta}) \xrightarrow{p} -\mathscr{F}_{(n)}(\theta_0).$$

Consequently,

$$l(\theta) \approx l(\widehat{\theta}) - \frac{1}{2}(\theta-\widehat{\theta})^\top\mathscr{F}_{(n)}(\theta_0)(\theta-\widehat{\theta}).$$

Taking $\theta = \theta_0$ and arguing as above, we obtain

$$\mathbb{E}_X\Big(l(\widehat{\theta})\Big) = \mathbb{E}_X(l(\theta_0)) + \frac{p}{2}. \tag{8.57}$$

However,

$$\mathbb{E}_X(l(\theta_0)) = \mathbb{E}_X\left(\sum_{j=1}^{n} \ln \psi(X_j;\theta_0)\right) = n\,\mathbb{E}_Z(\ln \psi(Z;\theta_0)) = m(\theta_0),$$

and substituting this into (8.57) yields

$$\mathbb{E}_X\Big(l(\widehat{\theta})\Big) \approx m(\theta_0) + \frac{p}{2}, \tag{8.58}$$

which indicates a bias gain of $p/2$ when using $l(\widehat{\theta})$ to estimate $m(\theta_0)$. Combining (8.56) and (8.58), we conclude that

$$\mathbb{E}_X\Big(l(\widehat{\theta})\Big) \approx \mathbb{E}_X\Big(m(\widehat{\theta})\Big) + p,$$

so that $l(\widehat{\theta})$, as an estimator of $\mathbb{E}_X(m(\widehat{\theta}))$, has asymptotic bias equal to p, the number of free parameters in the model. After an appropriate normalization, this leads to the definition of the *Akaike information criterion*:

$$\mathrm{AIC}(X) = -2\,l(X;\widehat{\theta}) + 2p. \tag{8.59}$$

As an illustration, consider the normal model of Example 8.5. In this case,

$$\begin{aligned}\mathrm{AIC}(X) &= -2\left(-\frac{n}{2}\ln(2\pi\widehat{\sigma}^2_{n^{-1}}) - \frac{1}{2\widehat{\sigma}^2_{n^{-1}}}\sum_{j=1}^{n}(X_j - \overline{X}_n)^2\right) + 2\cdot 2\\ &= -2\left(-\frac{n}{2}\ln(2\pi\widehat{\sigma}^2_{n^{-1}}) - \frac{n}{2}\right) + 4,\end{aligned}$$

so that

$$\mathrm{AIC}(X) = n\ln(2\pi\widehat{\sigma}^2_{n^{-1}}) + n + 4.$$

Thus, according to this criterion, the goodness of fit of the model is entirely determined by $\widehat{\sigma}^2_{n^{-1}}$, the ML estimator of the variance, which is consistent with the computation of the asymptotic covariance matrix of the ML estimator of the model parameter $(\theta_1,\theta_2) = (\mu,\sigma^2)$ in (8.48). □

Remark 8.35. (Kullback–Leibler divergence, Fisher information, and AIC) For any sufficiently regular parametric model with true parameter θ_0, the Kullback–Leibler divergence from Definition 8.2 admits the local expansion, as $\theta \to \theta_0$,

$$D^{KL}_{\theta_0}(\theta) = \frac{1}{2}(\theta-\theta_0)^{\top}\mathscr{F}_{(1)}(\theta_0)(\theta-\theta_0) + o\big(\|\theta-\theta_0\|^2\big),$$

where $\mathscr{F}_{(1)}(\theta_0)$ denotes the Fisher information matrix associated with a single observation $Y \sim \psi_\theta$. Indeed, we have

$$\left(\nabla_\theta D^{KL}_{\theta_0}\right)(\theta) = -\mathbb{E}_{\theta_0}\left(\nabla_\theta \ln\psi(Y;\theta)\right), \quad \left(\nabla_{\theta\theta} D^{KL}_{\theta_0}\right)(\theta) = -\mathbb{E}_{\theta_0}\left(\nabla_{\theta\theta}\ln\psi(Y;\theta)\right),$$

so that Corollary 8.15 gives

$$\left(\nabla_\theta D^{KL}_{\theta_0}\right)(\theta_0) = 0, \quad \left(\nabla_{\theta\theta} D^{KL}_{\theta_0}\right)(\theta_0) = \mathscr{F}_{(1)}(\theta_0),$$

as desired; for a full treatment of this local identification of the Fisher information matrix with the Hessian of the Kullback–Leibler divergence we refer to [CU14, Chapter 4]. On the other hand, if $\widehat{\theta}$ is the corresponding MLE based on an i.i.d. sample of size n, we may argue as in Remark 8.33 to see that

$$n(\widehat{\theta}-\theta_0)^{\top}\mathscr{F}_{(1)}(\theta_0)(\widehat{\theta}-\theta_0) \xrightarrow{d} \chi^2_p,$$

which implies

$$\mathbb{E}\Big((\widehat{\theta}-\theta_0)^{\top}\mathscr{F}_{(1)}(\theta_0)(\widehat{\theta}-\theta_0)\Big) = \frac{p}{n} + o\left(\frac{1}{n}\right)$$

and therefore

$$\mathbb{E}_{\theta_0}\Big(2nD^{KL}_{\theta_0}(\widehat{\theta})\Big) = p + o(1)\,.$$

Now recall from Remark 8.34 that Akaike's Information Criterion (AIC) is motivated by the problem of selecting, among competing parametric models, the one whose fitted distribution is closest to the true data–generating distribution in the sense that the Kullback–Leibler divergence $D_{\theta_0}^{KL}(\widehat{\theta})$ should be minimized. In this way, its final expression (8.59) may be interpreted as an approximately unbiased estimator of twice the expected Kullback–Leibler divergence between the fitted model and the true distribution, with the computation above indicating that the penalty term reflects the local curvature of the KL divergence as governed by Fisher information. □

Remark 8.36. Comparing Theorem 8.24 with Theorem 6.5, one observes that the "sample universality" so valued in the latter classical result is irretrievably lost. In essence, Theorem 8.24 states that for each choice of log-likelihood function as in (8.34), itself fully determined by the underlying density $\psi(\cdot, \theta)$ through (8.7) and (8.2), the maximum likelihood method produces an estimator of the form (8.9) to which a corresponding "limit theorem" applies, as in (8.36). Unlike the Central Limit Theorem, therefore, Theorem 8.24 is inherently model-dependent. □

9. THE METHOD OF LEAST SQUARES

If $\theta_2 = \sigma^2$ is known, maximizing l in (8.11) is equivalent to minimizing

$$\theta_1 \mapsto \frac{1}{2}\sum_j (x_j - \theta_1)^2,$$

which furnishes a variational characterization of the arithmetic mean $n^{-1}\sum_j x_j$. This is of course a manifestation of the Method of Least Squares, a celebrated procedure which provides a solution to the following kind of problem. Let us arrange the outcome of n measurements of p features (regressors, independent/explanatory variables, predictors, covariates, etc.) of a population by means of the $n \times (p+1)$-matrix

$$\mathfrak{x} = \left(\begin{array}{c|c} & \mathbf{x}_1 \\ \mathbf{1} & \vdots \\ & \mathbf{x}_n \end{array}\right)$$

where each

$$\mathbf{x}_j = (x_{j1}, \cdots, x_{jp}), \quad j = 1, \cdots, n,$$

is a row p-vector (representing the outcome of the j^{th} measurement) and $\mathbf{1}$ is the column n-vector whose entries all equal 1. If we suspect that these features relate to a response (regressand, dependent/explained variable, etc.) which has also been measured, thus yielding an n-vector $\mathbf{y}$, we may try to "predict" the response at some unknown feature by "best fitting" a (possibly non-linear) functional dependence, say $\mathbf{y} = F(\mathfrak{x})$, to the available data $(\mathfrak{x}, \mathbf{y})$[33]. The simplest choice is to postulate that F is *linear*, so that $\mathbf{y} = \mathfrak{x}\widehat{\beta}$, where $\widehat{\beta} = (\widehat{\beta}_0, \widehat{\beta}_1, \cdots, \widehat{\beta}_p)$ is determined by minimizing the corresponding quadratic objective function:

$$\widehat{\beta} = \operatorname{argmin}_\beta f(\beta), \quad f(\beta) = \frac{1}{2}\|\mathbf{y} - \mathfrak{x}\beta\|^2, \tag{9.1}$$

hence the "least squares" terminology.

This is a purely geometric problem (best fitting a hyperplane to a cloud of n points in $\mathbb{R}^p \times \mathbb{R} = \mathbb{R}^{p+1}$, where we usually assume that $p + 1 \ll n$), which can be solved by the methods of Calculus. Indeed, since

$$(\nabla f)(\beta) = -\mathfrak{x}^\top (\mathbf{y} - \mathfrak{x}\beta),$$

[33] In the modern language of Supervised Learning, the pair $(\mathfrak{x}, \mathbf{y})$ is referred to as the *training data set*, since it is used to instruct the learning algorithm on how to fit the model F [HTF09].

where $\top$ means transpose, we obtain

$$\widehat{\beta} = (\mathfrak{x}^\top \mathfrak{x})^{-1} \mathfrak{x}^\top \mathbf{y}, \tag{9.2}$$

where we assume that $\mathfrak{x}$ has full column-rank (this not only implies that $p+1 \leq n$ but also that the *Gram matrix* $\mathfrak{x}^\top \mathfrak{x}$ is symmetric and positive definite, hence invertible). Moreover, since

$$\nabla^2 f = \mathfrak{x}^\top \mathfrak{x},$$

we conclude that $\widehat{\beta}$ is the unique global minimum. Under these conditions, we then say that

$$\widehat{\mathbf{y}} = \mathfrak{x}\widehat{\beta} \tag{9.3}$$

is the *fitted vector* (that is, the vector of fitted values). Finally, we observe that $\widehat{\beta}$ is *linear* in $\mathbf{y}$ with coefficients depending on $\mathfrak{x}$.

9.1. **The statistical model behind the method of least squares.** In the examples below, we examine the statistical rationale underlying the purely data-driven minimization problem in (9.1). In line with the general estimation framework presented in Subsection 7.1, this requires introducing suitable assumptions on how the data array $(\mathbf{y}, \mathfrak{x})$ is drawn from an underlying population. These assumptions allow us to set up a statistical model in which β is treated as an unknown population parameter and $\widehat{\beta}$ is validated as an efficient estimator of β[34].

Example 9.1. (The general regression model) We start by viewing $(\mathfrak{x}, \mathbf{y})$ as the realization of a $\mathbb{R}^{n\times(p+1)+n}$-valued random vector $(\mathfrak{X}, \mathbf{Y})$ in $L^2(\Omega)$ and satisfying

- $\mathfrak{X} = (\mathbf{1}\ \mathbf{X})$, where $\mathbf{X}$ is a random $\mathbb{R}^{n\times p}$-valued vector whose realization is $\mathbf{x} = (\mathbf{x}_1, \cdots, \mathbf{x}_n)^\top$. In other words, $\mathfrak{X}$ is a random matrix whose first column is deterministic (non-random) and equals $\mathbf{1}$. Also, we assume that $\mathfrak{X}$ has full column-rank a.s.
- $\{(\mathbf{X}_{j\bullet}, \mathbf{Y}_j)\}_{j=1}^n$, where $\bullet \in \{1, \cdots, p\}$, are i.i.d. copies of the same $\mathbb{R}^p \times \mathbb{R}$-valued random vector, say $(\mathscr{X}, \mathscr{Y})$.

Under these conditions, we may initially impose the (not necessarily linear) *regression model*

$$\mathbf{Y} = F(\mathfrak{X}) + \mathbf{e}, \tag{9.4}$$

with the *regression function* $F : \mathbb{R}^{n(p+1)} \to \mathbb{R}^n$ being defined by

$$F(\mathfrak{x}) = \mathbb{E}\left(\mathbf{Y}|_{\mathfrak{X}=\mathfrak{x}}\right), \tag{9.5}$$

where we use here the notation of Subsection 3.1; see Remark 9.4 below for the justification of this choice of F, where it is shown that it minimizes the corresponding mean squared error:

$$\mathbb{E}(\|\mathbf{Y} - F(\mathfrak{X})\|^2) \leq \mathbb{E}(\|\mathbf{Y} - G(\mathfrak{X})\|^2),$$

for any $G : \mathbb{R}^{n(p+1)} \to \mathbb{R}^n$. Thus, we may view (9.4) as the definition of the random *error* $\mathbf{e}$, which by Proposition 3.14 may also be expressed as

$$\mathbf{e} = \mathbf{Y} - \mathbb{E}(\mathbf{Y}|\mathfrak{X}),$$

so that (3.14) easily implies *exogeneity*,

$$\mathbb{E}(\mathbf{e}|\mathfrak{X}) = 0, \tag{9.6}$$

and hence

$$\mathbb{E}(\mathbf{e}) = 0. \tag{9.7}$$

[34]For a critical discussion of model building and interpretation in Regression Analysis, with emphasis on the distinction between the information inherent in the data and the inferential consequences of assumptions about the sampling process that generated them, a perspective that extends to virtually any statistical analysis, see [Ber04].

More generally, again by (3.14),

$$\mathbb{E}(\mathfrak{X}^\top \mathbf{e}) = \mathbb{E}(\mathbb{E}(\mathfrak{X}^\top \mathbf{e}|\mathfrak{X})) = \mathbb{E}(\mathfrak{X}^\top \mathbb{E}(\mathbf{e}|\mathfrak{X})),$$

where we used Proposition 3.11 (7) in the last step, so that (9.6) applies to give

$$\mathbb{E}(\mathfrak{X}^\top \mathbf{e}) = 0, \tag{9.8}$$

so (9.4) decomposes $\mathbf{Y}$ as a sum of a term $F(\mathfrak{X})$ which is "explained" by $\mathfrak{X}$ and an error which has zero mean (conditioned to $\mathfrak{X}$) and is uncorrelated to (any function of) $\mathfrak{X}$[35]. Also, if we define the *error covariance function* by

$$\boldsymbol{\sigma}^2(\mathfrak{x}) := \mathrm{cov}(\mathbf{e}|_{\mathfrak{X}=\mathfrak{x}}) \overset{(9.6)}{=} \mathbb{E}(\mathbf{e}\otimes\mathbf{e}|_{\mathfrak{X}=\mathfrak{x}}), \quad \mathfrak{x}\in\mathbb{R}^{n(p+1)},$$

then

$$\mathrm{cov}(\mathbf{e}) \overset{(9.7)}{=} \mathbb{E}(\mathbf{e}\otimes\mathbf{e}) \overset{(3.14)}{=} \mathbb{E}(\mathbb{E}(\mathbf{e}\otimes\mathbf{e}|\mathfrak{X})),$$

and using Proposition 3.14,

$$\mathrm{cov}(\mathbf{e}) = \mathbb{E}(\boldsymbol{\sigma}^2(\mathfrak{X})).$$

In other words, the unconditioned error covariance equals the expected value of the conditioned error covariance. □

We now specialize the general setup above to the cases which appear more frequently in applications.

Example 9.2. (The linear regression model) The simplest of all choices for the regression function above is $F(\mathfrak{x}) = \mathfrak{x}\beta$, which gives rise to the *linear regression model*

$$\mathbf{Y} = \mathfrak{X}\beta + \mathbf{e}, \tag{9.9}$$

where

$$\begin{aligned} \beta &= \mathrm{argmin}_{\beta'\in\mathbb{R}^{p+1}} \frac{1}{n}\mathbb{E}(\|\mathbf{Y}-\mathfrak{X}\beta'\|^2) \\ &= \mathrm{argmin}_{\beta'\in\mathbb{R}^{p+1}} \mathbb{E}\left((\mathscr{Y}-\widetilde{\mathscr{X}}^\top\beta')^2\right), \quad \widetilde{\mathscr{X}} = (1,\mathscr{X}), \end{aligned} \tag{9.10}$$

provides the best linear fitting for $\mathbf{Y}$ (or $\mathscr{Y}$) in the L^2 sense. In this setting, (9.8) should be interpreted as the "projection condition" that $\mathbf{e} = \mathbf{Y} - \mathfrak{X}\beta$ should be "orthogonal" (again in the L^2 sense) to any potential linear fitting; see Remark 9.24 for an elaboration of this viewpoint. Also, (9.6) is automatically satisfied due to Remark 3.9. Now, if we apply the usual first order test from Calculus to (9.10) we find that

$$\beta = \mathbb{E}(|\widetilde{\mathscr{X}}|^2)^{-1}\mathbb{E}(\mathscr{Y}\,\widetilde{\mathscr{X}}^\top),$$

but this does not say much about the true nature of β because the joint distribution of $(\mathscr{X},\mathscr{Y})$ remains unknown, which makes the expectations intractable. Thus, this population parameter should somehow be estimated from data (a realization $(\mathbf{x},\mathbf{y})$ of $(\mathbf{X},\mathbf{Y})$) with $\widehat{\beta}$ in (9.2) being the most obvious candidate for an estimator. As it is always the case with any estimator, its efficiency only gets validated by the establishment of good inferential properties (say, by confirming that its mse is minimized within a given class of estimators and/or that it is consistent and asymptotically normal, etc.; see the general discussion in Subsection 7.1), so with this purpose in mind it is convenient to add to (9.6) the assumption of *spherical error*, which means that there exists $\sigma > 0$ such that

$$\mathrm{cov}(\mathbf{e}|_{\mathfrak{X}=\mathfrak{x}}) = \sigma^2\mathrm{Id}_n, \quad \text{independently of } \mathfrak{x}. \tag{9.11}$$

Thus,

$$\mathbb{E}(\mathbf{Y}|_{\mathfrak{X}=\mathfrak{x}}) = \mathfrak{x}\beta \quad \text{and} \quad \mathrm{cov}(\mathbf{Y}|_{\mathfrak{X}=\mathfrak{x}}) = \sigma^2\mathrm{Id}_n \tag{9.12}$$

[35]The implications of this remarkable decomposition to Regression Theory (and to Econometrics, in particular) are discussed at length in [AP09, Chapter 3].

summarize the assumptions of the linear regression model [36]. In the language of Example 9.1, (9.11) means that the random matrix $\boldsymbol{\sigma}^2(\mathfrak{X})$ is actually *constant* and equals $\sigma^2\mathrm{Id}_n$, an artifact also known as *homoscedasticity*. □

Example 9.3. (The linear regression model with a normal error) In the setting of the linear regression model (9.9), the "empirical" quadratic minimization in (9.1) may be justified via MLE under a normality assumption on the error[37]. Precisely, and in alignment with (9.6) and (9.11), let us further assume that the error $\mathbf{e}$ is such that $\{\mathbf{e}_j|_{\mathfrak{X}_j=\mathfrak{x}_j}\}_{j=1}^n$ is independent and distributed according to

$$\mathbf{e}_j|_{\mathfrak{X}_j=\mathfrak{x}_j} \sim \mathcal{N}(0,\sigma^2), \tag{9.13}$$

or equivalently,

$$\mathbf{e}|_{\mathfrak{X}=\mathfrak{x}} \sim \mathcal{N}(\vec{0},\sigma^2\mathrm{Id}_n), \tag{9.14}$$

by Proposition 4.11. It then follows from (9.9) that $\{\mathbf{Y}_j|_{\mathfrak{X}_j=\mathfrak{x}_j}\}_{j=1}^n$ is independent with

$$\mathbf{Y}_j|_{\mathfrak{X}_j=\mathfrak{x}_j} \sim \mathcal{N}\left(\sum_{k=0}^{p}\mathfrak{x}_{jk}\beta_k,\sigma^2\right). \tag{9.15}$$

In this way, we obtain an identifiable statistical model, conditionally on the observed value $\mathfrak{x}$ of $\mathfrak{X}$, in which β appears as the unknown parameter. Actually, this construction fits the extended notion of a statistical model in Remark 7.4, since the conditional distributions in (9.15) vary across observations as their expectations depend on covariates (the same observation affects the general linear model specified by (9.12)). In particular, we may apply MLE to (9.15), as in Definition 8.3, to find the corresponding estimator. Indeed, (9.15) can be succinctly written as

$$\mathbf{Y}|_{\mathfrak{X}=\mathfrak{x}} \sim \mathcal{N}(\mathfrak{x}\beta,\sigma^2\mathrm{Id}_n),$$

so the corresponding likelihood function is

$$L(\mathbf{y};\beta) = \left(2\pi\sigma^2\right)^{-n/2} e^{-\frac{\|\mathbf{y}-\mathfrak{x}\beta\|^2}{2\sigma^2}}. \tag{9.16}$$

Since the corresponding log-likelihood function to be maximized is

$$l(\mathbf{y};\beta) = -\frac{n}{2}\ln(2\pi\sigma^2) - \frac{\|\mathbf{y}-\mathfrak{x}\beta\|^2}{2\sigma^2}, \tag{9.17}$$

we see that, up to irrelevant constants (depending on σ^2, here assumed known), solving this maximization problem is equivalent to finding $\widehat{\beta}$ as in (9.1), thus confirming that MLE implies OLS under the stated assumptions. □

Although the terminology *ordinary least squares (OLS)* is sometimes reserved for the specialized linear model with normal errors in Example 9.3, we will use it more broadly to denote the framework defined by the empirical estimator $\widehat{\beta}$ in (9.1) and the assumptions in (9.12). In particular, we refer to $\widehat{\beta}$ as the OLS estimator.

Remark 9.4. The appearance of the regression function F in (9.5) may be justified by the fact that

$$\mathbb{E}(\|\mathbf{Y}-F(\mathfrak{X})\|^2) = \inf_G \mathbb{E}(\|\mathbf{Y}-G(\mathfrak{X})\|^2),$$

[36] Although the stronger assumption of the independence of $\{\mathbf{e}|_{\mathfrak{X}_j=\mathfrak{x}_j}\}_{j=1}^n$ may eventually be useful (as in Examples 9.25 and 9.26, for instance), we stress that only uncorrelatedness, as expressed by (9.11), is imposed at this point, as this already allows us to derive some nice inferential properties for $\widehat{\beta}$; cf. Propositions 9.6 and 9.7 and Remark 9.20. In any case, if $\mathbf{e}|_{\mathfrak{X}=\mathfrak{x}}$ is normally distributed, as in Example 9.3, then these assumptions (uncorrelatedness and independence) are equivalent indeed (by Corollary 4.12).

[37] As it is well-known, this connection between OLS and the normal distribution has been first observed by Gauss and Laplace [Sti90, Chapter 4].

for any $G : \mathbb{R}^{n(p+1)} \to \mathbb{R}^n$ measurable. To check this, first note that

$$\begin{aligned}\mathbb{E}(|\mathbf{Y} - G(\mathfrak{X})|^2) &= \mathbb{E}(\|\mathbf{Y} - F(\mathfrak{X}) + F(\mathfrak{X}) - G(\mathfrak{X})\|^2) \\ &= \mathbb{E}(\|\mathbf{Y} - F(\mathfrak{X})\|^2) + \mathbb{E}(\|F(\mathfrak{X}) - G(\mathfrak{X})\|^2) \\ &\quad +2\mathbb{E}(\langle \mathbf{Y} - F(\mathfrak{X}), F(\mathfrak{X}) - G(\mathfrak{X})\rangle).\end{aligned}$$

Also, by Proposition 3.11 (2) and Proposition 3.14,

$$\begin{aligned}\mathbb{E}(\langle \mathbf{Y} - F(\mathfrak{X}), F(\mathfrak{X}) - G(\mathfrak{X})\rangle) &= \mathbb{E}(\mathbb{E}(\langle \mathbf{Y} - F(\mathfrak{X}), F(\mathfrak{X}) - G(\mathfrak{X})\rangle|\mathfrak{X})) \\ &= \mathbb{E}(\langle \mathbf{Y} - F(\mathfrak{X}), F(\mathfrak{X}) - G(\mathfrak{X})\rangle|_{\mathfrak{X}=\mathfrak{x}}) \\ &= \langle F(\mathfrak{x}) - G(\mathfrak{x}), \mathbb{E}(\mathbf{Y}|_{\mathfrak{X}=\mathfrak{x}}) - F(\mathfrak{x})\rangle \\ &= 0.\end{aligned}$$

Thus,

$$\mathbb{E}(\|\mathbf{Y} - G(\mathfrak{X})\|^2) \geq \mathbb{E}(\|\mathbf{Y} - F(\mathfrak{X})\|^2),$$

with the equality holding if and only if $G(\mathfrak{X}) = F(\mathfrak{X})$ a.s. □

Remark 9.5. It has become clear from the preceding discussion that the vector $\mathfrak{X}$ can be treated in two distinct ways. One may regard it as random, in which case it directly contributes to the randomness of $\mathbf{Y}$. Alternatively, one may condition on specific realizations of $\mathfrak{X}$, thereby fixing it and allowing the randomness of $\mathbf{Y}$ to arise solely from the error term. Either viewpoint may be adopted depending on the context. From a conceptual standpoint, the distinction between these two perspectives lies in the scope of the randomness. In the fixed-design approach, common in controlled experiments, $\mathfrak{X}$ is treated as non-random and inference is made conditionally on its observed values. By contrast, in random-design settings, typical of observational studies and econometrics, $\mathfrak{X}$ is itself a random vector, and the regression coefficients describe relationships at the population level rather than merely conditional ones. While the algebraic form of estimators such as $\widehat{\beta} = (\mathfrak{x}^\top \mathfrak{x})^{-1}\mathfrak{x}^\top \mathbf{y}$ remains unchanged, the interpretation of their variance and the probabilistic framework underlying inference differ substantially. In experimental fields such as agriculture, where the investigator typically controls the values of $\mathbf{x}$ and subsequently records $\mathbf{y}$, the former perspective is more natural. By contrast, in disciplines such as Econometrics, where such control is neither possible nor desirable, $\mathfrak{X}$ is more appropriately regarded as random. As already pointed out, the algebraic form of the relevant inferential statistics is largely unaffected by the choice of viewpoint, particularly under a suitable normality assumption [RS08, Chapter 10]. Since normality will be assumed in most of what follows, it is convenient to omit explicit conditioning on $\mathfrak{X} = \mathfrak{x}$, with the understanding that $\mathfrak{X} = (\mathbf{1}\ \mathbf{X})$ should still be regarded as random whenever foundational issues are under consideration (as in Remark 9.16, for instance). Accordingly, at various points we shall use the same symbol to denote both a random variable and its observed value, in order to streamline the notation and avoid redundant symbols, an approach intended to maintain clarity while keeping the exposition concise. □

9.2. **Inference and goodness of fit for OLS.** With a statistical model for OLS at hand, we now proceed to the pertinent inferential analysis. We start by observing that although the normality assumption for the error in (9.14) is essential for interpreting OLS via MLE, good statistical properties of the associated estimator $\widehat{\beta}$ may be derived under the much less stringent assumptions of the linear regression model in Example 9.2.

Proposition 9.6. *Under the conditions of Example 9.2 there hold*

$$\mathbb{E}(\widehat{\beta}) = \beta, \quad \mathrm{cov}(\widehat{\beta}) = \sigma^2(\mathfrak{x}^\top \mathfrak{x})^{-1}. \tag{9.18}$$

As a consequence, $\widehat{\beta}$ is unbiased and $\mathrm{mse}(\widehat{\beta}) = \sigma^2 \mathrm{tr}(\mathfrak{x}^\top \mathfrak{x})^{-1}$.

Proof. With the simplifying notation suggested by Remark 9.5, we are assuming that

$$\mathbb{E}(\mathbf{e}_j) = 0 \quad \text{and} \quad \text{cov}(\mathbf{e}_j, \mathbf{e}_k) = \sigma^2 \delta_{jk}, \quad j,k = 1, \cdots, n. \tag{9.19}$$

Hence, from (9.2),

$$\widehat{\beta} = (\mathfrak{x}^\top \mathfrak{x})^{-1} \mathfrak{x}^\top (\mathfrak{x}\beta + \mathbf{e}),$$

so that

$$\widehat{\beta} = \beta + (\mathfrak{x}^\top \mathfrak{x})^{-1} \mathfrak{x}^\top \mathbf{e}, \tag{9.20}$$

which gives

$$\mathbb{E}(\widehat{\beta}) = \mathbb{E}(\beta) + (\mathfrak{x}^\top \mathfrak{x})^{-1} \mathfrak{x}^\top \mathbb{E}(\mathbf{e}) = \beta + (\mathfrak{x}^\top \mathfrak{x})^{-1} \mathfrak{x}^\top (\vec{0}) = \beta.$$

Also,

$$\begin{aligned} \text{cov}(\widehat{\beta}) &= \mathbb{E}((\widehat{\beta} - \beta) \otimes (\widehat{\beta} - \beta)) \\ &= \mathbb{E}(((\mathfrak{x}^\top \mathfrak{x})^{-1} \mathfrak{x}^\top \mathbf{e}) \otimes ((\mathfrak{x}^\top \mathfrak{x})^{-1} \mathfrak{x}^\top \mathbf{e})) \\ &= (\mathfrak{x}^\top \mathfrak{x})^{-1} \mathfrak{x}^\top \mathbb{E}(\mathbf{e} \otimes \mathbf{e}) \mathfrak{x} (\mathfrak{x}^\top \mathfrak{x})^{-1} \\ &= (\mathfrak{x}^\top \mathfrak{x})^{-1} \mathfrak{x}^\top \text{cov}(\mathbf{e}) \mathfrak{x} (\mathfrak{x}^\top \mathfrak{x})^{-1}, \end{aligned}$$

and using that $\text{cov}(\mathbf{e}) = \sigma^2 \text{Id}$ by (9.19), the result follows. □

As a first check on the efficiency of OLS estimator $\widehat{\beta}$ in (9.2), let us see how it competes with a general *linear* estimator

$$\overline{\beta} = C\mathbf{Y},$$

where C is a $(p+1) \times n$ matrix which is allowed to depend on $\mathfrak{x}$ but not on $\mathbf{Y}$. This leads to a remarkable result confirming that $\widehat{\beta}$ attains the best performance (as measured by the mse) within a natural class of estimators.

Theorem 9.7. *(Gauss-Markov) Let $\overline{\beta}$ as above be unbiased with $\mathbf{e}$ satisfying (9.19). Then $\text{cov}(\overline{\beta}) \geq \text{cov}(\widehat{\beta})$.*

Proof. We write

$$\overline{\beta} = \left((\mathfrak{x}^\top \mathfrak{x})^{-1} \mathfrak{x}^\top + D\right) \mathbf{Y},$$

where D has the same properties as C. It follows that

$$\overline{\beta} = ((\mathfrak{x}^\top \mathfrak{x})^{-1} \mathfrak{x}^\top + D)(\mathfrak{x}\beta + \mathbf{e}) = \widehat{\beta} + D\mathfrak{x}\beta + D\mathbf{e}$$

with $\mathbb{E}(D\mathbf{e}) = D\mathbb{E}(\mathbf{e}) = 0$ (by Proposition 4.9 (3)) so that

$$\mathbb{E}(\overline{\beta}) = \beta + D\mathfrak{x}\beta,$$

and letting β vary we see that a vanishing bias for $\overline{\beta}$ implies $D\mathfrak{x} = 0$. Hence,

$$\text{cov}(\overline{\beta}) = \text{cov}(\widehat{\beta}) + \text{cov}(D\mathbf{e}) + 2\text{cov}(\widehat{\beta}, D\mathbf{e}).$$

Now note that $\text{cov}(D\mathbf{e}) = D\text{cov}(\mathbf{e})D^\top = \sigma^2 DD^\top$. Moreover, using (9.20) and the fact that β is non-random,

$$\begin{aligned} \text{cov}(\widehat{\beta}, D\mathbf{e}) &= \text{cov}(\beta, D\mathbf{e}) + \text{cov}((\mathfrak{x}^\top \mathfrak{x})^{-1} \mathfrak{x}^\top \mathbf{e}, D\mathbf{e}) \\ &= \text{cov}((\mathfrak{x}^\top \mathfrak{x})^{-1} \mathfrak{x}^\top \mathbf{e}, D\mathbf{e}) \\ &= (\mathfrak{x}^\top \mathfrak{x})^{-1} \mathfrak{x}^\top \text{cov}(\mathbf{e}) D^\top \\ &= \sigma^2 (\mathfrak{x}^\top \mathfrak{x})^{-1} (D\mathfrak{x})^\top \\ &= 0. \end{aligned}$$

Thus,

$$\mathrm{cov}(\overline{\beta}) = \mathrm{cov}(\widehat{\beta}) + \sigma^2 DD^\top,$$

and the result follows because $DD^\top \geq 0$. □

Example 9.8. If we take it for granted that $\mathbf{X}$ has no influence whatsoever on $\mathbf{Y}$ then we are actually dealing with the "intercept-only" case $\beta = (\beta_0, 0, \cdots, 0)$, so we must impose $\widehat{\beta} = (\widehat{\beta}_0, 0, \cdots, 0)$ and, as expected, (9.2) gives

$$\widehat{\beta}_0 = \overline{\mathbf{Y}} = \frac{1}{n}\sum_j \mathbf{Y}_j, \tag{9.21}$$

the sample mean of $\mathbf{Y}$; in the simple linear regression case of Example 9.9 below, this is immediate from (9.22). If $w = (w_1, \cdots, w_n) \in \mathbb{R}^n$ we may consider the more general linear combination in the entries of $\mathbf{Y}$ given by

$$\widehat{\beta}_0^w = \sum_j w_j \mathbf{Y}_j = \widehat{\beta}_0 + \sum_j \left(w_j - \frac{1}{n}\right)\mathbf{Y}_j,$$

so that

$$\begin{aligned}
\mathbb{E}(\widehat{\beta}_0^w) &= \beta_0 + \sum_j \left(w_j - \frac{1}{n}\right)\mathbb{E}(\mathbf{Y}_j) \\
&= \beta_0 + \sum_j \left(w_j - \frac{1}{n}\right)(\mathfrak{x}\beta)_j \\
&= \beta_0 + \beta_0 \sum_j \left(w_j - \frac{1}{n}\right),
\end{aligned}$$

and $\widehat{\beta}_0^w$ is unbiased if and only w is a weight vector, $\sum_j w_j = 1$. Thus, Gauss-Markov applies to ensure that $\widehat{\beta}_0$ attains the best performance among all these *weighted* estimators of β_0. In particular, Gauss-Markov may be regarded as a generalization of Example 7.17. □

Example 9.9. (Simple linear regression) If $p = 1$ in Example 9.2 then $\mathfrak{X} = (\mathbf{1}, \mathbf{X})$, where $\mathbf{X} = (\mathbf{X}_{11}, \cdots, \mathbf{X}_{n1})^\top$, so that

$$\mathfrak{X}^\top \mathfrak{X} = \begin{pmatrix} n & n\overline{\mathbf{X}} \\ n\overline{\mathbf{X}} & \|\mathbf{X}\|^2 \end{pmatrix}, \quad \overline{\mathbf{X}} = \frac{1}{n}\sum_j \mathbf{X}_{j1}.$$

Since $\mathbf{X}$ is supposed not to be a multiple of $\mathbf{1}$ a.s., Cauchy-Schwartz implies that

$$\det \mathfrak{X}^\top \mathfrak{X} = n\|\mathbf{X}\|^2 - n^2\overline{\mathbf{X}}^2 > 0,$$

so that $\mathfrak{X}^\top \mathfrak{X}$ is invertible and

$$(\mathfrak{X}^\top \mathfrak{X})^{-1} = \frac{1}{n\|\mathbf{X}\|^2 - n^2\overline{\mathbf{X}}^2}\begin{pmatrix} \|\mathbf{X}\|^2 & -n\overline{\mathbf{X}} \\ -n\overline{\mathbf{X}} & n \end{pmatrix}.$$

A little computation using (9.2) then gives

$$\widehat{\beta} := \begin{pmatrix} \widehat{\beta}_0 \\ \widehat{\beta}_1 \end{pmatrix} = \begin{pmatrix} \overline{\mathbf{Y}} - \widehat{\beta}_1\overline{\mathbf{X}} \\ S_{\mathbf{XY}}/S_{\mathbf{XX}} \end{pmatrix}, \tag{9.22}$$

where

$$\overline{\mathbf{Y}} = \frac{1}{n}\sum_j \mathbf{Y}_j$$

is the sample mean of $\mathbf{Y}$ and

$$S_{\mathbf{XY}} = \sum_j (\mathbf{X}_{j1} - \overline{\mathbf{X}})(\mathbf{Y}_j - \overline{\mathbf{Y}}), \quad S_{\mathbf{XX}} = \sum_j (\mathbf{X}_{j1} - \overline{\mathbf{X}})^2. \tag{9.23}$$

Thus, the second line in (9.22) is used to compute the *slope* $\widehat{\beta}_1$ from sample data whereas the first line determines the *intercept* $\widehat{\beta}_0$. In this case, the fitted value is realized as

$$\widehat{\mathbf{y}} = \widehat{\beta}_0 + \widehat{\beta}_1 \mathbf{x}. \tag{9.24}$$

Also, $\widehat{\beta}$ is unbiased with

$$\mathrm{cov}(\widehat{\beta}) = \frac{\sigma^2}{n\|\mathbf{x}\|^2 - n^2\overline{\mathbf{x}}^2} \begin{pmatrix} \|\mathbf{x}\|^2 & -n\overline{\mathbf{x}} \\ -n\overline{\mathbf{x}} & n \end{pmatrix},$$

under the assumptions of Proposition 9.6. □

We now discuss the construction of confidence intervals for the entries of the unknown parameter β. Here we remain in the setting of Example 9.3, so we assume that, conditionally on $\mathfrak{X} = \mathfrak{x}$, $\{\mathbf{e}_j\}$ is independent with $\mathbf{e}_j \sim \mathcal{N}(0, \sigma^2)$ as in (9.14). It follows from (9.18), (9.20) and Corollary 4.7 that

$$\widehat{\beta} - \beta \sim \mathcal{N}(\vec{0}, \sigma^2 \mathfrak{s}), \quad \mathfrak{s} := (\mathfrak{x}^\top \mathfrak{x})^{-1}, \tag{9.25}$$

so we may use the pivotal quantity

$$\frac{\widehat{\beta}_j - \beta_j}{\mathbf{s}_j} \sim \mathcal{N}(0,1), \quad \mathbf{s}_j := \sigma\sqrt{\mathfrak{s}_{jj}}, \tag{9.26}$$

to exhibit confidence intervals for the unknown parameter β_j in case σ is known. Precisely, in the notation of Subsection 7.3,

$$\beta_j \in \left[\widehat{\beta}_j \mp z_{1-\delta/2}\mathbf{s}_j\right] \text{ with prob.} = 1 - \delta. \tag{9.27}$$

Otherwise, we proceed as follows. We define the *residual*

$$\widehat{\mathbf{e}} := \mathbf{Y} - \widehat{\mathbf{Y}}, \tag{9.28}$$

where $\widehat{\mathbf{y}} = \mathfrak{x}\widehat{\beta}$ is the fitted vector as in (9.3). As we shall see, $\|\widehat{\mathbf{e}}\|^2/(n-p-1)$ qualifies as an appropriate estimator for the error variance σ^2.

