ACTEX Learning

Study Manual for Exam CS1 Actuarial Statistics

1st Edition

Yiping Guo, Ph.D., ASA
Gabriel Necoechea, Ph.D.

An IFoA & IAI Exam

ACTEX Learning

Study Manual for Exam CS1 Actuarial Statistics

1st Edition

Yiping Guo, Ph.D., ASA
Gabriel Necoechea, Ph.D.

Actuarial & Financial Risk Resource Materials
Since 1972

ACTEX as a Benefit Corporation

Benefit Corporations are businesses that meet the highest standards of verified social performance, transparency, and legal accountability to balance profit and purpose. Benefit Corporations seek to redefine success in business and build stronger communities and a more equitable, inclusive and sustainable economy through the creation of high-quality jobs with dignity and purpose. Benefit Corporations use profits and growth as a means to a greater end: positive impact for their stakeholders.

We believe in the power of community. Over the decades, we've assembled a diverse collective of professors and professional subject matter experts and empowered them to create the best educational materials. These materials offer students unrivaled access to affordable and comprehensive learning solutions that students can tailor to their unique learning styles.

Reaching our Goals!

Our community members actively contribute and collaborate in support of our shared vision. Some members contribute individually, some in work in small teams, but each works collectively for the whole. The members of our educational community include YOU

- Students & Professionals – giving back through communication of ideas with our authors and instructors and the broader community
- Professors & Instructors – pedagogy thought-leaders, supporters of expanded access
- Authors & Professional Subject-Matter Experts – recognized leaders in their field, aspirants for ever-better authorship and instruction
- Our Team – employees dedicated to our company vision and mission
- Professional Societies (SOA, CAS, IFoA, etc.) – visionaries for professional education
- Planet Earth - Our stewardship of the environment

Free Resources!

Part of our mission as a benefit corp is opening doors for aspring actuaries around the world. Scan the QR codes below to recieve access to exam formula sheets or career and study guides. All resources are completely free and just one way we've chosen to give back.

Formula Sheets

Actuarial Exam Tactics

The Actuarial Career: Getting Started

Preface

About this Study Manual

This study manual has been specifically written for students preparing for the IFoA and IAI Exam CS1. The CS1 exam syllabus spans topics in basic probability (random variables), classical statistical inference, linear regression and generalised linear models, Bayesian credibility and data analysis. The study manual will help you prepare for each and every one of these topics. We developed the text by carefully studying the syllabus objectives, past exam questions and solutions, and the IFoA CS1 Core Reading. Based on our study of these materials, we have written a comprehensive text that will offer a solid theoretical and conceptual foundation for your exam days. Included in this text are illustrative examples and end-of-section exercises that will develop your calculation skills and deepen your understanding. We have included the R code where its inclusion will directly prepare you for your exam. The five appendices can help fill in gaps in your background knowledge and point you in the direction of additional learning resources.

The first author, Yiping Guo, extends his gratitude to Gabriel Necoechea, the second author, for his meticulous proofreading of the chapters on statistical inference and Bayesian statistics, which significantly enhanced their clarity and quality. It is a truly pleasure and enjoyment working with him.

The second author, Gabriel Necoechea, would like to thank Matthew Naeger, Anna Melikyan, Albert Ofoe, and Clement Moki for reading an early draft of chapters on probability & regression and offering helpful suggestions. Very special thanks go to Yiping Guo, coauthor, for setting a high standard for quality, which you will see in his chapters of this study manual. Reading his sections motivated me to revise (and hopefully improve) my sections. His keen proofreading and helpful comments on my chapters also improved the quality of the text. I thoroughly enjoyed our collaboration.

We would like to thank Abraham Weishaus for generously allowing us to use some of his original exercises, which appear in his ASM Study Manual for SOA Exam P. Thanks are also due to Yijia Liu for his editorial work, which greatly improved the visual appeal of this study manual. We sincerely thank Bill Marella and Joana Amorim for their valuable support throughout the development of this project.

Getting the Most out of this Study Manual

We have written a comprehensive text that covers all aspects of the Exam CS1 syllabus. However, you should consider your own background when using this manual. Identify weak spots and focus there. In particular, do a few of the exercises but do not do every single exercise. Once a particular kind of exercise feels comfortable, move on to the next kind of exercise (e.g., move from calculating expectation to calculating variance).

At the same time, do not underestimate the importance of building strong fundamentals! Students at the level of Exam CS1 frequently make the mistake to think that, because the exam is quite challenging, the approach to take is to move very quickly and learn very complicated things. Instead, you should focus on moving slowly but steadily, building a strong grasp of things that seem easy *so that more complicated things become easy as a consequence of your strong*

fundamentals. Basically, instead of thinking how far you can *stretch* your mathematical understanding, you should try to *compress* the syllabus by seeing how the fundamental principles apply over and over and over again. We hope that the written text and illustrative examples will help you to achieve this.

On a related note, you may find weaknesses in your fundamentals, either in probability theory, calculus, or even algebra. The published examiners' reports for IFoA Exam CS1 caution that many students would be well served by strengthening their calculus, algebra, and basic probability skills. Do not underestimate the marks you can earn simply by knowing how to do a calculus calculation correctly! For this reason, you should hold yourself to a very high standard when evaluating your performance in examples and exercises. You can *forgive* yourself for making an algebra mistake, especially if your strategy/theory is sound, but you must acknowledge that your attention to detail was not as sharp as it needs to be on exam day! Aim to improve your accuracy of calculations during the course of your exam preparation.

The R programming component of Exam CS1 is sure to loom large in your mind, possibly causing you quite a bit of anxiety. Our recommendation is to not overemphasise this aspect of the exam. The R programming requirements of Exam CS1 are relatively modest. Most of the challenge with programming comes not from the complexity of programming, but rather from the time constraints under which that programming must be completed. We believe that your best strategy for defeating this time challenge is not to devote hours and hours of R programming practice, but to get better at identifying the theoretically sound strategy for setting up the problem. For example, many of the programming tasks on Exam CS1 will fall into one of the following four categories:

- perform data visualisation
- conduct hypothesis testing
- fit a (generalised) linear model
- simulate values from a random variable

The programming required to carry out any of these tasks is modest. What is more important is that you make reasonable assumptions and, with the output in hand, make thoughtful comments on that output. Therefore, while we do discuss exam-relevant R skills in the study manual, our philosophy in developing this study manual has been to prioritise the theoretical content that we think will set you up to receive high marks on the exam. We have included an appendix with a link to the R Formula and Review Sheet by ACTEX Learning. That PDF contains everything we think you need to know to succeed in the R programming tasks of Exam CS1. At the time of writing, the first 21 pages of that PDF are relevant to Exam CS1.

Yiping Guo, Ph.D., ASA and Gabriel Necoechea, Ph.D.

March, 2025

Contents

Chapter 2

Statistical Inference

2.1 Sampling Theory Related to the Normal Distributions

When the data are i.i.d. normal:

$$X_1, \cdots, X_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2),$$

we can derive the exact (not asymptotic) sampling distributions of many quantities of interest. Specifically, we will study chi-square distributions, t-distributions and F-distributions in greater depth, and discuss how those distributions are related to the normal distributions. The results discussed in this part will be extensively used when we study confidence intervals and hypothesis testing in later chapters.

2.1.1 Distributions of the Sample Mean $\bar{X}$

Under the normal assumption, we can show that the sample mean $\bar{X}$ also follows a normal distribution.

Theorem 2.1.1. If $X_1, \cdots, X_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$, then $\bar{X}$ follows a normal distribution:

$$\bar{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right), \tag{2.1.1}$$

Proof of Theorem 2.1.1. Recall that if $X_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$ is independent of $X_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$, then,

$$X_1 + X_2 \sim \mathcal{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2).$$

Therefore,

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \sim \frac{1}{n}\mathcal{N}\left(n\mu, n\sigma^2\right) = \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right).$$

□

2.1.2 The Chi-square Distribution of the Sampling Variance S^2

We have studied how to calculate $\mathbb{E}[S^2]$ for general i.i.d. samples. Under the normal assumption, we can further explicitly derive the exact distribution of S^2.

Before presenting the theorem, we first discuss some further results regarding chi-square distributions beyond previous chapters.

Definition 2.1.1 (Chi-Square Distributions)**.** Let $Z_1, \cdots, Z_n \overset{iid}{\sim} \mathcal{N}(0, 1)$. Then,

$$V = Z_1^2 + \cdots + Z_n^2 \sim \chi_n^2, \tag{2.1.2}$$

where n is the degrees of freedom.

This definition implies a nice additivity result for the chi-square variables, which will be used later when deriving confidence intervals in some complicated two-sample cases. The result is presented in the next theorem.

Theorem 2.1.2. If $V_1 \sim \chi^2_{n_1}$ and $V_2 \sim \chi^2_{n_2}$ are independent, then $V = V_1 + V_2 \sim \chi^2_{n_1+n_2}$.

Proof of Theorem 2.1.2. By definition, we write:

$$V_1 = Z_1^2 + \cdots + Z_{n_1}^2,\ V_1 = \tilde{Z}_1^2 + \cdots + \tilde{Z}_{n_2}^2,$$

where all of $Z_1, \cdots, Z_n, \tilde{Z}_1, \cdots, \tilde{Z}_{n_2}$ are independent implies by the independence between V_1 and V_2.

Then,

$$V = V_1 + V_2 = Z_1^2 + \cdots + Z_{n_1}^2 + \tilde{Z}_1^2 + \cdots + \tilde{Z}_{n_2}^2 \sim \chi^2_{n_1+n_2}.$$

□

When calculating different quantities about chi-square distribution, such as the moments, it is often useful to use its equivalence to Gamma distributions.

Theorem 2.1.3. The χ^2_n distribution is the Gamma$(n/2, 1/2)$ distribution.

Proof of Theorem 2.1.3. First, we show that the PDF of Z_1^2 and Gamma$(1/2, 1/2)$ are the same. Let $F(x)$ and $f(x)$ be the CDF of Z_1^2 respectively. Also, use $\Phi(x)$ and $\varphi(x)$ to denote the CDF and PDF of the standard normal variable, respectively.

For $x > 0$,

$$F(x) = \mathbb{P}(Z_1^2 < x) = \mathbb{P}(\sqrt{x} < Z_1 < \sqrt{x}) = \Phi(\sqrt{x}) - \Phi(\sqrt{-x}) = 2\Phi(\sqrt{x}) - 1,$$

where we used a property of the normal CDF that $\Phi(x) + \Phi(-x) = 1$ for $x \in \mathbb{R}$. Then, taking derivative on both sides,

$$f(x) = F'(x) = 2\varphi(\sqrt{x}) \cdot \frac{1}{2\sqrt{x}} = 2 \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{x}{2}} \cdot \frac{1}{2\sqrt{x}} = \frac{1}{\sqrt{2\pi}} x^{-\frac{1}{2}} e^{-\frac{x}{2}}.$$

For the Gamma$(1/2, 1/2)$ distribution, we plug in $\alpha = \lambda = 1/2$ to the general PDF formula and use the fact that $\Gamma(1/2) = \sqrt{\pi}$:

$$\frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x} \longrightarrow \frac{\sqrt{1/2}}{\Gamma(1/2)} x^{-\frac{1}{2}} e^{-\frac{x}{2}} = \frac{1}{\sqrt{2\pi}} x^{-\frac{1}{2}} e^{-\frac{x}{2}},$$

which is equivalent to the PDF $f(x)$ of Z_1^2. Then, because $V = Z_1^2 + \cdots + Z_n^2$ is the sum of n i.i.d. Gamma$(1/2, 1/2)$ random variables, we have $V \sim$ Gamma$(n/2, 1/2)$. □

Theorem 2.1.3 also helps us to memorise the mean and variance formulas for chi-square distributions. Recall that for $X \sim$ Gamma(α, λ), $\mathbb{E}[X] = \alpha/\lambda$ and $\text{Var}(X) = \alpha/\lambda^2$. Therefore, for $V \sim \chi^2_n$, it immediately implies that $\mathbb{E}[V] = n$ and $\text{Var}(V) = 2n$.

The next theorem presents the relationship between the chi-square distribution and sample variance S^2 for under normal assumptions.

Theorem 2.1.4. If $X_1, \cdots, X_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$, then S^2 has the following distribution:

$$S^2 \sim \frac{\sigma^2}{n-1} \chi^2_{n-1}, \text{ or equivalently, } \frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}. \tag{2.1.3}$$

Furthermore, S^2 is independent of $\bar{X}$.

The proof is beyond the scope of the CS1 exam. Here, we provide some intuitive explanations to help understand the results, so that we do not have to memorise blindly.

- $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$: Writing down the quantity on the left-hand side:

$$\frac{(n-1)S^2}{\sigma^2} = \frac{(n-1)}{\sigma^2} \cdot \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2 = \sum_{i=1}^{n} \left(\frac{X_i - \bar{X}}{\sigma} \right)^2$$

  If we replace the sample mean $\bar{X}$ by the population mean μ, the right-hand side becomes:

$$\sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma} \right)^2 = \sum_{i=1}^{n} Z_i^2 \sim \chi^2_n,$$

  where $Z_i \sim \mathcal{N}(0,1)$ represents the standard normal and the resulting χ^2 distribution is by definition.

