Mathematical Statistics

Vasileios Katsianos
WashU

Panagiotis Andreou
UNC Chapel Hill

August 2024

“*There are three kinds of lies: lies, damned lies and Statistics.*”

Mark Twain

Contents

1 Elements of Probability Distributions **9**

1.1 Discrete Distributions 9

1.2 Continuous Distributions 12

1.3 Definitions and Properties 15

2 Exponential Family of Distributions **17**

2.1 Introduction 17

2.2 One-parameter Exponential Family 17

2.3 Multiparameter Exponential Family 19

3 Point Estimation **25**

3.1 Introduction 25

3.2 Unbiased Estimators 25

3.3 Mean Squared Error 27

3.4 Sufficiency 28

3.5 Completeness 32

3.6 Minimal Sufficiency 40

3.7 Uniformly Minimum-Variance Unbiased Estimators 44

3.8 Cramér - Rao Inequality 51

3.9 Multivariate Cramér - Rao Inequality 58

3.10 Asymptotic Distribution of Estimators 63

3.11 Maximum Likelihood Estimation 71

3.12 Method of Moments Estimators 78

4 Confidence Intervals **83**

4.1 Introduction 83

4.2 Pivotal Quantity Method 84

4.3 CIs for a Normal Population 94

4.4 CIs for Two Independent Normal Populations 97

4.5 Asymptotic Confidence Intervals 99

5 Statistical Hypothesis Testing **103**
5.1 Introduction 103
5.2 Fundamental Neyman - Pearson Lemma 106
5.3 Monotone Likelihood Ratio Property 112
5.4 Generalized Likelihood Ratio Tests 118
5.5 Statistical Hypothesis Tests for a Normal Population 121

Bibliography **125**

Preface

In the field of statistics, we are usually interested in studying some phenomenon. To this end, we acquire a sample of observations which we consider as realizations from a probability distribution with one or more parameters. In frequentist statistics, we regard those parameters as some unknown constants. Our goal is to draw inferences about them and use those inferences to answer any questions we might have about the phenomenon under study. As one might surmise, the study of statistics requires a thorough knowledge of probability distributions and their properties. Chapter 1 of this textbook summarizes some useful elements of distribution theory and chapter 2 introduces an important family of probability distributions with many useful applications in statistics.

Suppose we are interested in ascertaining whether a new cholesterol drug is effective or not. We prescribe the drug to 100 volunteers with a family history of high cholesterol and measure whether their cholesterol levels have dropped after 3 months of taking the new drug. For $i = 1, 2, \ldots, 100$, we define:

$$X_i = \begin{cases} 1, & \text{cholesterol levels of volunteer } i \text{ dropped} \\ 0, & \text{cholesterol levels of volunteer } i \text{ didn't drop} \end{cases}.$$

Suppose that $X_i \sim \text{Bernoulli}(p)$ for $i = 1, 2, \ldots, 100$, where p is the unknown probability of success of the new cholesterol drug. A good first step in our statistical analysis would be to obtain a logical estimate of the unknown parameter p based on the obtained sample of observations $x_1, \ldots, x_{100}$. One might correctly deduce that the proportion of volunteers whose cholesterol levels dropped after being on the new cholesterol drug for 3 months is a good estimate of the probability p. If that proportion is "comfortably" larger than 50%, then that's a sign towards the effectiveness of the new cholesterol drug. Chapter 3 rigorously introduces some methods of specifying such estimates of unknown parameters and presents several criteria based on which different estimates of the same unknown parameter may be compared against each other to determine the "best" among them. These criteria mainly aim at providing some "guarantees" that the value of the point estimate is going to lie close to the true value of the unknown parameter with high probability.

Obtaining a point estimate of the unknown parameter is usually not enough, since its value depends on the sample we happened to collect and doesn't provide us with any information about how the values of the same estimate based on samples that other people might collect are distributed. In other words, we also want a measure of how far away the most probable values of the estimate could lie from our specific point estimate. Thus, we get the idea for the construction of an interval which contains all the most probable values of the estimate. That interval is constructed in such a way that it contains the true value of the unknown parameter with some specified level of "confidence". In our previous example, if we arrive at an interval whose lower endpoint lies above 0.5, then that provides us with strong evidence that the new cholesterol drug is actually effective. Chapter 4 presents different methodologies according to which such confidence intervals are constructed.

