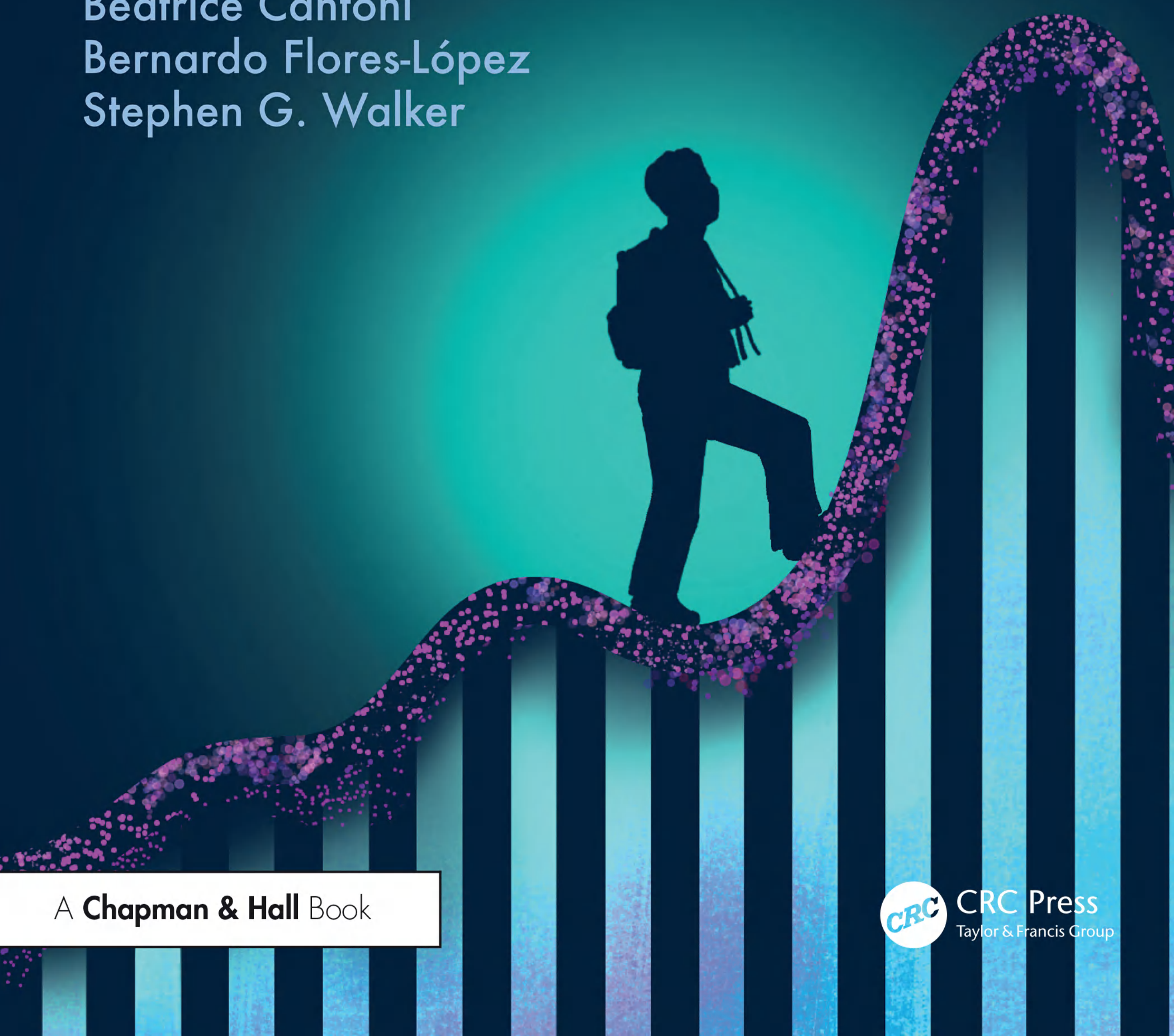
EXERCISES IN STATISTICAL REASONING

Michael R. Schwob
Yunshan Duan
Beatrice Cantoni
Bernardo Flores-López
Stephen G. Walker