Proposition 9.10. *$\widehat{\mathbf{e}} \sim \mathcal{N}(\vec{0}, \sigma^2 \mathrm{Id}_{n-p-1})$ and $\|\widehat{\mathbf{e}}\|^2/\sigma^2 \sim \chi^2_{n-p-1}$, $n \geq p+2$.*

Proof. We compute

$$\widehat{\mathbf{e}} = (\mathrm{Id} - \mathfrak{x}(\mathfrak{x}^\top\mathfrak{x})^{-1}\mathfrak{x}^\top)\mathbf{y} = Q(\mathfrak{x}\beta + \mathbf{e}),$$

where $Q = \mathrm{Id}_n - \mathfrak{x}(\mathfrak{x}^\top\mathfrak{x})^{-1}\mathfrak{x}^\top$ is an idempotent, symmetric matrix satisfying

$$Q\mathfrak{x} = 0, \tag{9.29}$$

which means that $\mathrm{rank}\, Q = n - p - 1 \geq 1$. Therefore, $\widehat{\mathbf{e}} = Q\mathbf{e}$ is normally distributed as in the statement (either by Proposition 4.9 (3) or by rotational invariance (Corollary 4.13)). Moreover, since

$$\frac{\|\widehat{\mathbf{e}}\|^2}{\sigma^2} = \left\langle \frac{\mathbf{e}}{\sigma}, Q\left(\frac{\mathbf{e}}{\sigma}\right)\right\rangle, \tag{9.30}$$

the last assertion follows from Proposition 4.27. □

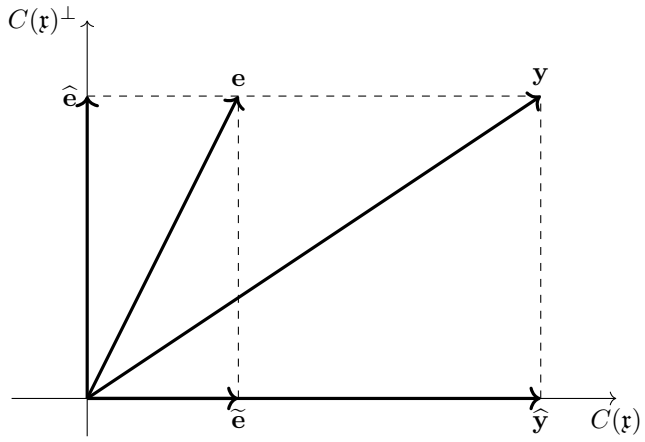


FIGURE 1. The geometry of the linear model

Remark 9.11. (The geometry of the linear model and regression diagnostics) The projection matrix[38] $H = \mathfrak{x}(\mathfrak{x}^\top\mathfrak{x})^{-1}\mathfrak{x}^\top$ appearing in the argument above is usually called the "hat matrix", as it projects $\mathbf{y}$ onto $\widehat{\mathbf{y}} = H\mathbf{y} \in C(\mathfrak{x}) \equiv \mathbb{R}^{p+1}$, the *design space*, which is the $(p+1)$-subspace of $\mathbb{R}^n$ spanned by the columns of the design matrix $\mathfrak{x}$. On the other hand, its complementary projection matrix $Q = \mathrm{Id}_n - H$ projects $\mathbf{y}$ (and also $\mathbf{e}$, because $\mathbf{y} - \mathbf{e} = \mathfrak{x}\beta \in C(\mathfrak{x})$) onto the residual $\widehat{\mathbf{e}} \in C(\mathfrak{x})^\perp \equiv \mathbb{R}^{n-p-1}$ lying in the orthogonal complement of $C(\mathfrak{x})$. This nice "orthogonal" geometry, which hinges on the general setting of Example 9.2, is depicted in Figure 1, where $\widetilde{\mathbf{e}} = H\mathbf{e}$. In particular, the orthogonal decomposition $\mathbf{y} = \widehat{\mathbf{y}} + \widehat{\mathbf{e}}$ clearly implies that the *sample correlation* between $\widehat{\mathbf{y}}$ and $\widehat{\mathbf{e}}$, defined by

$$\mathrm{corr}(\widehat{\mathbf{y}}, \widehat{\mathbf{e}}) = \frac{\widehat{\mathbf{e}}^\top\widehat{\mathbf{y}}}{\|\widehat{\mathbf{e}}\|\|\widehat{\mathbf{y}}\|}, \tag{9.31}$$

vanishes, which justifies the common practice of using a scatterplot of the residuals against the fitted values in order to identify patterns of goodness of fit (or lack thereof) of a given linear model, as far as linearity and homoscedasticity go [Far06, Section 6.1]. Now, if we further specialize to the setting of Example 9.3, which assumes $\mathbf{e} \sim \mathcal{N}(\vec{0}, \sigma^2\mathrm{Id}_n)$, then, by the projection property in (4.8),

$$\widehat{\mathbf{e}} \sim \mathcal{N}(\vec{0}, \sigma^2\mathrm{Id}_{n-p-1}), \tag{9.32}$$

so in particular $\widehat{\mathbf{e}}/|\widehat{\mathbf{e}}|$ is uniformly distributed in $\mathbb{S}^{n-p-2} \subset C(\mathfrak{x})^\perp$ by Remark 4.26. Since $\widehat{\mathbf{e}}$ is accessible from data and adjusted values, graphical methods (say, a Q-Q plot) may be used to confirm the empirical validity of (9.32), which somehow works as an indirect checking of the theoretical assumption on the normality of errors underlying Example 9.3; again, see [Far06, Section 6.1]. A further gauging of the goodness of fit of the model may be implemented after properly combining the residual and the fitted vector. The simplest way of doing this, which leads to a sharpening of (9.31), is to look at the joint distribution of $(\widehat{\mathbf{y}}, \widehat{\mathbf{e}})$ under error normality. Since

$$\begin{pmatrix} \widehat{\mathbf{y}} \\ \widehat{\mathbf{e}} \end{pmatrix} = \begin{pmatrix} H\mathbf{y} \\ Q\mathbf{y} \end{pmatrix} = \begin{pmatrix} H & 0 \\ 0 & Q \end{pmatrix}\begin{pmatrix} \mathbf{y} \\ \mathbf{y} \end{pmatrix},$$

we see that $(\widehat{\mathbf{y}}, \widehat{\mathbf{e}})$ is jointly normally distributed. To find the specific normal distribution we note that

$$\mathbb{E}\left(\begin{pmatrix} \widehat{\mathbf{y}} \\ \widehat{\mathbf{e}} \end{pmatrix}\right) = \begin{pmatrix} H & 0 \\ 0 & Q \end{pmatrix}\mathbb{E}\begin{pmatrix} \mathbf{y} \\ \mathbf{y} \end{pmatrix} = \begin{pmatrix} H & 0 \\ 0 & Q \end{pmatrix}\begin{pmatrix} \mathfrak{x}\beta \\ \mathfrak{x}\beta \end{pmatrix} = \begin{pmatrix} \mathfrak{x}\beta \\ 0 \end{pmatrix}$$

[38] Recall that this means that H is symmetric and idempotent, hence defining an orthogonal projection onto its range.

and

$$\operatorname{cov}\left(\begin{pmatrix}\widehat{\mathbf{y}}\\ \widehat{\mathbf{e}}\end{pmatrix}\right)=\begin{pmatrix}H & 0\\ 0 & Q\end{pmatrix}\operatorname{cov}\begin{pmatrix}\mathbf{y}\\ \mathbf{y}\end{pmatrix}\begin{pmatrix}H & 0\\ 0 & Q\end{pmatrix}^{\top}=\sigma^2\begin{pmatrix}H & 0\\ 0 & Q\end{pmatrix},$$

so that

$$\operatorname{cov}\left(\Lambda\begin{pmatrix}\widehat{\mathbf{y}}\\ \widehat{\mathbf{e}}\end{pmatrix}\right)=\sigma^2\Lambda\begin{pmatrix}H & 0\\ 0 & Q\end{pmatrix}\Lambda^{\top},$$

with Λ being any $2n\times 2n$ matrix. By choosing Λ orthogonal with the corresponding conjugation performing the appropriate diagonalization and viewing $(\widehat{\mathbf{y}},\widehat{\mathbf{e}})$ as an element of $\mathbb{R}^{p+1}\times\mathbb{R}^{n-p-1}=\mathbb{R}^n$, we find that there exists an orthogonal $n\times n$ matrix Λ' such that

$$\operatorname{cov}\left(\Lambda'\begin{pmatrix}\widehat{\mathbf{y}}\\ \widehat{\mathbf{e}}\end{pmatrix}\right)=\sigma^2\mathrm{Id}_n.$$

By Corollary 4.7, and viewing $\mathfrak{r}\beta$ as an element of $\mathbb{R}^{p+1}$,

$$\Lambda'\begin{pmatrix}\widehat{\mathbf{y}}\\ \widehat{\mathbf{e}}\end{pmatrix}\sim\mathcal{N}\left(\Lambda'\begin{pmatrix}\mathfrak{r}\beta\\ 0\end{pmatrix},\sigma^2\mathrm{Id}_n\right),$$

and therefore $\{\widehat{\mathbf{y}},\widehat{\mathbf{e}}\}$ is independent by Corollary 4.13. Further uses of the residual in the art of quantifying the goodness of fit of the linear model may be found in Remark 9.16 below. □

We now come back to the business of finding confidence intervals for the entries of β, this time with σ^2 regarded as unknown. In this case, the next result justifies the replacement of $\mathbf{s}_j$ in (9.26) by

$$\widehat{\mathbf{s}}_j:=\widehat{\sigma}\sqrt{\mathfrak{s}_{jj}},\quad \widehat{\sigma}:=\frac{\|\widehat{\mathbf{e}}\|}{\sqrt{n-p-1}}. \tag{9.33}$$

Note that with this notation, Proposition 9.10 says that

$$(n-p-1)\frac{\widehat{\sigma}^2}{\sigma^2}=\frac{\|\widehat{\mathbf{e}}\|^2}{\sigma^2}\sim\chi^2_{n-p-1}. \tag{9.34}$$

Proposition 9.12. *$\widehat{\sigma}^2$ is an unbiased and consistent estimator for σ^2 (as $n\to+\infty$ and p is held fixed). In particular, $\widehat{\sigma}$ is consistent for the error standard deviation σ. Moreover, $\{\widehat{\beta},\widehat{\sigma}^2\}$ is independent with*

$$\frac{\widehat{\beta}_j-\beta_j}{\widehat{\mathbf{s}}_j}\sim\mathfrak{t}_{n-p-1}.$$

Proof. From (9.34) and Corollary 4.21 we have $\mathbb{E}(\widehat{\sigma}^2)=\sigma^2$, that is, $\operatorname{bias}(\widehat{\sigma}^2)=0$. Also,

$$\operatorname{var}(\widehat{\sigma}^2)=\frac{2\sigma^4}{n-p-1}, \tag{9.35}$$

so that $\operatorname{mse}(\widehat{\sigma}^2)\to 0$ as $n\to+\infty$ and consistency follows from Proposition 7.11. Moreover, since

$$\frac{\widehat{\beta}_j-\beta_j}{\widehat{\mathbf{s}}_j}=\frac{\frac{\widehat{\beta}_j-\beta_j}{\mathbf{s}_j}}{\sqrt{\frac{\|\widehat{e}\|^2/\sigma^2}{n-p-1}}},$$

the last assertion follows from (9.26), (9.34) and Proposition 4.30 once one verifies that $\{\widehat{\beta},\|\widehat{\mathbf{e}}\|^2\}$ is independent. To check this, note that (9.20) gives

$$\mathfrak{x}^{\top}\mathfrak{x}\left(\frac{\widehat{\beta}-\beta}{\sigma}\right)=\mathfrak{x}^{\top}\left(\frac{\mathbf{e}}{\sigma}\right),$$

which together with (9.30), (9.29) and Proposition 4.28 implies that $\{\mathfrak{x}^\top\mathfrak{x}\widehat{\beta}, \|\widehat{\mathfrak{e}}\|^2\}$ is independent, from which the result follows (because $\mathfrak{x}^\top\mathfrak{x}$ is invertible). □

Thus, again using the notation of Subsection 7.3,

$$\beta_j \in \left[\widehat{\beta}_j \mp \mathfrak{t}_{n-p-1,1-\delta/2}\widehat{\mathfrak{s}}_j\right] \text{ with prob. } \approx 1-\delta, \tag{9.36}$$

a confidence interval estimate for β_j in case σ is unknown. Notice that if $n-p-1 \gg 0$ then we can replace $\mathfrak{t}_{n-p-1,1-\delta/2}$ by $z_{1-\delta/2}$ with a negligible error; this uses Remark 6.4.

Example 9.13. (Confidence region for the whole vector parameter β, with σ^2 unknown) Pick $\mathfrak{p}$ so that $\mathfrak{p}^\top\mathfrak{p} = \mathfrak{s}$ as in (9.25) and set $\mathfrak{n} = \sigma^{-1}(\mathfrak{p}^\top)^{-1}(\widehat{\beta}-\beta)$. It is immediate that $\mathfrak{n} \sim \mathcal{N}(\vec{0}, \mathrm{Id}_{p+1})$ so that

$$\frac{(\widehat{\beta}-\beta)^\top\mathfrak{s}(\widehat{\beta}-\beta)}{\sigma^2} = \mathfrak{n}^\top\mathfrak{n} \sim \chi^2_{p+1}$$

and is independent of $\widehat{\sigma}^2$. Therefore, by Propositions 9.12 and 4.33,

$$\frac{(\widehat{\beta}-\beta)^\top\mathfrak{s}(\widehat{\beta}-\beta)}{(p+1)\widehat{\sigma}^2} = \frac{\mathfrak{n}^\top\mathfrak{n}/(p+1)}{\widehat{\sigma}^2/\sigma^2} \sim \mathsf{F}_{p+1,n-p-1},$$

which gives the "confidence region" estimate

$$\beta \in \mathcal{U}_{n,p,\delta}(\widehat{\beta};\widehat{\sigma}^2) \text{ with prob. } 1-\delta, \tag{9.37}$$

where

$$\mathcal{U}_{n,p,\delta}(\widehat{\beta};\widehat{\sigma}^2) = \left\{\beta' \in \mathbb{R}^{p+1}; (\widehat{\beta}-\beta')^\top\mathfrak{s}(\widehat{\beta}-\beta') \le (p+1)\widehat{\sigma}^2\mathsf{f}_{p+1,n-p-1,1-\delta}\right\}. \tag{9.38}$$

In other words, the random ellipsoidal region $\mathcal{U}_{n,p,\delta}(\widehat{\beta};\widehat{\sigma}^2)$, which is fully specified by n, p, δ and the sample data $\widehat{\beta}$ and $\widehat{\sigma}^2$, covers the true vector parameter β with probability $1-\delta$. □

Example 9.14. (Simultaneous confidence band for the mean response) According to [CW09, page 27], "the primary goal in a regression analysis is to understand, as far as possible with the available data, how the conditional distribution of the response varies across sub-populations determined by the possible values of the predictors". Precisely, and using the notation of Example 9.1, if

$$\mathsf{R}^p = \{\mathsf{x}' = (1, x'_1, \cdots, x'_p); x'_j \in \mathbb{R}, j = 1, \cdots, p\}$$

and $\mathsf{x} \in \mathsf{R}^p$ is given, this amounts to checking how much information on the conditioned random variable $\mathscr{Y}|_{\widetilde{\mathscr{X}}=\mathsf{x}}$ may be extracted from the data array $(\mathbf{x}, \mathbf{y})$. We insist that x should be viewed as a future observation which has been consolidated from the same population *after* the data set has been drawn. From this perspective, the method of least squares in Example 9.2, according to which

$$\mathscr{Y}_\mathsf{x} := \mathscr{Y}|_{\widetilde{\mathscr{X}}=\mathsf{x}} = \mathsf{x}^\top\beta + \mathsf{e} \text{ with } \mathbb{E}(\mathsf{e}) = 0, \tag{9.39}$$

represents a first step toward this goal as it allows us to make use of the estimate $\widehat{\beta}$, which has been computed from $(\mathbf{x}, \mathbf{y})$, in order to retrieve information on the *mean response*

$$\mathsf{x}^\top\beta = \mathbb{E}(\mathscr{Y}_\mathsf{x}), \tag{9.40}$$

a collection of summaries which, as x varies, exhausts the realizations of the conditional expectation $\mathbb{E}(\mathscr{Y}|\widetilde{\mathscr{X}})$ (by Proposition 3.14). Indeed, if we further specialize to the setting of Example 9.3, where

$$\mathsf{e} \sim \mathcal{N}(0, \sigma^2), \tag{9.41}$$

then $\mathsf{x}^\top\widehat{\beta}$ may be used to properly estimate the population parameter in (9.40): from (9.25) we have

$$\mathsf{x}^\top(\widehat{\beta}-\beta) \sim \mathcal{N}(0, \sigma^2\mathsf{x}^\top\mathfrak{s}\mathsf{x}), \quad \mathfrak{s} = (\mathfrak{x}^\top\mathfrak{x})^{-1}, \tag{9.42}$$

that is,

$$\frac{\mathsf{x}^\top(\widehat{\beta}-\beta)}{\sigma\sqrt{\mathsf{x}^t\mathfrak{s}\mathsf{x}}}\sim\mathcal{N}(0,1),$$

so that, by Propositions 9.12 and 4.30,

$$\frac{\mathsf{x}^\top(\widehat{\beta}-\beta)}{\widehat{\sigma}\sqrt{\mathsf{x}^\top\mathfrak{s}\mathsf{x}}}\sim\mathfrak{t}_{n-p-1},$$

and hence,

$$\mathsf{x}^\top\beta\in\left[\mathsf{x}^\top\widehat{\beta}\mp\mathfrak{t}_{n-p-1,1-\delta/2}\widehat{\sigma}\sqrt{\mathsf{x}^\top\mathfrak{s}\mathsf{x}}\right]\ \text{with prob. } 1-\delta. \tag{9.43}$$

If we allow for a bit more of spread, a "simultaneous" version of this pointwise bound is also available, with the corresponding estimate holding for *any* $\mathsf{x}\in\mathsf{R}^n$, as follows. Starting with (9.37) and (9.38) and using the notation of Example 9.13,

$$\begin{aligned}1-\delta &= P\left(\frac{(\widehat{\beta}-\beta)^\top\mathfrak{s}(\widehat{\beta}-\beta)}{(p+1)\widehat{\sigma}^2}\leq\mathsf{f}_{p+1,n-p-1,1-\delta}\right)\\ &= P\left(\frac{\|\mathfrak{n}\|}{\widehat{\sigma}/\sigma}\leq\sqrt{(p+1)\mathsf{f}_{p+1,n-p-1,1-\delta}}\right),\end{aligned}$$

and recalling that, by Cauchy-Schwarz,

$$\sup_{\mathsf{x}\in\mathsf{R}^p}\frac{|(\mathfrak{p}\mathsf{x})^\top\mathfrak{n}|}{\|\mathfrak{p}\mathsf{x}\|\|\mathfrak{n}\|}=1,$$

we get

$$\begin{aligned}1-\delta &= P\left(\sup_{\mathsf{x}\in\mathsf{R}^p}\frac{|(\mathfrak{p}\mathsf{x})^\top\mathfrak{n}|}{(\widehat{\sigma}/\sigma)\sqrt{(\mathfrak{p}\mathsf{x})^\top\mathfrak{p}\mathsf{x}}}\leq\sqrt{(p+1)\mathsf{f}_{p+1,n-p-1,1-\delta}}\right)\\ &= P\left(\sup_{\mathsf{x}\in\mathsf{R}^p}\frac{|\mathsf{x}^\top(\widehat{\beta}-\beta)|}{\widehat{\sigma}\sqrt{\mathsf{x}^\top\mathfrak{s}\mathsf{x}}}\leq\sqrt{(p+1)\mathsf{f}_{p+1,n-p-1,1-\delta}}\right),\end{aligned}$$

so that

$$\mathsf{x}^\top\beta\in\left[\mathsf{x}^\top\widehat{\beta}\mp\sqrt{(p+1)\mathsf{f}_{p+1,n-p-1,1-\delta}}\,\widehat{\sigma}\sqrt{\mathsf{x}^\top\mathfrak{s}\mathsf{x}}\right]\forall\mathsf{x}\in\mathsf{R}^p\ \text{with prob. } 1-\delta, \tag{9.44}$$

As fully explained in [Liu10], this Scheffé-type simultaneous confidence band plays a fundamental role in the inference theory of OLS models; see also Example 12.15 for a generalization thereof. □

Example 9.15. (Simultaneous prediction band for the response) With the estimates for the mean response provided by (9.43) and (9.44) at hand, we may now pose ourselves the problem of "predicting" where the response itself,

$$\mathscr{Y}_\mathsf{x}=\mathsf{x}^\top\beta+\mathsf{e}\sim\mathcal{N}(\mathsf{x}^\top\beta,\sigma^2), \tag{9.45}$$

is likely to fall, where, as in Example 9.14, we should think of x as a new observation for the regressor which has taken place after the data $(\mathbf{x},\mathbf{y})$ has been gathered, hence the "prediction" terminology. In particular, e is independent of $\widehat{\beta}$, which has been constructed out of $(\mathbf{x},\mathbf{y})$, so that $\mathscr{Y}_\mathsf{x}$ is independent of $\widehat{\beta}$ as well. Combining this with (9.45), (9.42) and Proposition 4.8 (3) we see that

$$\mathscr{Y}_\mathsf{x}-\mathsf{x}^\top\widehat{\beta}\sim\mathcal{N}\left(0,\sigma^2\left(1+\mathsf{x}^\top\mathfrak{s}\mathsf{x}\right)\right),$$

which by the standard argument gives the sought-after pointwise "prediction interval" for the response,

$$\mathscr{Y}_\mathsf{x}\in\left[\mathsf{x}^\top\widehat{\beta}\mp\mathfrak{t}_{n-p-1,1-\delta/2}\widehat{\sigma}\sqrt{1+\mathsf{x}^\top\mathfrak{s}\mathsf{x}}\right]\ \text{with prob. } 1-\delta, \tag{9.46}$$

with this new terminology being adopted because the random variable $\mathscr{Y}_{\mathsf{x}}$ is *not* a pararameter, so this fails to be a confidence interval in the ordinary sense. In any case, upon comparison with (9.43) we see that when passing from the mean response to the response itself, the point estimate $\mathsf{x}^\top\widehat{\beta}$ remains the same but the dispersion gets expanded by a factor that makes it at least as large as $\mathsf{t}_{n-p-1,1-\delta/2}\widehat{\sigma}$, a lower bound which depends on the already observed data $(\mathbf{x},\mathbf{y})$ but *not* on the future observation x for the regressor. Following [Car86, Theorem 1] and [SA90, Theorem 1], we may also contemplate a "simultaneous" version of (9.46), which is obtained by means of an easy generalization of Scheffé's argument leading to (9.44). Indeed,

$$b = \begin{pmatrix} \widehat{\beta}-\beta \\ \mathsf{e} \end{pmatrix} \sim \mathcal{N}\left(\vec{0}, \begin{pmatrix} \sigma^2\mathfrak{s} & 0 \\ 0 & \sigma^2 \end{pmatrix}\right)$$

and

$$\overline{\mathfrak{s}} = \begin{pmatrix} \mathfrak{s} & 0 \\ 0 & 1 \end{pmatrix}$$

are such that

$$\sigma^{-2}b^\top\overline{\mathfrak{s}}b = \sigma^{-2}(\widehat{\beta}-\beta)^\top\mathfrak{s}(\widehat{\beta}-\beta) + \sigma^{-2}\|\mathsf{e}\|^2 \sim \chi^2_{p+2},$$

so that

$$\frac{b^\top\overline{\mathfrak{s}}b}{(p+2)\widehat{\sigma}^2} \sim \mathsf{F}_{p+2,n-p-1},$$

and hence

$$\begin{aligned} 1-\delta &= P\left(\frac{b^\top\overline{\mathfrak{s}}b}{(p+2)\widehat{\sigma}^2} \leq \mathsf{f}_{p+2,n-p-1,1-\delta}\right) \\ &= P\left(\frac{\|\overline{\mathfrak{n}}\|}{\widehat{\sigma}/\sigma} \leq \sqrt{(p+2)\mathsf{f}_{p+2,n-p-1,1-\delta}}\right), \end{aligned}$$

where $\overline{\mathfrak{n}} = \sigma^{-1}(\overline{\mathfrak{p}}^\top)^{-1}b$ with $\overline{\mathfrak{p}}$ satisfying $\overline{\mathfrak{p}}^\top\overline{\mathfrak{p}} = \overline{\mathfrak{s}}$; cf. the corresponding unbarred objects in Example 9.13. Thus, again using Cauchy-Schwarz,

$$\begin{aligned} 1-\delta &= P\left(\sup_{\mathsf{z}\in\mathsf{R}^{p+1}} \frac{|(\overline{\mathfrak{p}}\mathsf{z})^\top\overline{\mathfrak{n}}|}{(\widehat{\sigma}/\sigma)\sqrt{(\overline{\mathfrak{p}}\mathsf{z})^\top\overline{\mathfrak{p}}\mathsf{z}}} \leq \sqrt{(p+2)\mathsf{f}_{p+2,n-p-1,1-\delta}}\right) \\ &= P\left(\sup_{\mathsf{z}\in\mathsf{R}^{p+1}} \frac{|\mathsf{z}^\top b|}{\widehat{\sigma}\sqrt{\mathsf{z}^\top\overline{\mathfrak{s}}\mathsf{z}}} \leq \sqrt{(p+2)\mathsf{f}_{p+2,n-p-1,1-\delta}},\right), \end{aligned}$$

where $\mathsf{R}^{p+1} = \mathsf{R}^p \times \mathbb{R}$, so if we choose $\mathsf{z} = (\mathsf{x},-1)$, where $\mathsf{x} \in \mathsf{R}^p$ is arbitrary, we get $\mathsf{z}^\top b = \mathsf{x}^\top\widehat{\beta} - \mathscr{Y}_{\mathsf{x}}$ and $\mathsf{z}^\top\overline{\mathfrak{s}}\mathsf{z} = 1 + \mathsf{x}^\top\mathfrak{s}\mathsf{x}$, which finally gives

$$\mathscr{Y}_{\mathsf{x}} \in \left[\mathsf{x}^\top\widehat{\beta} \mp \sqrt{(p+2)\mathsf{f}_{p+2,n-p-1,1-\delta}}\,\widehat{\sigma}\sqrt{1+\mathsf{x}^\top\mathfrak{s}\mathsf{x}}\right] \forall \mathsf{x} \in \mathsf{R}^p \text{ with prob. } 1-\delta$$

as a simultaneous prediction band for the response. □

Remark 9.16. (Coefficient of determination and goodness of fit) As illustrated in Figure 1, where the residual vector $\widehat{\mathbf{e}}$ lies orthogonally to $C(\mathfrak{x})$, the residual sum of squares

$$SS_{\text{Res}} := \|\widehat{\mathbf{e}}\|^2 = \sum_j (\mathbf{Y}_j - \widehat{\mathbf{Y}}_j)^2$$

measures the amount of variation in $\mathbf{Y}$ that remains unexplained by the linear regression model (whose fitted values form the vector $\widehat{\mathbf{y}} \in C(\mathfrak{x})$). Since, by Remark 9.11 and Proposition 9.10,

$$(\mathbf{Y}-\widehat{\mathbf{Y}})^\top(\widehat{\mathbf{Y}}-\overline{\mathbf{Y}}\mathbf{1}) = \widehat{\mathbf{e}}^\top\widehat{\mathbf{Y}} - \widehat{\mathbf{e}}^\top\overline{\mathbf{Y}}\mathbf{1} = -\widehat{\mathbf{Y}}\sum_j \widehat{\mathbf{e}}_j = 0,$$

SS_{Res} also appears in the *orthogonal* decomposition

$$S_{\mathbf{YY}} = SS_{\mathrm{Reg}} + SS_{\mathrm{Res}}, \tag{9.47}$$

where

$$SS_{\mathrm{Reg}} = \sum_j (\widehat{\mathbf{Y}}_j - \overline{\mathbf{Y}})^2, \qquad S_{\mathbf{YY}} = \sum_j (\mathbf{Y}_j - \overline{\mathbf{Y}})^2$$

represent, respectively, the variation explained by the model and the total variation of $\mathbf{Y}$ about its mean $\overline{\mathbf{Y}}$ (notation consistent with (9.23)). Since $SS_{\mathrm{Res}} \leq S_{\mathbf{YY}}$, it is natural to define the *coefficient of determination*

$$R^2 = \frac{SS_{\mathrm{Reg}}}{S_{\mathbf{YY}}} = 1 - \frac{SS_{\mathrm{Res}}}{S_{\mathbf{YY}}}, \tag{9.48}$$

as a measure of the model's *goodness of fit*: the closer R^2 is to its maximal value 1, the better the model accounts for the observed variation. A simple justification of this interpretation may be given in the case of simple linear regression (Example 9.9). Indeed,

$$\begin{aligned}
\widehat{\mathbf{e}} &= \mathbf{Y} - (\widehat{\beta}_0 \mathbf{1} + \widehat{\beta}_1 \mathbf{X}) \\
&= \mathbf{Y} - \big((\overline{\mathbf{Y}} - \widehat{\beta}_1 \overline{\mathbf{X}})\mathbf{1} + \widehat{\beta}_1 \mathbf{X}\big) \\
&= \mathbf{Y} - \overline{\mathbf{Y}}\mathbf{1} - \widehat{\beta}_1 (\mathbf{X} - \overline{\mathbf{X}}\mathbf{1}) \\
&= \mathbf{Y} - \overline{\mathbf{Y}}\mathbf{1} - \frac{S_{\mathbf{XY}}}{S_{\mathbf{XX}}} (\mathbf{X} - \overline{\mathbf{X}}\mathbf{1}),
\end{aligned}$$

which yields

$$SS_{\mathrm{Res}} = \frac{S_{\mathbf{XX}} S_{\mathbf{YY}} - S_{\mathbf{XY}}^2}{S_{\mathbf{XX}}}.$$

Eliminating SS_{Res} from this expression and (9.48) gives

$$R^2 = \widehat{\rho}^2,$$

where

$$\widehat{\rho} = \frac{S_{\mathbf{XY}}}{\sqrt{S_{\mathbf{XX}} S_{\mathbf{YY}}}}$$

is the *sample correlation coefficient* of the pair $(\mathbf{X}, \mathbf{Y})$ underlying the model (cf. (9.31) and Example 9.2). It should also be mentioned that when comparing models with different numbers of regressors, it is often preferable to adjust for the associated degrees of freedom, leading to the *adjusted coefficient of determination*

$$R^2_{\mathrm{adj}} = 1 - \frac{SS_{\mathrm{Res}}/(n-p-1)}{S_{\mathbf{YY}}/(n-1)} = 1 - \frac{n-1}{n-p-1}(1 - R^2),$$

a statistic that penalizes model complexity and provides a fairer basis for comparison. It should be emphasized, however, that relying solely on R^2 or R^2_{adj} as measures of fit can be misleading, since no confidence levels are inherently attached to them (though this can be remedied; see [RS08, Section 10.5]). Consequently, they are treated here as mere point statistics which are best used in conjunction with confidence intervals for the parameters (as in (9.36)) and with the formal hypothesis-testing framework developed in Section 12. As an illustration of this latter procedure, Example 12.11 below presents the classical F-test which provides a much more rigorous assessment of the overall statistical significance of the fitted model by checking whether the observed ratio

$$\frac{SS_{\mathrm{Reg}}/p}{SS_{\mathrm{Res}}/(n-p-1)} = \frac{n-p-1}{p} \, \frac{SS_{\mathrm{Reg}}/S_{\mathbf{YY}}}{SS_{\mathrm{Res}}/S_{\mathbf{YY}}}$$

exceeds the corresponding critical value. □

Remark 9.17. (Regression to the mean, again) Using the notation of Example 9.9 and Remark 9.16, we see from (9.22) that the slope of the regression line is

$$\widehat{\beta}_1 = \frac{\sqrt{S_{\mathbf{yy}}}}{\sqrt{S_{\mathbf{xx}}}}\widehat{\rho}, \tag{9.49}$$

so that

$$\widehat{\beta}_0 = \overline{\mathbf{y}} - \widehat{\rho}\frac{\sqrt{S_{\mathbf{yy}}}}{\sqrt{S_{\mathbf{xx}}}}\overline{\mathbf{x}},$$

and from (9.24) we deduce that the fitted value is

$$\widehat{\mathbf{y}} = \overline{\mathbf{y}} - \widehat{\rho}\frac{\sqrt{S_{\mathbf{yy}}}}{\sqrt{S_{\mathbf{xx}}}}(\mathbf{x} - \overline{\mathbf{x}}).$$

Equivalently,

$$\frac{\widehat{\mathbf{y}} - \overline{\mathbf{y}}}{\sqrt{S_{\mathbf{yy}}}} = \widehat{\rho}\frac{\mathbf{x} - \overline{\mathbf{x}}}{\sqrt{S_{\mathbf{xx}}}}, \tag{9.50}$$

and we conclude that, unless the sampling informs us that $\mathbf{x}$ and $\mathbf{y}$ are perfectly correlated ($|\widehat{\rho}| = 1$), we should regard the appropriated standardization of $\widehat{\mathbf{y}}$ as being strictly smaller (in absolute value) than the standardization of $\mathbf{x}$, a circumstance which certainly indicates a "regression to the mean"; compare with Remark 4.18. □

Remark 9.18. ($\widehat{\beta}$ as a best unbiased estimator) It follows from (9.17) that the MLE for the parameter $\theta = (\beta, \sigma^2)$ of the regression linear model under error normality is $\widehat{\theta} = (\widehat{\beta}, \widehat{\sigma}^2_\bullet)$, where $\widehat{\beta}$ is the usual OLS estimator for β in (9.2), and $\widehat{\sigma}^2_\bullet = \|\widehat{\mathbf{e}}\|^2/n$ with $\widehat{\mathbf{e}} = \mathbf{Y} - \widehat{\mathbf{Y}}$ being the residual as in (9.28), so that (9.18), (9.35) and the independence of $\{\widehat{\beta}, \widehat{\sigma}^2_\bullet\}$ lead to

$$\mathrm{cov}(\widehat{\theta}) = \begin{pmatrix} \mathrm{cov}(\widehat{\beta}) & 0 \\ 0 & \mathrm{var}(\widehat{\sigma}^2_\bullet) \end{pmatrix} = \begin{pmatrix} \sigma^2(\mathfrak{x}^\top\mathfrak{x})^{-1} & 0 \\ 0 & 2(n-p-1)\sigma^4/n^2 \end{pmatrix},$$

where we used that $\widehat{\sigma}^2_\bullet = (n-p-1)\widehat{\sigma}^2/n$ with $\widehat{\sigma}$ as in (9.33). Also, again from (9.17) we easily compute

$$l_{\beta\beta} = -\frac{\mathfrak{x}^\perp\mathfrak{x}}{\sigma^2}, \quad l_{\sigma^2\sigma^2} = \frac{n}{2\sigma^4} - \frac{\|\mathbf{y} - \mathfrak{x}\beta\|^2}{\sigma^6}, \quad l_{\beta\sigma^2} = \frac{\mathfrak{x}^\perp(\mathbf{y} - \mathfrak{x}\beta)}{\sigma^2},$$

so it follows from $\sigma^{-2}(\mathbf{y} - \mathfrak{x}\beta) \sim \mathcal{N}(\vec{0}, \mathrm{Id}_n)$, (8.24) and Corollary 4.25 that

$$\mathscr{F}(\theta) = \begin{pmatrix} \mathscr{F}(\beta) & 0 \\ 0 & \mathscr{F}(\sigma^2) \end{pmatrix} = \begin{pmatrix} \sigma^{-2}\mathfrak{x}^\top\mathfrak{x} & 0 \\ 0 & n/2\sigma^4 \end{pmatrix}. \tag{9.51}$$

Regarding this analysis, we observe that:

- Since the Cramér-Rao lower bound is attained for $\widehat{\beta}$, Corollary 8.19 implies that it is the best *unbiased* estimator for β (under error normality), which should be compared with Theorem 9.7 (Gauss-Markov), where normality is relaxed to (9.19) but the competing unbiased estimators are required to be linear.
- We have
$$\mathrm{var}(\widehat{\sigma}^2_\bullet) = 2\frac{n-p-1}{n^2}\sigma^4 < \frac{2\sigma^4}{n} = \mathscr{F}(\sigma^2)^{-1},$$
a clear violation of the Cramér-Rao lower bound (8.27), but of course this poses no contradiction to Theorem 8.18 because $\widehat{\sigma}^2_\bullet$ is *not* unbiased.
- The unbiased estimator $\widehat{\sigma}^2$ for σ^2 satisfies $\mathrm{var}(\widehat{\sigma}^2) = 2\sigma^4/(n-p-1) > \mathscr{F}(\sigma^2)^{-1}$, corresponding to a strict inequality in (8.27).

- If we view $\widehat{\sigma}^2_\bullet \in \mathcal{E}_g$, where
$$g(\sigma^2) = \mathbb{E}(\widehat{\sigma}^2_\bullet) = \frac{n-p-1}{n}\sigma^2,$$
then the strict inequality in (8.26) holds.

Therefore, both $\widehat{\sigma}^2$ and $\widehat{\sigma}^2_\bullet$ fail to realize the best estimator for σ^2 in their natural classes, which suggests the existence of another estimator with a better performance in each case. □

Example 9.19. (AIC for the normal linear model) Starting from (8.59) we may now compute the AIC for the linear model discussed in Remark 9.18. Recalling that the ML estimator is $\widehat{\theta} = (\widehat{\beta}, \widehat{\sigma}^2_\bullet)$, we obtain

$$\begin{aligned}\mathrm{AIC}(\mathfrak{X}) &= -2\left(-\frac{n}{2}\ln(2\pi\widehat{\sigma}^2_\bullet) - \frac{1}{2\widehat{\sigma}^2_\bullet}\|\mathbf{y} - \mathfrak{x}\widehat{\beta}\|^2\right) + 2(p+2)\\ &= -2\left(-\frac{n}{2}\ln(\widehat{\sigma}^2_\bullet) - \frac{n}{2}\ln(2\pi) - \frac{n}{2}\right) + 2(p+2),\end{aligned}$$

and therefore

$$\mathrm{AIC}(\mathfrak{X}) = n\ln(\widehat{\sigma}^2_\bullet) + 2(p+2) + n\ln(2\pi) + n.$$

Note that the last two terms do not depend on the model (only on the observed data) and may be dropped for comparison purposes, as it is customary in model selection [KK08, HTF09, BA13]. Regarding this computation, we add two further remarks:

- In this normal linear model, the maximized log-likelihood term in the AIC represents, up to an additive constant, the Kullback–Leibler divergence $D^{KL}_{\widehat{\theta}_{\mathfrak{X}}}(\widehat{\theta})$ between the empirical distribution $\widehat{\theta}_{\mathfrak{X}}$ of the sample and the fitted model $\widehat{\theta}$. By contrast, the penalty term $2(p+2)$ is independent of the observed sample and arises as an asymptotic correction accounting for the gap between empirical and population Kullback–Leibler risks induced by the estimation of $p+2$ parameters; see Remark 8.35 for a geometric interpretation of this phenomenon.
- Once the linear model, and hence the number of parameters, is fixed, the normal likelihood induces no further notion of goodness of fit beyond the residual sum of squares: up to additive constants, the maximized log-likelihood, and hence the first term of the AIC, depends on the data solely through $\widehat{\sigma}^2_\bullet$. This connects the information-theoretic interpretation of AIC with classical goodness-of-fit diagnostics: in the normal linear model, all likelihood-based criteria reduce, up to additive and penalization terms, to functions of the residual sum of squares, reflecting its role as a sufficient statistic for model adequacy; cf. Remark 9.16.

For a non-asymptotic alternative to information-criterion–based model selection, grounded in concentration inequalities as introduced in Section 5, see [Mas07]. □

Remark 9.20. (Asymptotic normality of the OLS estimator) The "small sample" computations leading to the confidence interval estimates (9.27) and (9.36) rely heavily on the normality of the error and should be compared to the corresponding "small sample" estimates for the population mean μ in (7.36) and (7.37), respectively. Similarly to what occurred there, under the more general assumptions in (9.19) we must resort to the fundamental limit theorems in Section 6 in order to establish the asymptotic normality of the LSM estimator $\widehat{\beta}$ from which "large sample" estimates should be retrieved. We use the assumptions underlying the linear regression model in Example 9.2 and conveniently decompose the Gram matrix as

$$\mathfrak{x}^\top\mathfrak{x} = \sum_{j=1}^{n}\mathfrak{x}_j^\top\mathfrak{x}_j,$$

with a similar expression holding for $\mathfrak{X}^\top \mathbf{e}$. Thus,

$$\sqrt{n}\left(\widehat{\beta}-\beta\right)=\left(\frac{1}{n}\sum_{j=1}^{n}\mathfrak{X}_j^\top\mathfrak{X}_j\right)^{-1}\sqrt{n}\left(\frac{1}{n}\sum_{j=1}^{n}\mathfrak{X}_j^\top\mathbf{e}_j\right).$$

From LLN (Theorem 6.2) we know that

$$\frac{1}{n}\sum_{j=1}^{n}\mathfrak{X}_j^\top\mathfrak{X}_j\xrightarrow{p}\mathfrak{C}:=\mathbb{E}(\mathfrak{X}_j^\top\mathfrak{X}_j),$$

a symmetric and positive definite $(p+1)\times(p+1)$ random matrix. Also, since $\mathbb{E}(\mathfrak{X}_j^\top\mathbf{e})=\vec{0}$ by (9.8) and

$$\begin{aligned}\operatorname{cov}(\mathfrak{X}_j^\top\mathbf{e}_j)&\overset{(3.15)}{=}\mathbb{E}(\operatorname{cov}(\mathfrak{X}_j^\top\mathbf{e}_j|\mathfrak{X}))+\operatorname{cov}(\mathbb{E}(\mathfrak{X}_j^\top\mathbf{e}_j|\mathfrak{X}))\\&=\mathbb{E}(\mathfrak{X}_j^\top\operatorname{cov}(\mathbf{e}_j|\mathfrak{X})\mathfrak{X}_j)\\&\overset{(9.11)}{=}\sigma^2\mathfrak{C},\end{aligned}$$

we may use CLT (Theorem 6.5) to check that

$$\sqrt{n}\left(\frac{1}{n}\sum_{j=1}^{n}\mathfrak{X}_j^\top\mathbf{e}_j\right)\xrightarrow{d}\mathcal{N}\left(\vec{0},\sigma^2\mathfrak{C}\right).$$

Combining these calculations with Theorem 2.25 we conclude that, as $n\to+\infty$ and p is held fixed,

$$\sqrt{n}\left(\widehat{\beta}-\beta\right)\xrightarrow{d}\mathcal{N}\left(\vec{0},\sigma^2\mathfrak{C}^{-1}\right), \tag{9.52}$$

so that $\widehat{\beta}$ is asymptotically normal (and hence consistent) with asymptotic covariance $\sigma_\beta=\sigma^2\mathfrak{C}^{-1}$, which should be reliably estimated in order to obtain the desired confidence regions. We refer to [Ame85, Hay11] for full accounts of the estimation theory of the linear regression model, including the justification of our somewhat sloppy use of LLN and CLT above (recall that the conditional distributions across observations in a linear model are independent but not identically distributed as they depend on the observed covariates). □

Remark 9.21. (Asymptotic normality for the linear regression model under error normality) Using the results of Remark 9.18, notably the computation of the corresponding Fisher information matrix in (9.51), we find that the ML estimator $(\widehat{\beta},\widehat{\sigma}_\bullet^2)$ for the parameter (β,σ^2) in the linear regression model (under error normality) satisfies, as $n\to+\infty$ and p is held fixed,

$$\sqrt{n}\left(\begin{pmatrix}\widehat{\beta}\\ \widehat{\sigma}_\bullet^2\end{pmatrix}-\begin{pmatrix}\beta\\ \sigma^2\end{pmatrix}\right)\xrightarrow{d}\mathcal{N}\left(\begin{pmatrix}0\\0\end{pmatrix},\begin{pmatrix}\sigma^2\mathfrak{s}_\bullet & 0\\ 0 & 2\sigma^4\end{pmatrix}\right),$$

where

$$\mathfrak{s}_\bullet=\left(\lim_{n\to+\infty}\frac{\mathfrak{x}^\top\mathfrak{x}}{n}\right)^{-1}. \tag{9.53}$$

which establishes its asymptotic normality; here we use the generalization of Theorem 8.24 described in Remark 8.28. In particular, we have the asymptotic relations

$$\widehat{\beta}\approx\mathcal{N}\left(\beta,\frac{\sigma^2\mathfrak{s}_\bullet}{n}\right)\text{ and }\widehat{\sigma}_\bullet^2\approx\mathcal{N}\left(\sigma^2,\frac{2\sigma^4}{n}\right).$$

If, as usual, we appeal to consistency, then the first one yields

$$\beta_j\in\left[\widehat{\beta}_j\mp z_{1-\delta/2}\frac{\widehat{\mathbf{s}}_{\bullet j}}{\sqrt{n}}\right]\text{ with prob. }\approx1-\delta,\quad\widehat{\mathbf{s}}_{\bullet j}=\widehat{\sigma}_\bullet\sqrt{\mathfrak{s}_{\bullet jj}},\quad\widehat{\sigma}_\bullet=\frac{\|\widehat{\mathbf{e}}\|}{\sqrt{n}}, \tag{9.54}$$

which aligns with (9.27) because $\widehat{\mathbf{s}}_{\bullet j}/\sqrt{n} \approx \mathbf{s}_j$ by (9.53) and Proposition 9.12, while the second one gives

$$\sigma^2 \in \left[\widehat{\sigma}^2_\bullet \mp z_{1-\delta/2}\sqrt{\frac{2}{n}}\widehat{\sigma}^2_\bullet\right] \text{ with prob. } \approx 1-\delta,$$

a large sample confidence interval for σ^2. □

9.3. **Regularization in high dimension, sparsity and the LASSO.** The presence of the root-squared diagonal terms $\sqrt{\mathfrak{s}_{jj}}$, where $\mathfrak{s} = (\mathfrak{x}^\top\mathfrak{x})^{-1}$, in the confidence interval estimates above tends to increase the spread when $\mathfrak{x}^\top\mathfrak{x}$ is ill-conditioned, for instance when the ratio between its extremal eigenvalues is excessively large. In addition, the analysis above relies crucially on the assumption that $p+1 \leq n$, which makes it inapplicable in the high-dimensional regime where $p \gg n$ and $\mathfrak{x}^\top\mathfrak{x}$ is no longer invertible. A possible way to address this limitation is to adopt the regularized regression estimator

$$\widehat{\beta}_\lambda = \operatorname{argmin}_\beta \widehat{\mathscr{L}_\lambda}(\beta),$$

where

$$\widehat{\mathscr{L}_\lambda}(\beta) = \tfrac{1}{2}\|\mathbf{y}-\mathfrak{x}\beta\|^2 + \lambda\|\beta\|^2, \quad \lambda > 0.$$

This leads to the explicit solution

$$\widehat{\beta}_\lambda = (\mathfrak{x}^\top\mathfrak{x} + 2\lambda I)^{-1}\mathfrak{x}^\top\mathbf{y},$$

which is well defined even if $\mathfrak{x}^\top\mathfrak{x}$ does not have full column rank. This justifies the designation *ridge regularization* for this approach [HTW15, HTF09, Wai19, Led22]. Moreover, under the conditions of Proposition 9.6 one obtains

$$\mathbb{E}(\widehat{\beta}_\lambda) = (\mathfrak{x}^\top\mathfrak{x} + 2\lambda I)^{-1}\mathfrak{x}^\top\mathfrak{x}\beta$$

and

$$\operatorname{cov}(\widehat{\beta}_\lambda) = \sigma^2(\mathfrak{x}^\top\mathfrak{x} + 2\lambda I)^{-1}\mathfrak{x}^\top\mathfrak{x}(\mathfrak{x}^\top\mathfrak{x} + 2\lambda I)^{-1}.$$

Although $\widehat{\beta}_\lambda$ is not unbiased, these expressions show that there exists $\lambda_0 > 0$ such that $\operatorname{mse}(\widehat{\beta}_\lambda) < \operatorname{mse}(\widehat{\beta})$ for $0 < \lambda < \lambda_0$ [The74]; see also Remark 7.28, where a similar effect is described for the variance estimators $\widehat{\sigma}^2_c$, $c > 0$. Since ridge regression and its many variants are widely employed in practice, this confirms that a small amount of bias is acceptable when it comes with a significant reduction in variance. As explained below, starting with Remark 9.22, this principle connects naturally with the geometric viewpoint developed earlier in Remark 9.11; see also Figure 1.