  When we have the sample mean $\bar{X}$ as in S^2, the argument above does not hold. However, it is not far away from the truth. Although each of the summand $((X_i - \bar{X})/\sigma)^2$ is dependent with each other due to the common random variable $\bar{X}$, it can be shown that,

$$\sum_{i=1}^{n} \left(\frac{X_i - \bar{X}}{\sigma} \right)^2 = \sum_{i=1}^{n-1} Z_i^2 \sim \chi^2_{n-1}.$$

  Intuitively speaking, due to $\bar{X}$, we lose one degree of freedom. To illustrate, think about when $n = 3$, and we can calculate X_3 from X_1, X_2 and $\bar{X}$ through $X_3 = 3\bar{X} - X_1 - X_2$, and the same for X_1 and X_2. The formal proof relies on linear algebra and properties of multivariate normal distributions, so we refer interested readers to advanced statistical textbooks.

- $\bar{X}$ and S^2 are independent: The key is to show that, very surprisingly, each of $X_i - \bar{X}$ is independent of $\bar{X}$. This is a very counter-intuitive result since $\bar{X}$ also appears in $X_i - \bar{X}$. The formal proof, again requires knowledge of linear algebra and multivariate normal distributions or advanced statistical inference theory (Basu's Theorem). We skip the technical details here.

  This result immediately applies that $\bar{X}$ and S^2 are independent since each summand in

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$$

  is a function of $X_i - \bar{X}$.

  Here, we discuss a simpler case to demonstrate the independence between $X_i - \bar{X}$ and $\bar{X}$. Consider adjusting $\bar{X} \to \bar{X} + c$ by setting $X_i \to X_i + c$ for all i, where $c \in \mathbb{R}$. As a result, $X_i - \bar{X} \to (X_i + c) - (\bar{X} + c) = X_i - \bar{X}$ remains unchanged. That says, $X_i - \bar{X}$ is independent of $\bar{X}$, as $X_i - \bar{X}$ is invariant from the change of $\bar{X}$.

  It is also worth noting that the independence between $\bar{X}$ and S^2 only holds when the underlying distribution is normal, and not for any other distributions. To see why this property does not hold in non-normal cases, imagine $X \sim \text{Poi}(\lambda)$, and larger $\bar{X}$ would likely imply larger S^2, as $\mathbb{E}[X] = \text{Var}(X) = \lambda$ for Poisson distributions.

The **chi-square distribution** of S^2 allows us to investigate its further properties with more ease. For example, $\text{Var}(S^2)$ in general has a messy formula. However, when under the normal assumption, it can be easily derived.

Theorem 2.1.5. If $X_1, \cdots, X_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$, then we have:

$$\text{Var}(S^2) = \frac{2\sigma^4}{n-1}. \tag{2.1.4}$$

Proof of Theorem 2.1.5. From Theorem 2.1.4, we have:

$$S^2 \sim \frac{\sigma^2}{n-1}\chi^2_{n-1}.$$

Taking variance on both sides, we get:

$$\text{Var}(S^2) = \frac{\sigma^4}{(n-1)^2} \cdot 2(n-1) = \frac{2\sigma^4}{n-1},$$

where we use the result that $\text{Var}(X) = 2n$ for $X \sim \chi^2_n$. □

2.1.3 The t-Distributions and the t-Statistic

Recall that from Theorem 2.1.1 when $X_1, \cdots, X_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$, $\bar{X} \sim \mathcal{N}(\mu, \sigma^2/n)$. When dealing with normal random variable, it is common to *standardise* it by subtracting the mean and being divided by the standard deviation. Thus, for $\bar{X}$, the following is the so-called z-statistic, as the resulting distribution is standard normal:

$$z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1). \tag{2.1.5}$$

This quantity can be useful for making inference (specifically, confidence intervals or hypothesis testing) on μ, where we compare the observed sample mean $\bar{X}$ with the assumed value of the true mean μ.

However, this z-statistic has little practical use since we never know the value of σ. In practice, we replace the unknown standard deviation σ by the sample standard deviation $S = \sqrt{S^2}$, calculated by taking square root of the sample variance. The resulting quantity is called the t-statistic, and as the name suggests, it follows a **t-distribution**:

Theorem 2.1.6. If $X_1, \cdots, X_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$, then

$$t = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}. \tag{2.1.6}$$

You can think about this as one way that t-distributions are defined, which is exactly the case in the history of statistics. In general, t-distributions are defined in a similar fashion to Theorem 2.1.6:

Definition 2.1.2 (t-Distributions)**.** Let $Z \sim \mathcal{N}(0, 1)$ and $V \sim \chi^2_n$. Then,

$$T = \frac{Z}{\sqrt{V/n}}, \tag{2.1.7}$$

where n is the degrees of freedom.

Based on Definition 2.1.2, the proof of Theorem 2.1.6 is straightforward:

Proof of Theorem 2.1.6. Recall that $\bar{X} \sim \mathcal{N}(\mu, \sigma^2/n)$ and $(n-1)S^2/\sigma^2 \sim \chi^2_{n-1}$. Therefore, we have the following representations:

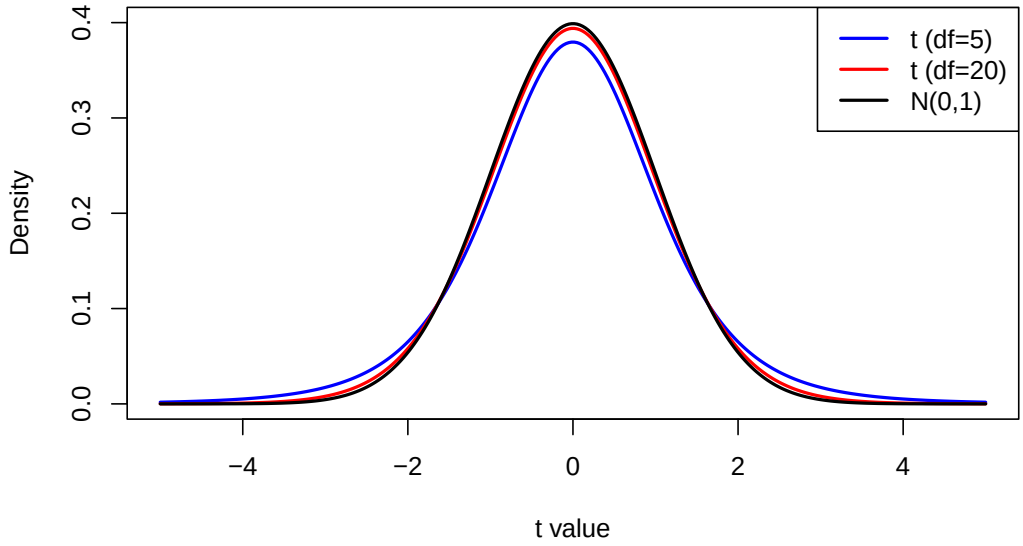
$$\bar{X} = \mu + \frac{\sigma}{\sqrt{n}} Z, \quad S^2 = \frac{\sigma^2}{n-1} V,$$

where $Z \sim \mathcal{N}(0,1)$ and $V \sim \chi^2_{n-1}$. Plugging into the expression of the t-statistic, we get:

$$t = \frac{\bar{X} - \mu}{S/\sqrt{n}} = \frac{\sigma Z/\sqrt{n}}{\sqrt{\sigma^2 V/(n(n-1))}} = \frac{Z}{\sqrt{V/(n-1)}} \sim t_{n-1}.$$

□

The t-distributions have similar symmetric bell-shapes to the normal distributions. To illustrate, we plot the PDF curves for the t-distributions with 5 degrees of freedom, 20 degrees of freedom, and for the standard normal distribution.



There are two important properties related of the t-distributions that we can observe from the plot.

1. The t-distributions have *heavier tails* than the normal distributions:

   The heavy-tail property of the t-distributions implies that we are expected to observe more extreme values than the normal distributions. Given the stronger tolerance to extreme observations, the t-distributions are commonly used in robust statistics and quantitative risk management.

   A more mathematical way to see the heavy-tail property is to compare the PDF of the t-distributions[i] to the normal:

$$t_n: \ f(x) = \frac{\Gamma((n+1)/2)}{\sqrt{\pi n}\Gamma(n/2)} \cdot \left(1 + \frac{x^2}{n}\right)^{-(n+1)/2}, \quad \mathcal{N}(0,1): \ g(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}.$$

[i] You are not required to memorise the PDF of the t-distributions for the exam, and we only present it for demonstrating the argument here.

Without bothering with the messy normalising constants, the PDF of the t-distributions behave like a power function x^{-n}, but by contrast, the PDF of the normal distributions behave like an exponential function e^{-x^2}. Since power functions decay slower than exponential functions (in the sense that exponential functions are at a smaller scale than the power function when $x \to \infty$)[ii], the t-distributions should have heavier tails.

2. A t-distribution converges to the standard normal distribution as the degrees of freedom $n \to \infty$.

   Proof. This follows directly from the Law of Large Numbers. Define

   $$T_n = \frac{Z}{\sqrt{V_n/n}} \sim t_n,$$

   where $Z \sim \mathcal{N}(0,1)$ and $V_n \sim \chi^2_n$. By the definition of chi-square distributions, we have $V_n = Z_1^2 + \cdots + Z_n^2$ for $Z_1, \cdots, Z_n \overset{iid}{\sim} \mathcal{N}(0,1)$. Therefore, by the Law of Large Numbers,

   $$\frac{V_n}{n} = \frac{Z_1^2 + \cdots + Z_n^2}{n} \longrightarrow \mathbb{E}[Z_1^2] = 1.$$

   Last,

   $$T_n = \frac{Z}{\sqrt{V_n/n}} \longrightarrow Z \sim \mathcal{N}(0,1).$$

   (To make the final step more rigorous, we actually need to apply the so-called *Slutsky's theorem*. We omit the technical details here, but the main argument should be straightforward to understand.)

   Alternatively, we can show that the PDF of the t-distributions converge to the standard normal PDF (ignoring the normalising constants):

   $$\lim_{n\to\infty} \left(1 + \frac{x^2}{n}\right)^{-(n+1)/2} = \lim_{n\to\infty} \left(1 + \frac{x^2}{n}\right)^{-n/2} = e^{-\frac{x^2}{2}},$$

   where we use the result from Calculus that $(1 + x/n)^n \to e^x$ as $n \to \infty$ for $x \in \mathbb{R}$. □

2.1.4 The F-Distributions and the F-Statistic

Another distribution related to the normal distributions is the **F-distribution**. The name "F" represents the one of the most prominent statisticians, Ronald A. Fisher, who more or less invented the F-distributions.

Unlike the z-statistic and t-statistic which are used mainly for making inference on the mean of a single normal distribution, the F-statistic, as defined below, cares about the ratio of the variances of two independent normal distributions. To set things up, suppose we have $X_1, \cdots, X_m \overset{iid}{\sim} \mathcal{N}(\mu_1, \sigma_1^2)$, $Y_1, \cdots, Y_n \overset{iid}{\sim} \mathcal{N}(\mu_2, \sigma_2^2)$ and all of $X_1, \cdots, X_m$ and $Y_1, \cdots, Y_n$ are independent. Say we are interested in the variance ratio σ_1^2/σ_2^2, and most relevant quantity we can obtain from the sample is the ratio of sample variances S_1^2/S_2^2. This is exactly what a F-statistic tries to characterise and where the F-distributions arise.

First, we give the definition of the F-distributions:

[ii] In case the readers are familiar with algorithm analysis or the big-O notations, a rigorous way to express that $g(x)$ decays faster than $f(x)$ as $x \to \infty$ is $f(x) = o(g(x))$.

Definition 2.1.3 (F-Distributions)**.** Let $V_1 \sim \chi^2_m$ and $V_2 \sim \chi^2_n$. Furthermore, assume that V_1 and V_2 are independent. Then,

$$F = \frac{V_1/m}{V_2/n} \sim F_{m,n}, \tag{2.1.8}$$

where m and n are the two parameters to control the degrees of freedom of the F-distributions.

Then, we define the F-statistic based on the sample variances ratio S_1^2/S_2^2 and show that it follows the F-distribution.

Theorem 2.1.7. Assume that $X_1, \cdots, X_m \overset{iid}{\sim} \mathcal{N}(\mu_1, \sigma_1^2)$, $Y_1, \cdots, Y_n \overset{iid}{\sim} \mathcal{N}(\mu_2, \sigma_2^2)$. Furthermore, all of $X_1, \cdots, X_m$ and $Y_1, \cdots, Y_n$ are independent, then

$$F = \frac{S_1^2/S_2^2}{\sigma_1^2/\sigma_2^2} \sim F_{m-1,n-1}, \tag{2.1.9}$$

where the sample variances are defined as:

$$S_1^2 = \frac{1}{m-1}\sum_{i=1}^{m}(X_i - \bar{X})^2, \quad S_2^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2.$$

Proof of Theorem 2.1.7. Recall that $S_1^2 \sim \chi^2_m$ and $S_1^2 \sim \chi^2_m$. Therefore, we have the following representations:

$$S_1^2 = \frac{\sigma_1^2}{m-1}V_1, \quad S_2^2 = \frac{\sigma_2^2}{n-1}V_2,$$

where $V_1 \sim \chi^2_{m-1}$ and $V_2 \sim \chi^2_{n-1}$. Plugging into the expression of the F-statistic, we get:

$$F = \frac{S_1^2/S_2^2}{\sigma_1^2/\sigma_2^2} = \frac{V_1/(m-1)}{V_2/(n-1)} \sim F_{m-1,n-1}.$$

□

There are several important properties of the F-distributions that might be useful in the exam.