A Chapman & Hall Book

CRC Press
Taylor & Francis Group

Exercises in Statistical Reasoning

Students cultivate learning techniques in school that emphasize procedural problem solving and rote memorization. This leads to efficient problem solving for familiar problems. However, conducting novel research is an exercise in creative problem solving that is at odds with a procedural approach; it requires thinking deeply about the topic and crafting solutions to unique problems. It is not easy to move from a topic-based, carefully curated curriculum to the daunting world of independent research, where solutions are unknown and may not even exist. In developing this book, we considered our experience as graduate students who faced this transition.

Exercises in Statistical Reasoning is a collection of exercises designed to strengthen creative problem-solving skills. The exercises are designed to encourage readers to understand the key points of a problem while seeking knowledge, rather than separating out these two activities. To complete the exercises, readers may need to reference the literature, which is how research-based knowledge is often acquired.

Features of the Exercises

- The exercises are self-contained, though several build upon concepts from previous problems.
- Each exercise opens with a brief introduction that emphasizes the relevance of the content. Then, the problem statement is presented as a series of intermediate questions.
- For each exercise, we suggest one possible solution, though many may exist.
- Following each solution, we discuss the historical background of the content and points of interest.
- For many exercises, a brief demonstration is provided illustrating relevant concepts.

There is an abundance of high-quality textbooks that cover a vast range of statistical topics. However, there is also a lack of texts that focus on the development of problem-solving techniques that are required for conducting novel statistical research. We believe that this book helps fill the gap. Any reader familiar with graduate-level classical and Bayesian statistics may use this book. The goal is to provide a resource that such students can use to ease their transition to conducting novel research.

Michael R. Schwob, Yunshan Duan, Beatrice Cantoni, and Bernardo Flores-López are Ph.D. candidates in the Department of Statistics and Data Sciences at The University of Texas at Austin. Michael will join the Department of Statistics at the Virginia Polytechnic Institute and State University as a faculty member in Fall 2025.

Stephen G. Walker is a professor in the Department of Mathematics and the Department of Statistics and Data Sciences at The University of Texas at Austin. He is the holder of the Paul D. and Betty Robertson Meek and American Petrofina Foundation Centennial Professorship in Business in the McCombs School of Business.

Exercises in Statistical Reasoning

Michael R. Schwob, Yunshan Duan,
Beatrice Cantoni, Bernardo Flores-López,
and Stephen G. Walker

CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an **informa** business
A CHAPMAN & HALL BOOK

First edition published 2025
by CRC Press
2385 Executive Center Drive, Suite 320, Boca Raton, FL 33431, U.S.A.

and by CRC Press
4 Park Square, Milton Park, Abingdon, Oxon, OX14 4RN

CRC Press is an imprint of Taylor & Francis Group, LLC

ISBN: 978-1-032-79710-6 (hbk)
ISBN: 978-1-032-78486-1 (pbk)
ISBN: 978-1-003-49347-1 (ebk)

DOI: 10.1201/9781003493471

Typeset in URW Palladio font
by KnowledgeWorks Global Ltd.

Publisher's note: This book has been prepared from camera-ready copy provided by the authors.

Dedication

Michael R. Schwob: "To my parents, whose love and support have fueled my pursuit of knowledge and happiness."

Yunshan Duan: "致我的家人，感谢你们始终如一的支持与鼓励。"

Beatrice Cantoni: "A Enrico e Marialuisa, anime curiose."

Bernardo Flores-López: "Para Migue; que nunca nadie te recuerde ausente."

Stephen G. Walker: "Ne chiege kod nyithinda mabeyo."