Finally, we are interested in checking the validity of hypotheses such as whether the unknown parameter takes a specific set of values based on the evidence contained in our sample. For example, we might be interested in knowing whether the probability of success of the new cholesterol drug is greater than 0.5 or not, i.e. whether the drug is effective or not. Chapter 5 sets the foundations of the framework for conducting such hypothesis tests in frequentist statistics and introduces several methods for obtaining decision rules based on the observed sample in such settings.

Bill Katsianos
Panos Andreou

Chapter 1

Elements of Probability Distributions

1.1 Discrete Distributions

Definition 1.1. (Probability Mass Function - PMF)

$$f_X(x) = \mathbb{P}(X = x), \quad x \in S = \{x_0, x_1, \dots\}$$

Proposition 1.1. (Properties of PMFs)

i. $0 \leqslant f_X(x) \leqslant 1$, $x \in S = \{x_0, x_1, \dots\}$;

ii. $\sum_{x \in S} f_X(x) = 1$.

Definition 1.2. (Cumulative Distribution Function - CDF)

$$F_X(x) = \mathbb{P}(X \leqslant x) = \sum_{y \leqslant x} \mathbb{P}(X = y) = \sum_{y \leqslant x} f_X(y), \quad x \in \mathbb{R}$$

Definition 1.3. (Expected Value) If $\sum_{x \in S} |x| f_X(x) < \infty$, then:

$$\mathbb{E}(X) = \sum_{x \in S} x f_X(x).$$

Definition 1.4. (Indicator Random Variable)

$$X = \mathbb{1}_A(Y) = \begin{cases} 1, & Y \in A \\ 0, & Y \notin A \end{cases}$$

It holds that $\mathbb{E}(X) = 1 \cdot \mathbb{P}(Y \in A) + 0 \cdot \mathbb{P}(Y \notin A) = \mathbb{P}(Y \in A)$.

Definition 1.5. (Variance) If $\sum_{x \in S} x^2 f_X(x) < \infty$, then:

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

Theorem 1.1. (Law of the Unconscious Statistician)

$$\mathbb{E}\left[g(X)\right] = \sum_{x \in S} g(x) f_X(x)$$

Definition 1.6. (Independence)

$$\begin{aligned} X, Y \text{ independent} \quad &\Leftrightarrow \quad \mathbb{P}(X \in A, Y \in B) = \mathbb{P}(X \in A)\mathbb{P}(Y \in B), \quad \forall A, B \subseteq \mathbb{R} \\ &\Leftrightarrow \quad f_{X,Y}(x, y) = f_X(x) f_Y(y), \quad \forall x \in S_X, \quad \forall y \in S_Y. \end{aligned}$$

Definition 1.7. (Moment Generating Function - MGF)

$$M_X(t) = \mathbb{E}\left(e^{tX}\right) = \sum_{x \in S} e^{tx} f_X(x)$$

Notable Discrete Distributions

Bernoulli Distribution - Bernoulli(p), $p \in (0, 1)$: Success/failure in 1 trial

$$f_X(x) = p^x (1 - p)^{1-x}, \quad x \in \{0, 1\},$$

$$\mathbb{E}(X) = p, \quad \mathrm{Var}(X) = p(1 - p),$$

$$M_X(t) = pe^t + 1 - p, \quad t \in \mathbb{R},$$

$$X, Y \sim \text{Bernoulli}(p) \text{ independent } \Rightarrow X + Y \sim \text{Bin}(2, p).$$

Binomial Distribution - Bin(N, p), $N \in \mathbb{N}$, $p \in (0, 1)$: Number of successes in N trials

$$f_X(x) = \binom{N}{x} p^x (1 - p)^{N-x}, \quad x \in \{0, 1, \ldots, N\},$$

$$\mathbb{E}(X) = Np, \quad \mathrm{Var}(X) = Np(1 - p),$$

$$M_X(t) = \left(pe^t + 1 - p\right)^N, \quad t \in \mathbb{R},$$

$$X \sim \text{Bin}(N, p), Y \sim \text{Bin}(M, p) \text{ independent } \Rightarrow X + Y \sim \text{Bin}(N + M, p).$$