Remark 9.22. (Dichotomy between model interpretability and prediction accuracy in linear models) In the classical regime $p < n$, the least squares solution decomposes the response vector $\mathbf{y}$ into two orthogonal pieces: the projection $\widehat{\mathbf{y}} = H\mathbf{y}$ onto the column space $C(\mathfrak{x})$, and the residual $\widehat{\mathbf{e}} = (\mathrm{Id}_n - H)\mathbf{y}$ lying in its orthogonal complement $C(\mathfrak{x})^\perp$. These two components correspond to two distinct, complementary features of the OLS. The projection onto $C(\mathfrak{x})$ carries the *predictive* content, since it represents the systematic variation in the response explained by the regressors. The orthogonal complement, by contrast, provides the basis for *inference* and *interpretability*: it isolates the random fluctuation not captured by the model, and this separation underpins our ability to quantify uncertainty, construct confidence intervals, and perform tests of significance (as in Section 12 below). In particular, each estimated coefficient inherits a transparent meaning: the expected change in the response for a unit change in the corresponding predictor, holding others fixed. Thus, prediction is geometrically tied to the column space, while interpretability rests on the existence of its orthogonal counterpart.

When the number of predictors increases until $p \geq n$, however, the elegant dichotomy in Remark 9.22 breaks down. The column space of $\mathfrak{x}$ expands to fill all of $\mathbb{R}^n$, so that every response vector lies within it and residuals vanish. In such cases, prediction not only persists but may interpolate the training data (the rows of $\mathfrak{x}$) exactly, yet the residual space disappears. Without an orthogonal complement, the classical geometric foundation for

inference and interpretability collapses, and the usual tools based on unexplained variation cease to apply. In this regime, regularization enters as a new source of geometry. As we have seen, instead of relying on a residual subspace, ridge regression constrains the parameter vector itself, producing numerically stable and statistically robust estimates by shrinking the coefficients smoothly toward zero, which reduces variance at the expense of a small bias. In this sense, modern high-dimensional regression can be seen as a migration of geometry: from projections in sample space, where prediction and inference were cleanly separated, to constraints in parameter space, where stability and interpretability are achieved by shrinkage.

A modern expression of this principle is given in Example 9.27, which discusses the prediction properties of the LASSO procedure, introduced in [Tib96] and now widely used in Data Science [HTF09, JWHT13, Wai19, Led22]. In contrast to ridge, the LASSO not only shrinks coefficients but also drives many of them exactly to zero, effectively selecting a subset of variables. This sparsity reintroduces a strong element of interpretability: the model highlights which predictors truly matter, while ignoring the rest. To emphasize the underlying geometric migration, this material is preceded by prediction bounds for the classical low-dimensional regime ($p \ll n$), presented in Examples 9.23, 9.25, and 9.26, where the error distribution is considered under increasingly relaxed assumptions.

Example 9.23. (High probability bounds for the prediction error under normality) In addition to the parameter recovery methods already discussed (based on the construction of confidence intervals for the unknown parameter β), we may also look at

$$\mathfrak{x}\widehat{\beta} - \mathfrak{x}\beta = \mathfrak{x}(\mathfrak{x}^t\mathfrak{x})^{-1}\mathfrak{x}^\top \mathbf{e}, \tag{9.55}$$

where we assume as always that $p \leq n$ and $\mathfrak{x}$ has full column rank and hence $\mathfrak{x}^\top\mathfrak{x}$ is invertible[39]. In the notation of Remark 9.11,

$$\mathfrak{x}\widehat{\beta} - \mathfrak{x}\beta = H\mathbf{e} = \mathbf{e} - \widehat{\mathbf{e}} \in C(\mathfrak{x}), \tag{9.56}$$

the difference between the true error and the residual. In other words, rather than paying attention to the projection of $\mathbf{e}$ onto $C(\mathfrak{x})^\perp$ under $Q = \mathrm{Id}_n - H$, which defines the residual $\widehat{\mathbf{e}}$, we now focus on its projection onto $C(\mathfrak{x})$ under H; in Figure 1, $\mathfrak{x}\widehat{\beta} - \mathfrak{x}\beta$ is represented by $\widetilde{\mathbf{e}}$, so that

$$\|\widetilde{\mathbf{e}}\|^2 = \|\mathfrak{x}\widehat{\beta} - \mathfrak{x}\beta\|^2 \tag{9.57}$$

is usually termed the *prediction error* Under the normality assumption $\mathbf{e} \sim \mathcal{N}(\vec{0}, \sigma^2\mathrm{Id}_n)$, it follows from (9.56) and rotational invariance that

$$\sigma^{-2}\|\mathfrak{x}\widehat{\beta} - \mathfrak{x}\beta\|^2 \sim \chi^2_p,$$

so if

$$\widetilde{\mathrm{mse}}(\mathfrak{x}\widehat{\beta}) = \frac{\mathrm{mse}(\mathfrak{x}\widehat{\beta})}{n} \tag{9.58}$$

is the *average prediction risk* then

$$\widetilde{\mathrm{mse}}(\mathfrak{x}\widehat{\beta}) = \frac{\sigma^2 p}{n}, \tag{9.59}$$

where we used Corollary 4.21[40]. Of course, Markov's inequality (2.16) allows us to pass from this "expectation bound" to the corresponding "high probability bound",

$$P\left(\frac{\|\mathfrak{x}\widehat{\beta} - \mathfrak{x}\beta\|^2}{n} \leq \frac{\sigma^2}{\delta}\frac{p}{n}\right) \geq 1 - \delta, \quad \delta > 0. \tag{9.60}$$

[39]Here and in the rest of this subsection we will assume, without loss of generality, that the intercept vanishes, so that $\beta_0 = 0$, $\mathfrak{x} = (\mathfrak{x}_1, \cdots, \mathfrak{x}_p)$, where $\mathfrak{x}_j$ is the j^{th} column of $\mathfrak{x}$, $j = 1, \cdots, p$, etc.

[40]A justification for adopting (9.59) as a measure of accuracy for the prediction error in (9.57) appears in Remark 9.24 below.

As expected, this analysis only provides satisfactory prediction results for the classical OLS if either σ^2, which is assumed known, is very small or $p \ll n$. □

Remark 9.24. We have seen in Example 9.2 that statistical reasoning demands that the OLS estimator should be specified by solving the minimization problem

$$\widehat{\beta} = \text{argmin}_\beta \widehat{\mathscr{L}}(\beta), \quad \widehat{\mathscr{L}}(\beta) = \frac{1}{n}\|\mathbf{y} - \mathfrak{x}\beta\|^2. \tag{9.61}$$

At least if n is large, we may argue[41] that this is the "empirical" version of the more fundamental minimization problem

$$\beta_{\text{m}} := \text{argmin}_\beta \mathscr{R}(\beta)$$

with

$$\mathscr{R}(\beta) = \mathbb{E}(\|\mathbf{Y} - \mathfrak{X}\beta\|^2)$$

being the associated *risk function*. Since $\mathfrak{X}\beta_{\text{m}}$ geometrically corresponds to the orthogonal projection of $\mathbf{Y}$ onto the subspace generated by the columns of $\mathfrak{X}$, we easily see that $\mathbf{e}_{\text{m}} := \mathbf{Y} - \mathfrak{X}\beta_{\text{m}}$ satisfies

$$\mathbb{E}(\mathfrak{X}^\top \mathbf{e}_{\text{m}}) = 0, \tag{9.62}$$

so that

$$\begin{array}{rcl} \mathscr{R}(\widehat{\beta}) & = & \mathscr{R}(\beta_{\text{m}} + \widehat{\beta} - \beta_{\text{m}}) \\ & = & \mathbb{E}(\|\mathbf{Y} - \mathfrak{X}(\beta_{\text{m}} + \widehat{\beta} - \beta_{\text{m}})\|^2) \\ & = & \mathbb{E}(\|\mathbf{e}_{\text{m}} - \mathfrak{X}(\widehat{\beta} - \beta_{\text{m}})\|^2) \\ & \overset{(9.62)}{=} & \mathbb{E}(\|\mathbf{e}_{\text{m}}\|^2) + \mathbb{E}(\|\mathfrak{X}(\widehat{\beta} - \beta_{\text{m}})\|^2), \end{array}$$

which gives

$$\mathbb{E}(\|\mathfrak{X}(\widehat{\beta} - \beta_{\text{m}})\|^2) = \mathscr{R}(\widehat{\beta}) - \mathscr{R}(\beta_{\text{m}}).$$

If we replace β_{m} by β (to comply with the notation of Example 9.23) and condition on $\mathfrak{X} = \mathfrak{x}$ we see that

$$\widetilde{\text{mse}}(\mathfrak{x}\widehat{\beta}) = \frac{\mathscr{R}(\widehat{\beta}) - \mathscr{R}(\beta)}{n},$$

which justifies the terminology employed in (9.58). □

Example 9.25. (High probability bounds for the prediction error without normality) The calculations leading to the expectation and high probability bounds in (9.59) and (9.60) rely heavily on the usual normality assumption on the error. It turns out that we may still obtain a quite effective high probability bound for the prediction error $\|\mathfrak{x}\widehat{\beta} - \mathfrak{x}\beta\|^2$ in (9.57) by merely assuming that, besides (9.11), the errors $\{\mathbf{e}_j\}_{j=1}^n$ are assumed to be independent and *sub-Gaussian* in the sense that

$$\mathbb{E}\left(e^{\mathbf{e}_j u}\right) \leq e^{\sigma^2 u^2/2}, \quad u \in \mathbb{R}, \tag{9.63}$$

so that $\mathbf{e}_j \in \mathsf{SubG}(\sigma)$ as in Definition 5.1. The key point is that (9.55) leads to

$$\|\mathfrak{x}\widehat{\beta} - \mathfrak{x}\beta\|^2 = \|D(D^\top D)^{-1}D^\top U^\top \mathbf{e}\|^2, \tag{9.64}$$

where $UDV^\top$ is a singular value decomposition for $\mathfrak{x}$ (in particular, U and V are both orthogonal). Using that D is diagonal, it is not hard to check that, under these conditions,

$$D(D^\top D)^{-1}D^\top = \begin{pmatrix} I_{p\times p} & \\ & 0_{(n-p)\times(n-p)} \end{pmatrix},$$

[41] Say, by "freezing" β and applying the LLN.

which gives

$$\|\mathfrak{x}\widehat{\beta} - \mathfrak{x}\beta\|^2 = \sum_{j=1}^{p} |(U^\top \mathbf{e})_j|^2. \tag{9.65}$$

Using that

$$(U^\top \mathbf{e})_j = \sum_k U_{kj}\mathbf{e}_k, \quad \sum_k U_{kj}^2 = 1,$$

(9.63) and the independence one easily verifies that

$$\mathbb{E}\left(e^{(U^\top \mathbf{e})_j u}\right) \leq e^{\sigma^2 u^2/2},$$

that is, each $\sigma^{-1}(U^\top \mathbf{e})_j \in \mathsf{SubG}(1)$, and from (9.65) we find that $\sigma^{-2}\|\mathfrak{x}\widehat{\beta} - \mathfrak{x}\beta\|^2 \in \mathsf{SubE}(\nu, 1)$ is sub-exponential as in Definition 5.4; see Remarks 5.12 and 5.13. Using the concentration inequalities in Proposition 5.7 we thus conclude that

$$\begin{aligned} P\left(\frac{\|\mathfrak{x}\widehat{\beta} - \mathfrak{x}\beta\|^2}{n} \leq t\sigma^2 \frac{p}{n}\right) &= P\left(\sigma^{-2}\|\mathfrak{x}\widehat{\beta} - \mathfrak{x}\beta\|^2 \leq pt\right) \\ &\geq 1 - 2e^{-t/2}, \end{aligned} \tag{9.66}$$

for $t \geq \nu^2$, which morally corresponds to (9.60) under the replacement $t \to \delta^{-1}$. □

Example 9.26. (High probability bounds for the prediction error without normality, again) An estimate similar to (9.66) may be obtained under the more general assumptions of Example 9.2), where no further knowledge of the error distribution is available besides (9.11). As we shall see, this ignorance will be counterbalanced by a precise control on the spectrum of the modified Gram matrix $\widehat{\mathbf{\Sigma}} := \mathfrak{x}^\top \mathfrak{x}/n$, which is known to be positive definite. As in Remark 9.24, we identify β_{m} to β and explore the variational characterization of $\widehat{\beta}$ in (9.61) to get $\widehat{\mathscr{L}}(\widehat{\beta}) \leq \widehat{\mathscr{L}}(\beta)$, which means that

$$\frac{\|\mathbf{y} - \mathfrak{x}\widehat{\beta}\|^2}{n} \leq \frac{\|\mathbf{e}\|^2}{n}.$$

If we set $\mathbf{y} = \mathfrak{x}\beta + \mathbf{e}$ in the right-hand side, expand the square and cancel out the terms which are quadratic in the errors we get

$$\frac{\|\mathfrak{x}\widehat{\beta} - \mathfrak{x}\beta\|^2}{n} \leq 2\frac{(\mathfrak{x}^\top \mathbf{e})^\top(\widehat{\beta} - \beta)}{n} \leq 2\frac{\|\mathfrak{x}^\top \mathbf{e}\|}{n}\|\widehat{\beta} - \beta\|, \tag{9.67}$$

where Cauchy-Schwarz has been used in the last step. If $\lambda_{\min}(\widehat{\mathbf{\Sigma}}) \leq \lambda_{\max}(\widehat{\mathbf{\Sigma}})$ stand for the (positive) extremal eigenvalues of $\widehat{\mathbf{\Sigma}}$ then we have

$$\frac{\|\mathfrak{x}\widehat{\beta} - \mathfrak{x}\beta\|^2}{n} = \langle \widehat{\mathbf{\Sigma}}(\widehat{\beta} - \beta), \widehat{\beta} - \beta\rangle \geq \lambda_{\min}(\widehat{\mathbf{\Sigma}})\|\widehat{\beta} - \beta\|^2, \tag{9.68}$$

which may be viewed as a control on the sample correlation between the columns of $\mathfrak{x}$ (because $n\widehat{\mathbf{\Sigma}}_{jk} = \mathfrak{x}_j^\top \mathfrak{x}_k = \|\mathfrak{x}_j\|\|\mathfrak{x}_k\|\mathrm{corr}(\mathfrak{x}_j, \mathfrak{x}_k)$; cf (9.31)), so if we combine these estimates we get

$$\frac{\|\mathfrak{x}\widehat{\beta} - \mathfrak{x}\beta\|^2}{n} \leq 4\frac{\|\mathfrak{x}^\top \mathbf{e}\|^2}{n^2\lambda_{\min}(\widehat{\mathbf{\Sigma}})}.$$

On the other hand, again using our standing assumptions (including (9.11)) we compute

$$\begin{aligned}\mathbb{E}\left(\|\mathfrak{r}^\top \mathbf{e}\|^2\right) &= \mathbb{E}\left(\mathrm{tr}\left((\mathfrak{r}^\top \mathbf{e})(\mathfrak{r}^\top \mathbf{e})^\top\right)\right)\\ &= \mathrm{tr}\,\mathrm{cov}\,(\mathfrak{r}^\top \mathbf{e})\\ &= \sigma^2 \mathrm{tr}(\mathfrak{r}^\top \mathfrak{r})\\ &\leq \sigma^2 pn\lambda_{\max}(\widehat{\boldsymbol{\Sigma}}),\end{aligned}$$

which gives the expectation bound

$$\widetilde{\mathrm{mse}}(\mathfrak{r}\widehat{\beta}) \leq 4\sigma^2\lambda(\widehat{\boldsymbol{\Sigma}})\frac{p}{n}, \quad \lambda(\widehat{\boldsymbol{\Sigma}}) := \frac{\lambda_{\max}(\widehat{\boldsymbol{\Sigma}})}{\lambda_{\min}(\widehat{\boldsymbol{\Sigma}})}, \tag{9.69}$$

from which we obtain the high probability bound

$$P\left(\frac{\|\mathfrak{r}\widehat{\beta} - \mathfrak{r}\beta\|^2}{n} \leq \frac{4\sigma^2}{\delta}\lambda(\widehat{\boldsymbol{\Sigma}})\frac{p}{n}\right) \geq 1-\delta, \quad \delta > 0, \tag{9.70}$$

again via Markov. □

Although its derivation requires only mild assumptions on the error distribution, the high-probability bound in (9.70) remains essentially similar to (9.66) and (9.60). In particular, its explicit dependence on the dimensional ratio p/n shows that, without further control of the error variance σ^2 and of the condition number $\lambda(\widehat{\boldsymbol{\Sigma}})$, the linear model can be trusted only when $p \ll n$. Outside this regime, for instance when $p < n$ but $p \approx n$ with n large, OLS faces at least two well-known deficiencies: *high variability* (while $\mathfrak{r}\widehat{\beta}$ is unbiased, variance estimates such as (9.69) fail to provide reliable control), and *low interpretability* (the sheer number of predictors obscures the identification of variables truly relevant for explaining the response). A natural remedy is to introduce a penalization term into the classical model, as in Example 9.3 on ridge regression; see [Led22, Introduction] for a useful overview of this approach[42]. As the next example shows, the situation becomes even more delicate in the high-dimensional regime $p \gg n$, where in particular the key correlation assumption in (9.68) breaks down, since $\mathfrak{r}^\top\mathfrak{r}$ is no longer invertible.

Example 9.27. (High dimensionality, sparsity and the LASSO) The discussion in the previous paragraph suggests regularizing a suitable multiple of the least squares objective function in order to restore interpretability in case $p \gg n$. When employing the L^1 norm of the vector parameter β, this gives rise to the *LASSO estimator*

$$\widehat{\beta}_L = \mathrm{argmin}_{\beta'} f_L(\beta'), \quad \widetilde{\mathscr{L}_L}(\beta') = \frac{1}{2n}\|\mathbf{y} - \mathfrak{r}\beta'\|^2 + \lambda\|\beta'\|_1,$$

where $\lambda > 0$ is a tuning parameter to be chosen later and

$$\|\beta'\|_1 = \sum_{j=1}^{p} |\beta'_j|.$$

Since $\widetilde{\mathscr{L}_L}(\widehat{\beta}_L) \leq \widetilde{\mathscr{L}_L}(\beta)$, where β is the true parameter appearing in the model equation $\mathbf{y} = \mathfrak{r}\beta + \mathbf{e}$, we thus get with a help from Hölder inequality,

$$\begin{aligned}\frac{1}{n}\|\mathfrak{r}\widehat{\beta}_L - \mathfrak{r}\beta)\|^2 &\leq \frac{2}{n}(\mathfrak{r}^\top \mathbf{e})^\top(\widehat{\beta}_L - \beta) + 2\lambda\left(\|\beta\|_1 - \|\widehat{\beta}_L\|_1\right)\\ &\leq \frac{2}{n}\|\mathfrak{r}^\top \mathbf{e}\|_\infty\|\widehat{\beta}_L - \beta\|_1 + 2\lambda\left(\|\beta\|_1 - \|\widehat{\beta}_L\|_1\right),\end{aligned}$$

[42]This kind of regularization has become a cornerstone of Supervised Learning, where it is crucial to determine on which side of the threshold $p \approx n$ a given problem lies [Don00, HTF09, BC11, BVDG11, HTW15, FBG+16, Ver18, Wai19, Led22].

an estimate which should be compared to (9.67), with its right-hand side effectively disentangling the contributions coming from the "effective error" $2\|\mathfrak{x}^\top \mathbf{e}\|_\infty/n$ and the penalization. Now, sparsity enters the game precisely to handle this latter term, as it contemplates the belief, substantiated by an "omniscient oracle", that a considerable portion of regressors may be dispensed with, so the corresponding parameter entries may be set to vanish. Precisely, there exists $S \subsetneq \{1, \cdots, p\}$ with $s := \sharp S \ll n$ such that $\beta_j = 0$ exactly when $j \notin S$. Thus, if β_S is the "restriction" of β to S, so that $\beta = \beta_S + \beta_{S^c}$, and setting $\widehat{\delta} = \widehat{\beta}_L - \beta$, we have

$$\begin{aligned}\|\beta\|_1 - \|\widehat{\beta}_L\|_1 &= \|\beta_S\|_1 - \|\beta + \widehat{\delta}\|_1 \\ &= \|\beta_S\|_1 - \|\beta_S + \widehat{\delta}_S + \widehat{\delta}_{S^c}\|_1 \\ &= \|\beta_S\|_1 - \|\beta_S + \widehat{\delta}_S\|_1 - \|\widehat{\delta}_{S^c}\|_1 \\ &\leq \|\widehat{\delta}_S\|_1 - \|\widehat{\delta}_{S^c}\|_1,\end{aligned}$$

which gives

$$\frac{1}{n}\|\mathfrak{x}\widehat{\delta}\|^2 \leq \frac{2}{n}\|\mathfrak{x}^\top \mathbf{e}\|_\infty \|\widehat{\delta}\|_1 + 2\lambda\left(\|\widehat{\delta}_S\|_1 - \|\widehat{\delta}_{S^c}\|_1\right),$$

so if we further assume that the tuning parameter dominates the "effective error" according to

$$\frac{2}{n}\|\mathfrak{x}^\top \mathbf{e}\|_\infty \leq \lambda \tag{9.71}$$

we end up with

$$\frac{1}{n}\|\mathfrak{x}\widehat{\delta}\|^2 \leq \lambda\left(3\|\widehat{\delta}_S\|_1 - \|\widehat{\delta}_{S^c}\|_1\right). \tag{9.72}$$

As a direct consequence of this basic inequality we see that

$$\widehat{\delta} \in \mathcal{C}(S) := \left\{\beta' \in \mathbb{R}^{p+1}; \|\beta'_{S^c}\|_1 \leq 3\|\beta'_S\|_1\right\},$$

which suggests that the appropriate replacement for (9.68) is to assume, for some $\kappa > 0$, that

$$\frac{1}{n}\|\mathfrak{x}\beta'\|^2 \geq \kappa\|\beta'\|^2, \quad \beta' \in \mathcal{C}(S). \tag{9.73}$$

Under this *restricted eigenvalue (RE)* condition,

$$\begin{aligned}\frac{1}{n}\|\mathfrak{x}\widehat{\delta}\|^2 &\overset{(9.72)}{\leq} 3\lambda\|\widehat{\delta}_S\|_1 \\ &\leq 3\lambda\sqrt{s}\|\widehat{\delta}\| \\ &\overset{(9.73)}{\leq} \frac{3\lambda\sqrt{s}}{\sqrt{\kappa n}}\|\mathfrak{x}\widehat{\delta}\|,\end{aligned}$$

which finally gives the bound

$$\frac{\|\mathfrak{x}\widehat{\beta}_L - \mathfrak{x}\beta\|^2}{n} \leq \frac{9\lambda^2 s}{\kappa}. \tag{9.74}$$

In order to estimate in terms of λ the probability of the event in (9.71), to which the validity of (9.74) is conditioned, let us assume for simplicity that $\mathbf{e} \sim \mathcal{N}(\vec{0}, \sigma^2 \mathrm{Id}_{n\times n})$. By the projection property in (4.8),

$$2\frac{\mathfrak{x}_k^\top \mathbf{e}}{n} \sim \mathcal{N}\left(0, 4\frac{\sigma^2}{n}\left\|\frac{\mathfrak{x}_k}{\sqrt{n}}\right\|^2\right), \quad k = 1, \cdots, p,$$

so if the columns of the design matrix are normalized so that

$$\left\|\frac{\mathfrak{x}_k}{\sqrt{n}}\right\| \leq C,$$

the standard Gaussian concentration inequality in (5.3) leads to

$$\begin{aligned} P\left(2\left\|\frac{\mathfrak{r}^\top \mathbf{e}}{n}\right\|_\infty \leq \lambda\right) &\geq 1 - 2pe^{-\frac{n\lambda^2}{8C^2\sigma^2}} \\ &= 1 - 2e^{-\frac{n\lambda^2}{8C^2\sigma^2}+\ln p}. \end{aligned}$$

This gives

$$P\left(2\left\|\frac{\mathfrak{r}^\top \mathbf{e}}{n}\right\|_\infty \leq \lambda\right) \geq 1 - 2e^{-\frac{t^2}{2}}$$

if

$$\lambda^2 = 8C^2\sigma^2\left(\frac{\ln p}{n} + \frac{t^2}{2n}\right),$$

so with this choice of λ, (9.74) immediately yields the bound

$$\frac{\|\mathfrak{x}\widehat{\beta}_L - \mathfrak{x}\beta\|^2}{n} \leq \frac{72C^2\sigma^2}{\kappa}\frac{s}{n}\left(\ln p + \frac{t^2}{2}\right) \tag{9.75}$$

with at least the same probability. Upon comparison with (9.70) and not taking into account certain structural constants, we have been able to replace the dimensional ratio p/n by $s\ln p/n$, which is linear in the "sparsity index" $s = \|\beta\|_0$ and scales logarithmically with p, added to another term which is driven by the "oracle rate" $sn^{-1} = o(1)$. Thus, it suffices to take $n \gg s\ln p > s$ in order to have LASSO's prediction nearly as accurate as if $S = \operatorname{supp}\beta$, whose elements classify the relevant regressors, was known a priori. We mention that similar estimates hold true under much weaker assumptions on the error[43] and even for other kinds of penalizations; we refer to [BVDG11, Chapter 6], [HTW15, Chapter 11], [Wai19, Chapter 7] and [Led22, Chapter 6] for such generalizations and, more importantly, for the heuristics behind the crucial RE condition in (9.73) above. Finally, the practical question remains of fine-tuning the parameter λ so as to obtain the right balance between variability and interpretability. In this regard, the feasibility of the most adopted procedure, cross-validation, is theoretically confirmed in [CLC21], where it is shown that, under suitable conditions, its use only adds to the right-hand side of (9.75) a multiplicative factor which is $O(\sqrt{\ln pn})$, hence negligible for most realistic purposes. □

10. The exponential family and generalized linear models

The linear model, introduced in Example 9.3, has historically served as the canonical tool for regression analysis. At its core, it assumes that a response vector $\mathbf{y} \in \mathbb{R}^n$ can be represented as

$$\mathbf{y} = \mathbf{x}\beta + \mathbf{e}, \qquad \mathbf{e} \sim \mathcal{N}(0, \sigma^2\mathrm{Id}_n), \tag{10.1}$$

where $\mathbf{x} \in \mathbb{R}^{n\times p}$ is the design matrix of predictors[44], $\beta \in \mathbb{R}^p$ is the parameter vector, and $\mathbf{e}$ is a homoscedastic normal error. As discussed in Section 9, these assumptions lead to tractable maximum likelihood estimation (which coincides with least squares), exact inference based on normal theory leading to a high degree of interpretability through parameter recovery, and elegant prediction properties for the mean response. In practice, however, empirical data rarely conforms to the normal-homoscedastic paradigm as outcomes may be binary (such as success or failure in a Bernoulli trial), counts (as in Poisson processes), or strictly positive, highly skewed data (for example, waiting times). In such cases, the linear model becomes conceptually inadequate, since it implicitly assumes additivity on the original scale of the response with a variance that is independent of the mean. These limitations motivate the consideration of a broader framework, where the key step lies in recognizing the role of a much wider family of distributions to which the response is supposed to follow.

[43] For instance, if the error is sub-Gaussian then the corresponding concentration inequalities in Section 5 might be useful.

[44] For simplicity, here we assume that no intercept is present and that $\mathbf{X}$ is fixed (Remark 9.5).

Definition 10.1. Let Y be a random variable whose pdf (or mdf), say $\psi(y;\theta)$, depends on an unknown parameter $\theta \in \mathbb{R}$. Then we say that ψ belongs to the *exponential family* if it takes the form

$$\psi(y;\theta) = \exp\left(\frac{\xi(\theta)y - b(\theta)}{\phi} + c(y,\phi)\right),$$

where $\phi > 0$ is the *dispersion parameter* and ξ, b and c are known functions. In this case, $\xi = \xi(\theta)$ is the *natural parameter* of Y.

We represent this by $Y \sim \mathsf{ExpFam}(\xi, \phi)$, leaving implicit the dependence on b and c. For simplicity, we assume that ϕ is known. To ensure that θ is well defined as a function of ξ, we assume throughout that $\xi' \neq 0$, with the prime meaning derivative with respect to θ. The next result shows that the expectation and variance of Y may be expressed as rational functions of derivatives of ξ and b up to second order.

Proposition 10.2. *Under the conditions above there hold*

$$\mathbb{E}(Y) = \frac{b'}{\xi'}, \quad \mathrm{var}(Y) = \frac{\phi(\xi' b'' - b'\xi'')}{(\xi')^3}. \tag{10.2}$$

Proof. The log-likelihood (for a single observation of Y) is

$$l = \frac{\xi y - b}{\phi} + c,$$

from which we find that

$$l' = \frac{\xi' y - b'}{\phi}. \tag{10.3}$$

Now, with this notation (8.23) says that $\mathbb{E}(l') = 0$, which immediately yields the expression for $\mathbb{E}(Y)$. On the other hand, (8.24) means that $\mathbb{E}((l')^2) = -\mathbb{E}(l''))$, which gives

$$\frac{1}{\phi^2}\left((\xi')^2\mathbb{E}(y^2) - 2\xi' b'\mathbb{E}(y) + (b')^2\right) = -\frac{1}{\phi}\left(\xi''\mathbb{E}(Y) - b''\right) = -\frac{1}{\phi}\left(\xi''\frac{b'}{\xi'} - b''\right).$$

Since $\xi' b'\mathbb{E}(y) = (b')^2$, we may rearrange terms in order to get

$$\frac{1}{\phi^2}(\xi')^2\mathrm{var}(Y) = \frac{1}{\phi^2}(\xi')^2\left(\mathbb{E}(Y^2) - \mathbb{E}(Y)^2\right) = \frac{1}{\phi}\frac{\xi' b'' - b'\xi''}{\xi'},$$

which completes the proof. □

Definition 10.3. If $Y \sim \mathsf{ExpFam}(\xi, \phi)$ then its *mean* and *variance* functions are respectively given by

$$\mu = \mathbb{E}(Y), \quad V(\mu) = \dot{\mu},$$

where the dot means derivative with respect to ξ.

Proposition 10.4. *(Mean-variance relationship) If* $Y \sim \mathsf{ExpFam}(\xi, \phi)$ *then*

$$\mathrm{var}(Y) = \phi V(\mu). \tag{10.4}$$

Proof. Immediate from (10.2) and the chain rule. □

Example 10.5. (Naturality) A distribution in the exponential family is called *natural* when $\xi(\theta) = \theta$, so that θ itself is the natural parameter to be estimated. In this case, (10.2) reduces to

$$\mu = \dot{b}, \quad \operatorname{var}(Y) = \phi\ddot{b}. \tag{10.5}$$

From this it is straightforward to verify that the specific form of the mean–variance relationship essentially determines the distribution within this subclass; see [Jor97, Theorem 2.11]. Now, if $Y_j \sim \mathsf{ExpFam}(\theta, \phi)$ is a random sample then the corresponding log-likelihood is

$$l(\mathbf{y}; \theta) = \frac{1}{\phi}\left(\theta \sum_j y_j - nb(\theta)\right) + \sum_j c(y_j, \phi),$$

so the score is

$$l_\theta(\mathbf{y}; \theta) = \frac{n}{\phi}\left(\overline{y}_n - b'(\theta)\right) \tag{10.6}$$

and the ML estimator $\widehat{\theta}$ is determined by the equation $b'(\widehat{\theta}) = \overline{Y}_n$. Since (10.4) clearly implies that $b'' > 0$, it follows that b' is strictly increasing and we get $\widehat{\theta} = (b')^{-1}(\overline{Y}_n)$. Thus, the MLE of the natural parameter θ depends only on the sample mean. Also, by means of (10.4), (10.5) and (10.6) we may rewrite the score in terms of μ as

$$l_\mu(\mathbf{y}; \mu) = \frac{d\theta}{d\mu} l_\theta(\mathbf{y}; \theta(\mu)) = \frac{n}{\phi V(\mu)}\left(\overline{y}_n - b'(\theta)(\mu)\right) = \frac{n}{\phi V(\mu)}\left(\overline{y}_n - \mu\right), \tag{10.7}$$

so the ML estimator $\widehat{\mu}$ of μ is the sample mean. Turning to asymptotics, from (10.6) and (8.24) we find that the Fisher information is

$$\mathcal{F}_{(n)}(\theta) = \frac{nb''(\theta)}{\phi}, \tag{10.8}$$

so that Theorem 8.37 and consistency give the large sample estimate

$$\widehat{\theta}_n \approx \mathcal{N}\left(\theta, \frac{\phi}{nb''(\widehat{\theta}_n)}\right).$$

For the mean parameter μ we may either start with (10.7) and (8.23), which directly gives the corresponding Fisher information

$$\mathcal{F}_{(n)}(\mu) = n/\phi V(\mu),$$

or use that $\mu' = V = b''$ together with the delta method, as explained in Remark 8.25, thus obtaining the large sample estimate

$$\widehat{\mu}_n \approx \mathcal{N}\left(\mu, \frac{\phi V(\widehat{\mu}_n)}{n}\right).$$

As usual, these asymptotic normality results immediately provide the basis for constructing large-sample confidence intervals for the parameters θ and μ (cf. Remark 8.26). □

Remark 10.6. (Kullback–Leibler divergence within an exponential family) Let $\{\psi(\cdot; \theta); \theta \in \mathbb{R}\}$ be a one-parameter exponential family as in Definition 10.1. Since, for any $\theta_0, \theta \in \mathbb{R}$,

$$\ln \frac{\psi(y; \theta_0)}{\psi(y; \theta)} = \frac{\big(\xi(\theta_0) - \xi(\theta)\big)y - \big(b(\theta_0) - b(\theta)\big)}{\phi},$$

the corresponding Kullback–Leibler divergence from Definition 8.2 is

$$\begin{aligned} D_{\theta_0}^{KL}(\theta) &= \mathbb{E}_{\theta_0}\left(\ln \frac{\psi(Y;\theta_0)}{\psi(Y;\theta)}\right) \\ &= \frac{1}{\phi}\left(\big(\xi(\theta_0)-\xi(\theta)\big)\mathbb{E}_{\theta_0}(Y) - \big(b(\theta_0)-b(\theta)\big)\right), \end{aligned}$$

so that Proposition 10.2 yields

$$D_{\theta_0}^{KL}(\theta) = \frac{1}{\phi}\left[\big(\xi(\theta_0)-\xi(\theta)\big)\frac{b'(\theta_0)}{\xi'(\theta_0)} - \big(b(\theta_0)-b(\theta)\big)\right].$$

In the natural case of Example 10.5, where $\xi(\theta)=\theta$, this expression simplifies to

$$D_{\theta_0}^{KL}(\theta) = \frac{1}{\phi}\Big(b(\theta)-b(\theta_0)-b'(\theta_0)(\theta-\theta_0)\Big).$$

In particular, by convexity of b, the Kullback–Leibler divergence is nonnegative and vanishes if and only if $\theta_0=\theta$. Moreover, by expanding b around $\theta=\theta_0$ and using (10.8) we find that

$$D_{\theta_0}^{KL}(\theta) = \frac{1}{2}\mathscr{F}_{(1)}(\theta_0)\,(\theta-\theta_0)^2 + o\left((\theta-\theta_0)^2\right). \tag{10.9}$$

Thus, in the natural parametrization of an exponential family, the Fisher information $\mathscr{F}_{(1)}(\theta_0)$ determines the local second-order behavior of the Kullback–Leibler divergence centered at θ_0, which is consistent with the general result in Remark 8.35. □

	θ	Likelihood $L(y;\theta)$	ξ	$\theta=\theta(\xi)$	ϕ	b	$\mu=\mathbb{E}(Y)$	$V(\mu)$
Binomial (Example 2.40)	p	$\binom{n}{y}p^y(1-p)^{n-y}$	$\ln\frac{p}{1-p}$	$\frac{1}{1+e^{-\xi}}$	1	$-n\ln(1-p)$	p	$\mu(1-\mu)$
Poisson (Example 2.41)	λ	$\frac{e^{-\lambda}\lambda^y}{y!}$	$\ln\lambda$	e^{ξ}	1	λ	λ	μ
Normal with σ known (Definition 4.1)	μ	$\frac{1}{\sqrt{2\pi}\sigma}\,e^{-\lvert y-\mu\rvert^2/2\sigma^2}$	μ	ξ	σ^2	$\frac{\mu^2}{2}$	μ	1
Gamma with λ known (Definition 4.19)	α	$\frac{\alpha^\lambda}{\Gamma(\lambda)}y^{\lambda-1}e^{-\alpha y}$	$-\alpha$	$-\xi$	1	$-\lambda\ln\alpha$	$\frac{\lambda}{\alpha}$	$\frac{\mu}{\alpha}$

TABLE 1. Examples of distributions in the exponential family

As shown in Table 1, most of the distributions considered so far can be expressed as members of the exponential family[45]. A distinctive role is played by the normal distribution, which is the only one in the table whose variance is entirely independent of the mean. This observation paves the way for a substantial enrichment of the class of regression models, while still preserving the desirable inferential properties of the classical linear model, as will be seen below.

Definition 10.7. A *generalized linear model (GLM)* for independent responses $\{Y_i\}_{i=1}^n$ consists of the following ingredients:

(1) **Random component:** each Y_i follows a one-parameter exponential family: $Y_i \sim \mathsf{ExpFam}(\xi_i,\phi)$, where ξ_i is the canonical parameter and ϕ a common dispersion parameter.

[45]A simple, commonly used example outside the exponential family is Student's t-distribution in Definition 4.29.

(2) **Systematic component:** a linear predictor

$$\eta_i = \mathbf{x}_i^\top \beta, \tag{10.10}$$

linking covariates $\mathbf{x}_i$ to coefficients β.

(3) **Link function:** a monotone differentiable map g connecting the mean $\mu_i := \mathbb{E}(Y_i)$ to the predictor,

$$g(\mu_i) = \eta_i.$$

When $g(\mu_i) = \xi_i$, the link is called *canonical*.

Thus, the GLM extends the linear model by allowing non-normal response distributions and by permitting nonlinear, yet monotone, transformations in the relationship between the mean response and the linear predictor. In particular, since the systematic component is linear in β and the non-linearity only affects the mean, GLMs remain interpretable in the sense of Remark 9.22, while still retaining much of its predictive power. From (10.4) it also follows that

$$\mathrm{Var}(Y_i) = \phi V(\mu_i),$$

which shows that *heteroscedasticity* (unequal variances across observations) is inherent to a GLM.

Example 10.8. Because of their flexibility, which balances mathematical rigor with empirical applicability, GLMs are widely used in both theory and applications [Agr15, DB18]. Here we restrict ourselves to three of the most prominent examples, corresponding to the first three rows of Table 1:

- *Logistic regression* arises when $Y_i \sim \mathsf{Ber}(p_i)$, a Bernoulli distribution, with the *logit link*

$$\eta_i = \mathrm{logit}(p_i) := \ln \frac{p_i}{1-p_i}. \tag{10.11}$$

  Equivalently, if we solve for $p_i = P(X_i = 1|x_i)$,

$$P(X_i = 1|x_i) = \mathrm{logit}^{-1}(\eta_i) = \frac{1}{1+e^{-\mathbf{x}_i^\top \beta}}. \tag{10.12}$$

- *Poisson regression* corresponds to $Y_i \sim \mathsf{Pois}(\lambda_i)$, a Poisson distribution, with the *log link*

$$\eta_i = \ln \mu_i.$$

- The classical *linear model* (from Example 9.3, with σ^2 known) assumes $Y_i \sim \mathcal{N}(\mu_i, \sigma^2)$ with the *identity link* $\eta_i = \mu_i$.