Contents

Preface

Students cultivate learning techniques in school that emphasize procedural problem solving and rote memorization. This leads to efficient problem solving for familiar problems. However, procedural problem solving is troublesome when one is confronted with an unfamiliar problem. Thus, when graduate students transition to conducting novel research, many are unequipped with the tools for autodidactic study. Research is an exercise in creative problem solving, which is at odds with a procedural approach; it requires thinking deeply about the topic and crafting solutions to unique problems. It is not easy to move from a topic-based, carefully curated curriculum to the daunting world of independent research, where solutions are unknown and may not even exist. In developing this book, we considered our experience as graduate students facing this transition.

This book is a collection of statistical exercises that train creative and holistic problem solving. These exercises appeared on statistics preliminary exams that demarcated the transition from coursework to research activity. They were designed to encourage exploratory thought and discourage rote mechanics and memorization. Exercises start with a seemingly trivial concept, which then develops into a significant topic in statistics. The path to some solutions may seem circuitous, while other paths may be elegantly simple. Each exercise is a practice of research methodology and encourages the reader to explore the true meaning and intuition behind many well-known statistical concepts.

Expected Audience

The expected readers are graduate or advanced undergraduate students that are interested in developing a research-oriented mindset. However, anyone with the appropriate prerequisites may benefit from this practice in creative problem solving. An understanding of graduate-level classical and Bayesian statistics is assumed. More specifically, we assume an intermediate understanding of the following topics: the exponential family, hypothesis testing, likelihood estimation, linear models, mathematical statistics, probability theory, Markov processes, and Markov chain Monte Carlo algorithms.

Structure of This Book

We grouped the exercises into eight chapters, which are named after particular statistical topics. Though the ordering of the chapters may suggest a linear reading order, each chapter is self-contained, so the reader may explore the topics in whichever order they wish. The chapters concerning probability, hypothesis testing, and asymptotics are divided into two sections: the classical perspective and the Bayesian perspective. We structured these chapters this way to explicitly compare the two frameworks.

The exercises are self-contained, though several build upon concepts from previous problems. Exercises that are particularly difficult have an asterisk next to their names. Many exercises bring several topics together to make progress on a single research problem. Hence, being widely read is important; understanding ideas beyond any specific research agenda will always be useful.

Each exercise opens with a brief introduction that emphasizes the relevance of the content. Then, the problem statement is presented as a series of intermediate questions; solutions to these questions typically depend on the previous steps. For each exercise, we suggest one possible solution, though many may exist. These solutions can be found at the end of each section. However, the reader needs to understand concepts and ideas in their own way to ensure that they will feel comfortable adapting to different problems that they have not yet seen. This is, in fact, the nature of research. Following each solution, we discuss the historical background of the content and points of interest. For many exercises, we provide a brief demonstration that illustrates relevant concepts. Finally, there are "Miscellaneous" points for several exercises that provide details on mathematical or numerical techniques used to complete the exercises.

The exercises in this book aim to check for an understanding of foundational statistical topics. Our approach does not heavily emphasize a theorem-proof structure nor does it focus on detailed regularity conditions upon which some technical result strictly holds. We do not argue over whether a function needs to be twice or thrice differentiable or debate the strict conditions under which a series expansion can be suitably stopped. Rather, this book explores how statistical concepts work, revisiting even the most basic ideas that may have been overlooked or omitted in previous coursework.

How to Use This Book

We recommend that readers mindfully work through this book, drafting solutions and deliberately noting any successes and obstacles. We encourage readers to attempt each exercise on their own before reviewing the proposed solution. The aim of this book is to teach problem-solving skills – not content. We do not present all the material needed to answer the questions. To complete an exercise, readers may need to reference the literature, which is how research-based knowledge is often acquired. In fact, the exercises are designed to encourage readers to understand the key points of a problem while seeking knowledge, rather than separating out these two activities. Finally, we highly encourage readers to look through the further reading for exercises of particular interest.