Geometric Distribution - Geom(p), $p \in (0, 1)$: Number of trials until the first success

$$f_X(x) = p(1 - p)^{x-1}, \quad x \in \{1, 2, \ldots\},$$

$$\mathbb{E}(X) = \frac{1}{p}, \quad \mathrm{Var}(X) = \frac{1}{p^2},$$

$$M_X(t) = \frac{pe^t}{1-(1-p)e^t}, \quad t < -\log(1-p),$$

$$X, Y \sim \mathrm{Geom}(p) \text{ independent } \Rightarrow X + Y \sim \mathrm{NegBin}(2, p).$$

Geometric Distribution - $\mathrm{Geom}(p)$, $p \in (0,1)$: Number of failures until the first success

$$f_X(x) = p(1-p)^x, \quad x \in \{0, 1, \dots\},$$

$$\mathbb{E}(X) = \frac{1-p}{p}, \quad \mathrm{Var}(X) = \frac{1-p}{p^2},$$

$$M_X(t) = \frac{p}{1-(1-p)e^t}, \quad t < -\log(1-p),$$

$$X, Y \sim \mathrm{Geom}(p) \text{ independent } \Rightarrow X + Y \sim \mathrm{NegBin}(2, p).$$

Negative Binomial Distribution - $\mathrm{NegBin}(N,p)$, $N \in \mathbb{N}$, $p \in (0,1)$: Number of trials until the N-th success

$$f_X(x) = \binom{x-1}{N-1} p^N (1-p)^{x-N}, \quad x \in \{N, N+1, \dots\},$$

$$\mathbb{E}(X) = \frac{N}{p}, \quad \mathrm{Var}(X) = \frac{N}{p^2},$$

$$M_X(t) = \left[\frac{pe^t}{1-(1-p)e^t}\right]^N, \quad t < -\log(1-p),$$

$$X \sim \mathrm{NegBin}(N,p), Y \sim \mathrm{NegBin}(M,p) \text{ independent } \Rightarrow X + Y \sim \mathrm{NegBin}(N+M, p).$$

Negative Binomial Distribution - $\mathrm{NegBin}(N,p)$, $N \in \mathbb{N}$, $p \in (0,1)$: Number of failures until the N-th success

$$f_X(x) = \binom{x+N-1}{N-1} p^N (1-p)^x, \quad x \in \{0, 1, \dots\},$$

$$\mathbb{E}(X) = N\frac{1-p}{p}, \quad \mathrm{Var}(X) = N\frac{1-p}{p^2},$$

$$M_X(t) = \left[\frac{p}{1-(1-p)e^t}\right]^N, \quad t < -\log(1-p),$$

$$X \sim \mathrm{NegBin}(N,p), Y \sim \mathrm{NegBin}(M,p) \text{ independent } \Rightarrow X + Y \sim \mathrm{NegBin}(N+M, p).$$

Poisson Distribution - $\mathrm{Poisson}(\lambda)$, $\lambda > 0$: Number of events in a fixed time interval

$$f_X(x) = e^{-\lambda}\frac{\lambda^x}{x!}, \quad x \in \{0, 1, \dots\},$$

$$\mathbb{E}(X) = \mathrm{Var}(X) = \lambda,$$

$$M_X(t) = e^{\lambda(e^t-1)},$$

$$X \sim \text{Poisson}(\lambda), Y \sim \text{Poisson}(\mu) \text{ independent } \Rightarrow X + Y \sim \text{Poisson}(\lambda + \mu).$$

1.2 Continuous Distributions

Definition 1.8. (Probability Density Function - PDF) Function $f_X : \mathbb{R} \to [0, \infty)$ such that:

$$\mathbb{P}(X \in A) = \int_A f_X(x)dx, \quad A \subseteq \mathbb{R}.$$

Definition 1.9. (Cumulative Distribution Function - CDF)

$$F_X(x) = \mathbb{P}(X \leqslant x) = \int_{-\infty}^{x} f_X(y)dy, \quad x \in \mathbb{R}$$