Note that the link is canonical in all these cases. □

We now turn to the most basic aspects of the estimation framework for GLMs. Since these models can be regarded as natural extensions of the normal linear model, it is reasonable to adopt maximum likelihood as the method for estimating β (cf. Example 9.3). The corresponding log-likelihood for n observations is

$$l(\mathbf{y};\beta) = \sum_i l_i(\mathbf{y};\beta), \quad l_i(\mathbf{y};\beta) = \frac{y_i\xi_i - b(\xi_i)}{\phi} + c(y_i,\phi), \tag{10.13}$$

so we should compute

$$\frac{\partial l_i}{\partial \beta_j} = \frac{\partial l_i}{\partial \xi_i}\frac{\partial \xi_i}{\partial \mu_i}\frac{\partial \mu_i}{\partial \eta_i}\frac{\partial \eta_i}{\partial \beta_j}, \quad j = 1, \cdots, p,$$

with the likelihood equations being obtained by summing up these terms over i and equating the result to zero (Definition 8.3). Now, (10.3) yields

$$\frac{\partial l_i}{\partial \xi_i} = \frac{\frac{\partial \xi_i}{\partial \theta_i} y_i - \frac{\partial b}{\partial \theta_i}}{\phi},$$

which together with (10.2) gives

$$\frac{\partial l_i}{\partial \xi_i} = \frac{\frac{\partial \xi_i}{\partial \theta_i}(y_i - \mu_i)}{\phi} \text{ and } \frac{\partial \xi_i}{\partial \mu_i} = \frac{\phi}{\frac{\partial \xi_i}{\partial \theta_i}\mathrm{var}(Y_i)}.$$

Also, (10.10) implies

$$\frac{\partial \eta_i}{\partial \beta_j} = x_{ij},$$

so if we put all the pieces of our computation together we obtain the following fundamental result.

Proposition 10.9. *The maximum likelihood estimator $\widehat{\beta}_{GLM}$ of a GLM satisfies the system of equations*

$$\mathbf{x}^\top D V^{-1}(\mathbf{y} - \boldsymbol{\mu}) = 0, \tag{10.14}$$

where $\boldsymbol{\mu} = (\mu_1, \cdots, \mu_n)^\top$, $D = \mathrm{diag}(\partial \mu_i / \partial \eta_i)$, a diagonal matrix whose entries depend on the specific shape of the link function of the model, and $V = \mathrm{diag}(\mathrm{var}(Y_i))$. As a consequence, if the link function is canonical then this reduces to

$$\mathbf{x}^\top(\mathbf{y} - \boldsymbol{\mu}) = 0, \tag{10.15}$$

Proof. The calculation above shows that the score components are

$$\frac{\partial l_i}{\partial \beta_j} = \frac{y_i - \mu_i}{\mathrm{var}(Y_i)} x_{ij} \frac{\partial \mu_i}{\partial \eta_i}, \tag{10.16}$$

so the defining condition for $\widehat{\beta}_{GLM}$, $\partial l_i / \partial \beta_j = 0$, is equivalent to (10.14). As for the last assertion, from $\eta_i = \xi_i$ we find that

$$\frac{\partial \mu_i}{\partial \eta_i} = \frac{\partial \mu_i}{\partial \xi_i} = \frac{\partial^2 b}{\partial \xi_i^2},$$

where we used (10.5) in the last step. Also, again by (10.5),

$$\mathrm{var}(Y_i) = \phi \frac{\partial^2 b}{\partial \xi_i^2}.$$

Together, these identities imply that $DV^{-1} = \phi^{-1}\mathrm{Id}_n$. □

Although the dependence on β (and hence on $\widehat{\beta}_{GLM}$) is not explicit in either (10.14) or (10.15), it is in fact present because $\mu_i = g^{-1}(\mathbf{x}_i^\top \beta)$. In general this dependence is non-linear, so the likelihood equations must be solved for β by means of an iterative method (usually, Newton-Raphson). As an illustration, in the logistic model the equations reduce to $\mathbf{x}^\top(\mathbf{y} - \mathbf{p}) = 0$, where $\mathbf{p} = (p_1, \ldots, p_n)^\top$. Here the non-linearity is entirely due to the inverse logit relation in (10.12). By contrast, if the GLM specializes to the classical linear model, then $\boldsymbol{\mu} = \mathbf{x}\beta$ (linearity) and (10.15) simplifies to $\mathbf{x}^\top(\mathbf{y} - \mathbf{x}\beta) = 0$, which directly yields the usual least squares estimator under the standard assumptions.

With the maximum likelihood framework established, we now briefly examine the asymptotic properties of $\widehat{\beta}_{GLM}$. From (10.16) and (8.24) we find that the Fisher information for the i^{th} observation is

$$\mathcal{F}_{jk}^{(i)} = \mathbb{E}\left(\frac{\partial l_i}{\partial \beta_j}\frac{\partial l_i}{\partial \beta_k}\right) = \frac{x_{ij}x_{ik}}{\mathrm{var}(Y_i)}\left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2,$$

so that, by independence, the Fisher information for the entire sample $Y^{[n]} = (Y_1, \ldots, Y_n)$ is

$$\mathcal{F} = \sum_i \mathcal{F}^{(i)} = \mathbf{x}^\top W \mathbf{x}, \qquad W = \mathrm{diag}\left(\frac{(\partial \mu_i / \partial \eta_i)^2}{\mathrm{var}(Y_i)}\right).$$

Therefore, applying Theorem 8.24 (see also its generalization in Remark 8.28) we deduce that as $n \to \infty$,

$$\widehat{\beta}_{GLM} \approx \mathcal{N}(\beta, (\mathbf{x}^\top W \mathbf{x})^{-1}). \tag{10.17}$$

As is customary, consistency permits the substitution of W by $\widehat{W} = W(\widehat{\beta}_{GLM})$, leading to the practical approximation

$$\widehat{\beta}_{GLM} \approx \mathcal{N}(\beta, (\mathbf{x}^\top \widehat{W} \mathbf{x})^{-1}), \tag{10.18}$$

which forms the basis for constructing large-sample confidence intervals for the components of β. In the linear model case (with σ^2 known) we have $\partial\mu_i/\partial\eta_i = 1$ and $\mathrm{var}(Y_i) = \sigma^{-2}$, so that $W = \sigma^{-2}\mathrm{Id}_n$ and (10.18) essentially reduces to (9.54).

With the appropriate care, most of the well-established estimation theory for the linear model can thus be carried over to this broader framework of GLMs. In particular, notions such as asymptotic efficiency, hypothesis testing, and likelihood-based inference retain essentially the same mathematical structure, even though the underlying distribution of the response is no longer normal [Agr15, DB18, GTP19]. This transfer of results is precisely what makes GLMs so attractive: they extend the familiar tools of linear regression to a far wider range of data types, while preserving a rigorous probabilistic foundation. As a consequence, GLMs provide a unified language for both theoretical developments and applied work, bridging the gap between classical models and modern data analysis.

Example 10.10. (AIC for GLMs) The AIC for a GLM with a canonical link may be computed explicitly. Indeed, since $\xi_i = \eta_i = \mathbf{x}_i^\top \beta$, it follows from (8.59) and (10.13) that

$$\mathrm{AIC} = -\frac{2}{\phi} \sum_i \Big(y_i \mathbf{x}_i^\top \widehat{\beta}_{GLM} - b(\mathbf{x}_i^\top \widehat{\beta}_{GLM}) \Big) + 2(p+2), \tag{10.19}$$

where $\widehat{\beta}_{GLM}$ denotes the corresponding MLE and p is the number of regressors (with the intercept excluded). In particular, this formula applies to the logistic model discussed in Example 10.8. Since $\theta_i = p_i$, the first row in Table 1 (with $n = 1$) yields

$$\theta_i(\xi_i) = \frac{1}{1 + e^{-\xi_i}},$$

and therefore

$$b(\xi_i) = -\ln(1 - \theta_i) = \ln(1 + e^{\xi_i}).$$

We thus conclude that

$$\mathrm{AIC} = 2 \sum_i \Big(\ln\Big(1 + e^{\mathbf{x}_i^\top \widehat{\beta}_{GLM}}\Big) - y_i \mathbf{x}_i^\top \widehat{\beta}_{GLM} \Big) + 2(p+1),$$

Comparing this expression with (10.19), we see that the penalty term $p + 2$ is replaced by $p + 1$, which is consistent with the fact that, in the logistic model, the variance function is $V(\mu) = \mu(1 - \mu)$. □

Example 10.11. (GLMs and Item Response Theory) A particularly fruitful domain where generalized linear models intersect with modern statistical methodology is *Item Response Theory* (IRT), which plays a central role in psychometrics and educational assessment [HSR91, DA13, VdL16]. Conceptually, IRT can be regarded as a GLM with a latent predictor, where the individual ability parameter γ_j functions as an unobserved covariate, typically following a centered normal, say $\gamma_j \sim \mathcal{N}(0, 1)$. For instance, in the classical *Rasch model*, the probability that an individual j with ability γ_j answers item i correctly is

$$P(Y_{ij} = 1|_{\gamma_j, b_i}) = \mathrm{logit}^{-1}(\gamma_j - b_i),$$

where b_i is the item difficulty parameter. This is directly analogous to the inverse logit link (10.12) in the GLM framework above, with linear predictor $\eta_{ij} = \gamma_j - b_i$. More generally, variants of the logistic regression model in (10.11) underlie both GLMs and IRT, with the key distinction being that in IRT, part of the predictor vector

corresponds to latent person parameters rather than observed covariates. In this way, the *2PL model* extends the Rasch model by introducing item discrimination a_i,

$$P(Y_{ij} = 1|_{\gamma_j, a_i, b_i}) = \text{logit}^{-1}\big(a_i(\gamma_j - b_i)\big), \tag{10.20}$$

while the *3PL model*, widely used in practice, adds a pseudo-guessing parameter c_i. We should also point out that from an asymptotic perspective, the connection between GLMs and IRT is especially revealing. Since IRT models are essentially Bernoulli GLMs with latent predictors, the same large-sample principles apply: maximum likelihood estimators of item parameters (difficulty b_i, discrimination a_i, and pseudo-guessing c_i) are consistent and asymptotically normal under standard regularity conditions, making sure that appropriate versions of Fisher's foundational conception in Theorem 8.24 remain operational in this broader context. In particular, the Fisher information for the IRT likelihood plays the same role as in the GLM framework, forming the basis for variance formulas and for the construction of confidence intervals and hypothesis tests. We illustrate these ideas by developing the corresponding asymptotic theory for the 2PL model in (10.20). To simplify matters, we estimate the respondent's ability γ_j under the assumption that the item parameters (a_i, b_i) are known [46]. As usual, we assume *local independence*, meaning that the N responses $Y_i = Y_{ij}$ are conditionally independent given γ_j, i.e. $\{Y_i|_{\gamma_j}\}_{i=1}^N$ is independent. Accordingly, and in alignment with (8.14), the corresponding log-likelihood is

$$l(\mathbf{y}; \gamma_j) = \sum_i \left(y_i \ln P_i(\gamma_j) + (1 - y_i) \ln Q_i(\gamma_j)\right), \tag{10.21}$$

where $P_i(\gamma_j)$ is a shorthand for the expressions in (10.20) and $Q_i(\gamma_j) = 1 - P_i(\gamma_j)$. Hence, the associated score function is

$$\begin{aligned} s(\mathbf{y}; \gamma_j) &= \sum_i \frac{\partial}{\partial \gamma_j} \left(y_i \ln P_i + (1 - y_i) \ln Q_i\right) \\ &= \sum_i \left\{ y_i \left(\frac{1}{P_i} a_i P_i Q_i \right) + (1 - y_i) \left(\frac{1}{Q_i} (-a_i P_i Q_i) \right) \right\}, \end{aligned}$$

which simplifies to

$$s(\mathbf{y}; \gamma_j) = \sum_i a_i (y_i - P_i(\gamma_j)). \tag{10.22}$$

The maximum likelihood estimator therefore satisfies

$$\sum_i a_i P_i(\widehat{\gamma}_j) = \sum_i a_i y_i, \tag{10.23}$$

a non-linear equation in $\widehat{\gamma}_j$ that must be solved numerically. The Fisher information follows from (10.22) and (8.15):

$$\begin{aligned} \mathscr{F}(\gamma_j) &= \mathbb{E}\left(\left(\sum_i a_i (Y_i - P_i) \right)^2 \right) \\ &= \sum_i \mathbb{E}\big(a_i^2 (Y_i - P_i)^2\big) + \sum_{i \neq k} \mathbb{E}(a_i a_k (Y_i - P_i)(Y_k - P_k)), \end{aligned}$$

[46] We are making two simplifying assumptions here. First, the item parameters are assumed to have been calibrated prior to analysis, so that their estimation uncertainty is ignored; such pre-calibration is routinely performed in large-scale assessments and adaptive testing systems (e.g., PISA, ENEM, TOEFL). Second, although the latent traits γ_j are modeled as random effects ($\gamma_j \sim \mathcal{N}(0, 1)$), we condition on the observed response patterns and treat each γ_j as an unknown constant when estimating individual abilities. Both assumptions are relaxed in more general formulations, where item and person parameters are estimated jointly and the latent distribution is integrated into the likelihood [BK04, VdL16]. The connection between such hierarchical treatments in IRT and generalized linear mixed models is discussed in Remark 10.12.

with the mixed terms vanishing due to local independence, Proposition 2.6, and the fact that $\mathbb{E}(Y_i) = P_i$ (recall that $Y_i \sim \mathsf{Ber}(P_i)$). Since $\mathrm{var}(Y_i) = P_iQ_i$, we thus obtain

$$\mathscr{F}(\gamma_j) = \sum_i a_i^2 P_i(\gamma_j)Q_i(\gamma_j), \tag{10.24}$$

and consequently the large-sample approximation

$$\widehat{\gamma}_j \approx \mathcal{N}\left(\gamma_j,\, \frac{1}{\sum_i a_i^2 P_i(\widehat{\gamma}_j)Q_i(\widehat{\gamma}_j)}\right), \tag{10.25}$$

which parallels the asymptotic result previously obtained for the GLM estimator $\widehat{\beta}_{\mathrm{GLM}}$. Thus, the GLM perspective not only clarifies the statistical structure of IRT but also provides a rigorous foundation for inference, ensuring that the asymptotic theory developed for GLMs can be effectively transplanted into psychometric applications. IRT, therefore, should not be seen as a distinct paradigm but as a specialized application of GLMs with latent predictors, offering a robust statistical framework for modeling educational and psychological measurement. □

Remark 10.12. (GLMMs as the bridge between GLMs and IRT). From a conceptual standpoint, the passage from generalized linear models (GLMs) to Item Response Theory (IRT) naturally goes through an intermediate class, namely, *generalized linear mixed models (GLMMs)* [Str13, JN21]. In a GLMM, the linear predictor of a GLM is extended by the inclusion of random effects, allowing part of the variation in the response to be attributed to unobserved random components. Formally, while a GLM is written as

$$g\left(\mathbb{E}(Y_i|_{\mathbf{X}=\mathbf{x}})\right) = \mathbf{x}_i^\top \beta, \quad i = 1, \cdots, n,$$

with Y_i following a member of the exponential family, a GLMM generalizes this expression to

$$g\left(\mathbb{E}(Y_i|_{\mathbf{X}=\mathbf{x},\Gamma=\gamma})\right) = \mathbf{x}_i^\top \beta + \mathbf{z}_i^\top \gamma,$$

where $\gamma \in \mathbb{R}^q$ comprises the random effects associated with an individual labeled by i, typically supposed to follow a centered normal distribution, and $\mathbf{z}$ is the associated $n \times q$ matrix design, which we assume fixed here. The latent ability parameter in IRT fulfills exactly this role: it acts as a random effect at the individual level, representing an unobserved source of variability across respondents. Accordingly, IRT models may be regarded as Bernoulli GLMMs in which the random component captures the heterogeneity among individuals that remains unobserved in the classical GLM framework. Thus, starting with linear models, at each step new layers of generality emerge—link functions, random components, latent traits—culminating in the IRT framework, where the random effect becomes not a nuisance term but the very object of substantive interpretation [DBW04].

11. SUFFICIENCY

In a statistical model, consider moving from the random sample

$$X = (X_1, \ldots, X_n), \quad X_j \sim \psi_\theta,$$

to an estimator $\widehat{\theta}$ defined through a statistic $h = h(X)$. A natural question then arises: how much information from the data has actually been retained in this transition? A complete answer would require a precise definition of the amount of information carried by the sample, which lies beyond the scope of these notes. A more modest but still important task is to verify whether the chosen statistic captures *all* the relevant information about the parameter θ, in the sense that no additional knowledge from the sample is required for its estimation. Put differently, the aim is to identify situations where the "extra randomness" in the sample X that is not reflected in $h(X)$ is unrelated to θ, and thus irrelevant for inference. This idea admits a neat probabilistic formulation in terms of conditional distributions, as introduced in Section 3.

Definition 11.1. A statistic $h = h(X)$ is said to be *sufficient* if, for any realization $\mathbf{x}$ of X, the conditional probability distribution $\psi_{\theta;X|h(X)=h(\mathbf{x})}$ evaluated at $\mathbf{x}$ does *not* depend on θ.

Using (3.3), it follows that sufficiency of h ensures the existence of a function $\xi = \xi(\mathbf{x})$ such that

$$\frac{\psi_{\theta;(h(X),X)}(h(\mathbf{x}),\mathbf{x})}{\psi_{\theta;h(X)}(h(\mathbf{x}))} = \xi(\mathbf{x}). \tag{11.1}$$

A key observation is that the inclusion of events $\{X = \mathbf{x}\} \subset \{h(X) = h(\mathbf{x})\}$ implies

$$\psi_{\theta;(h(X),X)}(h(\mathbf{x}),\mathbf{x}) = \psi_{\theta;X}(\mathbf{x}) = L(\mathbf{x};\theta), \tag{11.2}$$

the likelihood function. Substituting this into (11.1) yields a practical characterization of sufficiency: it occurs precisely when the dependence of $L(\mathbf{x};\theta)$ on θ is confined to a factor that depends on $\mathbf{x}$ only through the statistic h. This captures the essential content of the notion: all the information needed to estimate θ is already contained in the sufficient statistic, making further reference to the raw data X unnecessary.

Theorem 11.2. *(Fisher-Neyman factorization) h is sufficient if and only if the likelihood function factorizes as*

$$L(\mathbf{x};\theta) = \eta(h(\mathbf{x}),\theta)\xi(\mathbf{x}), \tag{11.3}$$

for positive functions η and ξ.

Proof. We have already seen that sufficiency implies (11.3). For the converse we first note that (11.2) leads to

$$\begin{aligned}\psi_{\theta;h(X)}(h(\mathbf{x})) &= \int_{\{\mathbf{x}':h(\mathbf{x}')=h(\mathbf{x})\}} \psi_{\theta;(h(X),X)}(h(\mathbf{x}'),\mathbf{x}')d\mathbf{x}' \\ &= \int_{\{\mathbf{x}':h(\mathbf{x}')=h(\mathbf{x})\}} \psi_{\theta;X}(\mathbf{x}')d\mathbf{x}'\end{aligned}$$

so we may again use (3.3) to compute:

$$\begin{aligned}\psi_{\theta;X|h(X)=h(\mathbf{x})} &= \frac{\psi_{\theta;(h(X),X)}(h(\mathbf{x}),\mathbf{x})}{\psi_{\theta;h(X)}(h(\mathbf{x}))} \\ &= \frac{\eta(h(\mathbf{x}),\theta)\xi(\mathbf{x})}{\int_{\{\mathbf{x}':h(\mathbf{x}')=h(\mathbf{x})\}} \eta(h(\mathbf{x}'),\theta)\xi(\mathbf{x}')d\mathbf{x}'} \\ &= \frac{\eta(h(\mathbf{x}),\theta)\xi(\mathbf{x})}{\eta(h(\mathbf{x}),\theta)\int_{\{\mathbf{x}':h(\mathbf{x}')=h(\mathbf{x})\}} \xi(\mathbf{x}')d\mathbf{x}'}.\end{aligned}$$

Thus,

$$\psi_{\theta;X|h(X)=h(\mathbf{x})} = \frac{\xi(\mathbf{x})}{\int_{\{\mathbf{x}':h(\mathbf{x}')=h(\mathbf{x})\}} \xi(\mathbf{x}')d\mathbf{x}'}$$

only depends on $\mathbf{x}$. □

Corollary 11.3. *A unique ML estimator is a function of a sufficient statistic. More generally, if a ML estimator exists then an ML estimator may be chosen so as to be a function of a sufficient statistic.*

Proof. Given that the ML estimator $\widehat{\theta}$ is obtained by maximizing the likelihood function $L(\mathbf{x};\theta)$ in θ (for each $\mathbf{x}$), this is an obvious consequence of (11.3). □

Example 11.4. (Sufficiency in a normal population) If $X_j \sim \mathcal{N}(\mu, \sigma^2)$ we know from Example 8.5 that

$$L(\mathbf{x};\theta) = (2\pi\theta_2)^{-n/2} e^{-\frac{1}{2\theta_2}\sum_{j=1}^{n}(x_j-\theta_1)^2}, \quad \mathbf{x} = (x_1, \cdots, x_n),$$

where $\theta = (\theta_1, \theta_2) = (\mu, \sigma^2) \in \Theta = \mathbb{R} \times \mathbb{R}_+$. We distinguish three cases:

- (θ_2 is known and θ_1 is the unknown parameter) Set

$$h_1(\mathbf{x}) = \frac{1}{n}\sum_j x_j$$

so that $\sum_j (x_j - h_1(x)) = 0$ implies

$$\begin{aligned}\sum_j (x_j-\theta_1)^2 &= \sum_j (x_j - h_1(\mathbf{x}) + h_1(\mathbf{x}) - \theta_1)^2 \\ &= \sum_j (x_j - h_1(\mathbf{x}))^2 + n(h_1(\mathbf{x}) - \theta_1)^2,\end{aligned}$$

which leads to the factorization

$$L(\mathbf{x};\theta_1) = \underbrace{(2\pi\theta_2)^{-n/2} e^{-\frac{\sum_j (x_j - h_1(\mathbf{x}))^2}{2\theta_2}}}_{\xi(\mathbf{x})} \underbrace{e^{-\frac{n(h_1(\mathbf{x})-\theta_1)^2}{2\theta_2}}}_{\eta(h_1(\mathbf{x}),\theta_1)}. \tag{11.4}$$

This shows that h_1 is a sufficient statistic for θ_1 (given θ_2).

- (θ_1 is known and θ_2 is the unknown parameter) Here,

$$h_2(\mathbf{x}) = \sum_j (x_j - \theta_1)^2$$

qualifies as a statistic and

$$L(\mathbf{x};\theta_2) = \underbrace{(2\pi\theta_2)^{-n/2} e^{-\frac{h_2(\mathbf{x})}{2\theta_2}}}_{\eta(h_2(\mathbf{x}),\theta_2)} \times \underbrace{1}_{\xi(\mathbf{x})}$$

shows that h_2 is a sufficient statistic for θ_2 (given θ_1).

- ($\theta = (\theta_1, \theta_2)$ is the unknown bi-dimensional parameter). Here we set

$$\widetilde{h}_2(\mathbf{x}) = \sum_j (x_j - h_1(\mathbf{x}))^2$$

so (11.4) gives

$$L(\mathbf{x};\theta) = \underbrace{(2\pi\theta_2)^{-n/2} e^{-\frac{\widetilde{h}_2(\mathbf{x}) + n(h_1(\mathbf{x})-\theta_1)^2}{2\theta_2}}}_{\eta((h_1(\mathbf{x}),\widetilde{h}_2(\mathbf{x}),\theta))} \times \underbrace{1}_{\xi(\mathbf{x})},$$

which shows that $H(\mathbf{x}) = (h_1(\mathbf{x}), \widetilde{h}_2(\mathbf{x}))$ is a sufficient statistic for θ. We thus see that the common practice, which has been extensively used in Subsection 7.3, of regarding H as a sufficient statistic when sampling from a normal population, is fully justified. □

Example 11.5. (Sufficiency in an exponential population) If $X_j \sim \mathrm{Exp}(\lambda)$ then from Example 8.7 we get

$$L(\mathbf{x};\lambda) = \underbrace{\lambda^n e^{-\lambda h(\mathbf{x})}}_{\eta(h(\mathbf{x}),\lambda)} \times \underbrace{1}_{\xi(\mathbf{x})},$$

where $h(\mathbf{x}) = \sum_j x_j$ is a sufficient statistic for λ. □

Example 11.6. (Sufficiency in a Bernoulli or Poisson population) If $X_j \sim \mathsf{Ber}(p)$ then Example 8.8 gives

$$L(\mathbf{x};p) = \underbrace{p^{h(\mathbf{x})}(1-p)^{n-h(\mathbf{x})}}_{\eta(h(\mathbf{x}),p)} \times \underbrace{1}_{\xi(\mathbf{x})},$$

which shows that $h(\mathbf{x}) = \sum_j x_j$ is a sufficient statistic for estimating p. On the other hand, if $X_j \sim \mathsf{Pois}(\rho)$ then, again by Example 8.8,

$$L(\mathbf{x};\rho) = \underbrace{\rho^{k(\mathbf{x})}e^{-n\rho}}_{\eta(k(\mathbf{x}),\rho)} \times \underbrace{(\Pi_j x_j!)^{-1}}_{\xi(\mathbf{x})},$$

which confirms that $k(\mathbf{x}) = \sum_j x_j$ is a sufficient statistic for estimating ρ. □

Example 11.7. (Sufficiency in the natural exponential family) If $Y_j \sim \mathsf{ExpFam}(\theta,\phi)$ as in Example 10.5 then

$$L(\mathbf{y};\theta) = \underbrace{e^{\frac{\theta h(\mathbf{y}) - nb(\theta)}{\phi}}}_{\eta(h(\mathbf{y}),\theta)} \times \underbrace{e^{\sum_j c(y_j,\phi)}}_{\xi(\mathbf{y})},$$

which shows that $h(\mathbf{y}) = \sum_j y_j$ is a sufficient statistic for estimating θ. □

As illustrated by the computations in Remark 8.26, at least in the regime of large samples it follows from (8.38) that the dependence on sample data of confidence intervals for ML estimators occurs only through the estimator itself. This general observation clearly aligns with Corollary 11.3 and is definitely confirmed by all the examples examined above, where a simple relationship of the given sufficient statistic with the corresponding ML estimator is manifest.

12. HYPOTHESIS TESTING

Our aim here is to discuss a bit more on the heuristics behind the choices of the rejection regions appearing in the F-tests implemented in Remark 7.37 and Example 7.40 above.

12.1. **A glimpse at the Neyman-Pearson setup.**

As usual, we are given a parametric statistic model

$$X_1, \cdots, X_n \sim \psi_\theta, \quad \theta \in \Theta \subset \mathbb{R}^p$$

as in Definition 7.2 and Remark 7.3, so that $(\Omega, \mathcal{F}, \{\mathcal{P}_\theta\}_{\theta\in\Theta})$ is the underlying family of probability spaces and $P_\theta = \psi_\theta dx$ is the common distribution of the components of the associated random vector $X = (X_1, \cdots, X_n) : \Omega \to \mathbb{R}^n$. Given *disjoint* subsets $\Theta_0, \Theta_{\mathrm{a}} \subset \Theta$ with $\Theta = \Theta_0 \cup \Theta_{\mathrm{a}}$, *hypothesis testing* concerns the prospect of using the available data in an observed value $\mathbf{x}$ of X to provide statistical evidence for deciding between the *null hypothesis*

$$H_0 : \quad \theta \in \Theta_0$$

and the *alternative hypothesis*

$$H_{\mathrm{a}} : \quad \theta \in \Theta_{\mathrm{a}}.$$

One adheres to the usual asymmetry in regarding H_0 as the *status quo* and then chooses a statistics $h = h(X) : \Omega \to \mathbb{R}$ and a *rejection region* $R \subset \mathbb{R}$ so that H_0 gets *rejected* if the realization $h(\mathbf{x})$ of $h(X)$ takes value in R. A pair $T = (h, R)$ as above is called a *test* for the given statistical model and we denote by $\mathscr{T}$ the collection of all such tests[47]. As we will see in Remark 12.3 below, the eventual implementation of a test $T \in \mathscr{T}$ necessarily involves the knowledge of the distribution of (a perhaps complicated function of) $h(X)$ under the null hypothesis.

In order to quantify the possible types of errors in making such a decision, we consider the *power function* $\pi : \Theta \to [0,1]$ of (h, R),

$$\pi(\theta) = \mathcal{P}_\theta(h(X) \in R).$$

[47]For instance, if $R = [r, +\infty)$ then we say that r is a *critical value* for the test.

We then see that restriction to Θ_0, namely,

$$\gamma(\theta) := \pi|_{\Theta_0}(\theta), \quad \theta \in \Theta_0,$$

quantifies the *type I error* of rejecting H_0 when it is true, whereas restriction to $\Theta_{\rm a}$,

$$\delta(\theta) := \pi|_{\Theta_{\rm a}}(\theta), \quad \theta \in \Theta_{\rm a},$$

is such that

$$1 - \delta(\theta) = \mathcal{P}_\theta(h(X) \notin R)$$

measures the *type II error* of *not* rejecting H_0 when it is false. Ideally, one would seek for an strategy minimizing *both* errors at a time, but simple examples show that this is doomed to fail in general. The standard way to overcome this is to search for a test which minimizes type II error under the constraint that type I error remains uniformly bounded from above by a fixed amount given in advance.

Definition 12.1. Given $\alpha \in (0,1)$ we say that a test $T \in \mathscr{T}$ has *confidence level* α if

$$\sup_{\theta \in \Theta_0} \gamma(\theta) = \alpha, \tag{12.1}$$

and we denote by $\mathscr{T}_\alpha$ the collections of all such tests.

Definition 12.2. A test $T \in \mathscr{T}_\alpha$ is *uniformly most powerful* (UMP) if it satisfies (with self-explanatory notation)

$$\delta(\theta) \geq \delta^*(\theta), \quad \theta \in \Theta_{\rm a},$$

for any $T^* \in \mathscr{T}_\alpha$.

Remark 12.3. Note that (12.1), which may be rewritten as

$$\sup_{\theta \in \Theta_0} \mathcal{P}_\theta(h(X) \in R) = \alpha, \tag{12.2}$$

allows us to explicitly determine the rejection region R from the confidence level α only in case a (perhaps approximate) knowledge of the distribution of $h(X)$ *under* H_0 is at hand, a procedure illustrated in the examples considered below. In other words, the ubiquitous "Problem of Distribution" in Parametric Statistics resurfaces in this setting as well, although here the relevant statistics $h(X)$ gets restricted to the parametric region where the null hypothesis holds true. □

The celebrated Neyman-Pearson lemma [CB21, Theorem 8.3.12] exhibits a UMP test in the simple hypotheses case, where both Θ_0 and $\Theta_{\rm a}$ contain a single element. Unfortunately, such a test may not exist even for one of the simplest *composite* hypotheses cases, namely, a "two-sided" test of the form $\Theta \subset \mathbb{R}$ some open interval, $\Theta_0 = \{\theta_0\}$ for some $\theta_0 \in \Theta$ and $\Theta_1 = \Theta \backslash \Theta_0$ (as in Remark 7.37, for instance); see [CB21, Example 8.3.19] and the surrounding discussion for more on this rather delicate point. Of course, we may always restrict further the class of contenders where the ideal test should be sought (consistent, unbiased, etc.) but it seems that none of these strategies produces a test with optimal performance in *all* cases.

12.2. **Testing via likelihood ratios.** The state of affairs indicated in the previous paragraph suggests that, instead of *systematically* trying to find the best test in a given context, one should proceed *heuristically* so as to single out a family of tests which are relatively easy to implement, reproduce most known composite tests for samples of any size and have nice asymptotic properties (see Remark 12.6 below).

Recall from Subsection 8.3 that the ML estimators, computed in terms of the likelihood function as in Definition 8.1, have many remarkable properties, including asymptotic normality. Moreover, given the available

information contained in the realization $\mathbf{x}$ of the random sample X, the discussion surrounding (8.8) justifies regarding $\sup_{\theta\in\Theta_0} L(\mathbf{x};\theta)$ as the best evidence in favor of H_0 and $\sup_{\theta\in\Theta_a} L(\mathbf{x};\theta)$ as the best evidence in favor of H_a, which suggests formulating a hypothesis test based on the *likelihood ratio*

$$\mathbf{x}\in\mathbb{R}^n \mapsto \frac{\sup_{\theta\in\Theta_0} L(\mathbf{x};\theta)}{\sup_{\theta\in\Theta_a} L(\mathbf{x};\theta)} \in [0,+\infty]. \tag{12.3}$$

At the risk of (over)simplifying the exposition, but at the same time remaining in a generality that will suffice for the applications we have in mind, we assume from now on that $\Theta_0 \subset \Theta$ has a negligible size (as a subset of Θ) and $\Theta_a = \Theta\backslash\Theta_0$. Typically, $\Theta \subset \mathbb{R}^p$ will be an open subset, Θ_0 the portion of an affine k-plane lying in Θ with $k < p$ and Θ_a its complement in Θ. In this setting it seems reasonable to replace Θ_a by Θ in the denominator of (12.3), so that under suitable regularity assumptions the likelihood ratio becomes

$$\mathbf{x}\in\mathbb{R}^n \mapsto \Lambda(\mathbf{x}) := \frac{L(\mathbf{x};\widehat{\theta}_0)}{L(\mathbf{x};\widehat{\theta})} \in [0,1], \tag{12.4}$$

where

$$\widehat{\theta} = \sup_{\theta\in\Theta} L(\mathbf{x};\theta)$$

is the MLE for θ (as in Definition 8.3). and

$$\widehat{\theta}_0 = \sup_{\theta\in\Theta_0} L(\mathbf{x};\theta)$$

is the *null* MLE for θ. This leads to a remarkable class of statistical tests.

Definition 12.4. Under the conditions above, the *likelihood ratio test* $T = (h, R) \in \mathscr{T}_\alpha$ is performed by choosing

$$h(\mathbf{x}) = -2\ln\Lambda(\mathbf{x})$$

and $R = [r, +\infty)$, where $r > 0$ is determined by

$$\sup_{\theta\in\Theta_0} \mathcal{P}_\theta(h(X) \geq r) = \alpha. \tag{12.5}$$

Remark 12.5. Clearly, (12.5) is a special case of (12.2), where the rejection region R now takes the form $[r, +\infty)$ for some critical value $r > 0$, and we are supposed to solve it for r given the confidence level α. But notice that, as already observed in Remark 12.3, this requires knowing the distribution of h under H_0. In this regard, if $h(X)$ is found to be Θ_0-*ancillary* in the sense that its distribution does *not* depend on $\theta \in \Theta_0$ then

$$\alpha = \mathcal{P}_\theta(h(X) \geq r) \text{ for any } \theta \in \Theta_0 \tag{12.6}$$

determines r as a function of α. □

Remark 12.6. It follows from (8.6) that the likelihood ratio statistics satisfies

$$h(X) \approx 2n\left(D^{KL}_{\widetilde{\theta}_X}(\widehat{\theta}_0) - D^{KL}_{\widetilde{\theta}_X}(\widehat{\theta})\right),$$

so that likelihood ratio tests admit a natural information-theoretic interpretation: they measures how much KL divergence to the empirical distribution $\widetilde{\theta}_X$ is reduced when passing from the null model to the full model. Again, taking into account that $\widetilde{\theta}_X$ is consistent (for the true parameter), we see that the likelihood ratio statistics asymptotes the gap, as measured by the Kullback-Leibler divergence, between the best approximations of the truth under the null and alternative. □

Although we will restrict ourselves to normal populations, the examples below will suffice to illustrate the remarkable flexibility of this construction. Moreover, in all these examples the likelihood ratio statistics is Θ_0-ancillary in the sense of Remark 12.5, so that (12.6) applies in order to solve for r in terms of α.