Why We Wrote This Book

We collaborated on this book because we shared a singular vision – that actively solving problems unlocks a far deeper understanding of statistical concepts than simply memorizing theorems, proofs, and definitions. The exercises in this book originally appeared as a series of preliminary exam questions for PhD students. The unique nature of the questions prompted us (the student authors) to compile a summary of solutions, along with points of interest containing further readings and research questions. The exam setter (Prof. Stephen Walker) was then brought in on the project to ensure that the solutions and questions were checked for full correctness. Hence, the original solutions were the students', the original questions were the professors', and all the rest a joint effort. To complete the book, we added brief backgrounds to each exercise and added figures to illustrate particular concepts of interest. Aside from correcting errors, the questions and answers remain as they were originally constructed.

There is an abundance of high-quality textbooks that cover a vast range of statistical topics. However, there is also a lack of texts that focus on the development of problem-solving techniques that are required for conducting novel statistical research. We believe that this book helps fill the gap.

A Perspective from Prof. Walker

Graduate students often learn advanced material and engage in research by participating in a structured master's or PhD program. The term "program" suggests a hint of homogeneity – that all students can learn in the same way. This assumption is likely not the case. In most graduate programs, the first year (possibly more) involves a regular schedule of courses, homework, and exams. If homework and exams shortly follow after a concept or idea is introduced, it is quite reasonable to ask, "where is the time to understand going to come from?" An alternative program design might check knowledge after students have had sufficient time to understand the material. From a practical experience, I attended an undergraduate program in which eight exams (three hours each) were administered over the final four days of the program, after 3 years. How was it possible to memorize all the material for these exams? The only way to deal with things was to understand the material. Some concepts took days to master, while others took weeks or even months. Given three years, there was more than enough time to deeply understand the content, and the most successful approach to this was to tackle relevant problems and seek help when getting stuck after sufficient time figuring it out. The idea of reading course materials may not be enough to fully understand or be able to move on. Hence, the book is designed to assist with the understanding through problem solving. This would involve self study and creating the time to do problems, not unlike my undergraduate and graduate programs.

A Perspective from the Students

The origins of this book can be traced back to Spring 2022, when we began to prepare for our preliminary exam. We initially struggled with the questions, finding them difficult to prepare for because they demanded a deeper level of understanding than the "reading, memorizing, and replicating" cycle that is prevalent in higher education. Success on the exam hinged on our ability to adapt to unfamiliar questions and apply our knowledge to new scenarios. Thus, we needed to move beyond memorization and cultivate the problem-solving skills necessary to tackle statistical problems that were new to us.

To prepare for the exam, we deeply engaged with questions from past exams every week. We intentionally approached the questions with a slow pace to foster a deeper understanding of the content and explore related concepts. After several weeks, all of us recognized the value of the questions and our more profound understanding of the material. After a couple of months, our problem-solving skills felt sharper than ever before.

Education often prioritizes breadth of knowledge over depth, leaving students with a superficial understanding of many topics. As a result, the learned content is quickly forgotten. The traditional learning-through-memorization paradigm indeed has a major flaw: its efficacy significantly decreases with time. An approach that aims to build problem-solving skills and a deeper understanding of content ages much better. The critical thinking and problem-solving skills that we strengthened have become essential tools in our broader academic and professional pursuits.

Acknowledgments

We are grateful to our professors and peers for their impacts on our academic journeys. In particular, we would like to thank Nhat Ho, Mevin Hooten, Peter Müller, Mary Parker, Abhra Sarkar, Purna Sarkar, James Scott, Cory Zigler, Tianqi Chen, Amber Day, Rosalía Hernández, Emily Hsiao, Khai Nguyen, and Jack Xiao. We also thank Angelica Faye Alcantara Mayor for designing the cover image and Denise Mayang Wong for designing the overall cover of our book.