Proposition 1.2. (Properties of PDFs and CDFs)

i. $f_X(x) \geqslant 0$, $x \in \mathbb{R}$;

ii. $\int_{\mathbb{R}} f_X(x)dx = 1$;

iii. $\int_a^b f_X(x)dx = \mathbb{P}(a < X < b) = \mathbb{P}(a \leqslant X \leqslant b) = \mathbb{P}(a < X \leqslant b) = \mathbb{P}(a \leqslant X < b)$;

iv. $\mathbb{P}(X = x) = 0$, $x \in \mathbb{R}$;

v. $f_X(x) = F'_X(x)$, $x \in \mathbb{R}$;

vi. F_X strictly increasing on the set $S = \{x \in \mathbb{R} : f_X(x) > 0\}$.

Definition 1.10. (Expected Value) If $\int_{\mathbb{R}} |x| f_X(x)dx < \infty$, then:

$$\mathbb{E}(X) = \int_{\mathbb{R}} x f_X(x)dx.$$

Proposition 1.3. If $X \geqslant 0$, i.e. $f_X(x) = 0 \; \forall x < 0$, then:

$$\mathbb{E}\left(X^k\right) = \int_0^\infty kx^{k-1} \left[1 - F_X(x)\right] dx, \quad k > 0.$$

In particular, it holds that:

$$\mathbb{E}\left(X\right) = \int_0^\infty \left[1 - F_X(x)\right] dx.$$

Definition 1.11. (Variance) If $\int_{\mathbb{R}} x^2 f_X(x)dx < \infty$, then:

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

Theorem 1.2. (Law of the Unconscious Statistician)

$$\mathbb{E}\left[g(X)\right] = \int_{\mathbb{R}} g(x) f_X(x) dx$$

Definition 1.12. (Independence)

$$\begin{aligned} X, Y \text{ independent} \quad &\Leftrightarrow \quad \mathbb{P}(X \in A, Y \in B) = \mathbb{P}(X \in A)\mathbb{P}(Y \in B), \quad \forall A, B \subseteq \mathbb{R} \\ &\Leftrightarrow \quad f_{X,Y}(x,y) = f_X(x) f_Y(y), \quad \forall x, y \in \mathbb{R} \end{aligned}$$

Definition 1.13. (Moment Generating Function - MGF)

$$M_X(t) = \mathbb{E}\left(e^{tX}\right) = \int_{\mathbb{R}} e^{tx} f_X(x) dx$$

Definition 1.14. (Gamma Function)

$$\Gamma(k) = \int_0^\infty x^{k-1} e^{-x} dx, \quad k > 0$$

Proposition 1.4. (Properties of the Gamma Function)

i. $\Gamma(k) = (k-1)\Gamma(k-1)$, $k > 1$;

ii. $\Gamma(k) = (k-1)!$, $k \in \mathbb{N}$.

Notable Continuous Distributions

Continuous Uniform Distribution - $\mathcal{U}(\vartheta_1, \vartheta_2)$, $\vartheta_1 < \vartheta_2$: Random number selection on the interval $[\vartheta_1, \vartheta_2]$

$$f_X(x) = \frac{1}{\vartheta_2 - \vartheta_1}, \quad F_X(x) = \frac{x - \vartheta_1}{\vartheta_2 - \vartheta_1}, \quad x \in [\vartheta_1, \vartheta_2],$$

$$\mathbb{E}(X) = \frac{\vartheta_1 + \vartheta_2}{2}, \quad \text{Var}(X) = \frac{(\vartheta_2 - \vartheta_1)^2}{12},$$

$$M_X(t) = \begin{cases} \frac{e^{\vartheta_2 t} - e^{\vartheta_1 t}}{(\vartheta_2 - \vartheta_1)t}, & t \neq 0 \\ 1, & t = 0 \end{cases},$$

$$X \sim \mathcal{U}(\vartheta_1, \vartheta_2) \;\Rightarrow\; U = \frac{X - \vartheta_1}{\vartheta_2 - \vartheta_1} \sim \mathcal{U}(0,1),$$

$$U \sim \mathcal{U}(0,1) \;\Rightarrow\; X = (\vartheta_2 - \vartheta_1)U + \vartheta_1 \sim \mathcal{U}(\vartheta_1, \vartheta_2).$$