Theorem 2.1.8. If $X \sim F_{m,n}$, then, $1/X \sim F_{n,m}$.

Proof of Theorem 2.1.8. The proof is trivial using the definition of the F-distributions. Let $V_1 \sim \chi^2_m$ and $V_2 \sim \chi^2_n$, then,

$$F = \frac{V_1/m}{V_2/n} \sim F_{m,n} \Longrightarrow \frac{1}{F} = \frac{V_2/n}{V_1/m} \sim F_{n,m}.$$

□

This result can be used to derive a more important property of the *F-scores*. An F-score is the quantity that represents the critical value of a F-distribution with a fixed right tail probability. More precisely, we use $F_{\alpha,m,n}$ to denote the value of c such that $\mathbb{P}(F > c) = \alpha$, where $F \sim F_{m,n}$. Theorem 2.1.8 has a direct implication of a nice symmetry property of F-scores, which will be useful when deriving the confidence interval for the variance ratio of two important normal samples.

Theorem 2.1.9. For $\alpha \in (0,1)$ and $m, n \in \mathbb{N}$, the following result for the F-scores holds:

$$F_{1-\alpha,m,n} = \frac{1}{F_{\alpha,n,m}} \tag{2.1.10}$$

Proof of Theorem 2.1.9. The proof is a direct application of Theorem 2.1.8. Let $f_1 = F_{1-\alpha,m,n}$ and $f_2 = F_{\alpha,n,m}$. We have:

$$\mathbb{P}(F_{m,n} > f_1) = 1 - \alpha \Longrightarrow \mathbb{P}(F_{m,n} < f_1) = \alpha.$$

By Theorem 2.1.8, we have $F_{m,n} \overset{d}{=} 1/F_{n,m}$, therefore,

$$\mathbb{P}(F_{m,n} < f_1) = \mathbb{P}\left(\frac{1}{F_{n,m}} < f_1\right) = \mathbb{P}\left(F_{n,m} > \frac{1}{f_1}\right),$$

which implies $f_2 = 1/f_1$, as desired. □

The last theorem can be useful in the context of regression analysis.

Theorem 2.1.10. If $T \sim t_n$, then, $T^2 \sim F_{1,n}$.

Proof of Theorem 2.1.10. The proof is also straightforward using the definition of the t-distributions, chi-square distributions and F-distributions. Let $Z \sim \mathcal{N}(0,1)$ and $V \sim \chi^2_n$, then,

$$T = \frac{Z}{\sqrt{V/n}} \sim t_n \Longrightarrow T^2 = \frac{Z^2/1}{V/n} \sim F_{1,n},$$

where we use the fact that $Z^2 \sim \chi^2_1$. □

This theorem can be applied in cases when we are asked to calculate the values of an F-statistic, but we are only given the value of the corresponding t-statistic, or vise versa.

Exercises

Exercise 2.1.1. A factory produces metal rods with diameters that follow a normal distribution with a mean of 20 mm and a variance of 2.5 mm^2. A quality control team selects a random sample of 16 rods and records their diameters.

(a) Determine the sampling distribution of the sample variance S^2.

(b) Find the mean and variance of S^2.

Solution.

(a) For a random sample of size n from a normal distribution $\mathcal{N}(\mu, \sigma^2)$, the sample variance S^2 follows:

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$$

Here, $n = 16$ and $\sigma^2 = 2.5$, so

$$\frac{15S^2}{2.5} \sim \chi^2_{15}$$

Thus, the sampling distribution of S^2 is:

$$S^2 \sim \frac{1}{6}\chi^2_{15}$$

(b) For a chi-squared distributed random variable $X \sim \chi^2_v$, the mean and variance are:

$$\mathbb{E}[X] = v, \quad \text{Var}(X) = 2v$$

Using $S^2 \sim \chi^2_{15}/6$ from (a), we get:

$$\mathbb{E}[S^2] = 6\mathbb{E}[\chi^2_{15}] = \frac{1}{6} \times 15 = 2.5,$$
$$\text{Var}(S^2) = \left(\frac{1}{6}\right)^2 \text{Var}(\chi^2_{15}) = \left(\frac{1}{6}\right)^2 \times (2 \times 15) = 0.8333.$$

Exercise 2.1.2. A coffee shop measures the daily average temperature of its brewed coffee. The temperature follows a normal distribution with a mean of 80.5°C and a standard deviation of 5.4°C. A random sample of 36 days is taken. What is the probability that the sample mean temperature falls between 79.2°C and 81.8°C?

Solution. The sample mean follows:

$$\bar{X} \sim \mathcal{N}\left(80.5, \left(\frac{5.4}{\sqrt{36}}\right)^2\right) = \mathcal{N}(80.5, 0.9^2).$$

Therefore, by standardizing

$$\begin{aligned}\mathbb{P}\left(79.2 \le \bar{X} \le 81.8\right) &= \mathbb{P}\left(\frac{79.2 - 80.5}{0.9} \le \bar{X} \le \frac{81.8 - 80.5}{0.9}\right) \\ &= \Phi(1.44) - \Phi(-1.44) = 0.8502.\end{aligned}$$

Exercise 2.1.3. A random sample of size 7 is taken from a normal population with variance 16. Find the probability that the sample variance falls between 12 and 20.

Solution. For a normal population with variance σ^2, the sample variance S^2 follows:

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$$

With $n = 7$ and $\sigma^2 = 16$,

$$\frac{6S^2}{16} \sim \chi^2_6.$$

Transforming the bounds:

$$\mathbb{P}(12 \le S^2 \le 20) = \mathbb{P}\left(\frac{6(12)}{16} \le \chi^2_6 \le \frac{6(20)}{16}\right) = \mathbb{P}(4.5 \le \chi^2_6 \le 7.5)$$

Using chi-squared tables:

$$\mathbb{P}(12 \le S^2 \le 20) = \mathbb{P}(\chi^2_6 \le 7.5) - \mathbb{P}(\chi^2_6 \le 4.5) = 0.66 - 0.15 = 0.51.$$

Exercise 2.1.4. If S_1 and S_2 are the standard deviations of independent random samples of sizes $n_1 = 60$ and $n_2 = 30$ from normal populations with variances $\sigma_1^2 = 10$ and $\sigma_2^2 = 15$, find $\mathbb{P}(S_1^2/S_2^2 > 1.16)$.

Solution. The ratio of two independent sample variances follows an F-distribution:

$$\frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim F_{n_1-1,n_2-1}$$

Substituting the given values:

$$\frac{S_1^2/10}{S_2^2/15} \sim F_{59,29}$$

Rearranging for S_1^2/S_2^2:

$$\mathbb{P}\left(\frac{S_1^2}{S_2^2} > 1.16\right) = \mathbb{P}\left(F_{59,29} > \frac{1.16 \times 15}{10}\right) = \mathbb{P}(F_{59,29} > 1.74)$$

Using F-distribution tables, $\mathbb{P}(F_{59,29} > 1.74) \approx 0.10$.

Exercise 2.1.5. A researcher conducting Monte Carlo simulations has access only to a random number generator that produces independent samples from a standard normal distribution $\mathcal{N}(0,1)$.

(a) Describe a procedure to generate random samples from a chi-square distribution with 4 degrees of freedom.

(b) Describe a procedure to generate random samples from a t-distribution with 6 degrees of freedom.

(c) Describe a procedure to generate random samples from an F-distribution with 4 and 8 degrees of freedom.

(d) Explain why the inverse CDF method is not suitable for (a)-(c).

Solution.

(a) Generate four independent standard normal variables $Z_1, Z_2, Z_3, Z_4 \sim \mathcal{N}(0,1)$ and compute:

$$X = Z_1^2 + Z_2^2 + Z_3^2 + Z_4^2$$

Then, $X \sim \chi_4^2$.

(b) Generate $Z \sim \mathcal{N}(0,1)$ and $Y \sim \chi_6^2$ using (a). Compute:

$$T = \frac{Z}{\sqrt{Y/6}}.$$

Then, $T \sim t_6$.

(c) Generate $X_1 \sim \chi_4^2$ and $X_2 \sim \chi_8^2$ using (a). Compute:

$$F = \frac{X_1/4}{X_2/8}.$$

Then, $F \sim F_{4,8}$.

(d) The inverse CDF method requires computing the quantile function (inverse of the cumulative distribution function). For (a)-(c), the CDFs of chi-square, t, and F-distributions involve complex integral expressions that do not have closed-form solutions, making direct inversion computationally infeasible.

Exercise 2.1.6. (**R Programming**)

(a) Simulate 1000 values from a $\mathcal{N}(0,1)$ distribution using an appropriate **R** command. Use 2025 as the random seed.

(b) Simulate 1000 values from a χ^2_5 distribution using an appropriate **R** command.

(c) Simulate 1000 values from a t_5 distribution **using your results from (a) and (b)**.

(d) Simulate 1000 values from a t_{25} distribution by directly calling an appropriate **R** command.

(e) Comparing two appropriate plots of the values simulated from parts (c) and (d). Explain your observations by referring to an important property of the t-distributions.

Solution.

(a)
```
set.seed(2025)
normal.sim = rnorm(1000)
```

(b)
```
chisq.sim = rchisq(1000,5)
```

(c) A random variable having the t_5 distribution is defined as:

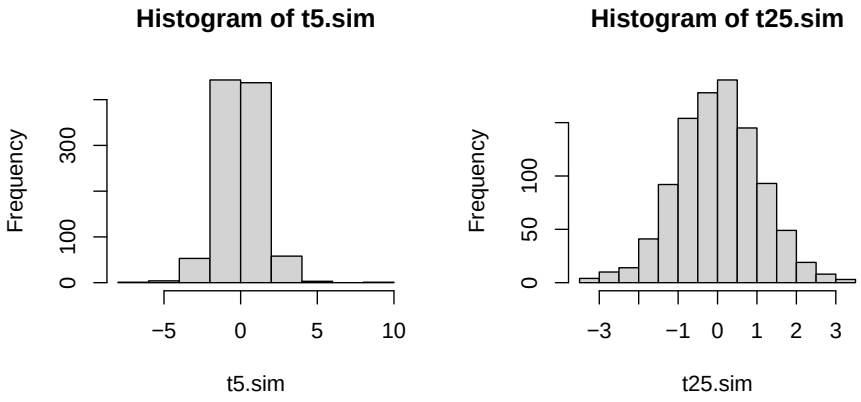
$$T = \frac{Z}{\sqrt{Y/5}}, \text{ where } Y = \chi^2_5.$$

Therefore,

```
t5.sim = normal.sim/sqrt(chisq.sim/5)
```

(d)
```
t25.sim = rt(1000,25)
```

(e)
```
par(mfrow=c(1,2)) hist(t5.sim) hist(t25.sim)
```



The histograms show that the t-distribution converges to the standard normal distribution as the degrees of freedom increases.

2.2 Point Estimation

2.2.1 Models, Parameters and Statistical Inference

2.2.1.1 What is Statistical Inference?

In previous chapters, we have laid the foundations of **probability theory**. Specifically, we have explored a variety of probability models and learned how to perform calculations with them. For instance, if we assume a random variable X follows a particular distribution, we can derive certain probabilistic quantities, such as $\mathbb{E}[X]$, $\text{Var}(X)$, or $\mathbb{P}(X > 0)$.

Statistical inference operates in the reverse direction of probability theory. Instead of starting with a known distribution to make statements about data, it begins with observed data and attempts to infer the underlying distribution. For example, consider the following integer-valued observations:

Value	1	2	3	4
# of Observations	100	101	102	97

Intuitively, one might guess that these data are sampled from a discrete uniform distribution over $\{1, 2, 3, 4\}$. In the upcoming chapters, we will discuss how to make such inferences more formally.

It is important to recognise that making inferences without any assumptions can be extremely challenging, or even impossible in many cases. For example, given six observations: $6, 10, 3, 2, 5, 5$, representing the number of insurance claims from different clients, it is impossible to determine the exact model from which these observations are drawn. However, some models might seem more plausible than others. For instance, a $\mathcal{N}(0, 1)$ distribution is unlikely since the observations are discrete and positive. Similarly, a $\text{Poi}(1000)$ distribution is improbable as it is unlikely to sample six small numbers from a Poisson distribution with a mean of 1000.

2.2.1.2 Parametric and Non-Parametric Models

One common strategy to address the inference challenge is **parametric modelling**. Let $\boldsymbol{X} = (X_1, \cdots, X_n)$ denote the observations, and X the underlying random variable. Parametric modelling assumes that X follows a family of distributions characterised by an unknown **parameter** (or multiple parameters) θ. The set of all possible values for θ is called the **parameter space**, denoted as Θ. Generally, a parametric model is expressed as:

$$\mathcal{M} = \{f(x|\theta) : \theta \in \Theta\}, \tag{2.2.1}$$

where $f(x|\theta)$ represents a family of distributions parametrised by θ.