1

Probability

Statisticians use an extensive amount of probability theory, both for constructing models and for finding properties of statistical procedures. For example, probability motivates the use of the sample mean for estimating a population mean; establishing properties of such estimators is done so by treating the sample as random variables. Many frequentist properties are based on the large sample behavior of an estimator. Thus, the convergence of random variables becomes important to statistical inference. In nonparametric problems, it is the convergence of random functions that is the focus of attention.

In Bayesian analysis, probability underpins the prior and posterior distributions, which are fundamental for Bayesian inference. Recently, many Bayesian models have been estimated using sampling-based techniques, such as Markov chain Monte Carlo samplers. As a consequence, much work on the convergence properties of Markov chains appears in the Bayesian literature. New sampling ideas include the use of diffusion processes, such as Langevin diffusion, which may be defined to have a specific stationary distribution. Other samplers include Hamiltonian Monte Carlo or particle filters, which are suitable for hidden Markov models.

Additionally, Bayesian nonparametrics is highly dependent on a strong knowledge of probability. Probability measures are constructed on spaces of functions, such as for density and distribution functions. For example, the Dirichlet process relies on a suitable definition of finite dimensional distributions but can also be understood directly via the use of stochastic processes behaving as a distribution function with probability one. Then, probability serves an essential role for deriving the correct posterior distribution and establishing conditions for consistency.

The questions in this chapter focus on the use of probability for statistics and stochastic processes. Special attention is given to the study of stochastic processes, which is often where statisticians meet probability for which they may not be so familiar.

Question 1.1.1 investigates the central limit theorem and its convergence in probability, which are useful for looking at sample means. Question 1.1.2 considers infinitely divisible random variables, which can be represented as the sum of n independent and identically distributed random variables for all n; such variables include the gamma, Gaussian, and Poisson. Finally, Question 1.1.3 is on the maximal inequality for a martingale sequence, which could be seen as an extension of the well-known Markov inequality.

The questions on stochastic processes mostly concern Markov processes. For example, Question 1.2.1 is on the simple random walk, whereby a walk on the integers goes up one or down one with equal probability. A primary interest is how long the process will take to return to the starting point while also reaching a given height. Question 1.2.2 includes a simple birth process, where any individual alive at a given generation gives birth to an independent set of a random number of offspring according to a probability mass function. The sum of these offspring determines the size of the next generation, and so on. The question of interest is the probability that the population goes extinct.

Two questions are concerned with Markov chains in discrete time, which are represented as transition matrices. Question 1.2.3 explains fundamental properties for a chain

DOI: 10.1201/9781003493471-1

with just two states; the eigenvalues of the transition matrix determine how the chain behaves and also establishes the rate of convergence to the stationary probability, assuming it exists. Question 1.2.4 explores a larger state space with focus placed on properties of the eigenvalues and the corresponding properties of the chain.

Finally, there are two questions on the Poisson process. Question 1.2.5 concerns the construction of the Poisson process via changes in the process during an arbitrarily small amount of time. Describing changes of random mechanisms over a small interval of time is often reasonable in practice, as this is where the information would be available. The subsequent plan is to find how the changes then look over a larger interval of time. Question 1.2.6 concerns the construction of the Poisson process via independent exponential random variables, which determine how long the process remains at any given height.

Q1.1 Questions – Probability Theory

Q1.1.1 – Types of convergence of a sample mean

Introduction. This question examines probabilistic properties of a sample mean. The first result is that the sample mean converges in distribution to a normal random variable when it is rescaled to have a fixed variance; this is regardless of the number of samples. The proof of this can be found using Laplace transforms. The next step is to show that the sample mean converges in probability to the mean of the population, which implies that a deterministic sequence of subsamples converges almost surely. The exercise does not include the almost sure convergence of $\overline{X}_n$ to the mean (and the conditions under which this happens) because it is quite difficult to prove.