Example 12.7. (z-test for the mean of a normal population with known variance) As in Example 8.5 we assume that $X_j \sim \mathcal{N}(\mu, \sigma^2)$, where $\theta_2 = \sigma^2$ is known. Thus, $\Theta = \{\theta_1 = \mu\} = \mathbb{R}$, $\Theta_0 = \{\mu_0\}$ for some $\mu_0 \in \mathbb{R}$, $\Theta_a = \mathbb{R}\backslash\{\mu_0\}$ and we want to test

$$H_0 : \mu = \mu_0 \quad \text{vs} \quad H_{\rm a} : \mu \neq \mu_0.$$

We recall that $\widehat{\theta}_1 = \overline{X}_n$ is the MLE for μ. Using (8.10) and (7.23) we compute

$$\begin{aligned} h(\mathbf{x}) &= -2\ln\frac{L(\mathbf{x};\mu_0)}{L(\mathbf{x};\widehat{\theta}_1)} \\ &= -2\ln\frac{(2\pi\sigma^2)^{-n/2}e^{-\frac{1}{2\sigma^2}\sum_j(x_j-\mu_0)^2}}{(2\pi\sigma^2)^{-n/2}e^{-\frac{1}{2\sigma^2}\sum_j(x_j-\widehat{\theta}_1)^2}} \\ &= -2\ln e^{-\frac{n}{2\sigma^2}(\widehat{\theta}_1-\mu_0)^2} \\ &= \left(\frac{\overline{x}_n-\mu_0}{\sigma/\sqrt{n}}\right)^2. \end{aligned}$$

Thus, under H_0 we see that

$$h(X) = Z(X)^2 \sim \chi^2_1,$$

where

$$Z(X) = \frac{\overline{X}_n - \mu_0}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1).$$

Now, by (12.5) the rejection interval $R = [r, +\infty)$ is determined by

$$\mathcal{P}_{\theta_0}(Z(X)^2 \geq r) = \alpha,$$

so we may take $r = \chi^2_{1,1-\alpha}$, the χ^2-quantile as in (7.62). Since

$$\alpha = \mathcal{P}_{\theta_0}(Z(X)^2 \geq r) = \mathcal{P}_{\theta_0}(-\sqrt{r} \leq Z(X) \leq \sqrt{r}),$$

we may also use the standard normal quantiles to make sure that if

$$Z(\mathbf{x}) \in (-\infty, -z_{1-\alpha/2}] \cup [z_{1-\alpha/2}, +\infty)$$

then H_0 gets rejected. □

Example 12.8. (t-test for the mean of a normal population with unknown variance) As in Example 8.5 we assume that $X_j \sim \mathcal{N}(\mu, \sigma^2)$, where $(\theta_1, \theta_2) = (\mu, \sigma^2)$ is unknown, so that $\Theta = \mathbb{R} \times \mathbb{R}_+$. Also, we fix $\mu_0 \in \mathbb{R}$ and set $\Theta_0 = \{\mu_0\} \times \mathbb{R}_+$, a half-line contained in Θ. As before, we want to test

$$H_0 : \mu = \mu_0 \quad \text{vs} \quad H_{\rm a} : \mu \neq \mu_0.$$

We recall that $\widehat{\theta} = (\widehat{\theta}_1, \widehat{\theta}_2)$, where $\widehat{\theta}_1 = \overline{X}_n$ and

$$\widehat{\theta}_2 = \frac{1}{n}\sum_j (X_j - \overline{\theta}_1)^2$$

is the MLE for θ_2, so that

$$\sup_{\theta\in\Theta} L(\mathbf{x};\theta) = L(\mathbf{x};\widehat{\theta}_1,\widehat{\theta}_2).$$

On the other hand, one has

$$\sup_{\theta\in\Theta_0} L(\mathbf{x};\theta) = L(\mathbf{x};\mu_0,\widehat{\theta}_{20}),$$

where the null MLE for θ_2 is

$$\widehat{\theta}_{20} = \frac{1}{n}\sum_j (x_j - \mu_0)^2.$$

Thus, the likelihood ratio is

$$\begin{aligned}\Lambda(\mathbf{x}) &= \left(\frac{\widehat{\theta}_{20}}{\widehat{\theta}_2)}\right)^{-n/2} \frac{e^{-\frac{1}{2\widehat{\theta}_{20}}\sum_j (x_j-\mu_0)^2}}{e^{-\frac{1}{2\widehat{\theta}_2}\sum_j (x_j-\widehat{\theta}_1)^2}} \\ &= \left(\frac{\widehat{\theta}_{20}}{\widehat{\theta}_2}\right)^{-n/2} \frac{e^{-n/2}}{e^{-n/2}} \\ &= \left(\frac{\widehat{\theta}_{20}}{\widehat{\theta}_2}\right)^{-n/2},\end{aligned}$$

so that, again using (7.23),

$$h(\mathbf{x}) = n \ln \left[1 + \frac{1}{n-1}\left(\frac{\overline{x}_n - \mu_0}{s_n(x)/\sqrt{n}}\right)^2\right].$$

Thus, under H_0 we see that

$$h(X) = n \ln \left[1 + \frac{1}{n-1} T_{n-1}(X)^2\right], \tag{12.7}$$

where

$$T_{n-1}(X) = \frac{\overline{X}_n - \mu_0}{S_n(X)/\sqrt{n}} \sim \mathsf{t}_{n-1},$$

or also

$$T_{n-1}(X)^2 \sim \mathsf{F}_{1,n-1}$$

by Corollary 4.34. Quite informally, we may expand (12.7) as $n \to +\infty$ to obtain

$$\begin{aligned}h(X) &= \ln \left[1 + \frac{1}{n-1} T_{n-1}(X)^2\right]^n \\ &= \ln \left[1 + \frac{n}{n-1} T_{n-1}(X)^2 + \cdots\right] \\ &= \frac{n}{n-1} T_{n-1}(X)^2 + \cdots \\ &\xrightarrow{p} \chi^2_1,\end{aligned}$$

where the dots represent lower order terms (which vanish as $n \to +\infty$) and we used Remark 6.4 in the last step. Thus, for large samples we may take $R = [\chi^2_{1,1-\alpha}, +\infty)$ as the "approximate" rejection interval. Otherwise, we use that

$$\alpha = \mathcal{P}_{\theta_0}(T_{n-1}(X)^2 \geq c_{n,r}) = \mathcal{P}_{\theta_0}(-\sqrt{c_{n,r}} \leq T_{n-1}(X)) \leq \sqrt{c_{n,r}}),$$

where

$$c_{n,r} = (n-1)(e^{r/n} - 1),$$

to reject H_0 if either

$$T_{n-1}(\mathbf{x})^2 \in [\mathsf{f}_{1,n-1,1-\alpha}, +\infty)$$

or equivalently

$$T_{n-1}(\mathbf{x}) \in (-\infty, \mathsf{t}_{n-1,\alpha/2}] \cup [\mathsf{t}_{n-1,1-\alpha/2}, +\infty),$$

which is a more familiar presentation of the test. □

Example 12.9. (F-test for the equality of variances of independent normal populations) We will use the notation of Example 7.36 and Remark 7.37 with the aim of testing

$$H_0 : \sigma_X^2 = \sigma_Y^2 \quad \text{vs} \quad H_{\text{a}} : \sigma_X^2 \neq \sigma_Y^2.$$

Since we assume independence of the samples, Example 8.5 implies that the corresponding likelihood function is

$$L(\mathbf{x}, \mathbf{y}; \theta) = \frac{1}{(2\pi)^{(m+n)/2}\theta_{2X}^{m/2}\theta_{2X}^{n/2}} e^{-\frac{1}{2}\left(\sum_{j=1}^{m}\frac{(x_j-\theta_{1X})^2}{\theta_{2X}}+\sum_{k=1}^{n}\frac{(y_k-\theta_{1Y})^2}{\theta_{2Y}}\right)},$$

where

$$\theta = (\theta_{1X}, \theta_{1Y}, \theta_{2X}, \theta_{2Y}) = (\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2) \in \mathbb{R}^2 \times \mathbb{R}_+^2,$$

so the MLE estimators are

$$\widehat{\theta}_{1X} = \overline{X}_m, \quad \widehat{\theta}_{1Y} = \overline{Y}_m$$

and

$$\widehat{\theta}_{2X} = \frac{m-1}{m}S_X^2 = \frac{1}{m}\sum_j (X_j - \widehat{\theta}_{1X})^2, \quad \widehat{\theta}_{2Y} = \frac{n-1}{n}S_Y^2 = \frac{1}{n}\sum_k (Y_k - \widehat{\theta}_{1Y})^2.$$

Hence,

$$\begin{aligned}
\sup_{\theta\in\Theta} L(\mathbf{x}, \mathbf{y}; \theta) &= L(\mathbf{x}, \mathbf{y}; \widehat{\theta}_{1X}, \widehat{\theta}_{1Y}, \widehat{\theta}_{2X}, \widehat{\theta}_{2Y}) \\
&= \frac{1}{(2\pi)^{(m+n)/2}\widehat{\theta}_{2X}^{m/2}\widehat{\theta}_{2Y}^{n/2}} e^{-\frac{1}{2}\left(\sum_{j=1}^{m}\frac{(x_j-\widehat{\theta}_{1X})^2}{\widehat{\theta}_{2X}}+\sum_{k=1}^{n}\frac{(y_k-\widehat{\theta}_{1Y})^2}{\widehat{\theta}_{2Y}}\right)} \\
&= \frac{e^{-(m+n)/2}}{(2\pi)^{(m+n)/2}\widehat{\theta}_{2X}^{m/2}\widehat{\theta}_{2X}^{n/2}}.
\end{aligned}$$

On the other hand, restriction to Θ_0 gives

$$L(\mathbf{x}, \mathbf{y}; \theta) = \frac{1}{(2\pi)^{(m+n)/2}\theta_2^{(m+n)/2}} e^{-\frac{1}{2\theta_2}\left(\sum_j (x_j-\theta_{1X})^2+\sum_k (y_k-\theta_{1Y})^2\right)},$$

where $\theta_2 = \sigma_X^2 = \sigma_Y^2$ is the common variance, so that maximization over all possible values of $(\theta_{1X}, \theta_{1Y}, \theta_2)$ is achieved at the null MLE $(\widehat{\theta}_{1X}, \widehat{\theta}_{1Y}, \widehat{\theta}_{20})$, where

$$\widehat{\theta}_{20} = \frac{1}{m+n}\left(\sum_j (x_j - \widehat{\theta}_{1X})^2 + \sum_k (y_k - \widehat{\theta}_{1Y})^2\right).$$

It follows that

$$\begin{aligned}
\sup_{\theta\in\Theta_0} L(\mathbf{x}, \mathbf{y}; \theta) &= L(\mathbf{x}, \mathbf{y}; \widehat{\theta}_{1X}, \widehat{\theta}_{1Y}, \widehat{\theta}_{20}) \\
&= \frac{1}{(2\pi)^{(m+n)/2}\widehat{\theta}_{20}^{(m+n)/2}} e^{-\frac{1}{2\widehat{\theta}_{20}}\left(\sum_j ((x_j-\widehat{\theta}_{1X})^2+\sum_k +(y_k-\widehat{\theta}_{1Y})^2\right)} \\
&= \frac{e^{-(m+n)/2}}{(2\pi)^{(m+n)/2}\widehat{\theta}_{20}^{(m+n)/2}},
\end{aligned}$$

so the likelihood radio is

$$\Lambda(\mathbf{x}, \mathbf{y}) = \left(\frac{\widehat{\theta}_{20}}{\widehat{\theta}_{2X}}\right)^{-m/2}\left(\frac{\widehat{\theta}_{20}}{\widehat{\theta}_{2Y}}\right)^{-n/2},$$

and from this we easily deduce that

$$h(\mathbf{x},\mathbf{y}) = \ln\left[c\left(1+a\left(\frac{S_X^2}{S_Y^2}\right)^{-1}\right)^m\left(1+b\frac{S_X^2}{S_Y^2}\right)^n\right],$$

where a, b and c are certain constants depending only on m and n with $ab=1$. Thus, as in Remark 7.37 we see that under H_0,

$$h(X,Y) = \ln\left(c\left(1+aU(X,Y)^{-1}\right)^m\left(1+bU(X,Y)\right)^n\right),$$

where

$$U(X,Y) = \frac{S_X^2}{S_Y^2} \sim \mathsf{F}_{m-1,m-1}.$$

Now note that $h(\mathbf{x},\mathbf{y}) \geq e^r$ if and only if $u := U(\mathbf{x},\mathbf{y})$ satisfies $f(u) \geq e^r/c$, where

$$f(u) := (1+au^{-1})^m(1+bu)^n, \quad u>0.$$

Also, a little Calculus shows that f is strictly convex with its unique minimal value achieved at $u_0 = ma/n = m/nb$ (this analysis uses that $ab=1$ in a crucial way). Hence, there exist a *maximal* $0<\underline{u}<u_0$ and a *minimal* $\overline{u}>u_0$ with the property that for any r such that

$$f\left(\frac{ma}{n}\right) = f\left(\frac{m}{nb}\right) = \left(1+\frac{n}{m}\right)^m\left(1+\frac{m}{n}\right)^n < \frac{e^r}{c}$$

there holds $f(u) \geq e^r/a$ whenever either $u \leq \underline{u}$ or $u \geq \overline{u}$. This means that we may reject H_0 if u falls outside $(\underline{u},\overline{u})$. More concretely, given a confidence level α small enough, we may reject H_0 if

$$u \in \left(0, \mathsf{f}_{m-1,n-1,\alpha/2}\right] \bigcup \left[\mathsf{f}_{m-1,n-1,1-\alpha/2}, +\infty\right),$$

which is consistent with (7.49). □

Example 12.10. (F-test for the equality of means of $p \geq 2$ independent normal populations with a common but unknown variance) We will use the notation from Example 7.40, so we have *independent* random samples $X_{jk} \sim \mathcal{N}(\mu_j,\sigma^2)$ for $j=1,\cdots,p$. Thus, $\Theta = \mathbb{R}^p \times \mathbb{R}_+$ with $\theta = (\theta_{11},\cdots,\theta_{1p},\theta_2)$, where $\theta_{1j} = \mu_j$ and $\theta_2 = \sigma^2$, the common variance. Also, $\Theta_0 = \{\theta \in \Theta; \theta_{11} = \cdots = \theta_{1p}\}$ and we want to test

$$H_0 : \mu_1 = \cdots = \mu_p \quad \text{vs} \quad H_{\mathrm{a}} : \mu_j \neq \mu_{j'} \text{ for some } j \neq j'.$$

Notice that this is precisely the one way ANOVA test in Example 7.40. If $\mathbf{x} = (\mathbf{x}_1,\cdots,\mathbf{x}_p)$ with $\mathbf{x}_j = (\mathbf{x}_{j1},\cdots,\mathbf{x}_{jn_j}) \in \mathbb{R}^{n_j}$ then the likelihood function is

$$\begin{aligned} L(\mathbf{x};\theta) &= \Pi_{j=1}^p \Pi_{k=1}^{n_j} (2\pi\theta_2)^{-n_j/2} e^{-\frac{1}{2\theta_2}\sum_{k=1}^{n_j}(x_{jk}-\theta_{1j})^2} \\ &= (2\pi\theta_2)^{-n/2} e^{-\frac{1}{2\theta_2}\sum_{j=1}^p\sum_{k=1}^{n_j}(\mathbf{x}_{jk}-\theta_{1k})^2}. \end{aligned}$$

By passing to the log likelihood function and maximizing over Θ in the usual way we see that the MLE for θ is $\widehat{\theta} = (\widehat{\theta}_{11},\cdots,\widehat{\theta}_{1p},\widehat{\theta}_2)$, where

$$\widehat{\theta}_{1j}(\mathbf{x}) = \overline{\mathbf{x}}_{j\bullet} = \frac{1}{n_j}\sum_{k=1}^{n_j}\mathbf{x}_{jk} \text{ and } \widehat{\theta}_2(\mathbf{x}) = \frac{1}{n}\sum_{j=1}^p\sum_{k=1}^{n_j}(\mathbf{x}_{jk}-\overline{\mathbf{x}}_{j\bullet})^2$$

are realizations of $\overline{X}_{j\bullet}$ and $S^2_{\text{Within}}(X)/n$, respectively. On the other hand, restricted to Θ_0 we have

$$L(\mathbf{x};\theta) = (2\pi\theta_2)^{-n/2} e^{-\frac{1}{2\theta_2}\sum_{j=1}^p\sum_{k=1}^{n_j}(\mathbf{x}_{jk}-\theta_0)^2},$$

where $\theta_0 = \mu_1 = \cdots = \mu_p$ is the common mean, so that maximization over Θ_0 gives the corresponding null MLEs

$$\widehat{\theta}_{00}(\mathbf{x}) = \mathbf{x}_{\bullet\bullet} = \frac{1}{n}\sum_{j=1}^{p} n_j \overline{\mathbf{x}}_{j\bullet} = \frac{1}{n}\sum_{j=1}^{p}\sum_{k=1}^{n_j} \mathbf{x}_{jk} \text{ and } \widehat{\theta}_{20}(\mathbf{x}) = \frac{1}{n}\sum_{j=1}^{p}\sum_{k=1}^{n_j} (\mathbf{x}_{jk} - \overline{\mathbf{x}}_{\bullet\bullet})^2,$$

which are realizations of $\overline{X}_{\bullet\bullet}$ and $S^2_{\text{Total}}(X)/n$, respectively. It follows that

$$\begin{aligned}\Lambda(\mathbf{x}) &= \frac{(2\pi\widehat{\theta}_{20}(\mathbf{x}))^{-n/2} e^{-\frac{1}{2\widehat{\theta}_{20}(\mathbf{x})}\sum_{j=1}^{p}\sum_{k=1}^{n_j}(\mathbf{x}_{jk}-\mathbf{x}_{\bullet\bullet})^2}}{(2\pi\widehat{\theta}_{2}(\mathbf{x}))^{-n/2} e^{-\frac{1}{2\widehat{\theta}_{2}(\mathbf{x})}\sum_{j=1}^{p}\sum_{k=1}^{n_j}(\mathbf{x}_{jk}-\mathbf{x}_{j\bullet})^2}} \\ &= \left(\frac{\widehat{\theta}_{20}(\mathbf{x})}{\widehat{\theta}_{2}(\mathbf{x})}\right)^{-n/2} \frac{e^{-n/2}}{e^{-n/2}},\end{aligned}$$

so that using (7.53),

$$h(\mathbf{x}) = n\ln\left(1 + \frac{p-1}{n-p}\,\frac{s^2_{\text{Between}}(\mathbf{x})/(p-1)}{s^2_{\text{Within}}(\mathbf{x})/(n-p)}\right).$$

Hence, as in Example 7.40 we see that under H_0,

$$h(X) = n\ln\left(1 + \frac{p-1}{n-p}V\right),$$

where

$$V = \frac{S^2_{\text{Between}}/(p-1)}{S^2_{\text{Within}}/(n-p)} \sim \mathsf{F}_{p-1,n-p}.$$

Again, we may expand this as $n \to +\infty$ to find that

$$\begin{aligned} h(X) &= \ln\left(1 + \frac{p-1}{n-p}V\right)^n \\ &= \ln\left(1 + \frac{n(p-1)}{n-p}V + \cdots\right) \\ &= (p-1)V + \cdots \\ &\xrightarrow{d} \chi^2_{p-1}, \end{aligned}$$

where we used Remark 6.4 in the last step. Thus, for large samples we may take $R = [\chi^2_{p-1,1-\alpha}, +\infty)$ as the rejection interval. Otherwise, we use that

$$\alpha = \sup_{\theta_0 \in \Theta_0} \mathcal{P}_{\theta_0}\left(V(\mathbf{x}) \geq \frac{n-p}{p-1}\left(e^{r/n} - 1\right)\right)$$

to reject H_0 if

$$V(\mathbf{x}) \in [\mathsf{f}_{p-1,n-p,1-\alpha}, +\infty)$$

as in (7.61). □

Example 12.11. (F-test for statistical significance of the linear regression model) We consider here the linear regression model in (9.15), whose likelihood function $L(\mathbf{y}; \beta, \sigma^2)$ is given by (9.16), in order to test the full "intercept-only" hypothesis appearing in Example 9.8:

$$H_0 : \beta_1 = \cdots = \beta_p = 0 \quad \text{vs} \quad H_{\text{a}} : \beta_j \neq 0 \text{ for some } j. \tag{12.8}$$

In other words, the null hypothesis here says that $\mathbf{X}$ has no influence whatsoever on $\mathbf{Y}$ so its rejection provides statistical evidence for employing the model as it is posed in Example 9.3 (that is, with the full "slope"

$(\beta_1, \cdots, \beta_n)$ included). We have $\theta = (\beta, \theta_2)$, where $\theta_2 = \sigma^2$, so the usual calculation implies that the corresponding MLE is $(\widehat{\beta}, \widehat{\theta}_2)$, where

$$\widehat{\theta}_2 = \frac{1}{n}|\mathbf{Y} - \mathfrak{x}\widehat{\beta}|^2 = \frac{1}{n}|\mathfrak{x}\beta + \mathbf{e} - \mathfrak{x}\widehat{\beta}|^2,$$

so that (9.56) gives

$$\widehat{\theta}_2 = \frac{|\widehat{\mathbf{e}}|^2}{n} = \frac{SS_{\text{Res}}}{n},$$

and we verify that

$$\sup_{\theta \in \Theta} L(\mathbf{y}; \beta, \theta_2) = L(\mathbf{y}; \widehat{\beta}, \widehat{\theta}_2) = (2\pi SS_{\text{Res}}/n)^{-n/2} e^{-n/2}.$$

On the other hand, under the null hypothesis,

$$L(\mathbf{y}; \beta_0, \theta_2) = (2\pi\theta_2)^{-n/2} e^{-\frac{|\mathbf{y} - \beta_0 \mathbf{1}|^2}{2\theta_2}},$$

so that the null MLE estimator for θ_2 is

$$\widehat{\theta}_{20} = \frac{1}{n}\|\mathbf{Y} - \widehat{\beta}_0 \mathbf{1}\|^2 \overset{(9.21)}{=} \frac{1}{n}\|\mathbf{Y} - \overline{\mathbf{Y}}\mathbf{1}\|^2 = \frac{SS_{\mathbf{YY}}}{n}, \tag{12.9}$$

which gives

$$\sup_{\theta \in \Theta_0} L(\mathbf{y}; \beta, \theta_2) = L(\mathbf{y}; \widehat{\beta}_0, \widehat{\theta}_{20}) = (2\pi SS_{\mathbf{YY}}/n)^{-n/2} e^{-n/2}.$$

It then follows that the likelihood ratio statistics is

$$h(\mathbf{y}) = \ln\left(\frac{\widehat{\theta}_{20}}{\widehat{\theta}_2}\right)^n = \ln\left(\frac{SS_{\mathbf{YY}}}{SS_{\text{Res}}}\right)^n = \ln\left(1 + \frac{SS_{\text{Reg}}}{SS_{\text{Res}}}\right)^n, \tag{12.10}$$

where we used that, as in (9.47),

$$SS_{\mathbf{YY}} = SS_{\text{reg}} + SS_{\text{Res}}. \tag{12.11}$$

We now proceed to the appropriate counting of degrees of freedom as we did in Example 12.10. We see from Propositions 9.10 and 9.12 that $\sigma^{-2} SS_{\text{Res}} \sim \chi^2_{n-p-1}$ is independent of SS_{Reg} and hence of $\widehat{\mathbf{Y}} = \mathfrak{x}\widehat{\beta}$, so SS_{Res} is independent of SS_{Reg} (recall that we are conditioning on $\mathfrak{X} = \mathfrak{x}$). On the other hand, under H_0 we have $\mathbf{Y} \sim \mathcal{N}(\beta_0 \mathbf{1}, \sigma^2 I_{n\times n})$ and $\widehat{\beta}_0 = \overline{\mathbf{Y}}$, which gives $\sigma^{-2} SS_{\mathbf{YY}} \sim \chi^2_{n-1}$ by Proposition 7.24. Thus, again under H_0, we get from (12.11) that $\sigma^{-2} SS_{\text{Reg}} \sim \chi^2_p$ and we conclude that

$$h(\mathbf{Y}) = \ln\left(1 + \frac{p}{n-p-1} W(\mathbf{Y})\right)^n,$$

where

$$W(\mathbf{Y}) = \frac{SS_{\text{Reg}}/p}{SS_{\text{Res}}/(n-p-1)} \sim \mathsf{F}_{p,n-p-1}.$$

As in Example 12.10, $W(\mathbf{Y}) \overset{d}{\to} \chi^2_p$ as $n \to +\infty$, so for large samples we may take $R = [\chi^2_{p,1-\alpha}, +\infty)$ as the rejection interval. Otherwise, we must reject H_0 if

$$W(\mathbf{y}) \in [\mathsf{f}_{p,n-p-1,1-\alpha}, +\infty)\,. \tag{12.12}$$

We mention that the theoretical procedure leading to the F-test above, based on a likelihood ratio test, is flexible enough to handle a general linear hypothesis test on the parameters, in which the null hypothesis may be expressed as $B\beta = c$, where B is a suitable $q \times (p+1)$ matrix and c is a q-vector; see [Ame85, Subsection 1.5] and [SL03, Chapter 4]. For instance, if $c = \vec{0}$ and B is suitably chosen then we can form the test

$$H_0^{\text{n}} : \beta_{q+1} = \cdots = \beta_p = 0 \quad \text{vs} \quad H_{\text{a}}^{\text{n}} : \beta_j \neq 0 \text{ for some } j \in \{q+1, \cdots, p\}, \tag{12.13}$$

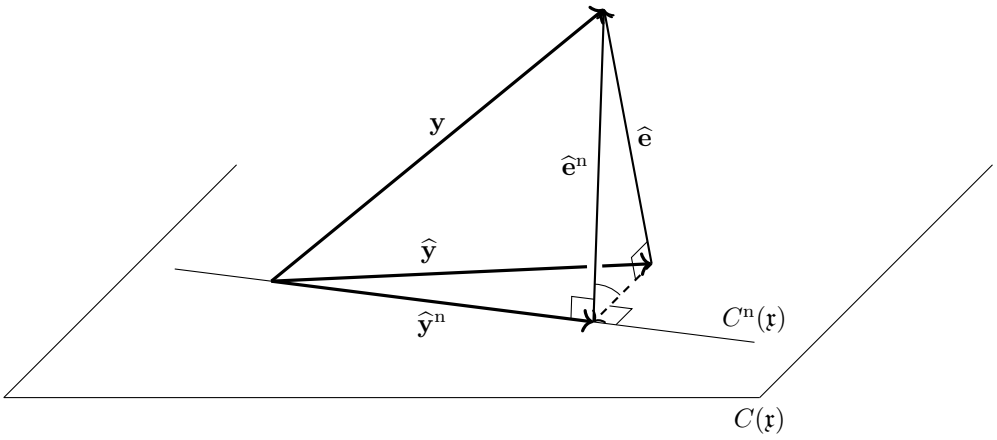


FIGURE 2. The geometry of nested models

which compares the full model and a new null model in which only the first q independent variables possibly appear as significant predictors; as indicated in Figure 2, the corresponding design spaces satisfy $C^{\rm n}(\mathfrak{X}) \subset C(\mathfrak{X})$ with $\dim C(\mathfrak{X})\backslash C^{\rm n}(\mathfrak{X}) = p - q$, this being the reason why the models are *nested*; see Remark 9.11. If we view $SS_{\mathbf{YY}}$ in (12.10) and (12.11) as the residual sum of squares (i.e. the norm squared residual) of the null model in (12.8) and proceed by analogy, it is not hard to check that the likelihood ratio statistics now is

$$h^{\rm n}(\mathbf{Y}) = \ln\left(\frac{SS^{\rm n}_{\rm Res}}{SS_{\rm Res}}\right)^n = \ln\left(1 + \frac{p-q}{n-p-1}W^{\rm n}(\mathbf{Y})\right)^n,$$

where

$$W^{\rm n}(\mathbf{Y}) = \frac{(SS^{\rm n}_{\rm Res} - SS_{\rm Res})/(p-q)}{SS_{\rm Res}/(n-p-1)}$$

and $SS^{\rm n}_{\rm Res}$ is the residual sum of squares of the null model in (12.13). Since the usual counting of degrees of freedom shows that $W^{\rm n}(\mathbf{Y}) \sim \mathsf{F}_{p-q,n-p-1}$ under $H^{\rm n}_0$, we find that the null hypothesis in (12.13) gets rejected if

$$W^{\rm n}(\mathbf{y}) \in [\mathsf{f}_{p-q,n-p-1,1-\alpha}, +\infty)\,,$$

the obvious extension of (12.12). Put in another way, if $SS^{\rm n}_{\rm Res}$ and $SS_{\rm Res}$ are close to each other, which intuitively means that the null model fits as well as the full model, then $W^{\rm n}(\mathbf{y})$ is small and hence $H^{\rm n}_0$ should *not* be rejected. In any case, the geometry backing not only this latter assertion but also the whole argument above is fully discernible from Figure 2, where $SS^{\rm n}_{\rm Res} = \|\widehat{\mathbf{e}}^{\rm n}\|^2$, $SS^{\rm n}_{\rm Res} - SS_{\rm Res} = \|\widehat{\mathbf{e}}^{\rm n} - \widehat{\mathbf{e}}\|^2$, the squared norm of the dashed vector, and so on. □

Remark 12.12. (p-value) As already observed, in all examples above the likelihood ratio statistics $h(X)$ is Θ_0-ancillary in the sense that its distribution does not depend on $\theta \in \Theta_0$. In those cases, an equivalent way of reporting the result of a likelihood ratio test is to look at the corresponding *p-value*

$$\mathfrak{p} = \mathcal{P}_\theta(h(X) \geq h(x)), \quad \theta \in \Theta_0,$$

where $h(x)$ is the observed value of $h(X)$. Thus, $\mathfrak{p}$ is the probability of finding, under H_0, an observed value at least as extreme as the one actually observed. With this terminology, H_0 gets rejected if $\mathfrak{p} \leq \alpha$, which is just a rephrasing of the rejection condition $h(x) \geq r$. Although this seems to be the preferred way of summarizing the outcome of a test in Applied Statistics, it is argued that the common misinterpretation of regarding $\mathfrak{p}$ as the probability that H_0 is true, thus erroneously accepting the validity of the alternative hypothesis (with high

probability) if $\mathfrak{p}$ is found to be sufficiently small, may be a source of confusion leading to "P-hacking", "the replication crisis", etc.; see [HB03, WL16, FP15, Gib21] for more on this quite controversial issue. □

In all examples above where we have been able to directly carry out the corresponding computation, the *asymptotic* likelihood radio statistics turned out to be χ^2_l-distributed, where $l = \dim\Theta - \dim\Theta_0$. In fact, this is a general phenomenon which substantially simplifies the implementation of the test for large samples.

Theorem 12.13. [Wil38] *Under the conditions above, and requiring suitable regularity assumptions on the underlying statistical model as usual, there holds $h(X) \overset{d}{\to} \chi^2_l$ as $n \to +\infty$ and under H_0.*

Proof. We only sketch the argument, which relies on the (multi-dimensional version) of the proof of Theorem 8.24 on the asymptotic normality of ML estimators (and the simplifying assumption that we may choose rectangular coordinates $(\theta_1, \cdots, \theta_p)$ on Θ so that Θ_0 is singled out by $\theta_{k+1} = \cdots = \theta_p = 0$). Now, under H_0 the true parameter value, say θ, lies in Θ_0. Moreover, if n is large enough then both

$$\widehat{\theta} = \mathrm{argmax}_{\theta\in\Theta} L(\mathbf{x};\theta) \quad \text{and} \quad \widehat{\theta}_0 = \mathrm{argmax}_{\theta\in\Theta_0} L(\mathbf{x};\theta)$$

are close to θ and hence close to each other, so we may expand about $\widehat{\theta}$,

$$\begin{aligned} h(X) &= -2\left(\ln L(X;\widehat{\theta}) - \ln L(X;\widehat{\theta}_0)\right) \\ &\approx \langle \widehat{\theta} - \widehat{\theta}_0, (-\nabla^2_{\theta\theta})(\ln L(X;\widehat{\theta})(\widehat{\theta} - \widehat{\theta}_0)\rangle, \end{aligned}$$

where we have discarded terms of order at least three and used that $\nabla_\theta \ln L(\mathbf{x};\widehat{\theta}) = 0$ by the definition of $\widehat{\theta}$; compare with (8.35). From (8.41) we also know that

$$(-\nabla^2_{\theta\theta})(\ln L(X;\widehat{\theta})) \overset{p}{\to} \mathscr{F}(\theta),$$

where $\mathscr{F}(\theta)$ is the Fisher information matrix of a single observation. Moreover,

$$W := \sqrt{n}\,\nabla_\theta \ln L(X;\theta) \overset{d}{\to} \mathcal{N}(\vec{0}, \mathscr{F}(\theta))$$

by (8.42), so that (8.36) may be rewritten as

$$\sqrt{n}(\widehat{\theta} - \theta) \overset{d}{\to} \mathscr{F}(\theta)^{-1} W.$$

Also, by examining the maximization problem restricted to Θ_0 which yields $\widehat{\theta}_0$, it is not hard to check that

$$\sqrt{n}(\widehat{\theta}_0 - \theta) \overset{d}{\to} \mathscr{G}(\theta) W,$$

where $\mathscr{G}(\theta)$ is a certain symmetric matrix with rank $k = \dim\Theta_0$ and satisfying

$$\mathscr{G}(\theta)\mathscr{F}(\theta)\mathscr{G}(\theta) = \mathscr{G}(\theta).$$

Thus, eliminating θ in the convergences above and using the expansion we get

$$h(X) \approx \langle W, (\mathscr{F}(\theta)^{-1} - \mathscr{G}(\theta))W\rangle$$

Now, if $\mathscr{F}(\theta) = B^2$ then using Corollary 4.7 we have that $Z = B^{-1}W \approx \mathcal{N}(\vec{0}, \mathrm{Id}_p)$, a standard normal vector, so that

$$\begin{aligned} h(X) &\approx \langle BZ, (\mathscr{F}(\theta)^{-1} - \mathscr{G}(\theta))BZ\rangle \\ &= \langle Z, B(\mathscr{F}(\theta)^{-1} - \mathscr{G}(\theta))BZ\rangle \\ &= \langle Z, (\mathrm{Id}_p - A\mathscr{G}(\theta)B)Z\rangle, \end{aligned}$$

where $B\mathscr{G}(\theta)B$ is easily seen to be idempotent with the same rank as $\mathscr{G}(\theta)$. Hence, $\mathrm{Id}_p - B\mathscr{G}(\theta)B$ is idempotent as well with rank $l = p - k = \dim\Theta - \dim\Theta_0$ and the result follows from Proposition 4.27. □

We refer to [Ame85, Subsection 4.5.1] and [Sha08, Theorem 6.5] for those interested in filling out the omitted details in the argument above. Also, it is worthwhile mentioning that Wilks' original proof in [Wil38] is equally elegant as it involves checking that $\phi_{h(X)}$, the characteristic function of $h(X)$, asymptotically approaches $\phi_{\chi^2_l}$. In any case, we stress that it is not required in Theorem 12.13 that $h(X)$ is Θ_0-ancillary so in a sense the result guarantees that this property gets restored in the asymptotic regime. Finally, we note that the material above by no means exhausts the rich literature on hypothesis testing and extensive treatments may be found elsewhere [Ame85, DM88, Wel96, LR05, Cas08, Sha08, Hay11, DS14].

We now discuss a non-conventional hypothesis testing and its connection with a major result in Differential Geometry, namely, Weyl's formula for the volume of tubes [Wey39, Gra03].

Example 12.14. (Testing for an additional term in the linear model and Weyl's formula for the volume of tubes) Let us consider a (possibly non-linear) perturbation of the linear model with a normal error from Example 9.3,

$$\mathbf{Y}_j = \mathfrak{X}_j\beta + cf_j(\mathbf{X}_j, \gamma) + \mathbf{e},$$

a setting first considered in a seminal paper by H. Hotelling [Hot39]. Here, f is known but $c \in \mathbb{R}$ and $\gamma \in \mathbb{R}^k$ are unknown parameters and our aim is to test

$$H_0 : c = 0 \quad \text{vs} \quad H_{\rm a} : c \neq 0,$$

so that not being able to reject H_0 indicates statistical evidence for ignoring f in the model design. In the geometric language of Remark 9.11, the null hypothesis says that $\mathbb{E}(\mathbf{Y}|_{\mathfrak{X}=\mathfrak{x}}) \in C(\mathfrak{x}) \equiv \mathbb{R}^{p+1}$, whereas the alternative adds a multiple of $f_\gamma = f(\cdot, \gamma)$ to this vector. Without loss of generality, we may assume that $f_\gamma \in C(\mathfrak{x})^\perp \equiv \mathbb{R}^{n-p-1}$ and, in order to make the model identifiable, that f_γ is not a multiple of $f_{\gamma'}$ if $\gamma \neq \gamma'$, but notice that the model still gets *non*-identifiable under H_0, so the usual dimensional counting that would allow us to determine the asymptotic distribution of the likelihood ratio statistics h via Theorem 12.13 does not apply. In particular, there is no point here in working with h so we turn our attention to the corresponding likelihood ratio $\Lambda = e^{-h/2}$, thus rejecting H_0 if this ratio, when observed, is conveniently small. Now, the full likelihood function is

$$L(\mathbf{y}; \beta, \theta_2, c, \gamma) = (2\pi\theta_2)^{-n/2} e^{-\frac{\|\mathbf{y}-\mathfrak{x}\beta - cf_\gamma\|^2}{2\theta_2}}, \quad \theta_2 = \sigma^2,$$

which under H_0 reduces to the usual likelihood function of the linear model treated in Example 12.11:

$$L(\mathbf{y}; \beta, \theta_2) = (2\pi\theta_2)^{-n/2} e^{-\frac{\|\mathbf{y}-\mathfrak{x}\beta\|^2}{2\theta_2}}.$$

Hence,

$$\sup_{\beta,\theta_2} L(\mathbf{y}; \beta, \theta_2) = (2\pi\widehat{\theta}_{20})^{-n/2} e^{-n/2},$$

where

$$\widehat{\theta}_{20} = \frac{1}{n}|\mathbf{Y} - \mathfrak{x}\widehat{\beta}|^2 = \frac{1}{n}|\widehat{\mathbf{e}}|^2$$

and $\widehat{\mathbf{e}}$ is the residual. It follows that the likelihood ratio is

$$\Lambda = \left(\frac{\widehat{\theta}_{20}}{\widehat{\theta}_2}\right)^{-n/2},$$

where

$$\widehat{\theta}_2 = \frac{1}{n}\|\mathbf{Y} - \mathfrak{x}\widehat{\beta} - \widehat{c}f_{\widehat{\gamma}}\|^2$$

and

$$(\widehat{\beta}, \widehat{c}, \widehat{\gamma}) = \text{argmax}_{\beta,c,\gamma} L(\mathbf{y}; \beta, c, \gamma) = \text{argmin}_{\beta,c,\gamma} \|\mathbf{y} - \mathfrak{x}\beta - cf_\gamma\|^2.$$

Now, using that $f_\gamma^\top \mathfrak{r}\beta = 0$ (in particular, $f_\gamma^t \mathfrak{r}\widehat{\beta} = 0$) we compute

$$\begin{aligned}(\widehat{\beta},\widehat{c},\widehat{\gamma}) &= \operatorname{argmin}_{\beta,c,\gamma}\|\mathbf{y}-\mathfrak{r}\beta\|^2 - 2cf_\gamma^\top\mathbf{y} + c^2\|f_\gamma\|^2\\ &= \operatorname{argmin}_{c,\gamma}\|\mathbf{y}-\mathfrak{r}\widehat{\beta}\|^2 - 2cf_\gamma^\top\mathbf{y} + c^2\|f_\gamma\|^2\\ &= \operatorname{argmin}_{c,\gamma}\|\mathbf{y}-\mathfrak{r}\widehat{\beta}\|^2 - 2cf_\gamma^\top(\mathbf{y}-\mathfrak{r}\widehat{\beta}) + c^2\|f_\gamma\|^2\\ &= \operatorname{argmin}_{c,\gamma}\|\widehat{\mathbf{e}}-cf_\gamma\|^2,\end{aligned}$$

and since

$$\widehat{c} = \operatorname{argmin}_c\|\widehat{\mathbf{e}}-cf_\gamma\|^2 = \frac{f_\gamma^\top\widehat{\mathbf{e}}}{\|f_\gamma\|^2},$$

we see that

$$\begin{aligned}\Lambda^{2/n} &= \inf_\gamma \frac{\|\widehat{\mathbf{e}}-\widehat{c}f_\gamma\|^2}{\|\widehat{\mathbf{e}}\|^2}\\ &= \inf_\gamma\left(1-\left(\frac{f_\gamma^\top\widehat{\mathbf{e}}}{\|f_\gamma\|\|\widehat{\mathbf{e}}\|}\right)^2\right)\\ &= 1-\sup_\gamma(\widetilde{f}_\gamma^\top\mathbf{U})^2.\end{aligned}$$

where, as γ varies, $\widetilde{f}_\gamma = f_\gamma/\|f_\gamma\|$ traces a subset $M_\gamma \subset \mathbb{S}^{n-p-2}$, the unit sphere in $C(\mathfrak{r})^\perp$, and $\mathbf{U} = \widehat{\mathbf{e}}/\|\widehat{\mathbf{e}}\|$ is a random vector also taking values in $\mathbb{S}^{n-p-2}$. Since

$$\widetilde{f}_\gamma^\top\mathbf{u} = \cos\operatorname{dist}(f_\gamma,\mathbf{u}),$$

where dist is the intrinsic distance in $\mathbb{S}^{n-p-2}$, we may choose as rejection region the "tubular neighborhood"

$$B_\rho(M_\gamma) = \{\vartheta\in\mathbb{S}^{n-p-2};\operatorname{dist}(\vartheta,M_\gamma)\le\rho\}$$

of radius $\rho>0$ around M_γ. Now, under H_0 we know from Remark 9.11 that $\mathbf{U}$ is uniformly distributed in $\mathbb{S}^{n-p-2}$ and we conclude that the significance level α of the test satisfies

$$\alpha = P(\mathbf{U}\in B_\rho(M_\gamma)) = \operatorname{vol}_{P_\mathbf{U}}(B_\rho(M_\gamma)),$$

where $\operatorname{vol}_{P_\mathbf{U}}$ is the (normalized) intrinsic volume (so that $\operatorname{vol}_{P_\mathbf{U}}(\mathbb{S}^{n-p-2})=1$). Thus, in order to determine the rejection "tube" associated to a given confidence level, an explicit formula for the volume of the tube is required, at least for ρ small enough. This turns out to be a rather formidable geometric problem which has been completely solved by H. Weyl [Wey39] in case M is a closed submanifold of a space form (a Riemannian manifold with constant sectional curvature)[48]. Unfortunately, this Weyl's formula does not directly apply to this problem (as M_γ may be only piecewise smooth or carry a boundary, etc.) so adjustments, mainly based on suitable approximations, are required [Nai90]. □

[48]In case $M\subset\mathbb{R}^l$ has dimension m, Weyl's formula says that

$$\operatorname{vol}(B_\rho(M)) = c_{m,l}\rho^{l-m}\sum_{q=0}^{[m/2]}\left(d_{m,l,q}\int_M\kappa_{2q}dM\right)\rho^{2q},$$

where $c_{m,l}$ and $d_{m,l,q}$ are positive constants (only depending on the indicated natural parameters) and κ_{2q} is a certain (universal) polynomial expression which is homogeneous of degree q in the curvature tensor of M. In particular, $\operatorname{vol}(B_\rho(M))$ depends only on ρ and the *intrinsic* geometry of M and not on the specific way it is embedded in $\mathbb{R}^l$. For a masterly account of this remarkable result and its many applications (including a proof, in full generality, of the Chern-Gauss-Bonnet formula in Riemannian Geometry) we refer to [Gra03]. Also, for a brief overview of the ubiquitous role the Gauss-Bonnet curvatures κ_{2q} play in Riemannian Geometry and related areas, see [Lab07] and the references therein.

Example 12.15. (Scheffé-type simultaneous band for the mean response in a normal linear model and the volume of tubes, again) One is often interested in obtaining simultaneous confidence bands for the mean response $\mathsf{x}^\top\beta$ in a normal regression model (as in Example 9.3) with x varying in some subset set $\mathscr{S} \subset \mathsf{R}^\mathsf{p}$, say $\mathscr{S}$ diffeomorphic to an interval or a rectangle and so on; here we retain the notation of Examples 9.13 and 9.14. Although the Scheffé-type band in (9.44) may be applied to this end, it certainly provides a wider band than required for a given confidence level, so a sensible strategy here is to seek for $c > 0$ satisfying

$$\mathsf{x}^\top\beta \in \left[\mathsf{x}^\top\widehat{\beta} \mp c\,\widehat{\sigma}\sqrt{\mathsf{x}^\top\mathfrak{s}\mathsf{x}}\right] \forall\, \mathsf{x} \in \mathscr{S} \text{ with prob. } 1-\delta. \tag{12.14}$$

Now, in terms of

$$\mathfrak{m} = \sigma\mathfrak{n} = (\mathfrak{p}^\top)^{-1}(\widehat{\beta}-\beta) \sim \mathcal{N}(\vec{0}, \sigma^2\mathrm{Id}_{p+1}), \tag{12.15}$$

$\varepsilon(\mathsf{x}) = \mathfrak{p}\mathsf{x}/|\mathfrak{p}\mathsf{x}| \in \mathbb{S}^p \subset \mathbb{R}^{p+1}$, $\mathsf{x} \in \mathscr{S}$, and $\mathfrak{u} = \mathfrak{m}/\|\mathfrak{m}\|$, a uniformly distributed random vector in $\mathbb{S}^p$, we have

$$\frac{\mathsf{x}^\top(\widehat{\beta}-\beta)}{\sqrt{\mathsf{x}^t\mathfrak{s}\mathsf{x}}} = \frac{(\mathfrak{p}^\top)^{-1}\mathsf{x}\mathfrak{m}}{\sqrt{(\mathfrak{p}\mathsf{x})^\top\mathfrak{p}\mathsf{x}}} = \left(\varepsilon(\mathsf{x})^\top\mathfrak{u}\right)\|\mathfrak{m}\|,$$

so if $T = \widehat{\sigma}/\|\mathfrak{m}\|$ then (12.14) expresses the corresponding tail probability as

$$\begin{aligned}\delta &= P\left(\sup_{\mathsf{x}\in\mathscr{S}}\left|\frac{\mathsf{x}^\top(\widehat{\beta}-\beta)}{\sqrt{\mathsf{x}^\top\mathfrak{s}\mathsf{x}}}\right| \geq c\widehat{\sigma}\right)\\ &= P\left(\sup_{\mathsf{x}\in\mathscr{S}}\left|\varepsilon(\mathsf{x})^\top\mathfrak{u}\right| \geq cT\right).\end{aligned}$$

Hence, using a notation similar to the one of the previous example, we see that the coverage probability in (12.14) is

$$1-\delta = \mathrm{vol}_{P_\mathfrak{u}}\left(B_{cT}(M_\varepsilon \cup M_{-\varepsilon})\right),$$

so that being able to compute the (normalized) volume of certain tubes around $M_\varepsilon \cup M_{-\varepsilon} \subset \mathbb{S}^p$ intervenes in determining the critical value c. Since T is independent of $\mathbf{u}$, this may be rewritten as

$$1-\delta = \int_0^{1/c} \mathrm{vol}_{P_\mathfrak{u}}\left(B_{ct}(M_\varepsilon \cup M_{-\varepsilon})\right)\psi_T(t)dt,$$

where ψ_T is the pdf of T. It follows from (9.34), (12.15) and the independence of $\mathbf{u}$ and $\widehat{\sigma}$ that $(p+1)T^2 \sim \mathsf{F}_{n-p-1,p+1}$, so (2.11) applies to give

$$\psi_T(t) = 2(p+1)t\psi_{\mathsf{F}_{n-p-1,p+1}}((p+1)t^2), \quad t \geq 0,$$

and the substitution $t = \cos\theta/c$ in the previous integral leads to

$$1-\delta = \int_0^{\pi/2} \mathrm{vol}_{P_\mathfrak{u}}\left(B_{\cos\theta}(M_\varepsilon \cup M_{-\varepsilon})\right)\psi(\theta)d\theta,$$

where

$$\psi(\theta) = \frac{2(p+1)\sin\theta\cos\theta}{c^2}\psi_{\mathsf{F}_{n-p-1,p+1}}\left(\frac{(p+1)\cos^2\theta}{c^2}\right).$$

To see how this implies Scheffé's original contribution, note that if $\mathscr{S} = \mathcal{R}^p$ then the volume function within this integral clearly equals 1 identically, so the substitution $\tau = (p+1)\cos^2\theta/c^2$ gives

$$1-\delta = \int_0^{(p+1)/c^2} \psi_{\mathsf{F}_{n-p-1,p+1}}(\tau)d\tau,$$

and using Corollary 4.34 with $\tau' = 1/\tau$,

$$1-\delta = \int_0^{c^2/(p+1)} \psi_{\mathsf{F}_{p+1,n-p-1}}(\tau')d\tau'.$$

But this means that

$$\frac{c^2}{p+1} = \mathsf{f}_{p+1,n-p-1,1-\delta},$$

which recovers (9.44) as promised. In general, when $\mathscr{S}$ is a proper subset of $\mathcal{R}^p$, "approximate" versions of Weyl's formula are needed in order to establish simultaneous confidence bands based on the computations above [SL94, Liu10]. □

13. A BRIEF OVERVIEW OF "CLASSICAL" PARAMETRIC ESTIMATION

Looking back, we can now outline some of the main mathematical foundations of the "classical" approach to Parametric Estimation Theory. Consider, for instance, a unidimensional statistical model with random sample

$$X_1, \dots, X_n \sim \psi_\theta, \quad \theta \in \mathbb{R}, \tag{13.1}$$

of size n. From this sample we form a statistic $h(X_1, \dots, X_n)$ with the aim of producing an estimator $\widehat{\theta}$ for θ, as in (7.2). Importantly, h should not depend on the unknown parameter θ.