The notation $f(x|\theta)$ used in this definition is actually extremely general. In elementary courses, $f(x|\theta)$ denotes the probability density function of a continuous distribution. Here, it can also refer to the probability mass function of a discrete distribution or more complicated distributions (neither discrete nor continuous). In advanced probability theory, all these concepts are considered special cases of a more general concept called *density* from measure theory.

Here are some common examples of parametric models and the corresponding parameter spaces:

- All Poisson distributions:
$$\left\{ f(x|\lambda) = \frac{e^{-\lambda}\lambda^x}{x!} : \lambda > 0 \right\};$$
- All normal distributions:
$$\left\{ f(x|\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} : \mu \in \mathbb{R}, \sigma > 0 \right\}.$$

In contrast, **non-parametric modelling** represents a family of distributions that cannot be parametrised by a finite number of parameters. Non-parametric models generally impose weaker assumptions and offer higher flexibility, though some inference tasks become more challenging.

A standard non-parametric model is "All distributions with finite variances". Although proving that this model cannot be parametrised is difficult, the key point is that it is too broad to be described by any parametric models.

In this manual, we will primarily focus on parametric models in the context of statistical inference but will also introduce a few simple non-parametric techniques, such as bootstrapping and chi-square goodness-of-fit tests.

2.2.1.3 Frequentist and Bayesian Inference

Under the parametric modelling framework, there are various approaches to perform statistical inference. The two most prominent approaches are **frequentist inference** and **Bayesian inference**.

Suppose for a set of data X, we have assumed a parametric model $\mathcal{M} = \{f(x|\theta) : \theta \in \Theta\}$. The fundamental difference between frequentist and Bayesian inference is straightforward:

- Frequentist: Treat the parameter θ as an unknown constant. The goal is to estimate θ and infer the distribution $f_\theta(x)$.
- Bayesian: Treat the parameter θ as a random variable. Use the data to update the distribution of θ.

The debate between frequentist and Bayesian methods has persisted for a long time. Frequentist techniques often require large samples of identical random experiments, which may not always be feasible, but they work well in many scenarios. In the next chapters, we will delve into three main areas of frequentist inference:

- Point estimation: Use a number $\hat{\theta}$ to estimate θ.
- Interval estimation: Use an interval $(L_\theta(\boldsymbol{X}), U_\theta(\boldsymbol{X}))$ to estimate θ, considering the estimating uncertainty.
- Hypothesis testing: Given a hypothesis H_0 on θ or the model, use the data $\boldsymbol{X}$ to test whether the hypothesis is likely to be true.

Conversely, Bayesian methods do not rely on large samples and resemble human reasoning more closely. However, their effectiveness depends on the appropriate choice of priors and computational challenges. We will explore Bayesian inference in Chapter 4 and discuss its applications in credibility theory.

2.2.1.4 Introduction to Point Estimation

In this section, we will study **point estimation**, perhaps one of the most intuitive and simplest form of statistical inference. As the name suggests, point estimation uses a single number to estimate each parameter in a model.

Throughout the discussion, we assume that the data $\boldsymbol{X} = (X_1, \cdots, X_n)$ are i.i.d. samples from a statistical model $f(x|\theta)$. Recall that $f_\theta(x)$ represents a family of distributions parametrised by θ, therefore, once we derive an appropriate value of θ from the data $\boldsymbol{X}$, we can specify the distribution exactly and make further probabilistic calculations.

Formally, a **point estimator** of the parameter θ is a function of the data $\boldsymbol{X}$:

$$\hat{\theta}(\boldsymbol{X}) = \hat{\theta}(X_1, \cdots, X_n). \tag{2.2.2}$$

Although $\theta(\boldsymbol{X})$ is a scalar, it is a random variable, depending on the random sample $\boldsymbol{X}$. That is, given different observations, we will get different estimate of $\hat{\theta}$. To make the distinction clearer, suppose $\boldsymbol{x} = (x_1, \cdots, x_n)$ is a realisation of $\boldsymbol{X} = (X_1, \cdots, X_n)$, which means $\boldsymbol{x}$ are non-random constants. Then, a **point estimate** of the parameter θ is a function of the particular observation $\boldsymbol{x}$:

$$\hat{\theta}(\boldsymbol{x}) = \hat{\theta}(x_1, \cdots, x_n), \tag{2.2.3}$$

where $\hat{\theta}(\boldsymbol{x})$ is now a non-random constant. In practice or the exam, it is unlikely that you need to distinguish these two concepts very rigorously, since we often only have one set of data and simply need to calculate a single estimate. For simplicity, we usually use $\hat{\theta}$ to denote the parameter estimate (or estimator) of θ.

2.2.2 Method of Moments Estimation

Many common distributions have their means as one of the parameters, for example, μ in $\mathcal{N}(\mu, \sigma^2)$ and λ in $\text{Poi}(\lambda)$. Therefore, a natural way to estimate the parameter θ is to match the sample mean to the population mean (either θ itself or a function of θ). Since the mean is also called the first moment of a random variable, this method is called the **method of moments estimation** (MME), in the case of only one parameter.

We use some simple examples to demonstrate the procedure of the MME.

Example 2.2.1 (Poisson MME). Suppose $X_1, \cdots, X_n \overset{\text{iid}}{\sim} \text{Poi}(\lambda)$. Derive the MME $\hat{\lambda}$ for λ.[iii]

Solution. For a Poisson random variable $X \sim \text{Poi}(\lambda)$, it has mean $\mathbb{E}[X] = \lambda$. The MME for λ is obtained by simply matching the population mean λ to the sample mean $\bar{X}$:

$$\hat{\lambda} = \bar{X}.$$

For most distributions, the population mean is not the parameter itself, but a function of the parameter. In these cases, we can match the sample mean $\bar{X}$ to the population mean (a function of θ) to solve for $\hat{\theta}$.

[iii] In some other texts, people might prefer to use the notation $\tilde{\theta}$ to denote method of moments estimator and use $\hat{\theta}$ to denote method of maximum likelihood estimator, to be introduced later in this chapter. In this manual, we do not make this distinction and use $\hat{\lambda}$ to denote a point estimator in general.

Example 2.2.2 (Geometric MME). Suppose $X_1, \cdots, X_n \overset{\text{iid}}{\sim} \text{Geo}(p)$, with the probability mass function:

$$p(x) = p(1-p)^{x-1}, \ x \geq 1.$$

Derive the MME $\hat{p}$ for p.

Solution. For a geometric random variable $X \sim \text{Geo}(p)$ starting from 1, we have shown that it has mean $\mathbb{E}[X] = 1/p$. The MME for p is obtained by matching the population mean $1/p$ to the sample mean $\bar{X}$:

$$\mathbb{E}[X] = \frac{1}{p} = \bar{X} \Longrightarrow \hat{p} = \frac{1}{\bar{X}}.$$

In general, the MME can be applied for distributions with more than one parameters. However, for more than two parameters, we often prefer the maximum likelihood estimation, and the corresponding MME is rarely used in practice or tested in the exam.

Here, we briefly discuss the case of two parameters. In this case, we need to match both the first moments and the second moments. By such, we obtain two equations, which normally allow us to solve for the two parameters. For clarity, we use $\hat{\mu}_1 = (X_1 + \cdots + X_n)/n$ and $\hat{\mu}_2 = (X_1^2 + \cdots + X_n^2)/n$ to denote the first two sample moments, respectively.

Example 2.2.3 (Normal MME). Suppose $X_1, \cdots, X_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$. Derive the MME for μ and σ^2.

Solution. For a normal random variable $X \sim \mathcal{N}(\mu, \sigma^2)$, it has mean $\mathbb{E}[X] = \mu$, and the second moment $\mathbb{E}[X^2] = \text{Var}(X) + (\mathbb{E}[X])^2 = \mu^2 + \sigma^2$. The MME for μ and σ^2 is obtained by matching the first two moments:

$$\mathbb{E}[X] = \mu = \hat{\mu}_1, \ \mathbb{E}[X^2] = \mu^2 + \sigma^2 = \hat{\mu}_2.$$

$$\Longrightarrow \hat{\mu} = \hat{\mu}_1, \ \hat{\sigma^2} = \hat{\mu}_2 - \hat{\mu}_1^2.$$

Example 2.2.4 (Gamma MME). Suppose $X_1, \cdots, X_n \overset{\text{iid}}{\sim} \text{Gamma}(\alpha, \lambda)$ with the probability density function:

$$f(x) = \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x}.$$

Derive the MME for α and λ.

Solution. For a Gamma random variable $X \sim \text{Gamma}(\alpha, \lambda)$, it has mean $\mathbb{E}[X] = \alpha/\lambda$, and the second moment

$$\mathbb{E}[X^2] = \text{Var}(X) + (\mathbb{E}[X])^2 = \frac{\alpha}{\lambda^2} + \frac{\alpha^2}{\lambda^2}.$$

The MME for α and λ is obtained by matching the first two moments:

$$\mathbb{E}[X] = \frac{\alpha}{\lambda} = \hat{\mu}_1,$$

$$\mathbb{E}[X^2] = \frac{\alpha}{\lambda^2} + \frac{\alpha^2}{\lambda^2} = \hat{\mu}_2.$$

To solve this system of equations, we first rearrange the first equation to $\alpha = \lambda\hat{\mu}_1$ and substitute into the second equation:

$$\frac{\lambda\hat{\mu}_1 + \lambda^2\hat{\mu}_1^2}{\lambda^2} = \hat{\mu}_2 \Longrightarrow \hat{\lambda} = \frac{\hat{\mu}_1}{\hat{\mu}_2^2 - \hat{\mu}_1^2}.$$

Substituting this back to the first equation, we obtain:

$$\hat{\alpha} = \hat{\lambda}\hat{\mu}_1 = \frac{\hat{\mu}_1^2}{\hat{\mu}_2^2 - \hat{\mu}_1^2}.$$

It should be noted that the MME have various limitations and are rarely used in practice or tested in the exam. Although it can give reasonable estimators for some simple distributions we have seen before, it is challenging to generalise this approach to more complex probabilistic models, such as regression or more advanced neural networks. Here we use two single-parameter examples to demonstrate some scenarios that the MME can fail to work:

- The population mean does not depend on the parameter:

  Suppose $X_1, \cdots, X_n \overset{\text{iid}}{\sim} U(-\theta, \theta)$. Here, $\mathbb{E}[X] = 0$, which is independent of the parameter θ. In this case, no matter what samples are drawn, the MME is always 0 and is not a proper estimator.

- The population mean does not exist:

  Suppose $X_1, \cdots, X_n$ are i.i.d. samples from a distribution with the probability density function:

  $$f(x) = \frac{1}{x^2}, \ x \geq 1.$$

  You can easily check this is a valid density function by $\int_1^\infty 1/x^2 dx = (-1/x)|_1^\infty = 1$, but the expected value is infinity:

  $$\mathbb{E}[X] = \int_1^\infty x \cdot \frac{1}{x^2} dx = \log x \Big|_1^\infty = \infty.$$

  In this case, the MME approach fails.

2.2.3 Maximum Likelihood Estimation

2.2.3.1 General Framework

A more desirable method of point estimation is to utilise the full information of the underlying distributions, instead of only a finite number of moments as in the MME. In this subsection, we will discuss one such more general point estimating method, the **maximum likelihood estimation**, known as the MLE. The MLE is so popular and dominant that it is basically the only point estimating method used in a wide range of areas in statistics and modern machine learning.

The idea of MLE is actually quite simple. Say we have i.i.d. samples $\boldsymbol{X} = (X_1, \cdots, X_n)$. Assume that all data are sampled from a certain type (or family) of distributions $f_\theta(x)$, where θ is the undetermined parameter. Our goal is to find the best point estimator $\hat{\theta}$ so that the model $f_{\hat{\theta}}(x)$ assigns the highest likelihood to the observed sample $\boldsymbol{X}$, among all possible θ for $f_\theta(x)$.

A simple example will make this point clearer. Suppose you have a potentially unfair coin with head probability p. You tossed the coin 10 times and got X heads, where the number of heads X is modelled by a binomial distribution $\text{Bin}(10, p)$. When you got 5 heads, that is, $X = 5$, the coin seems fair and you might guess $p = 0.5$; by contrast, $p = 0.1$ is not convincing as it is unlikely to obtain 5 heads out of 10 tosses given such a low head probability. We can also perform simple calculations to make this point concrete:

$$\mathbb{P}(X = 5 | p = 0.5) = \binom{10}{5} \cdot 0.5^{10} = 0.246, \quad \mathbb{P}(X = 5 | p = 0.1) = \binom{10}{5} \cdot 0.1^5 \cdot 0.9^5 = 0.001.$$

To formalise the procedure, we first need to define the terminology "**likelihood** function" for a set of observations. We focus on the case of single parameter, and will brief discuss the multi-parameter case towards the end of this part.

Definition 2.2.1 (Likelihood function)**.** Let $\boldsymbol{X} = (X_1, \cdots, X_n)$ be i.i.d. samples. Then, the likelihood function is defined as:

$$\mathcal{L}(\theta) = \prod_{i=1}^{n} f(x_i|\theta), \tag{2.2.4}$$

where $f(x|\theta)$ is the probability density function if X is continuous or the probability mass function if X is discrete.

We make several remarks on this definition.