For a random sequence $\{X_n\}$, there are associated sequences: the sequence of probabilities on events $\{P_n(A_n)\}$ and the sequence of distribution functions $\{F_{X_n}(x)\}$. These different types of sequences lead to the different types of convergence for $\{X_n\}$.

Question. Suppose $X_1, \ldots, X_n$ are independent and identically distributed random variables with mean zero and variance 1. Define

$$\overline{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

(i) Show that $\mathrm{E}(\overline{X}_n) = 0$ and $\mathrm{Var}(\overline{X}_n) = 1/n$.

(ii) By considering the Laplace transform $\phi_n(t) = \mathrm{E}\{\exp\left(t\sqrt{n}\overline{X}_n\right)\}$ with $t > 0$, show why $Z_n = \sqrt{n}\overline{X}_n$ is approximately a standard normal random variable for large n.

(iii) Prove the Markov inequality: if Y is a positive continuous random variable, then $\mathrm{P}(Y > a) < \mathrm{E}(Y)/a$ for all $a > 0$.

(iv) For all sequences $\varepsilon_n \to 0$ for which $n\varepsilon_n^2 \to \infty$, show that $\mathrm{P}(|\overline{X}_n| > \varepsilon_n) \to 0$.

(v) Show there exists a deterministic sequence $\{n_i\}_{i \geq 1}$ for which $|\overline{X}_{n_i}| \to 0$ almost surely as $i \to \infty$.

Q1.1.2 – Infinite divisibility of random variables

Introduction. Introduced by Bruno de Finetti in 1929, the concept of infinitely divisible distributions plays an important role in modern probability theory and financial modeling, where it has found applications in limit theorems and Lévy and additive processes. Many known distributions are infinitely divisible, such as the gamma, Gaussian, and Poisson. Thus, the study of infinitely divisible distributions and their decompositions are of broad interest. This question concerns the use of moment generating functions to determine whether distributions are infinitely divisible. In particular, the exercise investigates the gamma and Poisson distributions.

Question. A random variable X is infinitely divisible if, for every integer $n = 1, 2, 3, \ldots$, it is possible to write

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

where the $\{X_i\}$ are independent and identically distributed. The distribution of the $\{X_i\}$ may depend on n. For example, if X is standard normal, then each X_i is normal with mean 0 and variance $1/n$.

(i) Suppose $X \sim \mathrm{Ga}(a, 1)$. Find the moment generating function (Laplace transform) for X; i.e., $\phi_X(\theta) = \mathrm{E}\left(e^{-\theta X}\right)$ for $\theta > 0$.

(ii) If X is infinitely divisible, explain why $\phi_X(\theta) = \{\phi_{X_i}(\theta)\}^n$ for each $i = 1, \ldots, n$.

(iii) Determine whether the gamma variable is infinitely divisible, and if so, find the distribution for each X_i.

(iv) Show that $Z = \mathrm{Pois}(\lambda)$ is infinitely divisible.

(v) If X is $\mathrm{Ga}(a, 1)$, Z is $\mathrm{Pois}(\lambda X)$, and Y is $\mathrm{Ga}(a + Z, 1 + \lambda)$, show that Y is marginally infinitely divisible.

Q1.1.3* – A maximal inequality for a martingale sequence

Introduction. This question outlines a proof for the Doob martingale inequality, which provides an upper-bound for the probability that the maximum value of a martingale sequence exceeds any given value over a particular interval of (discrete) time. This inequality may be viewed as an extension of the Markov inequality.

Question. Suppose $\{M_n : n \in N\}$ is a non-negative martingale such that $\mathrm{E}\left(M_{n+1} \mid M_{1:n}\right) = M_n$. Define the events

$$E_n = \{M_1 < \epsilon, \ldots, M_{n-1} < \epsilon, M_n > \epsilon\}$$

for $n = 1, \ldots, K$.

(i) Show that the set of events $\{E_n : n = 1, \ldots, K\}$ is mutually disjoint.