A first step in evaluating the efficiency of $\widehat{\theta}$ is to compute (or at least reliably approximate) its mean squared error $\mathrm{mse}(\widehat{\theta})$. As illustrated above in the case $\widehat{\theta} = \widehat{\sigma}_c^2$, this requires computing the associated variance, a generally demanding task that becomes substantially simpler if ψ_θ is assumed to be normal. Although restrictive, this assumption is sometimes heuristically justified by appealing to the Central Limit Theorem (Theorem 6.5) together with the "hypothesis of elementary errors" [Fis11, Chapter 3]. If $\mathrm{mse}(\widehat{\theta})$ is sufficiently small, guaranteeing good performance of $\widehat{\theta}$, one may still need further information about the sampling distribution of the estimator, depending on the inferential goal.

For instance, as discussed in Subsection 7.3 for the sample mean, if the aim is to provide "small-sample" confidence intervals for θ, then for any n one should determine the explicit distribution $\psi_\theta^{(n)}$ of $h(X_1, \dots, X_n)$ (or of a suitable pivotal quantity). This is often delicate, even under normality, especially when h depends nonlinearly on the sample (as in the case of the sample variance). A landmark contribution here is Student's determination of the distribution of his pivotal quantity T_{n-1} in (7.26), which yields small-sample confidence intervals for the population mean μ. This, however, was achieved only under the assumption that the original sample is normal. Student's work profoundly influenced R. Fisher, who not only placed it on solid mathematical foundations through his "geometric method" (Remark 7.31), but also extended it to obtain the distribution of the correlation coefficient (Example 7.39)[49].

A less formidable task is to seek asymptotic information. A preliminary step is verifying that $\mathrm{mse}(\widehat{\theta}_n) \to 0$ as $n \to \infty$, which ensures that $\widehat{\theta}_n$ is consistent (see Proposition 7.11, itself a version of the LLN in this broader setting). To complement this asymptotic point estimate with a dispersion analysis, one typically looks for a distribution ψ such that a suitable standardization of $\widehat{\theta}_n$ converges in law to ψ. Ideally, this yields approximations for the distribution of $\sqrt{n}(\widehat{\theta}_n - \theta)$ that concentrate sharply around θ, allowing the construction of reliable confidence intervals from the tail probabilities of ψ in the large-sample regime. This approach, already illustrated in Subsection 7.3 for $\widehat{\theta}_n = \overline{X}_n$, is formalized in Theorem 8.24, which establishes the asymptotic normality of a broad class of consistent ML estimators. These estimators are thereby shown to be asymptotically efficient: as $n \to \infty$, their dispersion (measured by standard deviation) achieves the Cramér–Rao lower bound (Theorem 8.18); see also Remark 8.26[50]. Full expositions of these large-sample methods are available in [NM94, Leh99, LC06, Fer17, VdV00].

[49] Because of its intuitive geometric appeal, Fisher's approach is rarely reproduced in modern textbooks, where it is typically replaced by analytical methods involving Jacobian manipulations of the underlying coordinate transformations.

[50] The same dichotomy between small and large sample regimes reappears in Section 12, where hypothesis testing is treated.

It is remarkable that this modern strategy essentially mirrors the program laid out by R. Fisher in his foundational paper [Fis22], written a century ago. There Fisher declared that "the object of statistical methods is the reduction of data," and identified as the first major challenge the "Problems of Specification," namely, the choice of an appropriate statistical model (as in (13.1))[51]. He then turned to the "Problems of Estimation," concerned with selecting a statistic designed to estimate the parameters of the population. To judge the quality of an estimator, he proposed three criteria: the "Criterion of Consistency" (the estimator approaches the true parameter in the long run[52], the "Criterion of Efficiency" (for large samples, the estimator with smallest dispersion, what we would now phrase as variance achieving the Cramér–Rao lower bound, where Fisher information plays a key role; cf. Remark 8.25), and the "Criterion of Sufficiency" (the statistic should summarize all the relevant information in the sample). Finally, in what he called the "Problems of Distribution," Fisher emphasized the importance of computing the exact sampling distributions of estimators, to fully elucidate their theoretical properties. As Fisher himself noted, estimators that satisfy both consistency and efficiency criteria may still differ in finite-sample performance, which justifies sufficiency as a decisive tiebreaker. While the asymptotic aspects of estimation can often be handled analytically, the Problems of Distribution involve small-sample results of great mathematical difficulty[53]. In this same paper Fisher also introduced the method of maximum likelihood, offering for the first time a systematic procedure for constructing estimators within a given model. This represented a decisive step toward establishing the conceptual framework on which the frequentist approach to Statistical Estimation still rests[54].

14. A GLIMPSE AT THE BAYESIAN PATHWAY

As a way of comparison with the frequentist approach developed above, we now comment on the *Bayesian approach* to estimation, where it is assumed that the parameter θ in the i.i.d. measurements $X_j \sim \psi(\cdot;\theta)$ is random with a pdf ψ_ϑ, so that probabilities are assigned to parameters as well as to observations (here, we pretend that the random variable ϑ on Θ is the "capital" version of θ). It follows from Theorem 3.7 (Bayes rule) that

$$\psi_{\vartheta|_{X=\mathbf{x}}}(\theta) = \frac{\psi_{X|_{\vartheta=\theta}}(\mathbf{x})\psi_\vartheta(\theta)}{\psi_X(\mathbf{x})}, \quad \psi_X(\mathbf{x}) = \int_\Theta \psi_{X|_{\vartheta=\theta}}(\mathbf{x})\psi_\vartheta(\theta)d\theta,$$

where $X = (X_1, \cdots, X_n)$ and $\mathbf{x} = (x_1, \cdots, x_n)$ is a realization of X. In the Bayesian jargon, $\psi_\vartheta(\theta)$ is the *prior*, reflecting our knowledge of the underlying parameter θ previous to any measurement (and hence viewed as a *hypothesis*) and $\psi_{X|_{\vartheta=\theta}}(\mathbf{x})$ is the *likelihood*, which indicates the compatibility of the *evidence* X with the given hypothesis. Note that

$$\psi_{X|_{\vartheta=\theta}}(\mathbf{x}) = L(\mathbf{x};\theta),$$

the likelihood function in (8.2), hence the terminology; here we are momentarily coming back to the "frequentist" setting of Section 13 and thus regarding θ as deterministic (i.e. non-random). The prior and the likelihood combine to yield the *posterior* $\psi_{\vartheta|_{X=\mathbf{x}}}(\theta)$ through the proportionality

$$\psi_{\vartheta|_{X=\mathbf{x}}}(\theta) \propto L(\mathbf{x};\theta)\psi_\vartheta(\theta), \tag{14.1}$$

which provides an update of the probability distribution of the hypothesis as more observed evidence becomes available.

---

[51] According to Fisher, "these are entirely a matter for the practical statistician."

[52] Fisher's notion of consistency differs from the modern one presented here.

[53] This helps to explain Fisher's deep admiration for Student's work; see Remark 7.29. The same difficulty also reappears in hypothesis testing, as discussed in Remark 12.3.

[54] As noted by the leading historian of statistics S. Stigler in [Sti05]: "The paper is an astonishing work: It announces and sketches out a new science of statistics, with new definitions, a new conceptual framework and enough hard mathematical analysis to confirm the potential and richness of this new structure."

Example 14.1. For a normal sample $X_j \sim \mathcal{N}(\mu, \sigma^2)$, with σ known, we find that the likelihood is

$$L(\mathbf{x}; \mu) = \frac{1}{(2\pi)^{n/2}\sigma^n} e^{-\frac{1}{2\sigma^2}\sum_j (x_j - \mu)^2}.$$

Now assume that the prior, which expresses our initial degree of belief on the unknown parameter μ, follows the normal $\mathcal{N}(\mu_{\rm pr}, \sigma^2_{\rm pr})$, so that

$$\psi_\vartheta(\mu) = \frac{1}{\sqrt{2\pi}\sigma_{\rm pr}} e^{-\frac{(\mu-\mu_{\rm pr})^2}{2\sigma^2_{\rm pr}}}.$$

A direct computation using (14.1) confirms that the posterior also follows a normal, namely,

$$\psi_{\vartheta|_{X=\mathbf{x}}} \sim \mathcal{N}(\mu_{\rm pos}, \sigma^2_{\rm pos}),$$

where

$$\mu_{\rm pos} = (1-\lambda)\mu_{\rm pr} + \lambda \frac{\sum_j x_j}{n}, \quad \lambda = \frac{\sigma^2_{\rm pr}}{\sigma^2_{\rm pr} + \sigma^2/n}, \tag{14.2}$$

and

$$\sigma^2_{\rm pos} = \frac{\sigma^2_{\rm pr}\sigma^2/n}{\sigma^2_{\rm pr} + \sigma^2/n}.$$

Hence, the Bayesian recipe confines the posterior mean $\mu_{\rm pos}$ somewhere between the prior mean $\mu_{\rm pr}$ and the realization $\sum_j x_j/n$ of the sample mean, with a higher degree of belief than before (since $\sigma_{\rm pos} < \min\{\sigma_{\rm pr}, \sigma\}$). We thus see that the data-gathering provided by the sample has the net effect of fine-tuning our initial subjective knowledge regarding μ. □

Example 14.2. (Laplace's rule of succession) What chances are that the sun will rise tomorrow given that it has been so for the last n days? To ponder on this, consider a Bernoulli sample $X_j \sim \mathsf{Ber}(p)$ assigning probability p to a successful outcome corresponding to the event $\{1\}$. The question above is a special case (with $s = n$) of the general problem of computing

$$P(X_{n+1} = 1|_{X^{(n)}=s}) = P_{X_{n+1}|_{X^{(n)}=s}}(\{1\}), \quad X^{(n)} = X_1 + \cdots + X_n,$$

the probability that success occurs at the $(n+1)^{\rm th}$ outcome given that it has occurred s times previously; here we use the notation of (3.2). From Example 8.8 we know that the likelihood is

$$L(\mathbf{x}; p) = p^s(1-p)^{n-s},$$

where $s = x_1 + \cdots + x_n$ is the realization of $X^{(n)}$. The simplest choice for the prior distribution of $\mathfrak{p}$, the random variable associated to the Bayesian parameter p, appeals to the "Principle of Insufficient Reason": we declare that

$$\psi_{\mathfrak{p}}(p) = \mathbf{1}_{[0,1]}(p),$$

the *uniform distribution* supported on the unit interval $[0, 1]$. Using (14.1) we see that the posterior is

$$\begin{aligned} \psi_{\mathfrak{p}|_{X^{(n)}=s}}(p) &= \frac{p^s(1-p)^{n-s}\mathbf{1}_{[0,1]}(p)}{\int_0^1 p^s(1-p)^s dp} \\ &= \frac{(n+1)!}{s!(n-s)!} p^s(1-p)^{n-s}\mathbf{1}_{[0,1]}(p), \end{aligned}$$

the Beta distribution $\mathsf{Beta}(s+1, n-s+1)$; cf. Definition 4.35. It follows that

$$\begin{aligned} P_{X_{n+1}|_{X^{(n)}=s}}(\{1\}) &= \mathbb{E}(\mathfrak{p}|_{X^{(n)}=s}) \\ &= \int_0^1 p\psi_{\mathfrak{p}|_{X^{(n)}=s}}(p)dp, \end{aligned}$$

and using the previous expression for the posterior we get

$$P(X_{n+1}=1|_{X^{(n)}=s})=\frac{s+1}{n+2},$$

which in particular gives $(n+1)/(n+2)$ as the solution for Laplace's sunrise problem[55]. □

Example 14.3. If $\Theta\subset\mathbb{R}^q$ has a finite volume then the "Principle of Insufficient Reason" leads to

$$\psi_\vartheta(\theta)=\frac{1}{\mathrm{vol}(\Theta)}\mathbf{1}_\Theta(\theta),\quad \theta\in\Theta,$$

as the choice for the prior, so the corresponding posterior is

$$\psi_{\vartheta|X=\mathbf{x}}(\theta)=\frac{L(\mathbf{x};\theta)\mathbf{1}_\Theta(\theta)}{\int_\Theta L(\mathbf{x};\theta)d\theta},$$

a suitable normalization of the likelihood. This confirms that, viewed as a function of θ, the likelihood in general does not qualify as a pdf, which is consistent with the fact that the prescription for the MLE estimator in Definition 8.3 is insensitive to replacing L by cL, $c>0$ a constant. In other words, any multiple of the likelihood function carries the same information as far as selecting the ML estimator is concerned. □

The examples above illustrate the Bayesian credo according to which probability is nothing but a measure of our degree of belief on the underlying parameter θ, which thus should be random in nature. In any case, we may proceed with the corresponding estimation theory as follows. Given $\widetilde{\theta}$ define its *Bayes risk* as

$$\mathscr{R}(\widetilde{\theta})=\mathbb{E}_{\psi_\vartheta}(\mathscr{L}(\widetilde{\theta},\theta)),$$

where $\mathscr{L}$ is a (previously chosen) *loss function* (for instance, the quadratic loss $\mathscr{L}(\widetilde{\theta},\theta)=|\widetilde{\theta}-\theta|^2$ gives rise to the Bayesian analogue of the mse in (7.3), but be aware of a crucial difference: here we average against the prior $\psi_\vartheta(\theta)d\theta$ in alignment with the Bayesian philosophy according to which θ is random, whereas there we integrate against dP_θ since θ is regarded as deterministic (i.e. non-random); see Remark 7.3.

Definition 14.4. A *Bayes estimator* is any $\widehat{\theta}$ that minimizes the Bayes risk.

Notice that this only depends on the prior distribution (and the given loss function) and hence involves no observation. In any case, given this setup we are now in a position to implement the Bayesian updating paradigm relying on the subsequent measurement $X=x$ via (14.1). This leads to the following result, which provides a method for constructing Bayes estimators by solving a minimization problem formulated in terms of the posterior distribution $\psi_{\vartheta|X=x}$.

Theorem 14.5. *Assume that for almost all* $\mathbf{x}$ *there exists* $\widehat{\theta}(\mathbf{x})$ *minimizing*

$$\widetilde{\theta}\mapsto\mathbb{E}_{\psi_{\vartheta|X=\mathbf{x}}}(\mathscr{L}(\widetilde{\theta}(\mathbf{x}),\theta)),$$

where $\widetilde{\theta}$ *runs over the set of estimators with a finite risk. Then* $\widehat{\theta}=\widehat{\theta}(X)$ *is a Bayes estimator.*

[55] In the end of [Lap98, Chapter III], Laplace takes n corresponding to five thousand years and finds that "it is a bet of 1820214 to one that it will rise again tomorrow". But as Laplace himself recognizes in the sequel, this should be taken with a salt of grain, specially in regard to the choice of prior, as possibilities other than the uniform are certainly available; see Example 14.7.

Proof. By assumption we have, for almost all $\mathbf{x}$ and any $\widetilde{\theta}$,

$$\mathbb{E}_{\psi_{\vartheta|X=\mathbf{x}}}(\mathscr{L}(\widetilde{\theta}(\mathbf{x}),\theta)) \geq \mathbb{E}_{\psi_{\vartheta|X=\mathbf{x}}}(\mathscr{L}(\widehat{\theta}(\mathbf{x}),\theta)).$$

By Proposition 3.14, this may be expressed in terms of conditional expectations as

$$\mathbb{E}_{\psi_\vartheta}(\mathscr{L}(\widetilde{\theta}(X),\vartheta)|\mathcal{F}_X) \geq \mathbb{E}_{\psi_\vartheta}(\mathscr{L}(\widehat{\theta}(X),\vartheta)|\mathcal{F}_X).$$

By applying $\mathbb{E}_{\psi_\vartheta}$ to both sides and using Proposition 3.11 (2) we conclude that $\mathscr{R}(\widetilde{\theta}(X)) \geq \mathscr{R}(\widehat{\theta}(X))$. □

Corollary 14.6. *The Bayes estimator associated to the weighted quadratic loss $\mathscr{L}(\widetilde{\theta},\theta) := w(\theta)|\widetilde{\theta} - g(\theta)|^2$, $w > 0$, is given by*

$$\widehat{\theta}(\mathbf{x}) = \frac{\mathbb{E}_{\psi_{\vartheta|X=\mathbf{x}}}(wg)}{\mathbb{E}_{\psi_{\vartheta|X=\mathbf{x}}}(w)}.$$

Proof. The Cauchy-Schwartz inequality

$$E_{\psi_{\vartheta|X=\mathbf{x}}}(w(\theta)g(\theta))^2 < \mathbb{E}_{\psi_{\vartheta|X=\mathbf{x}}}(w(\theta))\mathbb{E}_{\psi_{\vartheta|X=\mathbf{x}}}(w(\theta)g(\theta)^2)$$

implies that

$$\begin{aligned}\mathbb{E}_{\psi_{\vartheta|X=\mathbf{x}}}(\mathscr{L}(\widetilde{\theta}(\mathbf{x}),\theta)) &= \mathbb{E}_{\psi_{\vartheta|X=\mathbf{x}}}(w(\theta))\widetilde{\theta}(\mathbf{x})^2 \\ &\quad -2\mathbb{E}_{\psi_{\vartheta|X=\mathbf{x}}}(w(\theta)g(\theta))\widetilde{\theta}(\mathbf{x}) + \mathbb{E}_{\psi_{\vartheta|X=\mathbf{x}}}(w(\theta)g(\theta)^2),\end{aligned}$$

viewed as a quadratic expression in $\widetilde{\theta}(\mathbf{x})$, has a negative discriminant and hence is minimized at $\widetilde{\theta}(\mathbf{x}) = \widehat{\theta}(\mathbf{x})$. □

We refer to [Rob07, Chapter 2] for an extensive discussion on this "decision-theoretic" approach to Bayes estimation.

Example 14.7. A much more flexible choice for the prior in the sunrise problem of Example 14.2 is

$$\psi_{\mathfrak{p}}(p) = \frac{\Gamma(\alpha+\beta)}{\Gamma(a)\Gamma(b)}p^{\alpha-1}(1-p)^{\beta-1}\mathbf{1}_{[0,1]}, \quad \alpha,\beta > 0,$$

the $\mathsf{Beta}(\alpha,\beta)$ distribution (the case $\alpha = \beta = 1$ corresponds to the uniform distribution). We thus calculate that the posterior is

$$\psi_{\mathfrak{p}|_{X^{(n)}=s}} \sim \mathsf{Beta}(\alpha+s, \beta+n-s),$$

so in this case we obtain

$$P(X_{n+1} = 1|_{X^{(n)}=s}) = \frac{s+\alpha}{n+\alpha+\beta}. \tag{14.3}$$

This illustrates how sensitive the Bayesian machinery is to the choice of the prior. Moreover, taking into account that the right-hand side of (14.3) equals the expected value of the posterior, if we choose the loss function to be $\mathscr{L}(\widetilde{\theta},\theta) = |\widetilde{\theta}-\theta|^2$ in Corollary 14.6 we find that the corresponding Bayes estimator is

$$\widehat{p}_{(n)}(X) = \frac{X^{(n)}+\alpha}{n+\alpha+\beta} = (1-\gamma)\frac{\alpha}{\alpha+\beta} + \gamma\overline{X}_n, \quad \gamma = \frac{n}{n+\alpha+\beta},$$

which interpolates between $\alpha/(\alpha+\beta)$, the expected value of $\psi_{\mathfrak{p}}$ (the natural estimator prior to any observation) and the sample mean $\overline{X}_n$ (the "frequentist" estimator that completely ignores the Bayesian paradigm incarnated in the prior). In particular, for large samples the prior mean plays a negligible role as $\widehat{p}_{(n)}(X)$ becomes

indistinguishable from the ML estimator $\overline{X}_n$. Finally, if we apply this same recipe to the normal setting of Example 14.1, it follows from (14.2) that the corresponding Bayes estimator is

$$\widehat{\mu}_{\mathrm{pos},n}(X) = (1-\lambda)\mu_{\mathrm{pr}} + \lambda\overline{X}_n, \quad \lambda = \frac{\sigma^2_{\mathrm{pr}}}{\sigma^2_{\mathrm{pr}} + \sigma^2/n}.$$

Again, this interpolates between the prior mean and the sample mean with

$$\widehat{\mu}_{\mathrm{pos},n}(X) \approx_{n\to+\infty} \overline{X}_n,$$

so the asymptotic behavior completely disregards the prior mean. □

Remark 14.8. (asymptotic efficiency of Bayes estimators) As illustrated in Example 14.7 above, Corollary 14.6 shows that the determination of a Bayes estimator for θ under a quadratic loss boils down to computing the expectation of the posterior, a quite feasible task in some cases. Similarly to the course of action taken in the "frequentist" setting, with those estimators at hand we may then examine their asymptotic efficiency. In the cases treated above, this may be easily reduced to CLT. Indeed, in the Bernoulli case we compute that

$$\sqrt{n}\left(\widehat{p}_{(n)}(X) - p\right) = \sqrt{n}\left(\overline{X}_n - p\right) + \frac{\sqrt{n}}{\alpha+\beta+n}\left(\alpha - (\alpha+\beta)\,\overline{X}_n\right),$$

where p is the true value of the unknown parameter, so that Theorem 2.25 and CLT apply to conclude that

$$\sqrt{n}\left(\widehat{p}_{(n)}(X) - p\right) \xrightarrow{d} \mathcal{N}(0, p(1-p)).$$

Similarly, in the normal setting,

$$\sqrt{n}\left(\widehat{\mu}_{\mathrm{pos},n}(X) - \mu\right) = \sqrt{n}\lambda\left(\overline{X}_n - \mu\right) + \sqrt{n}\,(1-\lambda)\left(\mu_{\mathrm{pr}} - \mu\right),$$

and since $\lambda \to 1$ and $\sqrt{n}(1-\lambda) = O(n^{-1/2}) \to 0$ we see that

$$\sqrt{n}\left(\widehat{\mu}_{\mathrm{pos},n}(X) - \mu\right) \xrightarrow{d} \mathcal{N}(0, \sigma^2).$$

Thus, in each case the limiting distribution of the appropriate standardization of the Bayes estimator is normal with a dispersion independent of the parameters of the prior distribution. This turns out to be a quite general phenomenon. Indeed, results in [LC06, Section 6.8] guarantee, under suitable regularity conditions and in the regime of large samples, that:

- the posterior distribution becomes asymptotically normal, and hence insensitive to the chosen prior, with a variance depending on the true value θ_0 of the unknown parameter only through its Fisher information:

$$\sqrt{n}\left(\psi_{\vartheta|_{X=\mathbf{x}}} - \theta_0\right) \xrightarrow{d} \mathcal{N}(0, \mathscr{F}(\theta_0)^{-1}). \tag{14.4}$$

- as a consequence, the limiting distribution associated to the Bayes estimator $\widehat{\theta}_n$ (under quadratic loss) is normal as well with the same asymptotic variance:

$$\sqrt{n}\left(\widehat{\theta}_n(X) - \theta_0\right) \xrightarrow{d} \mathcal{N}(0, \mathscr{F}(\theta_0)^{-1}). \tag{14.5}$$

  In particular, $\widehat{\theta}_n$ is asymptotically efficient.

We note that (14.5) follows from (14.4) and the fact that

$$\sqrt{n}\left(\widehat{\theta}_n(X) - \psi_{\vartheta|_{X=\mathbf{x}}}\right) \xrightarrow{d} 0.$$

Although in the long run these results eventually succeed in altogether eliminating the effect of the subjective choice of the prior, they remain a bit extraneous in regard to the Bayesian tenet according to which by its very nature the posterior is conditional on the sample, whose size has been fixed once and for all. □

APPENDIX A. BROWNIAN MOTION, ITÔ'S CALCULUS, AND SOME OF THEIR APPLICATIONS

In this rather long appendix, whose understanding requires only familiarity with the material above up to Subsection 4.1, we turn our attention to Brownian motion, an important example of a stochastic process, and its most basic properties. Although the main motivation here is to provide a proof of the Gaussian concentration inequality (5.18) with the sharp constant $C = 1/2$, which is presented in Section A.4, we also include a few other applications of the associated Itô's calculus, a cornerstone in the modern theory of stochastic processes.

A.1. **Brownian motion: its construction and basic regularity properties.**

Since Brownian motion is the prototypical example of a stochastic process, we start by recalling the definition of this fundamental concept.

Definition A.1. A stochastic process on a probability space $(\Omega, \mathcal{F}, P)$ is a one-parameter family of random variables $X_t : \Omega \to \mathbb{R}^n$, $t \geq 0$.

The map $\omega \in \Omega \mapsto X_t(\omega) \in \mathbb{R}^n$ allows us to think of Ω as a subset of $(\mathbb{R}^n)^{[0,+\infty)}$. Thus, to each $\omega \in \Omega$ the process defines a path in $\mathbb{R}^n$. In general, the regularity of the process is expressed in terms of the regularity of these paths. For instance, we say that the process is continuous if $X_t(\omega)$ is continuous for almost any $\omega \in \Omega$. Here we only deal with processes which are at least continuous. In any case, this pathwise description of stochastic processes motivates the following definition.

Definition A.2. Given a stochastic process X_t, its *probability distributions* in $\mathbb{R}^{nk}$, $k = 1, 2, \cdots$, are given by

$$\mu^X_{t_1,\cdots,t_k}(F_1 \times \cdots \times F_k) = P(X_{t_1} \in F_1, \cdots, X_{t_k} \in F_k),$$

where $t_i \geq 0$ and $F_i \in \mathcal{B}^n$, $i = 1, \cdots, k$.

In other words, $\mu^X_{t_1,\cdots,t_k} = P_{(X_{t_1},\cdots,X_{t_k})}$, the joint distribution of $(X_{t_1}, \cdots, X_{t_k})$; cf. Definition 2.8. The next result shows that a stochastic process can be reconstructed from their probability distributions given that a couple of compatibility conditions are satisfied.

Theorem A.3. *(Kolmogorov's extension) Assum that for any $t_1, \cdots, t_k \geq 0$ there exists a probability measure $\nu_{t_1,\cdots,t_k}$ in $\mathbb{R}^{nk}$ such that:*

- *(K_1) $\nu_{t_{\tau(1)},\cdots,t_{\tau(k)}}(F_1 \times \cdots \times F_k) = \nu_{t_1,\cdots,t_k}(F_{\tau^{-1}(1)}, \cdots, F_{\tau^{-1}(k)})$, for any permutation τ.*
- *(K_2) $\nu_{t_1,\cdots,t_k}(F_1 \times \cdots \times F_k) = \nu_{t_1,\cdots,t_k,t_{k+1},\cdots,t_{k+m}}(F_1 \times \cdots \times F_k \times \mathbb{R}^n \times \cdots \times \mathbb{R}^n)$, for any $m \geq 1$.*

Then there exists a probability space $(\Omega, \mathcal{F}, P)$ and a stochastic process $X_t : \Omega \to \mathbb{R}^n$ such that

$$\nu_{t_1,\cdots,t_k} = \mu^X_{t_1,\cdots,t_k},$$

for $(t_1, \cdots, t_k)$.

Proof. See [Tao11, Section 2.4]. □

Now we can construct Brownian motion in $\mathbb{R}^n$ following an approach due to Kolmogorov; for other possibilities see [SP14]. If $0 \leq t_1 < \cdots < t_k$ and $y = (y_1, \cdots, y_k) \in \mathbb{R}^{nk}$ define, if $t_1 > 0$,

$$\nu_{t_1,\cdots,t_k}(F_1 \times \cdots \times F_k) = \frac{1}{(2\pi)^{nk/2}\sqrt{\det C}} \int_{F_1 \times \cdots \times F_k} e^{-\frac{1}{2}\langle C^{-1}y, y\rangle} dy,$$

where each $F_i \in \mathcal{B}^n$ and C is $nk \times nk$-matrix whose ij-block is $C_{ij} = t_i \wedge t_j I_n$[56]. If $t_1 = 0$ we use instead $\delta_{\vec{0}} \otimes \nu_{t_2,\cdots,t_k}$, where $\delta_{\vec{0}}$ is the Dirac measure centered at the origin. This may be extended to all $(t_1, \cdots, t_k)$ so that (K_1) is satisfied. Moreover, (K_2) is satisfied as well because of Proposition 4.2. Thus, by means of Theorem A.3 we establish the following foundational existence result.

Theorem A.4. *There exists a probability space $(\Omega, \mathcal{F}, P)$ and a stochastic process $b_t : \Omega \to \mathbb{R}^n$ so that*

$$P(b_{t_1} \in F_1, \cdots, b_{t_k} \in F_k) = \frac{1}{(2\pi)^{nk/2}\sqrt{\det C}} \int_{F_1 \times \cdots \times F_k} e^{-\frac{1}{2}\langle C^{-1}x, x\rangle} dx.$$

This is called Brownian motion *(*BM*) in $\mathbb{R}^n$ (starting at $\vec{0}$).*

The next proposition lists the characterizing properties of BM.

Proposition A.5. *BM in $\mathbb{R}^n$ satisfies the following properties:*

(1) *$b_0 = 0$ a.s.;*

(2) *it has stationary normal increments, i.e. for any $0 \leq s < t$, $h \geq -s$, $b_{t+h} - b_{s+h}$ and $b_t - b_s$ are identically distributed with*

$$b_t - b_s \sim \mathcal{N}(0, (t-s)\mathrm{Id}_n); \tag{A.1}$$

(3) *it has independent increments, that is, for any $0 = t_0 < t_1 \cdots < t_k$, $\{b_{t_1} - b_{t_0}, \cdots, b_{t_k} - b_{t_{k-1}}\}$ are independent random vectors;*

(4) *$t \mapsto b_t(\omega)$ is continuous for any ω.*

Proof. From the construction, (1) follows immediately. To approach (2) and (3) we take $u = (u_1, \cdots, u_k) \in \mathbb{R}^{nk}$, $v_j = u_j + \cdots + u_k$, $j = 1, \cdots, k$, and $b = (b_{t_1}, \cdots, b_{t_k})$, so if we use the language of characteristic functions in Definition 2.26 and the explicit computation of this object for normally distributed random vectors in Proposition 4.4 we have (recalling that $b_{t_0} = 0$)

$$\begin{aligned}
\phi_{(b_{t_1}-b_{t_0},\cdots,b_{t_k}-b_{t_{k-1}})}(v_1, \cdots, v_k) &= \mathbb{E}(e^{\mathbf{i}\sum_{j=1}^k \langle b_{t_j} - b_{t_{j-1}}, v_j\rangle}) \\
&= \mathbb{E}(e^{\mathbf{i}\langle b, u\rangle}) \\
&= \phi_b(u) \\
&= e^{-\frac{1}{2}\langle Cu, u\rangle}.
\end{aligned}$$

[56]Here, $a \wedge b = \min\{a, b\}$. Also, note that the symmetric matrix C is positive definite.

But

$$\begin{aligned}\langle Cu,u\rangle &= \sum_{j=1}^{k}\sum_{l=1}^{k}(t_j\wedge t_l)\langle u_j,u_l\rangle\\ &= t_k\|u_k\|^2+\sum_{j=1}^{k-1}t_j\langle u_j,u_j+2u_{j+1}+\cdots 2u_k\rangle\\ &= t_k\|u_k\|^2+\sum_{j=1}^{k-1}t_j(\|u_j+\cdots+u_k\|^2-\|u_{j+1}+\cdots+u_k\|^2)\\ &= \sum_{j=1}^{k}t_j\|u_j+\cdots+u_k\|^2-\sum_{j=1}^{k}t_{j-1}\|u_j+\cdots+u_k\|^2\\ &= \sum_{j=1}^{k}(t_j-t_{j-1})\|v_j\|^2,\end{aligned}$$

so that

$$\phi_{(b_{t_1}-b_{t_0},\cdots,b_{t_k}-b_{t_{k-1}})}(v_1,\cdots,v_k)=\Pi_{j=1}^{k}e^{-\frac{1}{2}(t_j-t_{j-1})\|v_j\|^2}. \tag{A.2}$$

Notice that for $0\le s<t$ this specializes to

$$\phi_{b_t-b_s}(v)=e^{-\frac{1}{2}(t-s)\|v\|^2},\quad v\in\mathbb{R}^n,$$

so that Corollary 4.6 applies to ensure that (A.1) holds, which proves (2). As for (3), note that (A.2) may be rewritten as

$$\phi_{(b_{t_1}-b_{t_0},\cdots,b_{t_k}-b_{t_{k-1}})}(v_1,\cdots,v_k)=\Pi_{j=1}^{k}\phi_{b_{t_j}-b_{t_{j-1}}}(v_j),$$

so we may proceed as in the last step of the proof of Proposition 4.11 and use the standard Fourier inversion formula to confirm that the joint distribution of the random vector of increments decomposes as

$$\psi_{(b_{t_1}-b_{t_0},\cdots,b_{t_k}-b_{t_{k-1}})}(x_1,\cdots,x_k)=\Pi_{j=1}^{k}\psi_{b_{t_j}-b_{t_{j-1}}}(x_j),$$

which proves (3) by Proposition 2.15. The proof of (4) is presented in the next section; see Proposition A.10. □

Proposition A.6. *If $t\le s$ then $\mathbb{E}(\|b_s-b_t\|^2)=n(s-t)$.*

Proof. We have seen that $\mathbb{E}(b_t)=0$ and $\mathrm{cov}(b_s,b_t)=s\wedge tI_n$. Thus, if $b_t=(b_t^{(1)},\cdots,b_t^{(n)})$ is the coordinate expression of b_t we have

$$\begin{aligned}\mathbb{E}(\langle b_s,b_t\rangle) &= \sum_i\mathbb{E}\left(b_s^{(i)}b_t^{(i)}\right)\\ &= \sum_i\mathrm{cov}(b_s,b_t)_{ii}\\ &= ns\wedge t.\end{aligned}$$

If $t\le s$ we then have

$$\begin{aligned}\mathbb{E}(\|b_s-b_t\|^2) &= \mathbb{E}(\|b_s\|^2-2\langle b_s,b_t\rangle+\|b_t\|^2)\\ &= n(s-2t+t),\end{aligned}$$

as desired. □

Remark A.7. It follows from (A.1) that $\mathrm{cov}(b_t)_{ij} = t\delta_{ij}$, so that by Proposition 4.11 we see that the coordinate components $\{b_t^{(i)}\}_{i=1}^n$ of b_t form an independent family of BMs in $\mathbb{R}$. Conversely, we may first construct BM b_t in $\mathbb{R}$ by using the Kolmogorov's argument above (note that in this case the matrix C has a much simpler structure) and then take n *independent* copies of b_t, say $\{b_t^{(1)}, \cdots b_t^{(n)}\}$, in order to exhibit BM in $\mathbb{R}^n$ as $(b_t^{(1)}, \cdots, b_t^{(n)})$. □

We now turn to the basic regularity properties of Brownian motion and we start by proving Proposition A.5, (4). To simplify matters, we only consider the case $n = 1$; cf. Remark A.7. The proof is based on the following general regularity result for stochastic processes. We recall that saying that X_t' is a *modification* of X_t means that $P(X_t = X_t') = 1$ for any t.

Theorem A.8. *(Kolmogorov's continuity) If $X_t : \Omega \to \mathbb{R}$ is a stochastic process satisfying*

$$\mathbb{E}(|X_s - X_t|^\alpha) \leq C|s-t|^{\beta+1}, \quad s, t \geq 0,$$

then there exists a modification X_t' of X_t whose paths are locally γ-Hölder continuous, where $0 < \gamma < \beta/\alpha$. In particular, $X_t' \in C^0$.

Proof. See [LG13, Section 2.2]. □

The regularity of BM now follows from the following fact.

Proposition A.9. *BM in $\mathbb{R}$ satisfies*

$$\mathbb{E}(|b_s - b_t|^{2k}) = \frac{(2k)!}{2^k k!}|s-t|^k, \quad k \geq 1.$$

Proof. Since $b_s - b_t \sim \mathcal{N}(0, s-t)$, this follows from the discussion in Example 4.9. □

Proposition A.10. *Eventually passing to a modification, BM is locally $(\frac{1}{2} - \epsilon)$-Hölder continuous, for any $\epsilon > 0$.*

Proof. Apply the results above with $\alpha = 2k$ and $\beta = k - 1$ and send $k \to +\infty$. □

This is in a sense the best regularity we can have. To check this we need a definition.

Definition A.11. If $X_t : \Omega \to \mathbb{R}$ and $p > 0$, we define its p^{th} *variation* by

$$\langle X \rangle_t^{(p)}(\omega) = \lim_{\Delta t_k \to 0} \sum_{t_k \leq t} |X_{t_{k+1}}(\omega) - X_{t_k}(\omega)|^p,$$

where $\Delta t_k = t_{k+1} - t_k = t/k$ and the limit is taken in probability.

It turns out that the quadratic variation of BM can be explicitly computed.

Proposition A.12. *BM b_t in $\mathbb{R}$ satisfies*

$$\langle b \rangle_t^{(2)} = t,$$

with convergence in L^2-mean.

Proof. Note that

$$\begin{aligned}\mathbb{E}\left(\left(\sum_{t_k\leq t}(b_{t_{k+1}}-b_{t_k})^2-t\right)^2\right) &= \mathbb{E}\left(\left(\sum_{t_k\leq t}(b_{t_{k+1}}-b_{t_k})^2-(t_{k+1}-t_k)\right)^2\right)\\ &= I+II,\end{aligned}$$

where

$$I=\mathbb{E}\left(\sum_{t_k\leq t}\left((b_{t_{k+1}}-b_{t_k})^2-(t_{k+1}-t_k)\right)^2\right)$$

and

$$II=2\sum_{t_j<t_k\leq t}\underbrace{\mathbb{E}\left(\left((b_{t_{j+1}}-b_{t_j})^2-(t_{j+1}-t_j)\right)\left((b_{t_{k+1}}-b_{t_k})^2-(t_{k+1}-t_k)\right)\right)}_{III}$$

By Proposition A.6 we know that

$$\mathbb{E}\left((b_{t_{j+1}}-b_{t_j})^2\right)=t_{j+1}-t_j,$$

which implies that

$$III=\operatorname{cov}\left((b_{t_{j+1}}-b_{t_j})^2,(b_{t_{k+1}}-b_{t_k})^2\right).$$

But by Proposition A.5, (3),

$$b_{t_{j+1}}-b_{t_j}\perp b_{t_{k+1}}-b_{t_k}\Rightarrow(b_{t_{j+1}}-b_{t_j})^2\perp(b_{t_{k+1}}-b_{t_k})^2,$$

and hence $II=0$ by Corollary 2.7. On the other hand,

$$\begin{aligned}I &= \sum_{t_k\leq t}\left(\mathbb{E}\left((b_{t_{k+1}}-b_{t_k})^4\right)-2(t_{k+1}-t_k)\mathbb{E}((b_{t_{k+1}}-b_{t_k})^2)+(t_{k+1}-t_k)^2\right)\\ &\overset{\text{Prop.}A.9}{=} \sum_{t_k\leq t}\left(3(t_{k+1}-t_k)^2-2(t_{k+1}-t_k)^2+(t_{k+1}-t_k)^2\right)\\ &= 2\sum_{t_k\leq t}(t_{k+1}-t_k)^2\\ &\leq 2\frac{t^2}{k}\to 0,\end{aligned}$$

as $k\to+\infty$. □

Proposition A.13. *One has $\langle b\rangle_t^{(1)}=+\infty$ a.s. In other words, the total variation of b_t blows up in any interval.*

Proof. For $\omega\in\Omega$ we have

$$\begin{aligned}\sum_{t_k\leq t}(b_{t_{k+1}}(\omega)-b_{t_k}(\omega))^2 &\leq \sum_{t_k\leq t}(b_{t_{k+1}}(\omega)-b_{t_k}(\omega))\sup_{t_k}\sum_{t_k\leq t}|b_{t_{k+1}}(\omega)-b_{t_k}(\omega)|\\ &\leq \langle b\rangle_t^{(1)}(\omega)\sup_{t_k}\sum_{t_k\leq t}|b_{t_{k+1}}(\omega)-b_{t_k}(\omega)|.\end{aligned}$$

From Proposition A.12, and possibly passing to a subequence along the given partitions of $[0,t]$, we may assume that the left-hand side converges to t a.s. But the supremum goes to 0 as $b_t(\omega)$ is uniformly continuous in $[0,t]$, which yields a contradiction if $\langle b\rangle_t^{(1)}(\omega)$ is finite. □

A similar argument yields the following result.