- It should be noted that **likelihood function** $\mathcal{L}(\theta)$ is a function of the parameter θ rather than the data $\boldsymbol{X}$. This is because when evaluating the likelihood, we assume that the data is fixed but the parameter is yet to be determined (not random, but an unknown constant). The goal of the MLE is to use this function to choose the best parameter θ for a fixed set of data.
- The definition $\mathcal{L}(\theta) = \prod_{i=1}^{n} f(x_i|\theta)$ is a product of individual contributions to the total likelihood. This representation replies on the assumption that all samples are independent. Without the independence assumption, we can only define the likelihood function as $\mathcal{L}(\theta) = f(\boldsymbol{x}|\theta)$, where here f is the joint distribution of $\boldsymbol{X}$. Throughout this chapter, we always assume the independence unless explicitly stated otherwise. The non-independent scenario is much more complicated and beyond the scope of the CS1 syllabus.
- Technically speaking, the random variable $\boldsymbol{X}$ can be more complex and neither continuous or discrete. In this case, $f(x|\theta)$ is neither a density function nor a mass function. Consider a random variable X that has $1/2$ probability taking the value of 0, and has $1/2$ probability sampled from an exponential distribution $\text{Exp}(\lambda)$. Here, we can treat X as a "weighted mixture" of a discrete variable (a constant at 0) and a continuous variable (the exponential). In this case,

$$f(x|\theta) = \begin{cases} 0.5 & x = 0 \\ 0.5 \cdot \lambda e^{-\lambda x} & x > 0 \end{cases}$$

  In the CS1 exam, you will mostly encounter the purely discrete or continuous cases, but we mention the general setting here for completeness.

The MLE aims at finding the optimal value of θ (if it exists) to maximise the likelihood function $\mathcal{L}(\theta)$. However, it turns out that directly maximising $\mathcal{L}(\theta)$ is challenging in most cases. Therefore, in practice, we instead maximise the so-called **log-likelihood** function, simply defined as the logarithm of $\mathcal{L}(\theta)$:

Definition 2.2.2 (Log-likelihood function)**.** Let $\boldsymbol{X} = (X_1, \cdots, X_n)$ be i.i.d. samples. Then, the log-likelihood function is defined as:

$$\ell(\theta) = \log \mathcal{L}(\theta) = \sum_{i=1}^{n} \log f(x_i|\theta). \tag{2.2.5}$$

Because the log function is strictly increasing, a value maximises the likelihood if and only if it maximises the log-likelihood.

Why does taking the log significantly simplify the calculation of the MLE? Maximising the likelihood function $\mathcal{L}(\theta)$ directly is often difficult because it involves differentiating a product of multiple functions, which is computationally complex. Taking the logarithm to obtain the log-likelihood function $\ell(\theta) = \log \mathcal{L}(\theta)$ simplifies the calculations in two ways simultaneously:

- First, it transforms the product of individual likelihoods $\prod f(\cdot)$ into a sum of log-likelihoods $\sum \log f(\cdot)$, making differentiation much easier.
- Second, individual likelihoods are generally products of multiple terms, so taking the log of each likelihood also reduces these from a product form to a sum form, further simplifying the differentiation.

2.2.3.2 Examples

Without any doubts, the MLE is unarguably one of the most important statistical techniques. In the CS1 exam, it is almost surely that the MLE will be tested in multiple questions. Therefore, we use this dedicated subsection to discuss how to derive the MLE in many examples, ranging from the simplest standard distributions, to some more complicated cases.

We first state the standard procedure of performing an MLE in the single parameter cases:

1. For a set of i.i.d. samples $X_1, \cdots, X_n$, write down the likelihood function $\mathcal{L}(\theta)$ and corresponding log-likelihood function $\ell(\theta)$.
2. Find the first-order derivative $\ell'(\theta)$ and set it to 0.
3. Solve $\ell'(\theta) = 0$ and obtain the MLE $\hat{\theta}$.
4. Check that the second-order derivative at the MLE is negative $\ell''(\hat{\theta}) < 0$, to make sure that $\hat{\theta}$ is indeed a maximiser of $\ell(\theta)$.

It is worth noting that this is only the common strategy for maximising $\ell(\theta)$. There exist more complicated scenarios in which this procedure cannot work:

- $\ell(\theta)$ is not differentiable:

  One example of this is the so-called shifted Laplace distribution, with the density function:

  $$f(x) = \frac{1}{2} e^{-|x-\mu|}.$$

  The corresponding log-likelihood function $\ell(\mu)$ involves absolute values of μ, which is not differentiable. In cases like this, we have to use more advanced optimization techniques to find $\hat{\theta}$, which is beyond the scope of the CS1 syllabus.

- $\ell'(\theta) = 0$ has no solutions or multiple solutions:

  In these cases, it means that $\ell(\theta)$ has no or multiple stationary points. Further investigations of the second-order derivatives are required to study the monotonicity of $\ell(\theta)$ to find the global maximiser $\hat{\theta}$. We will later demonstrate this case in Example 2.2.9 using uniform distributions.

- $\ell'(\theta) = 0$ cannot be solved analytically:

  In the CS1, we mostly study cases where $\ell'(\theta) = 0$ can be solved explicitly and only has a unique solution. However, in practice, particularly in modern machine learning, there are very few cases that the MLE can be found analytically, and certain numerical optimization techniques are necessary.

There are a vast range of numerical methods to solve the MLE numerically. We will not cover any of these techniques as they are beyond the scope of the syllabus. Instead, we only briefly mention two important techniques for interested reader:

- ▷ **Newton-Raphson method**: A numerical method to find the roots of an equation, where $\ell'(\theta) = 0$ is our target equation here. This method is commonly used in parameter estimation of the generalised linear models.
- ▷ **Expectation-Maximisation (EM) algorithm**: An numerical approach to find the MLE without directly solving $\ell'(\theta) = 0$. It iteratively finds an approximation $\tilde{\ell}(\theta)$ for $\ell(\theta)$ and solves $\tilde{\ell}'(\theta) = 0$. This method is particularly useful in clustering algorithms in machine learning.

We will start with some simple examples to demonstrate the standard procedure. You might notice that the MLE leads to the same estimator as the MME for many distributions, particular for those with a parameter directly linked to the mean.

Example 2.2.5 (Bernoulli & Binomial MLE). Suppose $X_1, \cdots, X_n \overset{\text{iid}}{\sim} \text{Bernoulli}(p)$. Derive the MLE of p.

Solution.

$$\begin{aligned}
\mathcal{L}(p) &= \prod_{i=1}^{n} p^{X_i}(1-p)^{1-X_i} \\
\implies \ell(p) &= \sum_{i=1}^{n} [X_i \log p + (1 - X_i)\log(1-p)] \\
&= \log p \cdot \sum_{i=1}^{n} X_i + \log(1-p) \cdot \left(n - \sum_{i=1}^{n} X_i\right). \\
\implies \ell'(p) &= \frac{1}{p}\sum_{i=1}^{n} X_i - \frac{1}{1-p}\left(n - \sum_{i=1}^{n} X_i\right) = 0 \\
\implies (1-p)\sum_{i=1}^{n} X_i &= p\left(n - \sum_{i=1}^{n} X_i\right) \\
\implies \hat{p} &= \frac{\sum_{i=1}^{n} X_i}{n} = \bar{X}.
\end{aligned}$$

We finally check the second-order derivative:

$$\implies \ell''(\hat{p}) = \underbrace{-\frac{\sum_{i=1}^{n} X_i}{\hat{p}^2}}_{<0} \underbrace{-\frac{n - \sum_{i=1}^{n} X_i}{(1-\hat{p}^2)}}_{<0} < 0.$$

Therefore, the MLE of p is the sample mean $\hat{p} = \bar{X}$.

It is worth noting that this question is equivalent to "For $X \sim \text{Bin}(n, p)$, find the MLE of p", since the sum of n i.i.d. Bernoulli variables is a binomial variable.

Example 2.2.6 (Geometric MLE). Suppose $X_1, \cdots, X_n \overset{\text{iid}}{\sim} \text{Geo}(p)$, with the probability mass function:

$$p(x) = p(1-p)^{x-1}, \ x \geq 1.$$

Derive the MLE of p.

Solution.

$$\begin{aligned}
\mathcal{L}(p) &= \prod_{i=1}^{n} p(1-p)^{X_i-1} \\
\Longrightarrow \ell(p) &= \sum_{i=1}^{n} \left[\log p + (X_i - 1)\log(1-p)\right] \\
&= n\log p - \log(1-p)\cdot\left(\sum_{i=1}^{n} X_i - n\right) \\
\Longrightarrow \ell'(p) &= \frac{n}{p} - \frac{\sum_{i=1}^{n} X_i - n}{1-p} = 0 \\
\Longrightarrow (1-p)n &= p\sum_{i=1}^{n} X_i - pn \\
\Longrightarrow \hat{p} &= \frac{n}{\sum_{i=1}^{n} X_i} = \frac{1}{\bar{X}}.
\end{aligned}$$

We finally check the second-order derivative:

$$\Longrightarrow \ell''(\hat{p}) = \underbrace{-\frac{n}{\hat{p}^2}}_{<0} \underbrace{-\frac{\sum_{i=1}^{n} X_i - n}{(1-\hat{p}^2)}}_{<0} < 0.$$

Therefore, the MLE of p is the inverse of the sample mean $\hat{p} = 1/\bar{X}$.

Example 2.2.7 (Poisson MLE). Suppose $X_1, \cdots, X_n \overset{\text{iid}}{\sim} \text{Poi}(\lambda)$. Derive the MLE of λ.

Solution.

$$\begin{aligned}
\mathcal{L}(\lambda) &= \prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{X_i}}{X_i!} \\
\Longrightarrow \ell(\lambda) &= \sum_{i=1}^{n} \left(-\lambda + X_i \log\lambda - \log X_i!\right).
\end{aligned}$$

Since $\ell(\lambda)$ is a function of λ, we can treat any terms that do not involve λ as constants, and these terms will become zero during the differentiation. This trick can sometimes heavily simplify the calculation and will be frequently used in many places.

$$\begin{aligned}
\Longrightarrow \ell(\lambda) &= -n\lambda + \log\lambda \cdot \sum_{i=1}^{n} X_i + \text{const.} \\
\Longrightarrow \ell'(\lambda) &= -n + \frac{\sum_{i=1}^{n} X_i}{\lambda} = 0 \\
\Longrightarrow \hat{\lambda} &= \frac{\sum_{i=1}^{n} X_i}{n} = \bar{X}.
\end{aligned}$$

We finally check the second-order derivative:

$$\Longrightarrow \ell''(\hat{\lambda}) = -\frac{\sum_{i=1}^{n} X_i}{\hat{\lambda}^2} < 0.$$

Therefore, the MLE of λ is the sample mean $\hat{\lambda} = \bar{X}$.

Example 2.2.8 (Exponential MLE). Suppose $X_1, \cdots, X_n \overset{\text{iid}}{\sim} \text{Exp}(\lambda)$, with the probability density function:

$$f(x) = \lambda e^{-\lambda x}.$$

Derive the MLE of λ.

Solution.

$$\begin{aligned}
\mathcal{L}(\lambda) &= \prod_{i=1}^{n} \lambda e^{-\lambda X_i} \\
\implies \ell(\lambda) &= \sum_{i=1}^{n} (\log \lambda - \lambda X_i) = n \log \lambda - \lambda \sum_{i=1}^{n} X_i \\
\implies \ell'(\lambda) &= \frac{n}{\lambda} - \sum_{i=1}^{n} X_i = 0 \\
\implies \hat{\lambda} &= \frac{n}{\sum_{i=1}^{n} X_i} = \frac{1}{\bar{X}}.
\end{aligned}$$

We finally check the second-order derivative:

$$\implies \ell''(\hat{\lambda}) = -\frac{n}{\hat{\lambda}^2} < 0.$$

Therefore, the MLE of λ is the inverse of the sample mean $\hat{\lambda} = 1/\bar{X}$.

We next look at the uniform distributions, which relies on a different approach to find the MLE.

Example 2.2.9 (Uniform MLE). Suppose $X_1, \cdots, X_n \overset{\text{iid}}{\sim} U(0, \theta)$. Derive the MLE of θ.

Solution.

$$\begin{aligned}
\mathcal{L}(\theta) &= \prod_{i=1}^{n} \frac{1}{\theta} = \theta^{-n} \\
\implies \ell(\theta) &= -n \log \theta.
\end{aligned}$$

Without further calculations, we can observe that $\ell(\theta)$ is decreasing with θ. Therefore, $\ell'(\theta)$ is always negative and $\ell'(\theta) = 0$ has no solutions.

In this case, to maximise $\ell(\theta) = -n \log \theta$, we need to minimise $\log \theta$ and thus minimise θ. Notice that $0 \leq X_i \leq \theta$ for all $i = 1, \cdots, n$, we must have $\theta \geq X_1, \cdots, \theta \geq X_n$. This implies that $\theta \geq \max(X_1, \cdots, X_n)$[iv]. Therefore, the smallest possible value of θ is $\max(X_1, \cdots, X_n)$.

Therefore, the MLE of θ is $\hat{\theta} = \max(X_1, \cdots, X_n)$.

For distributions with two parameters θ_1, θ_2, finding the MLE is essentially the same as the single-parameter cases. Instead of finding the first-order derivative $\ell'(\theta)$, we now need to calculate the partial derivatives $\partial \ell / \partial \theta_1$ and $\partial \ell / \partial \theta_2$ and set both to 0. We use the most important two-parameter distribution, the normal distribution, to illustrate.