Exponential Distribution - $\text{Exp}(\lambda)$, $\lambda > 0$: Time between 2 events

$$f_X(x) = \lambda e^{-\lambda x}, \quad F_X(x) = 1 - e^{-\lambda x}, \quad x > 0,$$

$$\mathbb{E}(X) = \frac{1}{\lambda}, \quad \mathrm{Var}(X) = \frac{1}{\lambda^2},$$

$$M_X(t) = \frac{\lambda}{\lambda - t}, \quad t < \lambda,$$

$$X \sim \mathrm{Exp}(\lambda) \;\Rightarrow\; cX \sim \mathrm{Exp}\left(\lambda/c\right), \quad c > 0,$$

$$X \sim \mathrm{Exp}(\lambda), Y \sim \mathrm{Exp}(\mu) \text{ independent } \;\Rightarrow\; \min\{X, Y\} \sim \mathrm{Exp}(\lambda + \mu),$$

$$X, Y \sim \mathrm{Exp}(\lambda) \text{ independent } \;\Rightarrow\; X + Y \sim \mathrm{Gamma}(2, \lambda).$$

Gamma Distribution - $\mathrm{Gamma}(k, \lambda)$, $k > 0$, $\lambda > 0$

$$f_X(x) = \frac{\lambda^k}{\Gamma(k)} x^{k-1} e^{-\lambda x}, \quad x > 0,$$

$$\mathbb{E}(X) = \frac{k}{\lambda}, \quad \mathrm{Var}(X) = \frac{k}{\lambda^2},$$

$$M_X(t) = \left(\frac{\lambda}{\lambda - t}\right)^k, \quad t < \lambda,$$

$$X \sim \mathrm{Gamma}(k, \lambda) \;\Rightarrow\; cX \sim \mathrm{Gamma}\left(k, \lambda/c\right), \quad c > 0,$$

$$X \sim \mathrm{Gamma}(k, \lambda), Y \sim \mathrm{Gamma}(\ell, \lambda) \text{ independent } \;\Rightarrow\; X + Y \sim \mathrm{Gamma}(k + \ell, \lambda).$$

Normal Distribution - $\mathcal{N}\left(\mu, \sigma^2\right)$, $\mu \in \mathbb{R}$, $\sigma^2 > 0$

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

$$\mathbb{E}(X) = \mu, \quad \mathrm{Var}(X) = \sigma^2,$$

$$M_X(t) = \exp\left\{\mu t + \frac{1}{2}\sigma^2 t^2\right\}, \quad t \in \mathbb{R},$$

$$X \sim \mathcal{N}\left(\mu, \sigma^2\right) \;\Rightarrow\; Z = \frac{X - \mu}{\sigma} \sim \mathcal{N}(0, 1),$$

$$Z \sim \mathcal{N}(0, 1) \;\Rightarrow\; X = \sigma Z + \mu \sim \mathcal{N}\left(\mu, \sigma^2\right),$$

$$X \sim \mathcal{N}\left(\mu_1, \sigma_1^2\right), Y \sim \mathcal{N}\left(\mu_2, \sigma_2^2\right) \text{ independent } \;\Rightarrow\; X + Y \sim \mathcal{N}\left(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2\right).$$

Beta Distribution - $\mathrm{Beta}(\vartheta_1, \vartheta_2)$, $\vartheta_1 > 0$, $\vartheta_2 > 0$

$$f_X(x) = \frac{\Gamma(\vartheta_1 + \vartheta_2)}{\Gamma(\vartheta_1)\Gamma(\vartheta_2)} x^{\vartheta_1 - 1}(1 - x)^{\vartheta_2 - 1}, \quad x \in (0, 1),$$

$$\mathbb{E}(X) = \frac{\vartheta_1}{\vartheta_1 + \vartheta_2}, \quad \mathrm{Var}(X) = \frac{\vartheta_1\vartheta_2}{(\vartheta_1 + \vartheta_2 + 1)(\vartheta_1 + \vartheta_2)^2},$$

$$X \sim \mathrm{Beta}(\vartheta_1, \vartheta_2) \;\Rightarrow\; 1 - X \sim \mathrm{Beta}\left(\vartheta_2, \vartheta_1\right),$$

$$X \sim \mathrm{Beta}(\vartheta, 1) \;\Rightarrow\; Y = -\log X \sim \mathrm{Exp}(\vartheta),$$

$$X \sim \text{Beta}(1, \vartheta) \;\Rightarrow\; Y = -\log(1 - X) \sim \text{Exp}(\vartheta),$$

$$Y \sim \text{Exp}(\vartheta) \;\Rightarrow\; X_1 = e^{-Y} \sim \text{Beta}(\vartheta, 1) \text{ and } X_2 = 1 - e^{-Y} \sim \text{Beta}(1, \vartheta).$$