Proposition A.14. *The paths of b_t are nowhere γ-Hölder continuous for $\gamma > 1/2$.*

Proof. Assume $|b_{s'} - b_{t'}| \leq K|s' - t'|^{\gamma}$, $0 \leq t' \leq s' \leq t$. It follows that

$$\sum_{t_k \leq t} (b_{t_{k+1}}(\omega) - b_{t_k}(\omega))^2 \leq K^2 t \sup_k |t_{k+1} - t_k|^{2\gamma - 1}.$$

As above, we may assume that the left-hand side converges to t a.s. But the supremum goes to 0 if $\gamma > 1/2$, which gives a contradiction. □

Remark A.15. (The Wiener space) From Proposition A.5 (4) we may conveniently identify the sample space Ω underlying the construction of Brownian motion in Theorem A.4 to $C_{\vec{0}}$, the space of continuous functions $\omega : [0, +\infty) \to \mathbb{R}^n$ with $\omega(0) = \vec{0}$ by the rule $\omega(t) = b_t(\omega)$. It then follows from Proposition A.14 that the support of the underlying probability measure P fails to contain any ω which is sufficiently regular (i.e. γ-Hölder continuos for $\gamma > 1/2$). From this perspective, we call $C_{\vec{0}}$ endowed with the induced probability measure, still denoted by P, as the *Wiener space* (starting at $\vec{0}$) and any ω lying in the support of P (the *Wiener measure*) as a *Brownian path* (again, starting at $\vec{0}$).

A.2. **Martingales.** We now isolate another central notion in the theory.

Definition A.16. Let $b_t : \Omega \to \mathbb{R}^n$ be BM in $\mathbb{R}^n$ (starting at x) and let $\mathcal{F}_t$ be the σ-algebra generated by $\{b_{t'}\}_{t' \leq t}$. A *martingale* (rel. to b_t) is a stochastic process $M_t : \Omega \to \mathbb{R}^n$ such that

- M_t is $\mathcal{F}_t$-measurable for any $t > 0$;
- $\mathbb{E}(\|M_t\|) < +\infty$;
- $\mathbb{E}(M_s|\mathcal{F}_t) = M_t$ whenever $t \leq s$.

Proposition A.17. *b_t is a martingale.*

Proof. If $t \leq s$ write

$$\mathbb{E}(b_s|\mathcal{F}_t) = \mathbb{E}(b_s - b_t|\mathcal{F}_t) + \mathbb{E}(b_t|\mathcal{F}_t).$$

Proposition A.5, (3), implies that $b_s - b_t \perp \mathcal{F}_t$. Hence, by Proposition 3.11, (4), $\mathbb{E}(b_s - b_t|\mathcal{F}_t) = \mathbb{E}(b_s - b_t) = 0$. On the other hand, since b_t is (obviously) $\mathcal{F}_t$-measurable, Proposition 3.11, (3), implies that $\mathbb{E}(b_t|\mathcal{F}_t) = b_t$. □

For our purposes, a basic property of a martingale is that its expectation is preserved in time. This confirms that martingales are "pure fluctuation" processes.

Proposition A.18. *If M_t is a martingale then $\mathbb{E}(M_t) = \mathbb{E}(M_s)$, for any s, t.*

Proof. By Proposition 3.11, (2), if $t \leq s$ we have

$$\mathbb{E}(M_t) = \mathbb{E}(\mathbb{E}(M_s|\mathcal{F}_t)) = \mathbb{E}(M_s).$$

□

A.3. **Itô's integral and Itô's formula.** Consider a partition $0 = t_0 < t_1 \cdots < t_k = t$. Let b_t be BM in $\mathbb{R}$ with $b_0 = 0$. By Proposition A.5, (3), $b_{t_{j+1}} - b_{t_j} \perp b_{t_j}$ and hence

$$\mathbb{E}\left(\sum_j b_{t_j}(b_{t_{j+1}} - b_{t_j})\right) = 0. \tag{A.3}$$

On the other hand,

$$\begin{aligned}\mathbb{E}\left(\sum_j b_{t_{j+1}}(b_{t_{j+1}} - b_{t_j})\right) &\overset{(A.3)}{=} \mathbb{E}\left(\sum_k (b_{t_{j+1}} - b_{t_j})^2\right) \\ &= \sum_j (t_{j+1} - t_j) \\ &= t,\end{aligned}$$

where we used Proposition A.6 in the second step. This simple computation, which reflects the already known fact that db_t can not be interpreted as a classical Lebesgue-Stieltjes integrator since b_t has infinite total variation by Proposition A.13, illustrates the difficulty of making sense of stochastic integrals like $\int_0^t b_s db_s$ by standard methods. Put in another way, each choice of $\{\widehat{t}_j\}$ such that $t_j \leq \widehat{t}_j \leq t_{j+1}$ yields its own output for the "approximate" stochastic integral $\sum_j b_{\widehat{t}_j}(b_{t_{j+1}} - b_{t_j})$. Among the many possibilities available, Itô's integration corresponds to choosing the first option (A.3) above. The basics of this kind of stochastic integration may be found in many sources [KS12, LG13, Bau14] and our presentation below follows [Oks13] closely, a text very much oriented to applications (in particular, to Mathematical Finance) to which we refer for the detailed proofs of most of the results on Itô's calculus described in the sequel.

Recall that a filtration $\mathcal{F}_t$ of a σ-algebra is a nested family of σ-subalgebras of $\mathcal{F}$. Here we consider the filtration $\mathcal{F}_t = \mathcal{F}_{\{b_{t'}\}_{t' \leq t}}$.

Definition A.19. We say that a process $f : \mathbb{R}_+ \times \Omega \to \mathbb{R}$ is *adapted* if $\omega \mapsto f(t, \omega)$ is $\mathcal{F}_t$-measurable for any t.

Definition A.20. For $0 \leq S < T$ we denote by $\mathcal{A}(S, T)$ the class of all processes $f : \mathbb{R}_+ \times \Omega \to \mathbb{R}$ such that:

(1) f is $\mathcal{B} \times \mathcal{F}$-measurable;
(2) f if *adapted* to $\mathcal{F}_t$, the filtration defined by b_t;
(3) $\mathbb{E}(\int_S^T f(t, \omega)^2 dt) < +\infty$.

Definition A.21. We say that $f \in \mathcal{A}(S, T)$ is *elementary* if

$$f(t, \omega) = \sum_j f_j(\omega) \mathbf{1}_{[t_j, t_{j+1})}(t).$$

Note that if f is elementary then f_j is $\mathcal{F}_{t_j}$-measurable.

Definition A.22. (Itô's integral for elementary processes) If $f \in \mathcal{A}(S,T)$ is elementary we define

$$\int_S^T f(t,\omega)db_t(\omega) = \sum_j f_j(\omega)(b_{t_{j+1}}(\omega) - b_{t_j}(\omega)).$$

Notice that this depends measurably on ω and hence defines a random variable.

Proposition A.23. *(Itô's isometry) If $f \in \mathcal{A}(S,T)$ is elementary then*

$$\mathbb{E}\left(\left(\int_S^T f(t,\omega)db_t(\omega)\right)^2\right) = \mathbb{E}\left(\int_S^T f(t,\omega)^2 dt\right).$$

Proof. Let $f = \sum_j f_j \mathbf{1}_{[t_j,t_{j+1})}$. Since f_j is $\mathcal{F}_{t_j}$-measurable, Proposition A.5, (3), implies that $f_j \perp b_{t_{j+1}} - b_{t_j}$. Hence, $f_j^2 \perp (b_{t_{j+1}} - b_{t_j})^2$ and we have

$$\mathbb{E}\left(f_j^2(b_{t_{j+1}} - b_{t_j})^2\right) = \mathbb{E}(f_j^2)\mathbb{E}((b_{t_{j+1}} - b_{t_j})^2) = \mathbb{E}(f_j^2)(t_{j+1} - t_j),$$

where we used Proposition A.6 in the last step. On the other hand, if $j < k$ we have $f_j f_k(b_{t_{j+1}} - b_{t_j}) \perp b_{t_{k+1}} - b_{t_k}$ and hence,

$$\mathbb{E}\left(f_j f_k(b_{t_{j+1}} - b_{t_j})(b_{t_{k+1}} - b_{t_k})\right) = \mathbb{E}\left(f_j f_k(b_{t_{j+1}} - b_{t_j})\right)\mathbb{E}(b_{t_{k+1}} - b_{t_k}) = 0$$

It follows that

$$\begin{aligned}
\mathbb{E}\left(\left(\int_S^T f(t,\omega)db_t(\omega)\right)^2\right) &= \sum_{jk} \mathbb{E}\left(f_j f_k(b_{t_{j+1}} - b_{t_j})(b_{t_{k+1}} - b_{t_k})\right) \\
&= \sum_j \mathbb{E}(f_j^2)(t_{j+1} - t_j) \\
&= \mathbb{E}\left(\int_S^T f(t,\omega)^2 dt\right),
\end{aligned}$$

as desired. □

Proposition A.24. *(approximation) For any $f \in \mathcal{A}(S,T)$ there exists $\{f_i\}_{i=1}^{+\infty} \subset \mathcal{A}(S,T)$, f_i elementary, so that*

$$\lim_{i\to+\infty} \mathbb{E}\left(\int_S^T |f - f_i|^2 dt\right) = 0. \tag{A.4}$$

Proof. [Oks13, pg. 27-28]. □

Definition A.25. (Itô's integral in $\mathcal{A}(S,T)$) If $f \in \mathcal{A}(S,T)$ we define

$$\int_S^T f(t,\omega)db_t(\omega) \stackrel{L^2}{=} \lim_{i\to+\infty} \int_S^T f_i(t,\omega)db_t(\omega),$$

for some $\{f_i\}$ as in (A.4).

Notice that, by Proposition A.23, the limit exists and does not depend on the sequence $\{f_i\}$ chosen to approximate f.

We now list the basic properties of Itô's integral.

Proposition A.26. *The Itô's integral satisfies the following properties:*

(1) $\int_S^T f db_t = \int_S^U f db_t + \int_U^T f db_t$;

(2) $\int_S^T (af + bg)db_t = a\int_S^T f db_t + b\int_S^T g db_t$, $a, b \in \mathbb{R}$;

(3) $\mathbb{E}(\int_S^T f db_t) = 0$;

(4) $\int_S^T f db_t$ *is* $\mathcal{F}_T$*-measurable;*

(5) *(Itô's isometry) There holds*

$$\mathbb{E}\left(\left(\int_S^T f(t,\omega)db_t(\omega)\right)^2\right) = \mathbb{E}\left(\int_S^T f(t,\omega)^2 dt\right).$$

(6) *If*

$$\lim_{n\to+\infty} \mathbb{E}\left(\int_S^T \left(f_n(t,\omega) - f(t,\omega)\right)^2 dt\right) = 0$$

then

$$\int_S^T f_n db_t \stackrel{L^2}{\to} \int_S^T f db_t.$$

(7) *Any Itô's integral has a continuous modification.*

Proof. The proofs of (1)-(5) follow the same method, namely, we first check the property for elementary processes and then pass the limit. Also, (6) follows immediately from (5). Finally, the proof of (7) can be found in [Oks13, Theorem 3.2.5]. □

Example A.27. Let

$$f_n(s,\omega) = \sum_j b_{t_j}(\omega)\mathbf{1}_{[t_j,t_{j+1})}(s),$$

where $\Delta t_j = t_{j+1} - t_j = t/n$. We have

$$\begin{aligned}
\mathbb{E}\left(\int_0^t (f_n - b_s)^2 ds\right) &= \mathbb{E}\left(\sum_j \int_{t_j}^{t_{j+1}} (f_n - b_s)^2 ds\right) \\
&= \mathbb{E}\left(\sum_j \int_{t_j}^{t_{j+1}} (b_{t_j} - b_s)^2 ds\right) \\
&= \sum_j \int_{t_j}^{t_{j+1}} (s - t_j) ds \\
&= \sum_j \frac{1}{2}(t_{j+1} - t_j)^2 \overset{n\to+\infty}{\to} 0.
\end{aligned}$$

Thus, by Proposition A.26, (6),

$$\int_0^t b_s db_s = \lim_{n\to+\infty} \int_0^t f_n db_s = \lim_{\Delta t_j \to 0} \sum_j b_{t_j}(b_{t_{j+1}} - b_{t_j}).$$

But, since $b_0 = 0$,

$$\begin{aligned}
b_t^2 &= \sum_j (b_{t_{j+1}}^2 - b_{t_j}^2) \\
&= \sum_j (b_{t_{j+1}} - b_{t_j})^2 + 2\sum_j b_{t_j}(b_{t_{j+1}} - b_{t_j}),
\end{aligned}$$

that is,

$$\sum_j b_{t_j}(b_{t_{j+1}} - b_{t_j}) = \frac{1}{2} b_t^2 - \frac{1}{2}\sum_j (b_{t_{j+1}} - b_{t_j})^2.$$

By passing the limit and using Proposition A.12 to handle the last term in the right-hand side we conclude that

$$\int_0^t b_s db_s = \frac{1}{2} b_t^2 - \frac{t}{2}.$$

At least formally, we can rewrite this as

$$db_t^2 = 2 b_t db_t + dt.$$

Setting $f(x) = x^2$ we have

$$df(b_t) = f'(b_t) db_t + \frac{1}{2} f''(b_t) dt.$$

This rather special case of the famous Itô's formula illustrates the appearance of an extra term in the chain rule in the stochastic chain rule. In fact, if we interpret Proposition A.12 as saying that $db_t^2 = dt$, we have

$$df(b_t) = f'(b_t) db_t + \frac{1}{2} f''(b_t) db_t^2,$$

which means that we must expand up to second order in db_t to obtain the correct version of the chain rule. □

We now prove that Itô's integrals are martingales.

Proposition A.28. *If $f \in \mathcal{A}(0,t)$ consider the process*

$$M_t = \int_0^t f(\rho,\omega) db_\rho(\omega).$$

Then M_t is martingale.

Proof. If $t \leq s$ we have

$$\mathbb{E}\left(M_s|\mathbb{F}_t\right) = \mathbb{E}\left(M_t|\mathbb{F}_t\right) + \mathbb{E}\left(\int_t^s f(\rho,\omega) db_\rho(\omega)\right).$$

Since M_t is $\mathcal{F}_t$-measurable (Proposition A.26, (4)), we have $\mathbb{E}\left(M_t|\mathbb{F}_t\right) = M_t$. Moreover, $\int_t^s f(\rho,\omega) db_\rho(\omega)$ is 'independent' of $\mathcal{F}_t$ in the sense that

$$\mathbb{E}\left(\int_t^s f(\rho,\omega) db_\rho(\omega)\right) = 0, \tag{A.5}$$

which completes the proof except for the checking of (A.5), which needs to be carried out only for f of the type $f = \sum_j f_j \mathbf{1}_{[t_j,t_{j+1})}$. In this case,

$$\begin{aligned}
\mathbb{E}\left(\int_t^s f(\rho,\omega) db_\rho(\omega)\right) &= \mathbb{E}\left(\sum_j f_j (b_{t_{j+1}} - b_{t_j})|\mathcal{F}_t\right) \\
&\overset{\mathcal{F}_t \subset \mathcal{F}_{t_j} + \text{Prop.3.11,(5)}}{=} \sum_j \mathbb{E}\left(\mathbb{E}\left(f_j (b_{t_{j+1}} - b_{t_j})|\mathcal{F}_{t_j}\right)|\mathcal{F}_t\right) \\
&\overset{\text{Prop.3.11,(2)}}{=} \sum_j \mathbb{E}\left(f_j \mathbb{E}\left((b_{t_{j+1}} - b_{t_j})|\mathcal{F}_{t_j}\right)|\mathcal{F}_t\right),
\end{aligned}$$

and this vanishes because $\mathbb{E}\left((b_{t_{j+1}} - b_{t_j})|\mathcal{F}_{t_j}\right) = 0$. □

We now discuss a multi-dimensional version of Itô's integral which will suffice for our applications. We first recall from Remark A.7 that if $b_t = (b_t^{(1)}, \cdots b^{(n)})$ is Brownian motion in $\mathbb{R}^n$ then $\{b_t^{(i)}\}_{i=1}^n$ is an independent family of BMs on $\mathbb{R}$ (and conversely). We will use this to define the integral

$$\int_S^T v\, db_t = \int_S^T \begin{pmatrix} v_{11} & \cdots & v_{1n} \\ \vdots & \ddots & \vdots \\ v_{m1} & \cdots & v_{mn} \end{pmatrix} \begin{pmatrix} db_t^{(1)} \\ \vdots \\ db_t^{(n)} \end{pmatrix} \tag{A.6}$$

as a process in $\mathbb{R}^n$ for a suitable $v_{ij} = v_{ij}(t,\omega)$.

Definition A.29. Let $\mathcal{A}_{\mathcal{H}}(S,T)$ be the collection of functions $f : [0,+\infty) \times \Omega \to \mathbb{R}$ such that

(1) f is $\mathcal{B} \times \mathcal{F}$-measurable;
(2) there exists a filtration $\mathcal{H}_t \subset \mathcal{F}$ such that:
   - b_t is a martingale with respect to $\mathcal{H}_t$ (b_t is BM in $\mathbb{R}$);
   - $f(t,\cdot)$ is adapted to $\mathcal{H}_t$, $t > 0$.
(3) $\mathbb{E}(\int_S^T f(t,\omega)^2 dt) < +\infty$.

Since $\mathcal{F}_t \subset \mathcal{H}_t$ and $\mathbb{E}(b_s - b_t|\mathcal{H}_t) = 0$ we can proceed as before and define

$$\int_S^T f db_t, \quad f \in \mathcal{A}_{\mathcal{H}}(S,T),$$

which is a martingale (see Proposition A.28). Coming back to $b = (b^{(1)}, \cdots, b^{(n)}) \in \mathbb{R}^n$, let $\mathcal{F}_t^{(n)}$ be the σ-algebra generated by $b_{s_1}^{(1)}, \cdots, b_{s_n}^{(n)}$, where $s_k \leq t$, $k = 1, \cdots, n$. Using the componentwise independence mentioned above, we see that $t < s$ implies that $b_s^{(k)} - b_t^{(k)} \perp \mathcal{F}_t^{(n)}$, so that by a previous argument each $b_t^{(k)}$ is a martingale with respect to $\mathcal{F}_t^{(n)}$. This allows us to define integrals like

$$\int_S^T f(t, b_t^{(1)}, \cdots, b_t^{(n)}) db_t^{(k)}, \quad f \in \mathcal{A}_{\mathcal{F}^{(n)}}(S,T), \quad k = 1, \cdots, n.$$

Thus, if we set $\mathcal{A}^{m,n}_{\mathcal{F}^{(n)}}(S,T)$ to be the collection of all $\{v_{ij}\}_{i=1,\cdots,m;j=1,\cdots,n}$ such that $v_{ij} \in \mathcal{A}_{\mathcal{F}^{(n)}}(S,T)$ then the multi-dimensional Itô's integral in (A.6) above is well-defined and has the expected properties (in particular, it is a martingale).

The Itô's integral considered above turns out to be the main ingredient in defining an important class of stochastic processes which, as we shall see, are quite amenable to formal manipulations resembling those available from the ordinary calculus.

Definition A.30. Let b_t be BM in $\mathbb{R}$ (with probability space $(\Omega, \mathcal{F}, P)$). Then an *Itô process* (or *diffusion*) is a stochastic process in $(\Omega, \mathcal{F}, P)$ of the type

$$X_t = X_0 + \int_0^t u(s,\omega) ds + \int_0^t v(s,\omega) db_s, \quad v \in \mathcal{A}_{\mathcal{H}}(0,t), \tag{A.7}$$

for some $\mathcal{H}$ as in Definition A.29.

Formally, we can rewrite (A.7) as

$$dX_t = u dt + v db_t, \tag{A.8}$$

where u is the *drift coefficient* and v is the *diffusion coefficient*. Itô's formula in (A.10) below shows that reasonable functions of Itô's processes are Itô's processes as well. It provides the correct change of variables formula in the setting of Stochastic Calculus.

Proposition A.31. *If X_t is an Itô's process as in (A.8) and $g = g(t,\omega) \in C^{2,1}([0,+\infty) \times \mathbb{R})$ then $Y_t(t,\omega) = g(t, X_t(\omega))$ is an Itô's process as well. More precisely,*

$$dY_t(t, X_t) = \frac{\partial g}{\partial t}(t, X_t) dt + \frac{\partial g}{\partial x}(t, X_t) dX_t + \frac{1}{2} \frac{\partial^2 g}{\partial x^2}(t, X_t) dX_t^2, \tag{A.9}$$

where in handling the quadratic term dX_t^2 we should use the multiplication table

	dt	db_t
dt	0	0
db_t	0	dt

As a consequence,

$$dY_t = \left(\frac{\partial g}{\partial t} + u \frac{\partial g}{\partial x} + \frac{1}{2} v^2 \frac{\partial^2 g}{\partial x^2} \right) dt + v \frac{\partial g}{\partial x} db_t. \tag{A.10}$$

Proof. [Oks13, Theorem 4.1.2]. □

We now discuss the multi-dimensional version of this result. Let $b_t = (b_t^{(1)}, \cdots, b_t^{(n)})$ be BM in $\mathbb{R}^n$, $b_0 = 0$. We can consider a multi-dimensional Itô's process

$$dX_t = udt + vdb_t,$$

where $X = (X_1, \cdots, X_n)^\top \in \mathbb{R}^n$, $u = (u_1, \cdots, u_n)^\top \in \mathbb{R}^n$, $u_i \in \mathcal{A}_{\mathcal{F}^{(n)}}$, and $v \in \mathcal{A}^{n,m}_{\mathcal{F}^{(n)}}$. Thus,

$$dX_{ti} = u_i dt + \sum_{j=1}^{m} v_{ij} db_t^{(j)}, \quad i = 1, \cdots, n.$$

Proposition A.32. *If X is as above and $Y(t,\omega) = g(t, X_t(\omega))$, where $g : [0, +\infty) \times \mathbb{R}^n \to \mathbb{R}^p$, then*

$$dY_k = \frac{\partial g_k}{\partial t} dt + \sum_{i=1}^{n} \frac{\partial g_k}{\partial x_i} dX_i + \frac{1}{2} \sum_{i,j=1}^{n} \frac{\partial^2 g_k}{\partial x_i \partial x_j} dX_i dX_j,$$

where in handling the quadratic terms $dX_i dX_j$ we should use the multiplication table

	dt	$db_t^{(i)}$
dt	0	0
$db_t^{(j)}$	0	$\delta_{ij} dt$

As a consequence,

$$dY_k = \left(\frac{\partial g_k}{\partial t} + \sum_{i=1}^{n} \frac{\partial g_k}{\partial x_i} u_i + \frac{1}{2} \sum_{i,j=1}^{n} \underbrace{\sum_{l=1}^{m} v_{il} v_{jl}}_{=(v^* v)_{ij}} \frac{\partial^2 g_k}{\partial x_i \partial x_j} \right) dt + \sum_{i=1}^{n} \sum_{l=1}^{m} \frac{\partial g_k}{\partial x_i} v_{il} db^{(l)}. \tag{A.11}$$

Proof. [Oks13, Theorem 4.2.1]. □

Example A.33. (The geometric Brownian) We assume that $Y_t(t,\omega) = g(t, b_t(\omega))$ in (A.9), so that (A.10) becomes

$$dY_t = \left(\frac{\partial g}{\partial t} + \frac{1}{2} \frac{\partial^2 g}{\partial x^2} \right) dt + \frac{\partial g}{\partial x} db_t. \tag{A.12}$$

If $Y_t = e^{\sigma b_t}$, $\sigma \neq 0$, we then get

$$dY_t = \frac{\sigma^2}{2} Y_t dt + \sigma Y_t db_t, \tag{A.13}$$

which shows that, in the stochastic setting, the exponential fails to satisfy the self-reproducing property under differentiation. We may cancel out the annoying first term in the right-hand side above by considering the *geometric Brownian process*

$$Z_t = e^{(\mu - \frac{\sigma^2}{2})t + \sigma b_t}, \quad \mu \in \mathbb{R}, \tag{A.14}$$

which satisfies

$$dZ_t = \mu Z_t dt + \sigma Z_t db_t, \tag{A.15}$$

or equivalently,

$$\frac{dZ_t}{Z_t} = \mu dt + \sigma db_t. \tag{A.16}$$

As we shall see in Subsection A.6, this kind of process plays a central role in the Black-Scholes strategy in Finance. □

Example A.34. (Diffusion processes) We discuss here one of the most notable motivation behind Itô's construction of his integral. Roughly, this accomplishment allowed the proper interpretation of solutions of a large class of stochastic differential equations, which in particular led to a pathwise approach to diffusion processes. We start by recalling that a *transition function* is a map $\mathfrak{R} : [0, +\infty) \times \mathbb{R}^n \times \mathcal{B}_n \to [0, 1]$ such that:

- $x \mapsto \mathfrak{R}(t, x, B)$ is measurable;
- $B \mapsto \mathfrak{R}(t, x, B)$ is a probability measure on $\mathbb{R}^n$;
- $\mathfrak{R}(0, x, \cdot) = \delta_x$;
- The *Chapman-Kolmogorov equation* holds:

$$\mathfrak{R}(t + s, x, B) = \int_{\mathbb{R}^n} \mathfrak{R}(t, x, dy)\mathfrak{R}(s, y, B). \tag{A.17}$$

An application of Theorem A.3 guarantees the existence of a measurable space $(\Omega, \mathcal{F})$ and, for each $x \in \mathbb{R}^n$, a probability measure $\mathbb{P}^x$ on $(\Omega, \mathcal{F})$ and a stochastic process $X_t^x : (\Omega, \mathcal{F}, \mathbb{P}^x) \to \mathbb{R}^n$ such that $\mathfrak{R}(t, x, B) = \mathbb{P}^x(X_t^x \in B)$. Equivalently, $\mathfrak{R}(t, x, \cdot) = X_t^x{}_\sharp \mathbb{P}^x$. In particular, $X_0^x{}_\sharp \mathbb{P}^x = \delta_x$. Moreover, the following *Markov property* holds:

$$\mathbb{E}^x(f(X_{t+s}^x)|\mathcal{F}_s^X) = \mathbb{E}^x(f(X_t^x)), \quad \mathbb{P}^x \text{ a.s.}, \tag{A.18}$$

for any $x \in \mathbb{R}^n$ and any f as above. Intuitively, this means X_t^x is memoryless. Attached to any $\mathfrak{R}$ as above is the associated semigroup $t \mapsto \mathfrak{B}_t$ given by

$$(\mathfrak{P}_t f)(x) = \int_{\mathbb{R}^n} f(y)\mathfrak{R}(t, x, dy),$$

for f a bounded, measurable function on $\mathbb{R}^n$. Notice that $(\mathfrak{P}_t f)(x) = \mathbb{E}^x f(X_t^x)$. Now, any *Markov process* X_t^x as above has an *infinitesimal generator* L, which is a linear operator defined on the space of all functions f such that

$$\lim_{t\to 0} \frac{\mathfrak{P}_t f - f}{t}$$

exists. We then define, for any such f,

$$(Lf)(x) = \lim_{t\to 0} \frac{(\mathfrak{P}_t f)(x) - f(x)}{t}. \tag{A.19}$$

Under certain regularity assumptions, the generator is a *diffusion operator*, that is,

$$(Lf)(x) = \frac{1}{2}\sum_{ij} a_{ij}(x)\frac{\partial^2 f}{\partial x_i \partial x_j} + \sum_i u_i(x)\frac{\partial f}{\partial x_i}, \quad x \in \mathbb{R}^n, \tag{A.20}$$

where a is a symmetric, non-negative matrix. The problem now is how to recover X_t^x starting from L (or, more precisely, from the coefficients a and u defining it). It turns out that Itô's calculus may be used to solve this problem as follows. Write $a = v^\top v$ and form the stochastic differential equation

$$dX_t = u(X_t)dt + v(X_t)db^t.$$

Under mild conditions on the coefficients, it is shown that this equation has a unique solution X_t^x with $X_0^x = x$. This of course means that

$$X_t^x = x + \int_0^t u(X_s^x)ds + \int_0^t v(X_s^x)db^s,$$

so a notion of stochastic integral is required here in order to properly interpret the last term above. It is now immediate to check that X_t^x solves the problem in the sense that (A.17), (A.18) and (A.19) are satisfied if we set $\mathfrak{R}(t,x,B) = P(X_t^x \in B)$, where P is Wiener measure. Here we only check that (A.19) holds. From Itô's formula (A.11), for any $f : \mathbb{R}^n \to \mathbb{R}$ we have

$$f(X_t^x) = f(x) + \int_0^t (Lf)(X_s^x)ds + M_t^f,$$

where M_t^f is a martingale. By taking expectation we see that

$$(\mathfrak{P}_t f)(x) = \mathbb{E}^x(f(X_t^x)) = f(x) + \mathbb{E}^x\left(\int_0^t (Lf)(X_s^x)ds\right),$$

and (A.19) follows. In fact, this procedure of solving suitable stochastic differential equations yields a systematic way of associating a diffusion process X_t^x to any operator L as above. Indeed, the resulting process is completely characterized by the fact that for any f the martingale

$$M_t^{f,x} = f(X_t^x) - f(x) - \int_0^t (Lf)(X_s^x)ds$$

has quadratic variation given by

$$\langle M^{f,x}\rangle_t^{(2)} = \int_0^t \left(\sum_{ij} a_{ij}\frac{\partial f}{\partial x^i}\frac{\partial f}{\partial x^j}\right)(X_s^x)ds.$$

We then say that X_t^x is the *diffusion process* driven by L. To see that we are in the right track, let us take $L = \frac{1}{2}\Delta$ and let us set $\mathsf{X}^i = M^{x_i,x}$ for simplicity. It follows that

$$\begin{aligned}\langle \mathsf{X}^i, \mathsf{X}^j\rangle_t^{(2)} &:= \frac{1}{2}\left(\langle \mathsf{X}^i + \mathsf{X}^j\rangle_t^{(2)} - \langle \mathsf{X}^i\rangle_t^{(2)} - \langle \mathsf{X}^j\rangle_t^{(2)}\right)\\ &= \delta_{ij}t,\end{aligned}$$

so a celebrated result due to P. Lévy [KS12, Theorem 3.16] implies that $X_t = b_t$. Thus, as expected, BM is the diffusion process driven by the Laplacian $\frac{1}{2}\Delta$. Finally, we indicate how the diffusion process can be directly defined in terms of the coefficients defining L. It suffices to take $f(x) = x^i$ and observe that

$$N_t^i := X_t^i - x_i - \int_0^t u_i(X_s)ds \tag{A.21}$$

is a martingale with

$$\langle N^i, N^j\rangle_t^{(2)} = \int_0^t a_{ij}(X_s)ds.$$

In particular, if $L = \sum_i u_i\partial_i$ is a vector field then $\langle N^i, N^j\rangle_t^{(2)} = 0$ and hence $N_t = 0$. Thus, (A.21) says that L integrates to a determinstc flow (no fluctuation!). In this way we recover the classical result on integration of vector fields. In a nutshell, whereas standard calculus allows us to integrate vector fields (giving rise to deterministic dynamical systems), Itô's calculus allows us to integrate diffusion operators (giving rise to random dynamical systems). For a modern take on the theory of diffusion processes we refer to [Bau14] □

With the basics of Itô's calculus at hand, we present in the rest of this Appendix some of its most glamorous applications.

A.4. **The Gaussian concentration inequality (again).** We start by providing here an elegant proof of the optimal version of the Gaussian concentration inequality (5.18) which is attributed to B. Maurey in [Pis06, Chapter 2] and relies on the full power of Itô's Stochastic Calculus developed in the previous section.

For $0 \leq t \leq 1$ we consider the "reversed" heat semigroup $P_t = e^{\frac{1}{2}(1-t)\Delta}$, so that for any (smooth and Lipschitz) $F : \mathbb{R}^k \to \mathbb{R}$ there holds

$$\frac{\partial}{\partial t}(P_t F) + \frac{1}{2}\Delta(P_t F) = 0.$$

We now use Itô's formula (A.11) with $Y(t,\cdot) = (P_t F)(b_t)$, where b_t is a standard BM in $\mathbb{R}^k$ (recall that $b_t - b_{t'} \sim \mathcal{N}(\vec{0}, (t-t')\mathrm{Id}_k)$, $t' < t$). Since $u = 0$ and $v_{ij} = \delta_{ij}$ it simplifies to

$$dY_t = \langle (\nabla P_t F)(b_t), db_t \rangle,$$

and integrating this from $t = 0$ to $t = 1$,

$$\begin{aligned} F(b_1) &= (e^{\frac{1}{2}\Delta}F)(0) + \int_0^1 \langle (\nabla P_t F)(b_t), db_t \rangle \\ &= \mathbb{E}(F(b_1)) + \int_0^1 \langle (\nabla P_t F)(b_t), db_t \rangle, \end{aligned}$$

where we used Propositions A.18 and A.28 in the last step. We may now adapt the Cramér-Chernoff method in Section 5 to this setting: for $\tau > 0$ and $w \geq 0$,

$$P(|F(b_1) - \mathbb{E}(F(b_1))| > \tau) \leq 2e^{-w\tau}\mathbb{E}\left(e^{w\int_0^1 \langle (\nabla P_t F)(b_t), db_t\rangle}\right), \tag{A.22}$$

so it remains to estimate the expectation in the right-hand side.

The key observation at this point is that $\mathrm{Lip}(P_t F) = \mathrm{Lip}(F)$ and hence $|\nabla P_t F)| \leq \mathrm{Lip}(F)$ a.s. Now let $\pi = \{t_0 = 0 < t_1 < \cdots < t_n = 1\}$ be a partition of the interval $[0,1]$ with $|\pi| = \max_l |t_l - t_{l-1}|$ its width. As $|\pi| \to 0$ the Itô's integral within the expectation, by its very definition, may be arbitrarily approximated (say, in probability) by S_n, where

$$S_j = \sum_{l=1}^{j} \langle V_l, b_{t_l} - b_{t_{l-1}} \rangle, \quad 1 \leq j \leq n,$$

$V_l = (\nabla P_{t_{l-1}} F)(b_{t_{l-1}})$ is $\mathcal{F}_{t_{l-1}}$-measurable (where $\{\mathcal{F}_t\}$ is the filtration associated to b_t) and satisfies $|V_l| \leq \mathrm{Lip}(F)$. We have

$$S_j = S_{j-1} + \langle V_j, b_{t_j} - b_{t_{j-1}} \rangle,$$

a decomposition into independent factors, and since the inner product follows the normal $\mathcal{N}(0, |V_j|^2(t_j - t_{j-1}))$ by Proposition (4.8) (3), assuming of course that $V_j \neq \vec{0}$, we obtain

$$\begin{aligned} \mathbb{E}(e^{wS_j}) &= \mathbb{E}\left(e^{wS_{j-1}}\right)\mathbb{E}\left(e^{w\langle V_j, b_{t_j} - b_{t_{j-1}}\rangle}\right) \\ &\overset{(4.6)}{=} \mathbb{E}\left(e^{wS_{j-1}}\right) e^{\frac{1}{2}w^2|V_j|^2(t_j - t_{j-1})} \\ &\leq \mathbb{E}\left(e^{wS_{j-1}}\right) e^{\frac{1}{2}w^2\mathrm{Lip}(f)^2(t_j - t_{j-1})}. \end{aligned}$$

Note that this obviously remains true if $V_j = \vec{0}$. In any case, if we iterate this starting with $j = n$ we get

$$\mathbb{E}(e^{wS_n}) \leq e^{\frac{1}{2}w^2\mathrm{Lip}(f)^2},$$

where the right-hand side, remarkably enough, does not depend on π. Passing the limit as $|\pi| \to 0$ on the left-hand side we thus obtain

$$\mathbb{E}\left(e^{w\int_0^1 \langle (\nabla P_t F)(b_t), db_t\rangle}\right) \leq e^{\frac{1}{2}w^2\mathrm{Lip}(f)^2},$$

which is the same as saying that

$$\int_0^1 \langle (\nabla P_t F)(b_t), db_t \rangle \in \mathsf{SubG}(\mathrm{Lip}(f)). \tag{A.23}$$

Leading this to (A.22) we get

$$P(|F(b_1) - \mathbb{E}(F(b_1))| > \tau) \leq 2e^{\frac{1}{2}w^2 \mathrm{Lip}(F)^2 - w\tau},$$

so if we minimize the right-hand side over $w \geq 0$ we find that

$$P(|F(b_1) - \mathbb{E}(F(b_1))| > \tau) \leq 2e^{-\frac{\tau^2}{2\mathrm{Lip}(F)^2}}, \quad \tau > 0,$$

which is equivalent to (5.18) with the optimal constant $C = 1/2$ because the normal random vector appearing there has been chosen so that $X \sim \mathcal{N}(\vec{0}, \mathrm{Id}_k) \sim b_1$.

A.5. **The Feynman-Kac formula and the path integral representation of the heat kernel.** Let us consider

$$g(t, X_t) = e^{-\int_0^t V(X_t)dt} w(T-t, X_t), \quad 0 \leq t \leq T,$$

where $V = V(x)$, $x \in \mathbb{R}^n$, is a (well-behaved) potential function and we assume that X_t is an Itô's diffusion as in (A.8):

$$dX_t = u(X_t)dt + v(X_t)db_t,$$

where b_t is BM in $\mathbb{R}^n$. We compute that

$$\frac{\partial g}{\partial t} = -e^{-\int_0^t V(X_t)dt} \left(V(X_t)w(T-t, X_t) + \frac{\partial w}{\partial t}(T-t, X_t), \right),$$

$$\frac{\partial g}{\partial x_i} = e^{-\int_0^t V(X_t)dt} \frac{\partial w}{\partial x_i}(T-t, X_t),$$

and

$$\frac{\partial^2 g}{\partial x_i \partial x_j} = e^{-\int_0^t V(X_t)dt} \frac{\partial^2 w}{\partial x_i \partial x_j}(T-t, X_t),$$

so that Itô's formula in (A.11) applies to give

$$\begin{aligned} dg &= e^{-\int_0^t V(X_t)dt} \left(-\frac{\partial w}{\partial t}(T-t, X_t) + \sum_i u_i \frac{\partial w}{\partial x_i}(T-t, X_t) - \right. \\ &\quad \left. -V(X_t)w(T-t, X_t) + \frac{1}{2}\mathcal{L}' w(T-t, X_t) \right) dt \\ &\quad + e^{-\int_0^t V(X_t)dt} \sum_i \frac{\partial g}{\partial x_i} db_t^{(i)}, \end{aligned}$$

where

$$\mathcal{L}' w = \sum_{ij} (v^* v)_{ij} \frac{\partial^2 w}{\partial x_i \partial x_j}.$$

Putting all the pieces of this computation together we obtain a remarkable stochastic (or path integral) representation of solutions of certain heat-type equations.

Proposition A.35. *(Feynman-Kac formula I) If $w = w(t, x)$ satisfies the heat-type equation*

$$\begin{cases} \frac{\partial w}{\partial t} &= \frac{1}{2}\mathcal{L}' w + \langle u, \nabla w \rangle - Vw \\ w(0, x) &= f(x) \end{cases}$$

then the following holds:

$$w(t,x_0)=\mathbb{E}_{x_0}\left(e^{-\int_0^t V(X_t)dt}f(X_t)\right), \tag{A.24}$$

where $\mathbb{E}_{x_0}$ refers to the law P_{x_0} of BM in $\mathbb{R}^n$ starting at x_0.

Proof. From the computation above,

$$dg=e^{-\int_0^t V(X_t)dt}\sum_i\frac{\partial g}{\partial x_i}db_t^{(i)},$$

which shows that the process $g(t,X_t)$ is a martingale (with respect to $\mathcal{F}^{(n)}$); see Proposition A.28. Since

$$\mathbb{E}_{x_0}\left(g(t,X_t)\right)|_{t=0}=\mathbb{E}_{x_0}\left(w(T,X_0)\right)=w(T,x_0),$$

and

$$\mathbb{E}_{x_0}\left(g(t,X_t)\right)|_{t=T}=\mathbb{E}_{x_0}\left(e^{-\int_0^t V(X_s)ds}w(0,X_T)\right)=\mathbb{E}_{x_0}\left(e^{-\int_0^t V(X_s)ds}f(X_T)\right),$$

the result follows in view of Proposition A.18 and the fact that T is arbitrary. □

Corollary A.36. *(Exponential control) Under the conditions above, if $|f|\leq M$ and $V\geq c$, $c\in\mathbb{R}$, then*

$$|u(t,x_0)|\leq Me^{-ct}.$$

An important special case of Proposition A.24 occurs when $u=0$ and $v_{ij}=\delta_{ij}$, so that

$$\mathcal{L}'=\frac{1}{2}\Delta,$$

where Δ is the Laplacian. We then see that any solution of

$$\left\{\begin{array}{rcl}\frac{\partial w}{\partial t}&=&\frac{1}{2}\Delta w-Vw\\ w(0,x)&=&f(x)\end{array}\right.$$

satisfies

$$w(t,x)=\mathbb{E}_x\left(e^{-\int_0^t V(b_\tau)d\tau}f(b_t)\right). \tag{A.25}$$

On the other hand, we know from Analysis [Paz12] that this can be rewritten as

$$w(t,x)=(e^{t\mathcal{L}}f)(x)=\int_{\mathbb{R}^n}K_{\mathcal{L}}(t;x,y)f(y)dy, \tag{A.26}$$

where $e^{t\mathcal{L}}$ is the heat semigroup generated by $\mathcal{L}=\frac{1}{2}\Delta-V$ and $K_{\mathcal{L}}$ is the associated *heat kernel*, i.e. $K_{\mathcal{L}}$ satisfies

$$\left\{\begin{array}{rcl}\frac{\partial K_{\mathcal{L}}}{\partial t}&=&\frac{1}{2}\Delta K_{\mathcal{L}}-VK_{\mathcal{L}}\\ K_{\mathcal{L}}(0;x,y)&=&\delta(x-y)\end{array}\right.$$

This suggests the existence of a stochastic representation for $K_{\mathcal{L}}$, thus pointing toward a version of a Feynman-Kac formula working at the more fundamental level of heat kernels.

To find this representation we fix $t>0$ and consider the process $\{B_s\}_{0\leq s<t}$ satisfying

$$dB_s=db_s-\frac{B_s-y}{t-s}ds,\quad B_0=x,$$

or equivalently,

$$B_s=x+b_s-\int_0^s\frac{B_\tau-y}{t-\tau}d\tau.$$

We will now show that the law of this process can be computed in terms of $K_{\frac{1}{2}\Delta}$, the heat kernel of the Laplacian $\frac{1}{2}\Delta$, and P_x, the law of BM starting at x.