[iv] The quantity $\max(X_1, \cdots, X_n)$ is also called the largest order statistic of $X_1, \cdots, X_n$, denoted as $X_{(n)}$. The theory of order statistics is reviewed in the appendix of this manual.

Example 2.2.10 (Normal MLE). Suppose $X_1, \cdots, X_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$. Derive the MLE of μ and σ^2.

Solution.

$$\begin{aligned}
\mathcal{L}(\mu, \sigma^2) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma^2} e^{-\frac{(X_i-\mu)^2}{2\sigma^2}} \\
\implies \ell(\mu, \sigma^2) &= \sum_{i=1}^{n} \left[-\log\sqrt{2\pi} - \log\sigma^2 - \frac{(X_i-\mu)^2}{2\sigma^2} \right] \\
&= -n\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i-\mu)^2 + \text{const.} \\
\implies \frac{\partial \ell}{\partial \mu} &= \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\mu) = 0, \quad \frac{\partial \ell}{\partial \sigma^2} = -\frac{n}{\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(X_i-\mu)^2 = 0.
\end{aligned}$$

Solve this system of equations, we have:

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X}, \quad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2,$$

which give us the desired MLE of μ and σ^2. Note that $\hat{\sigma}^2$ is slightly different from the definition of the sample variance S^2, as denominator is n instead of $n-1$.

For the normal distributions, It can be shown that $\hat{\mu}$ and $\hat{\sigma}^2$ indeed maximise $\ell(\mu, \sigma^2)$. However, checking the optimality of the MLE in the two-parameter cases is more technical and requires knowledge of multivariable calculus. Instead of requiring the second-order derivative being negative, now we need to check that the **Hessian matrix** is negative definite. This is beyond the scope of the CS1 syllabus.

2.2.4 Evaluating Point Estimators

A point estimator is a random variable, and its realised value depends on the data observed. Consider a parameter θ in a continuous distribution and a corresponding point estimator $\hat{\theta}(\boldsymbol{X})$. We have $\mathbb{P}(\hat{\theta}(\boldsymbol{X}) = \theta) = 0$ since $\hat{\theta}(\boldsymbol{X})$ has a continuous distribution. This says, we will *almost surely*[v] obtain an estimate $\hat{\theta}$ different from its true value θ.

This fact does not mean point estimators are useless. For example, if an estimator gives us values $\hat{\theta}$ quite close to the true parameter θ, even under varying the observed data, we will probably argue that this is a decent estimator. Therefore, in the part, we will discuss some commonly used concepts for evaluating the quality of point estimators.

Definition 2.2.3 (Bias and Unbiasedness of Estimators). Let $\hat{\theta}$ be a point estimator of a parameter θ. Then, the **bias** of $\hat{\theta}$ is defined as:

$$\text{Bias}(\hat{\theta}) = \mathbb{E}[\hat{\theta}] - \theta. \tag{2.2.6}$$

If the bias is 0, that is, $\mathbb{E}[\hat{\theta}] = \theta$, we call that $\hat{\theta}$ is an **unbiased estimator**.

If an estimator is unbiased, it tells us that if we repeat the experiment infinitely many times, on average, the estimate $\hat{\theta}$ will be equal to the true value θ. Many common estimators are actually unbiased:

[v] In advanced probability theory, the word "almost surely" specifically means that an event A will happen with probability one, that is, $\mathbb{P}(A) = 1$.

- $\hat{\mu} = \bar{X}$ for $X \sim \mathcal{N}(\mu, \sigma^2)$. $\hat{\mu}$ is unbiased as $\mathbb{E}[\hat{\mu}] = \mathbb{E}[\bar{X}] = \mu$.
- $\hat{\lambda} = \bar{X}$ for $X \sim \text{Poi}(\lambda)$. $\hat{\lambda}$ is unbiased as $\mathbb{E}[\hat{\lambda}] = \mathbb{E}[\bar{X}] = \lambda$.

Unbiasedness is a good property for an estimator, but it does not tell the full story. Another measurement of quality is called the **variance**:

Definition 2.2.4 (Variance and Standard Error of Estimators)**.** Let $\hat{\theta}$ be a point estimator of a parameter θ. Then, the **variance** of $\hat{\theta}$ is defined as:

$$\text{Var}(\hat{\theta}) = \mathbb{E}[(\hat{\theta} - \mathbb{E}[\hat{\theta}])^2], \tag{2.2.7}$$

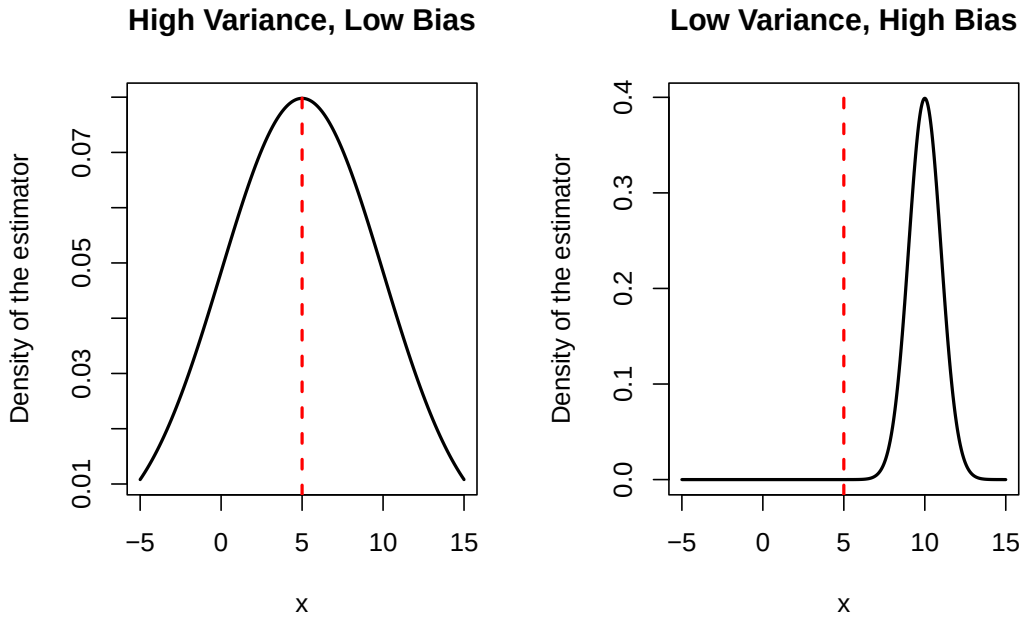
and the **standard error** of $\hat{\theta}$ is defined as:

$$\text{se}(\hat{\theta}) = \sqrt{\text{Var}(\hat{\theta})}. \tag{2.2.8}$$

Either a high bias or high variance might result in unsatisfactory estimating performance. The following two scenarios demonstrate the trade-off:

- Low bias but high variance: On average, $\hat{\theta}$ will be close to θ, but the estimate for a single experiment is not reliable as it can heavily deviate from θ. For example, if $\theta = 5$, you might expect the realisations of $\hat{\theta}$ to be like $12, 4, 2, 10, 8, \cdots$.
- Low variance but high bias: The realisation of $\hat{\theta}$ will have small variations for different samples, but on average, $\hat{\theta}$ is very different from θ. For the same example if $\theta = 5$, you might expect the realisations of $\hat{\theta}$ to be like $10, 9, 10, 11, 10, \cdots$.

The pictures below illustrate the two scenarios, where the black lines represent the density function of the estimator $\hat{\theta}$, and the red dotted lines represent the true value of θ, which is set to 5 here.



Therefore, to find a desirable estimator, simply minimising one of the bias and variance is not sufficient. Instead, we aim at obtain an estimator with a low bias and a low variance. This motivates the concept of **mean square error**:

Definition 2.2.5 (Mean Square Error)**.** Let $\hat{\theta}$ be a point estimator of a parameter θ. Then, the **mean square error** (MSE) of $\hat{\theta}$ is defined as:

$$\text{MSE}(\hat{\theta}) = \mathbb{E}[(\hat{\theta} - \theta)^2]. \tag{2.2.9}$$

It turns out that the MSE integrates the bias and variance of an estimator, as we will show now.

Theorem 2.2.1. Let $\hat{\theta}$ be a point estimator of a parameter θ. Then, its MSE can be decomposed as the square of the sum of the bias and variance:

$$\text{MSE}(\hat{\theta}) = (\text{Bias}(\hat{\theta}))^2 + \text{Var}(\hat{\theta}). \tag{2.2.10}$$

Proof of Theorem 2.2.1.

$$\text{MSE}(\hat{\theta}) = \mathbb{E}[(\hat{\theta}-\theta)^2] = \mathbb{E}[(\hat{\theta}-\mathbb{E}[\hat{\theta}])+(\mathbb{E}[\hat{\theta}]-\theta)]^2 = (\text{Bias}(\hat{\theta}))^2+\text{Var}(\hat{\theta})+2\mathbb{E}[(\hat{\theta}-\mathbb{E}[\hat{\theta}])(\mathbb{E}[\hat{\theta}]-\theta)].$$

The cross term can be shown to be 0:

$$\mathbb{E}[(\hat{\theta} - \mathbb{E}[\hat{\theta}])\underbrace{(\mathbb{E}[\hat{\theta}] - \theta)}_{\text{constant}}] = (\mathbb{E}[\hat{\theta}] - \theta) \cdot \underbrace{\mathbb{E}[\hat{\theta} - \mathbb{E}[\hat{\theta}]]}_{=\mathbb{E}[\hat{\theta}]-\mathbb{E}[\hat{\theta}]=0} = 0,$$

and this concludes the proof. □

In practice, the MSE is perhaps the most commonly used criterion to evaluate the quality of point estimators. In general, an estimator with a **lower MSE** is desired. For a parameter θ, an estimator $\hat{\theta}_1$ is said to be **more efficient** than another estimator $\hat{\theta}_2$ if $\text{MSE}(\hat{\theta}_1) < \text{MSE}(\hat{\theta}_2)$.

It should be noted that although unbiased estimators are commonly found in practice, they do not necessarily have the lowest MSE. In some cases, introducing a small bias can lead to a big variance decrease, giving an overall improvement of the MSE. This can be demonstrated in the following example:

Example 2.2.11 (Comparing estimators for the normal variance)**.** Suppose $X_1, \cdots, X_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$. We have two plausible estimators for σ^2:

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2, \quad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2,$$

where S^2 is the sample variance, and $\hat{\sigma}^2$ is the MLE or MME of σ^2.

(a) Compare the bias of S^2 and $\hat{\sigma}^2$.

(b) Compare the MSE of S^2 and $\hat{\sigma}^2$. Comment on your findings.

Solution.

(a) We have shown that $\mathbb{E}[S^2] = \sigma^2$, thus S^2 is an unbiased estimator of σ^2. This implies that $\hat{\sigma}^2$ is biased, with the bias equal to

$$\text{Bias}(\hat{\sigma}^2) = \mathbb{E}[\hat{\sigma}^2] - \sigma^2 = \mathbb{E}\left[\frac{n-1}{n} \cdot S^2\right] - \sigma^2 = -\frac{1}{n}\sigma^2.$$

(b) We first calculate the variances of both estimators. For S^2, we know that:

$$\text{Var}(S^2) = \frac{2\sigma^4}{n-1}$$

Using this result, we have:

$$\text{Var}(\hat{\sigma}^2) = \text{Var}\left(\frac{n-1}{n} \cdot S^2\right) = \frac{(n-1)^2}{n^2} \cdot \frac{2\sigma^4}{n-1} = \frac{2(n-1)\sigma^4}{n^2}$$

Therefore, the MSE can be calculated as:

$$\text{MSE}(S^2) = (\text{Bias}(S^2))^2 + \text{Var}(S^2) = 0 + \frac{2\sigma^4}{n-1} = \frac{2\sigma^4}{n-1}$$
$$\text{MSE}(\hat{\sigma}^2) = (\text{Bias}(\hat{\sigma}^2))^2 + \text{Var}(\hat{\sigma}^2) = \left(-\frac{\sigma^2}{n}\right)^2 + \frac{2(n-1)\sigma^4}{n^2} = \frac{(2n-1)\sigma^4}{n^2}$$

Finally, we can compare the two MSE:

$$\text{MSE}(\hat{\sigma}^2) - \text{MSE}(S^2) = \left[\frac{(2n-1)}{n^2} - \frac{2}{n-1}\right] \cdot \sigma^4 = \frac{1-3n}{n^2(n-1)} \cdot \sigma^4 < 0$$

This says, the MLE (or MME) $\hat{\sigma}^2$ has lower MSE than the unbiased estimator S^2, thus being more efficient. By trading off the bias and variance, we achieve an improvement of the MSE.

2.2.5 Introduction to Bootstrap Methods

We have studied the evaluation of point estimators, including measures like bias, variance, and the standard error of estimators for some common distributions. For example, the standard error of the sample mean for n i.i.d. normal data is easily calculated as $\text{se}(\bar{X}) = \sigma/\sqrt{n}$. However, in many practical scenarios:

- **No closed-form solutions exist**: For example, the standard error of the sample median has no simple formula.
- **The underlying population distribution is unknown**: Traditional methods often rely on assumptions about the distribution, which may not hold in real-world scenarios.