(ii) Show that

$$\{\max\{M_1,\ldots,M_K\} > \epsilon\} \equiv \bigcup_{n=1}^{K} E_n.$$

(iii) Using the martingale property, show that

$$\int_{E_n} M_K\, f(M_1,\ldots,M_K)\, dM_1 \ldots dM_K = \int_{E_n} M_n\, f(M_1,\ldots,M_K)\, dM_1 \ldots dM_K$$

for any $n \leq K$, where $f(M_1,\ldots,M_K)$ represents the joint density function for $\{M_1,\ldots,M_K\}$.

(iv) Show that $\int_{E_n} M_n\, f(M_1,\ldots,M_K)\, dM_1 \ldots dM_K > \epsilon\, \mathrm{P}(E_n)$.

(v) Show that $\mathrm{P}(\max\{M_1,\ldots,M_K\} > \epsilon) \leq \mathrm{E}(M_K)/\epsilon$.

S1.1 Solutions & Further Reading – Probability Theory

S1.1.1 – Types of convergence of a sample mean

(i) The mean and variance are

$$\mathrm{E}(\overline{X}_n) = \mathrm{E}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \mathrm{E}(X_1) = 0,$$
$$\mathrm{Var}(\overline{X}_n) = \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\mathrm{Var}(X_1) = \frac{1}{n},$$

which follow from the independence of the $\{X_i\}$. If the $\{X_i\}$ are not independent, then covariances would need to be included.

(ii) This can be solved through a Taylor expansion of e^x (ignoring higher order terms of $1/n$) and then taking an expectation. This is a standard practice for establishing the central limit theorem:

$$\begin{aligned}
\phi_n(t) &= \mathrm{E}\left\{\exp(t\sqrt{n}\overline{X}_n)\right\} = \mathrm{E}\left\{\exp\left(t\sqrt{n}\frac{1}{n}\sum_{i=1}^{n} X_i\right)\right\} \\
&= \mathrm{E}\left\{\exp\left(t\frac{1}{\sqrt{n}}\sum_{i=1}^{n} X_i\right)\right\} \\
&= \left[\mathrm{E}\left\{\exp\left(t\frac{1}{\sqrt{n}}X_i\right)\right\}\right]^n \\
&= \left\{\mathrm{E}\left(1 + tX_i/\sqrt{n} + \tfrac{1}{2}t^2X_i^2/n + o(1/n)\right)\right\}^n \\
&= \left(1 + 0 + \tfrac{1}{2}t^2/n + o(1/n)\right)^n \to \exp\left(\tfrac{1}{2}t^2\right),
\end{aligned}$$

where the fourth equality is obtained because the $\{X_i\}$ are independent and identically distributed. The final line uses the exponential approximation $(1 + x/n)^n \to e^x$. Note that $o(1/n)$ represents the terms for which $n\,o(1/n) \to 0$, so $(1 + a/n + 1/n^2)^n \to e^a$. By Lévy's continuity theorem, pointwise convergence of the Laplace transform of Z_n implies convergence in distribution to a standard normal random variable as $n \to \infty$.

(iii) The key starting point is $a\,1(Y > a) \leq Y\,1(Y > a) \leq Y$. Taking expectations with respect to Y, it is seen that $a\,\mathrm{P}(Y > a) \leq \mathrm{E}(Y)$, where $\mathrm{E}\{1(Y > a)\} = \mathrm{P}(Y > a)$.

(iv) To obtain a useful Markov inequality, consider

$$\mathrm{P}(|\overline{X}_n| > \epsilon_n) = \mathrm{P}\big(|Z_n| > \sqrt{n}\epsilon_n\big) = \mathrm{P}(Z_n^2 > n\epsilon_n^2) \leq \mathrm{E}(Z_n^2)/(n\epsilon_n^2),$$

which goes to 0 as $n\epsilon_n^2 \to \infty$.