We first note that the discussion above gives

$$(e^{\frac{1}{2}\tau\Delta}f)(x) = \int_{\mathbb{R}^n} K_{\frac{1}{2}\Delta}(\tau; x, y)f(y)dy = \mathbb{E}_x\left(f(b_\tau)\right), \quad \tau \geq 0. \tag{A.27}$$

In particular,

$$\int_{\mathbb{R}^n} K_{\frac{1}{2}\Delta}(t; x, y)dy = 1, \tag{A.28}$$

which also follows from Proposition 4.2 because, as is well-known,

$$K_{\frac{1}{2}\Delta}(t; x, y) = (2\pi t)^{-n/2}e^{-|x-y|^2/2t},$$

From this we see that

$$\nabla_x \ln K_{\frac{1}{2}\Delta}(t; x, y) = -\frac{x-y}{t},$$

and hence

$$dB_s = db_s + \nabla_x \ln K_{\frac{1}{2}\Delta}(t-s; B_s, y)ds, \quad s < t. \tag{A.29}$$

Thus, the Brownian bridge B_s is just the Brownian motion b_s with an added drift involving the logarithmic derivative of $K_{\frac{1}{2}\Delta}$. We note however that the drift is singular at $s = t$. Fortunately, careful first order estimates of $K_{\frac{1}{2}\Delta}$ [Hsu02, Section 5.5] allow us to bypass this difficulty and confirm not only that this is well defined for $s = t$ but also that $B_s \to y$ as $s \to t$. Thus, we call $\{B_s\}_{0\leq s\leq t}$ the *Brownian bridge* connecting x to y with lifetime t.

We should think of B_s as a process on the *bridge space* $C_{t;x,y} \subset C_x$ of all Brownian paths starting at x and conditioned to hit y at time t; cf. Remark A.15. To find the law B_s we first note that

$$\frac{\partial}{\partial t}\ln K_{\frac{1}{2}\Delta} = \frac{1}{2}\Delta \ln K_{\frac{1}{2}\Delta} + \frac{1}{2}\|\nabla \ln K_{\frac{1}{2}\Delta}\|^2,$$

so if we apply Itô's formula to

$$E_s = \ln \frac{K_{\frac{1}{2}\Delta}(t-s; B_s, y)}{K_{\frac{1}{2}\Delta}(t; x, y)}$$

we find that

$$dE_s = \langle F_s, dB_s\rangle - \frac{1}{2}\|F_s\|^2 ds,$$

where $F_s = \nabla_x K_{\frac{1}{2}\Delta}(t-s; B_s, y)$. Now define a measure Q in $C_{t:x,y}$ by

$$\frac{dQ}{dP_x}|_{\mathcal{F}_s} = \exp\left(\int_0^s \langle F_u, db_u\rangle - \frac{1}{2}\int_0^s \|F_u\|^2 du\right).$$

By Girsanov's theorem [Oks13, Theorem 8.6.4], under Q the process

$$B_s - \int_0^s F_u du$$

is a BM. Thus, we see that $P_{t;x,y} := Q$ is the law of the Brownian bridge B_s and there holds

$$\frac{dP_{t;x,y}}{dP_x}|_{\mathcal{F}_s} = \frac{K_{\frac{1}{2}\Delta}(t-s; b_s, y)}{K_{\frac{1}{2}\Delta}(t; x, y)}. \tag{A.30}$$

With these preliminaries at hand, finally we will be able to provide a path integral representation for $K_{\mathcal{L}}$.

Proposition A.37. *(Feynman-Kac formula II) One has*

$$K_{\mathcal{L}}(t;x,y) = K_{\frac{1}{2}\Delta}(t;x,y)\mathbb{E}_{t;x,y}\left(e^{-\int_0^t V(B_\tau)d\tau}\right).$$

In other words, if we define the conditional Wiener measure *on $C_{t;x,y}$ by*

$$\mu_{t;x,y} = K_{\frac{1}{2}\Delta}(t;x,y)P_{t;x,y} \tag{A.31}$$

then

$$K_{\mathcal{L}}(t;x,y) = \int_{C_{t;x,y}} e^{-\int_0^t V(B_\tau)d\tau} d\mu_{t;x,y}. \tag{A.32}$$

Proof. First we have from (A.27) with $\tau = 0$ that

$$e^{-\int_0^t V(b_\tau)d\tau} f(b_t) = \int_{\mathbb{R}^n} K_{\frac{1}{2}\Delta}(0;b_t,y)e^{-\int_0^t V(b_\tau)d\tau} f(y)dy,$$

and taking expectation we get

$$\begin{aligned}\mathbb{E}_x\left(e^{-\int_0^t V(b_\tau)d\tau} f(b_t)\right) &= \int_{\mathbb{R}^n} K_{\frac{1}{2}\Delta}(t;x,y)\mathbb{E}_x\left(\frac{K_{\frac{1}{2}\Delta}(0;b_t,y)}{K_{\frac{1}{2}\Delta}(t;x,y)}e^{-\int_0^t V(b_\tau)d\tau} f(y)\right)dy \\ &\overset{(A.30)\text{ with } s=t}{=} \int_{\mathbb{R}^n} K_{\frac{1}{2}\Delta}(t;x,y)\mathbb{E}_{t;x,y}\left(e^{-\int_0^t V(B_\tau)d\tau} f(B_t)\right)dy.\end{aligned}$$

On the other hand, we know from (A.25) and (A.26) that

$$\mathbb{E}_x\left(e^{-\int_0^t V(b_\tau)d\tau} f(b_t)\right) = \int_{\mathbb{R}^n} K_{\mathcal{L}}(t;x,y)f(y)dy.$$

Since $B_t = y$ and f is arbitrary, the result follows. □

Remark A.38. (The Laplacian heat kernel as a transition probability) It follows from the formalism above that the conditioned Wiener measure $\mu_{t;x,y}$ in (A.31) may be characterized by

$$\int_{\mathbb{R}^n}\left(\int_{C_{t;x,y}} F(\omega)d\mu_{t;x,y}(\omega)\right)dy = \int_{C_x} F(\omega)dP_x(\omega),$$

where F varies over the set all bounded functions on (C_x, P_x), the Wiener space starting at x; cf.Remark A.15. By taking $F \equiv 1$ we thus see that

$$\int_{\mathbb{R}^n}\left(\int_{C_{t;x,y}} d\mu_{t;x,y}(\omega)\right)dy = 1,$$

which is just a restatement of (A.28), as it follows either from (A.31) or from (A.32) with $V \equiv 1$ that

$$K_{\frac{1}{2}\Delta}(t;x,y) = \int_{C_{t;x,y}} d\mu_{t;x,y}(\omega),$$

the total measure of the Brownian bridge $C_{t;x,y}$ endowed with $\mu_{t;x,y}$. It then follows that:

- for each $t \geq 0$ and $x \in \mathbb{R}^n$ the function
$$y \mapsto K_{\frac{1}{2}\Delta}(t;x,y) = \mu_{t;x,y}(C_{t;x,y})$$
defines a probability density in $\mathbb{R}^n$;
- For each $U \in \mathcal{B}^n$ the quantity
$$P_{t;x}(U) := \int_U K_{\frac{1}{2}\Delta}(t;x,y)dy$$

may be interpreted as the probability that a Brownian path passes through U when $s = t$ given that it has started at x when $s = 0$. By shrinking U to $\{y\}$ we thus conclude that $K_{\frac{1}{2}\Delta}(t; x, y)$ may be viewed as the *transition probability* that a Brownian path hits y at $s = t$ given that it has started at x when $s = 0$.

Remark A.39. (Weyl's law and its modern incarnations in Index Theory) If $e^{t\mathcal{L}}$ is of trace class, its trace can be computed by integrating (A.32) along the diagonal $x = y$ of $\mathbb{R}^n \times \mathbb{R}^n$:

$$\operatorname{Tr} e^{t\mathcal{L}} = \int_{\mathbb{R}^n} \left(\int_{C_{t;x,x}} e^{-\int_0^t V(X_\tau)\, d\tau}\, d\mu_{t;x,x} \right) dx.$$

Defining a measure $d\mu_t$ on the space $C_t = \cup_{x\in\mathbb{R}^n} C_{t;x,x}$ of all *Brownian loops* in $\mathbb{R}^n$ with lifetime t by setting $d\mu_t = d\mu_{t;x,x}\, dx$, we obtain

$$\operatorname{Tr} e^{t\mathcal{L}} = \int_{C_t} e^{-\int_0^t V(X_\tau)\, d\tau}\, d\mu_t.$$

This formula continues to hold when $\mathbb{R}^n$ is replaced by a compact Riemannian manifold (M, g), with Δ_g denoting the Laplace–Beltrami operator. In this case,

$$\operatorname{Tr} e^{\frac{1}{2}t\Delta_g} = \int_{C_t} d\mu_t,$$

where the measure μ_t is defined locally in coordinate charts and then assembled in the usual way; see [Hsu02] for a detailed account of Brownian motion on Riemannian manifolds and its fundamental properties. As $t \to 0$, a typical Brownian loop in C_t contracts to its base point while remaining within a geodesic ball whose radius also vanishes with t [Hsu02, Lemma 7.7]. Consequently, the path integral on the right-hand side becomes localized around $M \subset C_t$. Coupled with the "principle of not feeling the curvature," which asserts that $K_{\frac{1}{2}\Delta_g}(t; x, x) \sim (2\pi t)^{-n/2}$ as $t \to 0$, we obtain

$$\operatorname{Tr} e^{\frac{1}{2}t\Delta_g} \sim (2\pi t)^{-n/2} \operatorname{vol}(M, g).$$

Since $\operatorname{Tr} e^{\frac{1}{2}t\Delta_g} = \sum_i e^{-\frac{1}{2}\lambda_i t}$, where $\{\lambda_i\}$ are the positive eigenvalues of Δ_g, this yields Weyl's celebrated result: the volume of (M, g) can be recovered from the asymptotic behavior of its spectrum. A more sophisticated version of this argument, involving the short-time asymptotics of the heat kernel associated with a supersymmetric Dirac operator on spinors, leads to a probabilistic proof of the Atiyah–Singer index theorem [Bis84, Hsu02]. Further developments along these lines, relying on refined Feynman–Kac representations of the heat kernel for certain Hodge Laplacians acting on sections of geometric vector bundles over Riemannian manifolds (possibly noncompact and with boundary), can be found in [dL17a, dL17b, dL20] and references therein. □

A.6. **The Black-Scholes strategy in Finance.** Here we derive the celebrated Black-Scholes option pricing formula[57]. From the outset, this involves a risky asset S_t, a *stock*, evolving in time according to a geometric Brownian as in Example A.33:

$$\frac{dS_t}{S_t} = \mu dt + \sigma db_t. \tag{A.33}$$

Here, $\mu > 0$ is the *mean rate of return* and $\sigma > 0$ is the *volatility*. Recall that

$$S_t = S_0 e^{(\mu - \frac{\sigma^2}{2})t + \sigma b_t}$$

[57] As it is well-knwon, this has been worth a Nobel Prize in 1997.

provides the explicit solution of (A.33). In particular,

$$\ln S_t = \ln S_0 + \left(\mu - \frac{\sigma^2}{2}\right)t + \sigma b_t \sim \mathcal{N}\left(\underbrace{\ln S_0 + \left(\mu - \frac{\sigma^2}{2}\right)t}_{=:m}, \underbrace{\sigma^2 t}_{=:\nu^2}\right),$$

so that, by Example 4.10,

$$S_t = e^{\ln S_t} \sim \mathcal{LN}\left(S_0 e^{\mu t}, S_0^2 e^{2\mu t}(e^{\sigma^2 t} - 1)\right) = \Lambda(m, \nu^2). \tag{A.34}$$

On the other hand, we have an investor's *portfolio* (A_t, B_t) whose value is

$$V_t = A_t S_t + B_t \gamma_t, \tag{A.35}$$

with the risk-less *bond* γ_t satisfying $d\gamma_t = r\gamma_t dt$, where $r > 0$ is the associated *interest rate*. The *option pricing problem* addressed by Black-Scholes consists in adjusting the trading strategy (A_t, B_t) to the underlying asset S_t by (deterministically!) finding a function u such that

$$V_t = u(t, S_t), \quad 0 \leq t \leq T, \tag{A.36}$$

where $T > 0$ is the *expiration time* for the option. The key point here is that $u = u(t, x)$ should satisfy a certain PDE. In order to find it, we start with (A.33) and apply Itô's formula to (A.36) to check that

$$dV_t = \left(\frac{\partial u}{\partial t} + \mu S_t \frac{\partial u}{\partial x} + \frac{\sigma^2 S_t^2}{2}\frac{\partial^2 u}{\partial x^2}\right)dt + \sigma S_t \frac{\partial u}{\partial x} db_t. \tag{A.37}$$

On the other hand, if we assume that our portfolio is *self-financing* in the sense that

$$dV_t = A_t dS_t + B_t d\gamma_t,$$

we get

$$dV_t = (\mu A_t S_t + r B_t \gamma_t)\, dt + \sigma A_t S_t db_t. \tag{A.38}$$

By comparing the diffusion and drift coefficients in the expressions for dV_t above we get

$$A_t = \frac{\partial u}{\partial x} \tag{A.39}$$

and hence

$$\frac{\partial u}{\partial t} + \frac{\sigma^2 S_t^2}{2}\frac{\partial^2 u}{\partial x^2} = r B_t \gamma_t. \tag{A.40}$$

Now note that from (A.35), (A.36) and (A.39),

$$B_t \gamma_t = V_t - A_t S_t = u - S_t \frac{\partial u}{\partial x}, \tag{A.41}$$

so if we replace this in the right-hand side of (A.40) and make $S_t = x$ we conclude that u must satisfy the *Black-Scholes equation*

$$\frac{\partial u}{\partial t} + \frac{\sigma^2 x^2}{2}\frac{\partial^2 u}{\partial x^2} + rx\frac{\partial u}{\partial x} - ru = 0. \tag{A.42}$$

Notice that the coefficients of this PDE depend on σ and r but not on μ. Also, as written the PDE fails to be of heat type because the coefficients of $\partial u/\partial t$ and $\partial^2 u/\partial x^2$ have the same sign. This suggests that we should try to solve it by imposing the "terminal condition"

$$u(T, S_T) = V_T. \tag{A.43}$$

In fact, the choice

$$V_T = \max\{0, S_T - K\}, \quad K > 0, \tag{A.44}$$

corresponds to the investor holding at time $t = 0$ the option (but not the obligation) of buying the stock by a fixed price K at the expiration time T. Hence, in this *European call*, if $S_T > K$ then the owner of the option will obtain the payoff $S_T - K$ whereas if $S_T \leq K$ the owner will not exercise his option, thus obtaining a null payoff.

In terms of the cumulative normal distribution

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-y^2/2} dy, \quad x \in \mathbb{R},$$

the Black-Scholes equation (A.42) with $u(T, S_T) = \max\{0, S_T - K\}$ may be explicitly solved as

$$u(t, x) = x\Phi(g(t, x)) - Ke^{-r(T-t)}\Phi(h(t, x)), \tag{A.45}$$

where

$$g(t, x) = \frac{\ln(x/K) + \left(r + \frac{1}{2}\sigma^2\right)(T - t)}{\sigma\sqrt{T - t}}$$

and

$$h(t, x) = g(t, x) - \sigma\sqrt{T - t} = \frac{\ln(x/K) + \left(r - \frac{1}{2}\sigma^2\right)(T - t)}{\sigma\sqrt{T - t}}.$$

Notice that, as expected, the solution depends on σ and r but not on μ. We conclude that the *Black-Scholes option pricing formula*

$$V_0 = u(0, S_0) = S_0\Phi(g(0, S_0)) - Ke^{-rT}\Phi(h(0, S_0))$$

provides the rational price to hold at the initial time $t = 0$ a European call option with price K. Also, from (A.39) and (A.41) we see that the corresponding self-financing strategy is

$$(A_t, B_t) = \left(\frac{\partial u}{\partial x}, \gamma_0^{-1} e^{-rt}\left(u - S_t \frac{\partial u}{\partial x}\right)\right).$$

The explicit solution (A.45) to (A.42)-(A.44) may be obtained in many ways. For instance, a "deterministic" approach may be pursued upon successive changes of variables so as to transform (A.42) into the standard heat equation, which can then be explicitly solved by the usual methods [KK12, Section 10.3]. Alternatively, we may appeal to the full power of the Stochastic Calculus as follows. Let us set $\theta = (\mu - r)/\sigma$ and consider the process

$$M_t = e^{-\theta b_t - \frac{1}{2}\theta^2 t}, \quad 0 \leq t \leq T.$$

From Itô's formula we obtain

$$dM_t = -\theta M_t db_t,$$

so M_t is a b_t-martingale. Hence, by Proposition A.18,

$$\mathbb{E}^P(M_T) = \mathbb{E}^P(M_0) = \mathbb{E}^P(1) = 1,$$

where P is Wiener measure. Thus, if we define a new measure $P^\bullet$ on Wiener space by requiring that $dP^\bullet = M_T dP$, it is immediate that $P^\bullet$ is a probability measure. A version of Girsanov's theorem [Oks13, Theorem 8.6.4] applies here and we conclude that $b_t^\bullet := b_t + \theta t$ is a standard Brownian with respect to $P^\bullet$ (so that $b_t^\bullet \sim_{P^\bullet} \mathcal{N}(0, t)$) and, moreover,

$$dS_t = rS_t dt + \sigma S_t db_t^\bullet. \tag{A.46}$$

Thus, we have been able to modify the drift of the dynamics of the stock (from μS_t to rS_t) at the cost of changing the underlying measure (from P to $P^\bullet$) and the driving Brownian (from b_t to $b_t^\bullet$). The reason for doing this is now obvious: in terms of the infinitesimal generator of (A.46), namely,

$$L = \frac{\sigma^2 x^2}{2}\frac{\partial^2}{\partial x^2} + rx\frac{\partial}{\partial x},$$

(A.42) may be rewritten as

$$\frac{\partial u}{\partial t} + Lu - ru = 0,$$

whose solution may be obtained by the method leading to the Feynman-Kac formula discussed in Section A.5 (with r playing the role of a constant potential). Indeed, if we apply Itô's formula to $v(t, S_t) := e^{r(T-t)}u(t, S_t)$, we easily see that

$$dv(t, S_t) = \sigma S_t e^{r(T-t)}\frac{\partial u}{\partial x}(t, S_t)db_t^\bullet,$$

which means that $v(t, S_t)$ is a $b_t^\bullet$-martingale. Thus, if we calculate the (identical!) expectations at the endpoints of the interval $[t, T]$ and use (A.44) we end up with

$$u(t, S_t) = e^{-r(T-t)}\mathbb{E}^{P^\bullet}(\max\{S_T - K, 0\}).$$

We now observe that, due to (A.46) and similarly to (A.34), we now have

$$S_T \sim_{P^\bullet} \mathcal{LN}\left(S_t e^{r(T-t)}, S_t^2 e^{2r(T-t)}(e^{\sigma^2(T-t)} - 1)\right), \tag{A.47}$$

or equivalently,

$$\ln S_T \sim_{P^\bullet} \mathcal{N}\left(\underbrace{\ln S_t + \left(\mu - \frac{\sigma^2}{2}\right)(T-t)}_{=:m}, \underbrace{\sigma^2(T-t)}_{=:\nu^2}\right) = \Lambda(m, \nu^2). \tag{A.48}$$

Now,

$$\begin{aligned}\mathbb{E}^{P^\bullet}(\max\{S_T - K, 0\}) &= \int_K^{+\infty}(S_T - K)dF_{S_T} \\ &= \int_K^{+\infty} S_T dF_{S_T} - K\int_K^{+\infty} dF_{S_T},\end{aligned}$$

where $dF_{S_T} = \psi_{S_T} dx$ is the cdf of S_T, and these integrals may be computed in terms of Φ by means of the recipe in Example 4.10. For instance,

$$\begin{aligned}
\int_K^{+\infty} dF_{S_T} &= 1 - \int_{-\infty}^{K} dF_{S_T} \\
&= 1 - F_{S_T}(K) \\
&= 1 - \Phi\left(\frac{\overbrace{\ln K - \ln S_t - \left(r - \frac{\sigma^2}{2}\right)(T-t)}^{=-m}}{\underbrace{\sigma\sqrt{T-t}}_{=\nu}} \right) \\
&= 1 - \Phi\left(-\frac{\ln S_t/K + \left(r - \frac{\sigma^2}{2}\right)(T-t)}{\sigma\sqrt{T-t}} \right) \\
&= \Phi\left(\frac{\ln S_t/K + \left(r - \frac{\sigma^2}{2}\right)(T-t)}{\sigma\sqrt{T-t}} \right) \\
&= \Phi(h(t, S_t)),
\end{aligned}$$

where we used (4.18), our choices for m and ν as in (A.48) and the fact that $\Phi(x) + \Phi(-x) = 1$. Similarly,

$$\begin{aligned}
\int_K^{+\infty} S_T dF_{S_T} &= \mathbb{E}^{P^\bullet}(S_T) - \int_{-\infty}^{K} S_T dF_{S_T} \\
&= S_t e^{r(T-t)} - \int_{-\infty}^{K} x \psi_{S_T}(x) dx \\
&= S_t e^{r(T-t)} - \frac{1}{\sqrt{2\pi}\sigma\sqrt{T-t}} \int_{-\infty}^{K} e^{-\frac{1}{2}\left(\frac{\ln x - \ln S_t - \left(r - \frac{\sigma^2}{2}\right)(T-t)}{\sigma\sqrt{T-t}} \right)^2} dx,
\end{aligned}$$

where we used (A.47) and (4.17). After an appropriate change of variables and using again that $\Phi(x)+\Phi(-x) = 1$, we get

$$\begin{aligned}
\int_K^{+\infty} S_T dF_{S_T} &= e^{r(T-t)} S_t \Phi\left(\frac{-\ln K + \ln S_t + \left(r - \frac{\sigma^2}{2}\right)(T-t) + \sigma^2(T-t)}{\sigma\sqrt{T-t}} \right) \\
&= e^{r(T-t)} S_t \Phi\left(g(t, X_t) \right).
\end{aligned}$$

Putting all the pieces of this computation together we find that

$$u(t, S_t) = S_t \Phi(g(t, S_t)) - K e^{-r(T-t)} \Phi(h(t, S_t)),$$

which matches (A.45) if we make $S_t = x$.

REFERENCES

[AB69] John Aitchison and James A C Brown. *The lognormal distribution, with special reference to its uses in Economics*. Cambridge University Press, 1969.

[Agr15] Alan Agresti. *Foundations of linear and generalized linear models*. John Wiley & Sons, 2015.

[Aka73] Hirotugu Akaike. Information theory and an extension of the maximum likelihood principle. In B. N. Petrov and F. Csáki, editors, *Second International Symposium on Information Theory*, pages 267–281. Akadémiai Kiadó, Budapest, 1973.

[Ame85] Takeshi Amemiya. *Advanced econometrics*. Harvard university press, 1985.

[And03] Theodore W Anderson. *An introduction to multivariate statistical analysis*. Wiley & Sons, 2003.

[AP09] Joshua D Angrist and Jörn-Steffen Pischke. *Mostly harmless econometrics: An empiricist's companion*. Princeton University Press, 2009.

[AS16] Noga Alon and Joel H Spencer. *The probabilistic method*. John Wiley & Sons, 2016.

[BA13] Kenneth P Burnham and David R Anderson. *Model selection and inference: a practical information-theoretic approach*. Springer Science & Business Media, 2013.

[Bau13] Johannes Bausch. On the efficient calculation of a linear combination of chi-square random variables with an application in counting string vacua. *Journal of Physics A: Mathematical and Theoretical*, 46(50):505202, 2013.

[Bau14] Fabrice Baudoin. *Diffusion processes and stochastic calculus*. EMS, 2014.

[BC11] Alexandre Belloni and Victor Chernozhukov. *High dimensional sparse econometric models: An introduction*. Springer, 2011.

[Ber04] Richard A Berk. *Regression analysis: A constructive critique*, volume 11. Sage, 2004.

[BHK20] Avrim Blum, John Hopcroft, and Ravindran Kannan. *Foundations of data science*. Cambridge University Press, 2020.

[Bis84] Jean-Michel Bismut. The Atiyah—Singer theorems: a probabilistic approach. I. the index theorem. *Journal of functional analysis*, 57(1):56–99, 1984.

[BK04] Frank B Baker and Seock-Ho Kim. *Item Response Theory: Parameter Estimation Techniques*. CRC Press, 2nd edition, 2004.

[BLM13] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence,(2013)*. OUP: Oxford, 2013.

[Bor75] Christer Borell. The Brunn-Minkowski inequality in Gauss space. *Inventiones Mathematicae*, 30(2):207–216, 1975.

[Bul03] Michael G Bulmer. *Francis Galton: pioneer of heredity and biometry*. JHU Press, 2003.

[BVDG11] Peter Bühlmann and Sara Van De Geer. *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media, 2011.

[CA06] Kai L Chung and Farid AitSahlia. *Elementary probability theory: with stochastic processes and an introduction to mathematical finance*. Springer Science & Business Media, 2006.

[Car86] Edward Carlstein. Simultaneous confidence regions for predictions. *The American Statistician*, 40(4):277–279, 1986.

[Cas08] George Casella. *Statistical design*. Springer, 2008.

[CB21] George Casella and Roger L Berger. *Statistical inference*. Cengage Learning, 2021.

[Che21] Louis H Chen. Stein's method of normal approximation: Some recollections and reflections. *The Annals of Statistics*, 49(4):1850–1863, 2021.

[CLC21] Denis Chetverikov, Zhipeng Liao, and Victor Chernozhukov. On cross-validated lasso in high dimensions. *The Annals of Statistics*, 49(3):1300–1317, 2021.

[Cow72] Ruth S Cowan. Francis galton's statistical ideas: the influence of eugenics. *Isis*, 63(4):509–528, 1972.

[CU14] Ovidiu Calin and Constantin Udrişte. *Geometric modeling in probability and statistics*. Springer, 2014.

[CW09] Ralph D Cook and Sanford Weisberg. *Applied regression including computing and graphics*. John Wiley & Sons, 2009.

[DA13] Rafael J De Ayala. *The theory and practice of item response theory*. Guilford Publications, 2013.

[DB18] Annette J Dobson and Adrian G Barnett. *An introduction to generalized linear models*. Chapman and Hall/CRC, 2018.

[DBW04] Paul De Boeck and Mark Wilson. *Explanatory Item Response Models: A Generalized Linear and Nonlinear Approach*. Springer, New York, 2004.

[DF87] Persi Diaconis and David Freedman. A dozen de Finetti-style results in search of a theory. *Annales de l'IHP Probabilités et Statistiques*, 23(S2):397–423, 1987.

[dL17a] Levi L de Lima. A Feynman–Kac formula for differential forms on manifolds with boundary and geometric applications. *Pacific Journal of Mathematics*, 292(1):177–201, 2017.

[dL17b] Levi L de Lima. A probabilistic proof of the Gauss-Bonnet formula for manifolds with boundary. *arXiv:1709.03772*, 2017.

[dL20] Levi L de Lima. Heat conservation for generalized Dirac Laplacians on manifolds with boundary. *Annali di Matematica Pura ed Applicata (1923-)*, 199(3):997–1021, 2020.

[DM88] Edward J Dudewicz and Satya Mishra. *Modern mathematical statistics*. John Wiley & Sons, Inc., 1988.

[Don00] David L Donoho. High-dimensional data analysis: The curses and blessings of dimensionality. *AMS math challenges lecture*, 1(2000):32, 2000.

[DS14] Morris H DeGroot and Mark J Schervish. *Probability and statistics*, volume 563. Pearson Education London, UK., 2014.

[EM77] Bradley Efron and Carl Morris. Stein's paradox in statistics. *Scientific American*, 236(5):119–127, 1977.

[Far06] Julian J Faraway. *Linear models with R*. Chapman and Hall/CRC, 2006.

[FBG+16] Arnoldo Frigessi, Peter Bühlmann, Ingrid K Glad, Sylvia Richardson, and Marina Vannucci. Some themes in high-dimensional statistics. In *Statistical Analysis for High-Dimensional Data: The Abel Symposium 2014*, pages 1–13. Springer, 2016.

[Fer17] Thomas S Ferguson. *A course in large sample theory*. Routledge, 2017.

[FG13] Bert E Fristedt and Lawrence F Gray. *A modern approach to probability theory*. Springer Science & Business Media, 2013.

[Fis15] Ronald A Fisher. Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. *Biometrika*, 10(4):507–521, 1915.

[Fis21] Ronald A Fisher. On the "probable error" of a coeficient of correlation deduced from a small sample. *Metron*, pages 3–32, 1921.

[Fis22] Ronald A Fisher. On the mathematical foundations of theoretical statistics. *Philosophical transactions of the Royal Society of London. Series A, containing papers of a mathematical or physical character*, 222(594-604):309–368, 1922.

[Fis25] Ronald A Fisher. Applications of "student's" distribution. *Metron*, pages 90–104, 1925.

[Fis11] Hans Fischer. *A history of the central limit theorem: from classical to modern probability theory*, volume 4. Springer, 2011.

[FK16] Alan Frieze and Michal Karonski. *Introduction to random graphs*. Cambridge University Press, 2016.

[FP15] Juliana C Ferreira and Cecilia M Patino. What does the p value really mean? *Jornal Brasileiro de Pneumologia*, 41(5):485, 2015.

[Gib21] Eric W Gibson. The role of p-values in judging the strength of evidence and realistic replication expectations. *Statistics in Biopharmaceutical Research*, 13(1):6–18, 2021.

[GKPS99] Mikhael Gromov, Misha Katz, Pierre Pansu, and Stephen Semmes. *Metric structures for Riemannian and non-Riemannian spaces*, volume 152. Springer, 1999.

[Gne18] Boris V Gnedenko. *Theory of probability*. CRC Press, 2018.

[Gor16] Prakash Gorroochurn. *Classic topics on the history of modern mathematical statistics: From Laplace to more recent times*. John Wiley & Sons, 2016.

[Gra03] Alfred Gray. *Tubes*, volume 221. Springer Science & Business Media, 2003.

[GTP19] Jeff Gill, Michelle Torres, and Silvia M T Pacheco. *Generalized linear models: a unified approach*, volume 134. Sage Publications, 2019.

[Gut06] Allan Gut. *Probability: a graduate course*. Springer, 2006.

[Hay11] Fumio Hayashi. *Econometrics*. Princeton University Press, 2011.

[HB03] Raymond Hubbard and María Jesús Bayarri. Confusion over measures of evidence (p's) versus errors (α's) in classical statistical testing. *The American Statistician*, 57(3):171–178, 2003.

[HCP22] Serim Hong, Carlos A Coelho, and Junyong Park. An exact and near-exact distribution approach to the Behrens–Fisher problem. *Mathematics*, 10(16):2953, 2022.

[Hil73] Victor Hilts. Statistics and social science. *Foundations of scientific method: the nineteenth century*, pages 206–233, 1973.

[HL51] Joseph L Hodges and Erich L Lehmann. Some applications of the Cramér-Rao inequality. In *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, volume 2, pages 13–23. University of California Press, 1951.

[HN64] Takeyuki Hida and Hisao Nomoto. Gaussian measure on the projective limit space of spheres. *Proceedings of the Japan Academy*, 40(5):301–304, 1964.

[Hot39] Harold Hotelling. Tubes and spheres in n-spaces, and a class of statistical problems. *American Journal of Mathematics*, 61(2):440–460, 1939.

[HSR91] Ronald K Hambleton, Hariharan Swaminathan, and H J Rogers. *Fundamentals of item response theory*, volume 2. Sage, 1991.

[Hsu02] Elton P Hsu. *Stochastic analysis on manifolds*. American Mathematical Soc., 2002.

[HTF09] Trevor Hastie, Robert Tibshirani, and Jerome H Friedman. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer, 2009.

[HTW15] Trevor Hastie, Robert Tibshirani, and Martin Wainwright. *Statistical learning with sparsity: the lasso and generalizations*. CRC press, 2015.

[JLR11] Svante Janson, Tomasz Luczak, and Andrzej Rucinski. *Random graphs*. John Wiley & Sons, 2011.

[JN21] Jiming Jiang and Thuan M Nguyen. *Linear and Generalized Linear Mixed Models and Their Applications*. Springer, Cham, 2021.

[Jor97] Bent Jorgensen. *The theory of dispersion models*. CRC Press, 1997.

[JS61] William James and Charles Stein. Estimation with quadratic loss. In *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 361–379. University of California Press, 1961.

[JWHT13] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An introduction to statistical learning*, volume 112. Springer, 2013.

[KC98] Seock-Ho Kim and Allan S Cohen. On the Behrens-Fisher problem: A review. *Journal of Educational and Behavioral Statistics*, 23(4):356–377, 1998.

[Ken46] Maurice G Kendall. *The advanced theory of statistics.* Charles Griffin and Co., 1946.

[Kit04] Charles Kittel. *Elementary statistical physics*. Courier Corporation, 2004.

[KK08] Sadanori Konishi and Genshiro Kitagawa. *Information criteria and statistical modeling*. Springer, 2008.

[KK12] Gopinath Kallianpur and Rajeeva L Karandikar. *Introduction to option pricing theory*. Springer Science & Business Media, 2012.

[KL51] Solomon Kullback and Richard A. Leibler. On information and sufficiency. *Annals of Mathematical Statistics*, 22(1):79–86, 1951.

[Kle13] Achim Klenke. *Probability theory: a comprehensive course*. Springer Science & Business Media, 2013.

[Kol18] Andrei N Kolmogorov. *Foundations of the theory of probability: Second English Edition*. Courier Dover Publications, 2018.

[Kre11] Ulrich Krengel. *Ergodic theorems*, volume 6. Walter de Gruyter, 2011.

[KS12] Ioannis Karatzas and Steven Shreve. *Brownian motion and stochastic calculus*, volume 113. Springer Science & Business Media, 2012.

[Lab07] Mohammed L Labbi. On Gauss-Bonnet curvatures. *SIGMA. Symmetry, Integrability and Geometry: Methods and Applications*, 3:118, 2007.

[Lap98] Pierre-Simon Laplace. *Pierre-Simon Laplace philosophical essay on probabilities*, volume 13. Springer Science & Business Media, 1998.

[LC06] Erich L Lehmann and George Casella. *Theory of point estimation*. Springer Science & Business Media, 2006.

[Led01] Michel Ledoux. *The concentration of measure phenomenon*, volume 89. American Mathematical Soc., 2001.

[Led06] Michel Ledoux. Isoperimetry and gaussian analysis. *Lectures on Probability Theory and Statistics: Ecole d'Eté de Probabilités de Saint-Flour XXIV—1994*, pages 165–294, 2006.

[Led22] Johannes Lederer. *Fundamentals of high-dimensional statistics*. Springer, 2022.

[Leh99] Erich L Lehmann. *Elements of large-sample theory*. Springer, 1999.

[LG13] Jean-Francois Le Gall. *Mouvement brownien, martingales et calcul stochastique*, volume 71. Springer, 2013.

[Liu10] Wei Liu. *Simultaneous inference in regression*. CRC Press, 2010.

[LR05] Erich L Lehmann and Joseph P Romano. *Testing statistical hypotheses*. Springer, 2005.

[LT13] Michel Ledoux and Michel Talagrand. *Probability in Banach spaces: isoperimetry and processes*. Springer Science & Business Media, 2013.

[Luk42] Eugene Lukacs. A characterization of the normal distribution. *The Annals of Mathematical Statistics*, 13(1):91–93, 1942.

[Luk70] Eugene Lukacs. *Characteristic functions*. Charles Griffin & Company, 1970.

[Mac81] Donald A MacKenzie. Statistics in Britain, 1865-1930: The social construction of scientific knowledge. *(No Title)*, 1981.

[Mas07] Pascal Massart. *Concentration inequalities and model selection: Ecole d'Eté de Probabilités de Saint-Flour XXXIII-2003*. Springer, 2007.

[McK73] Henry P McKean. Geometry of differential space. *The Annals of Probability*, 1(2):197–206, 1973.

[Mos85] Peter G Moschopoulos. The distribution of the sum of independent gamma random variables. *Annals of the Institute of Statistical Mathematics*, 37(1):541–544, 1985.

[Nai90] Daniel Q Naiman. Volumes of tubular neighborhoods of spherical polyhedra and statistical inference. *The Annals of Statistics*, pages 685–716, 1990.

[NM94] Whitney K Newey and Daniel McFadden. Large sample estimation and hypothesis testing. *Handbook of econometrics*, 4:2111–2245, 1994.

[Oks13] Bernt Oksendal. *Stochastic differential equations: an introduction with applications*. Springer Science & Business Media, 2013.

[ONe14] Ben ONeill. Some useful moment results in sampling problems. *The American Statistician*, 68(4):282–296, 2014.

[OP58] Ingram Olkin and John W Pratt. Unbiased estimation of certain correlation coefficients. *The Annals of Mathematical Statistics*, pages 201–211, 1958.

[Paz12] Amnon Pazy. *Semigroups of linear operators and applications to partial differential equations*, volume 44. Springer Science & Business Media, 2012.

[Pis06] Gilles Pisier. Probabilistic methods in the geometry of Banach spaces. In *Probability and Analysis: Lectures given at the 1st 1985 Session of the Centro Internazionale Matematico Estivo (CIME) held at Varenna (Como), Italy May 31–June 8, 1985*, pages 167–241. Springer, 2006.

[RC10] Christian P Robert and George Casella. *Introducing Monte Carlo methods with R*, volume 18. Springer, 2010.

[RF10] Halsey Royden and Patrick M Fitzpatrick. *Real analysis*. China Machine Press, 2010.

[Rob07] Christian P Robert. *The Bayesian choice: from decision-theoretic foundations to computational implementation*, volume 2. Springer, 2007.

[RP61] W D Ray and A E N T Pitman. An exact distribution of the Fisher–Behrens–Welch statistic for testing the difference between the means of two normal populations with unknown variances. *Journal of the Royal Statistical Society: Series B (Methodological)*, 23(2):377–384, 1961.

[RS08] Alvin C Rencher and Bruce Schaalje. *Linear models in statistics*. John Wiley & Sons, 2008.

[SA90] Seyed-Mahmoud Sadooghi-Alvandi. Simultaneous prediction intervals for regression models with an intercept. *Communications in Statistics-Theory and Methods*, 19(4):1433–1441, 1990.

[Saz81] Vjačeslav V Sazonov. *Normal approximation: some recent advances*, volume 879. Springer, 1981.

[Sch47] Henry Scheffé. A useful convergence theorem for probability distributions. *The Annals of Mathematical Statistics*, 18(3):434–438, 1947.

[Sha08] Jun Shao. *Mathematical statistics*. Springer Science & Business Media, 2008.

[Shi16] Takashi Shioya. *Metric measure geometry - Gromov's theory of convergence and concentration of metrics and measures*. European Mathematical Society, 2016.

[SIK86] Yosiyuki Sakamoto, Makio Ishiguro, and Genshiro Kitagawa. *Akaike information criterion statistics*. KTK Scientific Publishers, Tokyo, 1986.

[SL94] Jiayang Sun and Clive R Loader. Simultaneous confidence bands for linear regression and smoothing. *The Annals of Statistics*, pages 1328–1345, 1994.

[SL03] George A Seber and Alan J Lee. *Linear regression analysis*. John Wiley & Sons, 2003.

[SP14] René L Schilling and Lothar Partzsch. *Brownian motion: an introduction to stochastic processes*. Walter de Gruyter, 2014.

[ST78] Vladimir N Sudakov and Boris S Tsirel'son. Extremal properties of half-spaces for spherically invariant measures. *Journal of Soviet Mathematics*, 9(1):9–18, 1978.

[Sti90] Stephen M Stigler. *The history of statistics: The measurement of uncertainty before 1900*. Harvard University Press, 1990.

[Sti97] Stephen M Stigler. Regression towards the mean, historically considered. *Statistical methods in medical research*, 6(2):103–114, 1997.

[Sti05] Stephen Stigler. Fisher in 1921. *Statistical Science*, pages 32–49, 2005.
[Sto13] Jordan M Stoyanov. *Counterexamples in probability*. Dover Publications, Inc., 2013.
[Str13] Walter W Stroup. *Generalized Linear Mixed Models: Modern Concepts, Methods and Applications*. CRC Press, Boca Raton, FL, 2013.
[Stu08a] Student. Probable error of a correlation coefficient. *Biometrika*, pages 302–310, 1908.
[Stu08b] Student. The probrable error of a mean. *Biometrika*, pages 1–25, 1908.
[Tal96] Michel Talagrand. A new look at independence. *The Annals of probability*, pages 1–34, 1996.
[Tao11] Terence Tao. *An introduction to measure theory*, volume 126. American Mathematical Soc., 2011.
[Tao12] Terence Tao. *Topics in random matrix theory*, volume 132. American Mathematical Soc., 2012.
[The74] Chris M Theobald. Generalizations of mean square error applied to ridge regression. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 36(1):103–106, 1974.
[Tib96] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 58(1):267–288, 1996.
[Ton90] Yung L Tong. *The multivariate normal distribution*. Springer, 1990.
[VDH24] Remco Van Der Hofstad. *Random graphs and complex networks*. Cambridge university press, 2024.
[VdL16] Wim J Van der Linden. *Handbook of item response theory*, volume 1. CRC press New York, 2016.
[VdV00] Adrianus W Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge University Press, 2000.
[Ver18] Roman Vershynin. *High-dimensional probability: An introduction with applications in Data Science*, volume 47. Cambridge university press, 2018.
[Wai19] Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge university press, 2019.
[Wel96] Alan H Welsh. *Aspects of statistical inference*, volume 246. John Wiley & Sons, 1996.
[Wey39] Hermann Weyl. On the volume of tubes. *American Journal of Mathematics*, 61(2):461–472, 1939.
[Wil38] Samuel S Wilks. The large-sample distribution of the likelihood ratio for testing composite hypotheses. *The annals of mathematical statistics*, 9(1):60–62, 1938.
[Wil91] David Williams. *Probability with martingales*. Cambridge University Press, 1991.
[WL16] Ronald L Wasserstein and Nicole A Lazar. The ASA statement on *p*-values: context, process, and purpose. *The American Statistician*, 70(2):129–133, 2016.

UNIVERSIDADE FEDERAL DO CEARÁ, DEPARTAMENTO DE MATEMÁTICA, CAMPUS DO PICI, R. HUMBERTO MONTE, S/N, 60455-760, FORTALEZA/CE, BRAZIL (levi@mat.ufc.br).