The bootstrap method provides a powerful and flexible alternative for addressing these challenges. The key idea is to approximate the sampling distribution of an estimator $\hat{\theta}$ by **resampling** from the observed data, treating it as representative of the population. In practice, we are often limited to a single sample of data, but resampling allows us to artificially generate multiple datasets from the original sample. By calculating $\hat{\theta}$ for each resampled dataset, we can approximate the sampling distribution of $\hat{\theta}$ as if repeated random samples had been drawn from the population.

Let $\boldsymbol{X} = \{X_1, \cdots, X_n\}$ represent the observed data, and let $\hat{\theta}$ denote an estimator of interest (e.g., sample mean, median, or variance). Two primary types of bootstrapping are commonly used:

- **Non-parametric bootstrap**: The non-parametric bootstrap involves directly resampling the observed data with replacement, making no assumptions about the population

distribution. This method is particularly useful when we have no prior knowledge of the underlying distribution or when the distribution is complex and difficult to model parametrically.

Procedure:

1. Generate a bootstrap sample $\boldsymbol{X}^* = \{X_1^*, \cdots, X_n^*\}$ by sampling n observations from $\boldsymbol{X}$.
2. Compute the estimator $\hat{\theta}^*$ for each bootstrap sample.
3. Repeat this process B times to obtain $\{\hat{\theta}_1^*, \cdots, \hat{\theta}_n^*\}$, where B is often chosen to be a large constant, for example, $B = 1,000$.

To understand why this works, consider the **empirical distribution function** (EDF) of the observed data, $F_n(x)$, defined as:

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}(X_i < x) \tag{2.2.11}$$

The non-parametric bootstrap can be thought of as sampling directly from $F_n(x)$. It turns out that $F_n(x)$ is a good estimator of the true population CDF $F(x)$. Technically speaking, this approximation is guaranteed by the *Glivenko-Cantelli theorem*, which states that the EDF converges uniformly to the true distribution function almost surely. We omit the detail here.

- **Parametric bootstrap**: This method assumes that the population follows a specific parametric distribution $F_\theta(x)$ such as normal or Poisson. This requires additional distributional information about $F(x)$ other than that contained in the observations.

  Procedure:

  1. Fit the parametric model F_θ to the observed data $\boldsymbol{X}$ and estimating its parameter θ (using methods such as MME or MLE).
  2. Generate a bootstrap sample $\boldsymbol{X}^*$ by simulating n observations from the fitted model $F_{\hat{\theta}}$.
  3. Compute the estimator $\hat{\theta}^*$ for each bootstrap sample.
  4. Same as the non-parametric bootstrap, repeat this process B times to obtain $\{\hat{\theta}_1^*, \cdots, \hat{\theta}_B^*\}$.

Once the empirical distribution of the estimator $\hat{\theta}$ is obtained, it can be used for inference. The sample mean and standard deviation and of the bootstrap estimates provides an approximation of the mean and standard error, respectively:

$$\mathbb{E}[\hat{\theta}] \approx \bar{\theta^*} = \frac{1}{B}\sum_{i=1}^{B} \hat{\theta}_i^* \tag{2.2.12}$$

$$\text{se}(\hat{\theta}) \approx \sqrt{\frac{1}{B-1}\sum_{i=1}^{B} (\hat{\theta}_i^* - \bar{\theta^*})^2} \tag{2.2.13}$$

The bootstrap method can also be used to construct confidence intervals, particularly in situations where classical approaches are not applicable. Traditional methods for constructing

confidence intervals often rely on strong distributional assumptions, such as normality. In addition, classical methods may perform poorly when the sample size is small, as the theoretical properties of many estimators rely on asymptotic approximations such as the Central Limit Theorem. The bootstrap offers a flexible alternative by leveraging the empirical distribution obtained through resampling to approximate the sampling variability of the estimator directly.

One commonly used bootstrap approach for confidence interval construction is the **percentile method**. This method defines a $1-\alpha$ confidence interval using the $(\alpha/2)$-th and $(1-\alpha/2)$-th percentiles of the bootstrap estimates. Intuitively, these percentiles represent the range within which the parameter of interest is most likely to lie, based on the variability observed in the bootstrap samples. Detailed discussions on confidence interval construction, including practical examples and comparisons of methods, will follow in later sections.

2.2.6 Theoretical Properties of Maximum Likelihood Estimators

Maximum likelihood estimation is perhaps the most commonly used point estimating method in various scenarios. In this subsection, we will briefly discuss some of the nice properties of MLE. Most of the results require rigorous mathematical arguments to prove in full, so we will focus on the intuition and implications.

As usual, suppose we have a parameter of interest θ, and use $\hat{\theta}$ to denote its MLE. Under some technical conditions[vi], the MLE $\hat{\theta}$ has the following properties:

1. The MLE is **consistent**:
$$\hat{\theta} \xrightarrow{p} \theta$$
Here the notation $\xrightarrow{p}$ means "*converges in probability*". The precise definition of convergence in probability[vii] is beyond the scope of the CS1 syllabus. Intuitively speaking, it simply says that the MLE $\hat{\theta}$ will be very close to the true parameter value θ as $n \to \infty$.

2. The MLE is **invariant**:
$$\text{If } \hat{\theta} \text{ is the MLE of } \theta, \text{ then } g(\hat{\theta}) \text{ is the MLE of } g(\theta).$$
This property is particularly useful when we want to estimate some derived quantities, rather than simply the parameter itself. Consider the example of $X \sim \text{Bin}(n, p)$, here we imagine X as the number of wins in n games, where p is the winning probability for each single game. Suppose we want to use MLE to estimate the probability $\mathbb{P}(\text{Win three times in three games}) = p^3$. Since we know the MLE of p is $\hat{p} = X/n$. Then, the MLE of p^3 can be immediately calculated as $\hat{p}^3 = X^3/n^3$.

3. The MLE is **asymptotically normal**:
$$\hat{\theta} \overset{a}{\sim} \mathcal{N}(\theta, \text{CRLB})$$
The asymptotic variance "CRLB" is called the **Cramér-Rao Lower Bound**. This is a lower bound on the variance that any **unbiased** estimators can attain.

In general, since an estimator is a function (likely to be non-linear) of the data, its distribution can be very challenging to determine. The asymptotic normality guarantees that

[vi] We will not worry about the precise technical conditions here. Interested reader can refer to advanced probability theory or real analysis books. The only condition we need to keep in mind is that the range of the distribution should not involve the parameters. One such distribution is the uniform distribution $U(0, \theta)$.

[vii] For interested readers, a sequence of random variables $\{X_n\}$ converges to a random variable X in probability if for all $\varepsilon > 0$, $\mathbb{P}(|X_n - X| > \varepsilon) \xrightarrow{n\to\infty} 0$. Note that the random variable X can also be a constant, like θ in the definition of consistency.

any MLE (again, under some technical conditions) will have a normal distribution as the sample size $n \to \infty$.

4. The MLE is **asymptotically efficient** or **asymptotically optimal**:

   This is a direct consequence of the consistency and asymptotic normality described above. An estimator is called asymptotically efficient or asymptotically optimal if it is asymptotically unbiased, consistent, and has the limiting variance equal to the CRLB.

   This property is more like a concluding result. Roughly speaking, it says that among all well-behaved estimators (asymptotically unbiased and consistent), the MLE has the smallest possible variance.

We now discuss the CRLB in more detail. Say we only care about the unbiased estimators. In this case, minimising the MSE is equivalent to minimising the variance. The CRLB exactly represents the "best performance" that an estimator can achieve. More precisely, the result can described as follows:

Theorem 2.2.2. If $\hat{\theta}$ is an unbiased estimator of θ, then

$$\text{Var}(\hat{\theta}) \geq \text{CRLB}. \tag{2.2.14}$$

Until now, we have not discussed how to calculate the CRLB exactly. In most cases, the CRLB can be found analytically, as stated below:

Definition 2.2.6 (Cramér-Rao Lower Bound)**.** Let $\boldsymbol{X} = (X_1, \cdots, X_n)$ be i.i.d. samples from the distribution with density $f(x|\theta)$, where θ is the parameter of interest. Also, use $\ell(\theta) = \sum_{i=1}^{n} \log f(x_i|\theta)$ to denote the log-likelihood function. Then, the CRLB is defined as:

$$\text{CRLB} = \frac{1}{I(\theta)} \tag{2.2.15}$$

The denominator $I(\theta)$ is called the **Fisher information**, defined by the following equivalent ways (under some regularity conditions):

$$I(\theta) = \mathbb{E}[(\ell'(\theta))^2] = n\mathbb{E}\left[\left(\frac{\partial}{\partial\theta}\log f(X|\theta)\right)^2\right] \tag{2.2.16}$$

or

$$I(\theta) = -\mathbb{E}[\ell''(\theta)] = -n\mathbb{E}\left[\frac{\partial^2}{\partial\theta^2}\log f(X|\theta)\right], \tag{2.2.17}$$

where the expectation $\mathbb{E}[\cdot]$ is respect to the sample $\boldsymbol{X}$.

The proof of Theorem 2.2.2 and the equivalence between multiple expressions of $I(\theta)$ is beyond the scope of the exam. Here, we provide some intuition of the definition of CRLB and the Fisher information.

- **Fisher information**: Recall that when finding the MLE, we need to solve $\ell'(\theta) = 0$. If the data contain a lot of *information*, we should expect that the log-likelihood function $\ell(\theta)$ concentrates on the true value θ, so that we can locate the MLE with high certainty. This "concentration" implies that $|\ell'(\theta)|$ should be large for most of the values other than the true value θ (averaged over all possible samples $\boldsymbol{X}$), which means the Fisher information $I(\theta) = \mathbb{E}[(\ell'(\theta))^2]$ is large in this case.

  The intuition is similar for the second expression. Since more information means the "concentration" of $\ell(\theta)$ around θ, it implies that $\ell(\theta)$ has a spike. This further implies

that the curvature of $\ell(\theta)$, which is $\ell''(\theta)$, should be largely negative ($\ell'(\theta)$ decreases quickly). Therefore, the Fisher information $I(\theta) = -\mathbb{E}[\ell''(\theta)]$ is (positively) large in this case.

- **CRLB**: When the Fisher information $I(\theta)$ is large, we are more certain about the true value of θ, and thus the variance of the estimator is smaller. This is consistent with the expression of the CRLB, equal to $1/I(\theta)$.

The following two examples will help demonstrate the calculation of the CRLB and the asymptotic distribution of MLEs.

Example 2.2.12 (Bernoulli & Binomial CRLB). Suppose $X_1, \cdots, X_n \overset{\text{iid}}{\sim} \text{Bernoulli}(p)$.

(a) Derive the CRLB for unbiased estimators of p.

(b) Does the MLE $\hat{p} = \bar{X}$ attains the CRLB? Comment on the result.

Solution.

(a) $$f(X|p) = p^X(1-p)^{1-X}$$

$$\implies \log f(X|p) = X\log p + (1-X)\log(1-p)$$

$$\implies \frac{\partial}{\partial p}\log f(X|p) = \frac{X}{p} - \frac{1-X}{1-p} = \frac{X-p}{p(1-p)}$$

$$\implies I(p) = n\mathbb{E}\left[\left(\frac{\partial}{\partial p}\log f(X|p)\right)^2\right] = \frac{n}{p^2(1-p)^2}\cdot \underbrace{\mathbb{E}[X-p]^2}_{=\text{Var}(X)=p(1-p)} = \frac{n}{p(1-p)}$$

$$\implies \text{CRLB} = \frac{1}{I(p)} = \frac{p(1-p)}{n}$$

Note that the Fisher information can also be calculated using the second-order derivative:

$$\frac{\partial^2}{\partial p^2}\log f(X|p) = -\frac{X}{p^2} - \frac{1-X}{(1-p)^2} = -\frac{p^2 + (1-2p)X}{p^2(1-p^2)}$$

$$\implies I(p) = -n\mathbb{E}\left[\frac{\partial^2}{\partial p^2}\log f(X|p)\right] = n\cdot\frac{p^2+(1-2p)p}{p^2(1-p)^2} = \frac{n}{p(1-p)}$$

(b) Since

$$\text{Var}(\hat{p}) = \text{Var}(\bar{X}) = p(1-p)/n,$$

it attains the CRLB. It is worth noting that $\bar{X}$ attains the CRLB exactly, not asymptotically here.

Example 2.2.13 (Exponential CRLB). Suppose $X_1, \cdots, X_n \overset{\text{iid}}{\sim} \text{Exp}(\lambda)$.

(a) Derive the Fisher information $I(\lambda)$ contained in $X_1, \cdots, X_n$.

(b) Write down the asymptotic distribution of the MLE $\hat{\lambda} = 1/\bar{X}$.

(c) Assume that we observed $\bar{X} = 5$ and $n = 10$. Estimate the standard error of $\hat{\lambda}$.

Solution.

(a) $$f(X|\lambda) = \lambda e^{-\lambda X}$$
$$\Longrightarrow \log f(X|\lambda) = \log\lambda - \lambda X.$$
$$\Longrightarrow \frac{\partial}{\partial\lambda}\log f(X|\lambda) = \frac{1}{\lambda} - X.$$
$$\Longrightarrow I(\lambda) = n\mathbb{E}\left[\left(\frac{\partial}{\partial\lambda}\log f(X|\lambda)\right)^2\right] = n\cdot\underbrace{\mathbb{E}\left[X-\frac{1}{\lambda}\right]^2}_{=\mathrm{Var}(X)=\frac{1}{\lambda^2}} = \frac{n}{\lambda^2}.$$

Similarly, we can also find the Fisher information through the second-order derivative:

$$\frac{\partial^2}{\partial\lambda^2}\log f(X|\lambda) = -\frac{1}{\lambda^2}$$
$$\Longrightarrow I(\lambda) = -n\mathbb{E}\left[\frac{\partial^2}{\partial\lambda^2}\log f(X|\lambda)\right] = \frac{n}{\lambda^2}.$$

Notice that in this example, the second-order derivative does not involve the random variable X, so we can remove the expectation operation here.

(b) We first calculate the CRLB:

$$\mathrm{CRLB} = \frac{1}{I(\lambda)} = \frac{\lambda^2}{n}.$$

The asymptotic distribution of $\hat{\lambda} = 1/\bar{X}$ is then

$$\hat{\lambda} \overset{a}{\sim} \mathcal{N}(\lambda, \mathrm{CRLB}) = \mathcal{N}\left(\lambda, \frac{\lambda^2}{n}\right).$$

(c) From (b), we know that

$$\mathrm{se}(\hat{p}) \approx \sqrt{\mathrm{Var}(\hat{\lambda})} = \frac{\lambda}{\sqrt{n}}.$$

Substitute λ by the MLE $\hat{\lambda} = 1/\bar{X} = 1/5$, we have the estimated standard error equal to

$$\hat{\mathrm{se}}(\hat{\lambda}) = \frac{\hat{\lambda}}{\sqrt{n}} = \frac{1/5}{\sqrt{10}} \approx 0.063.$$

Note that, we use two different approximations in the calculations: 1) Use the CRLB to estimate $\mathrm{Var}(\hat{\lambda})$; 2) Use the MLE $\hat{\lambda}$ to estimate λ in the standard error.

Exercises

Exercise 2.2.1. Let $X_1, \cdots, X_n$ be random samples from the following probability density function:

$$f(x) = \frac{2(\theta - x)}{\theta^2},\ 0 < x < \theta,$$

where $\theta > 0$. Find the method of moment estimator of θ.

Solution. We first find the mean

$$\mathbb{E}[X] = \int_0^\theta \frac{2x(\theta - x)}{\theta^2} dx = \frac{1}{\theta^2}\left(\theta x^2 - \frac{2x^3}{3}\right)\Big|_0^\theta = \frac{\theta}{3}.$$

By matching the sample mean and theoretical mean, we have:

$$\bar{X} = \frac{\theta}{3} \Longrightarrow \hat{\theta} = 3\bar{X}.$$

Exercise 2.2.2. Let $X_1, \cdots, X_n$ be random samples from the following probability density function:

$$f(x) = \frac{1}{2\theta^3} x^2 e^{-x/\theta}, \ x > 0, \theta > 0.$$

Find the maximum likelihood estimator of θ.

Solution. The likelihood function is

$$\mathcal{L}(\theta) = \prod_{i=1}^n \frac{1}{2\theta^3} X_i^2 e^{-X_i/\theta},$$

and the log-likelihood function is

$$\ell(\theta) = -n\log 2 - 3n\log\theta - \frac{1}{\theta}\sum_{i=1}^n X_i + 2\sum_{i=1}^n \log X_i.$$

Setting the first-order derivative to 0:

$$\ell'(\theta) = -\frac{3n}{\theta} + \frac{1}{\theta^2}\sum_{i=1}^n X_i = 0$$

$$\Longrightarrow \hat{\theta} = \frac{1}{3n}\sum_{i=1}^n X_i = \frac{\bar{X}}{3}.$$

Exercise 2.2.3. Consider a random sample of n observations on X having the following PMF:

x	0	1	2
$p(x)$	$1-\theta$	$\theta/2$	$\theta/2$

(a) Find an unbiased estimators T_1 for θ based on the sample mean $\bar{X}$.

(b) Find an unbiased estimators T_2 for θ based on the sample mean $Y = \text{freq}(0) = \sum_{i=1}^n \mathbf{1}(X_i = 0)$ (the frequency of $x = 0$).

(c) Compare the two estimators above in terms of the mean square error (MSE).

Solution.

(a) We can first find that

$$\mathbb{E}[X] = 1 \cdot \frac{\theta}{2} + 2 \cdot \frac{\theta}{2} = \frac{3\theta}{2}.$$

Therefore, $\mathbb{E}[\bar{X}] = \mathbb{E}[X] = 3\theta/2$, a desirable unbiased estimator is

$$T_1 = \frac{2\bar{X}}{3}.$$

(b) Note that $Y \sim \text{Bin}(n, 1-\theta)$, which means $\mathbb{E}[Y] = n(1-\theta)$ and $\mathbb{E}[Y/n] = 1-\theta$. Thus, a desirable unbiased estimator is

$$T_2 = 1 - \frac{Y}{n}.$$

(c) Since both estimators T_1 and T_2 are unbiased, the MSE equals the variance of each estimator.

For T_1, we first calculate:

$$\text{Var}(X) = \frac{5\theta}{2} - \left(\frac{3\theta}{2}\right)^2 = \frac{5\theta}{2} - \frac{9\theta^2}{4}.$$

Therefore,

$$\text{Var}(T_1) = \text{Var}\left(\frac{2\bar{X}}{3}\right) = \frac{4}{9} \cdot \frac{\text{Var}(X)}{n} = \frac{10\theta - 9\theta^2}{9n}.$$

For T_2, as $Y \sim \text{Bin}(n, 1-\theta)$, we have $\text{Var}(Y) = \theta(1-\theta)/n$. Therefore,

$$\text{Var}(T_2) = \frac{\text{Var}(Y)}{n^2} = \frac{\theta(1-\theta)}{n}.$$

Noticing that

$$\text{Var}(T_1) = \frac{\theta(1-\theta)}{n} + \frac{\theta}{9n} > \frac{\theta(1-\theta)}{n} = \text{Var}(T_2),$$

we can conclude that $\text{MSE}(T_1) > \text{MSE}(T_2)$ and T_2 is a better estimator.

Exercise 2.2.4. If $\bar{X}_1$ and $\bar{X}_2$ are the sample means of independent random samples of sizes n_1 and n_2 from a normal population with the mean μ and the variance σ^2.

(a) Show that the estimator $T = w\bar{X}_1 + (1-w)\bar{X}_2$ is an unbiased estimator of μ.

(b) Find the value of w so that the variance of T is minimized, and calculate the minimized variance.

Solution.

(a) Since we have $\mathbb{E}[\bar{X}_1] = \mathbb{E}[\bar{X}_2] = \mu$, therefore,

$$\mathbb{E}[T] = \mathbb{E}[w\bar{X}_1 + (1-w)\bar{X}_2] = w\mu + (1-w)\mu = \mu,$$

so T is unbiased.

(b) First,

$$\text{Var}(T) = w^2\text{Var}(\bar{X}_1) + (1-w)^2\text{Var}(\bar{X}_2) + \underbrace{2w(1-w)\text{Cov}(\bar{X}_1, \bar{X}_2)}_{=0 \text{ by independence}}$$

$$= \frac{w^2\sigma^2}{n_1} + \frac{(1-w)^2\sigma^2}{n_2}$$
$$= \frac{\sigma^2}{n_1 n_2}\left[(n_1+n_2)w^2 - 2n_1\sigma^2 w + n_1\sigma^2\right].$$

This is a quadratic function and its value is minimized at

$$w = \frac{n_1}{n_1+n_2},$$

with the minimized variance being

$$\min_w \text{Var}(T) = \frac{\sigma^2}{n_1+n_2}.$$

Exercise 2.2.5. Let $X_1, \cdots, X_n$ be random samples from the following probability density function:

$$f(x) = \frac{x}{\theta^2}e^{-x/\theta},\ x>0, \theta>0.$$

(a) Determine the maximum likelihood estimator of θ.

(b) Find the Cramér-Rao Lower Bound for unbiased estimators of θ. Also determine the asymptotic distribution of the MLE. Comment on your findings.

(c) Suppose that the sample size $n = 25$ and the sample mean is $\bar{X} = 12$. Estimate the standard error of the MLE.

Solution.

(a)

$$\mathcal{L}(\theta) = \prod_{i=1}^{n} \frac{x}{\theta^2}e^{-x/\theta},$$

and the log-likelihood function is

$$\ell(\theta) = -2n\log\theta - \frac{1}{\theta}\sum_{i=1}^{n} X_i + \text{constant}.$$

The first-order derivative is :

$$\ell'(\theta) = -\frac{2n}{\theta} + \frac{1}{\theta^2}\sum_{i=1}^{n} X_i.$$

Setting the first-order derivative to 0:

$$\ell'(\theta) = -\frac{2n}{\theta} + \frac{1}{\theta^2}\sum_{i=1}^{n} X_i = 0$$

$$\Longrightarrow \hat{\theta} = \frac{\bar{X}}{2}.$$

(b) We find the second-order derivative of the log-likelihood function:

$$\ell''(\theta) = \frac{2n}{\theta^2} - \frac{2n\bar{X}}{\theta^3}.$$

Therefore, the Fisher information is

$$I(\theta) = -\mathbb{E}\left[\ell''(\theta)\right] = -\frac{2n}{\theta^2} + \frac{2n \cdot 2\theta}{\theta^3} = \frac{2n}{\theta^2}.$$

This suggests that the Cramér-Rao Lower Bound is

$$\text{CRLB} = \frac{1}{I(\theta)} = \frac{\theta^2}{2n}.$$

The asymptotic distribution of $\hat{\theta}$ is

$$\hat{\theta} \overset{a}{\sim} \mathcal{N}(\theta, \text{CRLB}) = \mathcal{N}\left(\theta, \frac{\theta^2}{2n}\right).$$

(c) Substituting the numbers, we get the MLE $\hat{\theta} = \bar{x}/2 = 6$, and the standard error of $\hat{\theta}$ is

$$\text{se}(\hat{\theta}) \approx \sqrt{\text{Var}(\hat{\theta})} = \frac{\hat{\theta}}{\sqrt{2n}} = \frac{6}{\sqrt{2 \cdot 25}} = 0.846.$$

Exercise 2.2.6. (**R Programming**)

An actuary is analysing the number of insurance claims reported by 8 small businesses in a given year. The observed claim counts are shown as follows:

$$5, 8, 6, 9, 7, 2, 4, 3.$$

The actuary assumes that the number of claims follows a Poisson distribution with an unknown mean θ.

(a) Write **R** code to determine the maximum likelihood estimate (MLE) of θ.

(b) Write **R** code to estimate the standard error of the MLE.

(c) Now, removing the Poisson assumption, estimate the standard error of the sample mean using bootstrap resampling (1000 times). Use a random seed of 2025 before starting the bootstrap resampling.

Solution.

(a) The MLE of a Poisson mean θ is simply the sample mean.

```
claim.counts <- c(5, 8, 6, 9, 7, 2, 4, 3)
theta.mle <- mean(claim.counts)
theta.mle
```

```
## [1] 5.5
```

(b)
```
se.mle <- sqrt(theta.mle / 8)
se.mle
```

```
## [1] 0.8291562
```

(c)
```
set.seed(2025)
B <- 1000
bootstrap.means <- numeric(B)
for (i in 1:B) {
sample.data <- sample(claim.counts, size = 8, replace = TRUE)
bootstrap.means[i] <- mean(sample.data) }
se.bootstrap <- sd(bootstrap.means)
se.bootstrap
```

```
## [1] 0.8343267
```

2.3 Interval Estimation

2.3.1 Introduction to Confidence Intervals

2.3.1.1 Motivation

Point estimators are never right, since they are derived from finite samples. Suppose two analysts both want to estimate the average height μ of a group of people assuming that the height follows a normal distribution $\mathcal{N}(\mu, \sigma^2)$:

- Analyst A: $\hat{\mu} = 170$ with sample size $n = 10,000$.
- Analyst B: $\hat{\mu} = 165$ with sample size $n = 100$.

Intuitively, the estimate obtained by Analyst A is more trustworthy given the significantly larger sample size but even this argument is only qualitative. Therefore, what would be desired is an inference statement like this:

"*I'm 95% confident that the parameter μ is between 165 and 168.*"

This can be obtained from the so-called *interval estimates*, which could not only contain the information of point estimates but also quantify *uncertainty* of the estimates. A *confidence interval* is one of the most common interval estimate constructed such that the corresponding interval estimator has a pre-specified probability, known as the *confidence level*, of containing the true value of the estimated parameter.

2.3.1.2 Definition

Now, we define a confidence interval mathematically.

Definition 2.3.1 (Confidence Intervals)**.** Let $\boldsymbol{X}$ be a random sample from a probability distribution with parameter θ[viii]. Then, a $1-\alpha$ confidence interval $(L_\theta(\boldsymbol{X}), U_\theta(\boldsymbol{X}))$ is a **random**

[viii] In general, there can be multiple parameters of interest $\boldsymbol{\theta}$. In this case, a confidence interval is called a *confidence set*. In CS1, we only study confidence intervals for a single parameter.