1.3 Definitions and Properties

Definition 1.15. (Covariance) If $\mathbb{E}(XY) < \infty$, then:

$$\text{Cov}(X, Y) = \mathbb{E}\left[(X - \mathbb{E}(X))\,(Y - \mathbb{E}(Y))\right] = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y).$$

Proposition 1.5. (Properties of the Expected Value)

i. $\mathbb{E}(aX + b) = a\mathbb{E}(X) + b$;

ii. $\mathbb{E}(aX + bY) = a\mathbb{E}(X) + b\mathbb{E}(Y)$;

iii. X, Y independent implies that $\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$;

iv. $a \leqslant X \leqslant b$ implies that $a \leqslant \mathbb{E}(X) \leqslant b$;

v. $X \leqslant Y$ implies that $\mathbb{E}(X) \leqslant \mathbb{E}(Y)$.

Proposition 1.6. (Properties of the Variance)

i. $\text{Var}(X) \geqslant 0$;

ii. $\text{Var}(aX + b) = a^2 \,\text{Var}(X)$;

iii. $\text{Var}(aX + bY) = a^2 \,\text{Var}(X) + b^2 \,\text{Var}(Y) + 2ab \,\text{Cov}(X, Y)$;

iv. X, Y independent implies that $\text{Var}(aX + bY) = a^2 \,\text{Var}(X) + b^2 \,\text{Var}(Y)$;

v. X, Y independent implies that $\text{Var}(XY) = \mathbb{E}\left(X^2\right)\mathbb{E}\left(Y^2\right) - [\mathbb{E}(X)\mathbb{E}(Y)]^2$.

Proposition 1.7. (Properties of the Covariance)

i. $\text{Cov}(X, a) = 0$;

ii. $\text{Cov}(X, X) = \text{Var}(X)$;

iii. $\text{Cov}(Y, X) = \text{Cov}(X, Y)$;

iv. $\text{Cov}(aX + b, cY + d) = ac \,\text{Cov}(X, Y)$;

v. $\text{Cov}(X + Y, Z + W) = \text{Cov}(X, Z) + \text{Cov}(X, W) + \text{Cov}(Y, Z) + \text{Cov}(Y, W)$;

vi. X, Y independent implies that $\text{Cov}(X, Y) = 0$.

Definition 1.16. (Conditional Expectation of X given $Y = y$)

$$m_{X|Y}(y) = \mathbb{E}(X \mid Y = y) = \begin{cases} \sum_{x \in S_X} x f_{X|Y}(x \mid y), & X \text{ discrete} \\ \int_{\mathbb{R}} x f_{X|Y}(x \mid y) dx, & X \text{ continuous} \end{cases}$$

Definition 1.17. (Conditional Expectation of X given Y)

$$\mathbb{E}(X \mid Y) = m_{X|Y}(Y)$$

Theorem 1.3. (Law of Iterated Expectations)

$$\mathbb{E}(X) = \mathbb{E}\left[\mathbb{E}(X \mid Y)\right] = \mathbb{E}\left[m_{X|Y}(Y)\right] = \begin{cases} \sum_{y \in S_Y} m_{X|Y}(y) f_Y(y), & Y \text{ discrete} \\ \int_{\mathbb{R}} m_{X|Y}(y) f_Y(y) dy, & Y \text{ continuous} \end{cases}$$

Proposition 1.8. (Properties of MGFs)

i. $M_X(t) = M_Y(t)\ \forall t \in \mathbb{R}$ if and only if X, Y identically distributed (belonging to the same family of distributions with the same parameter values);

ii. $M_{aX+b}(t) = e^{bt} M_X(at)$;

iii. $M_X^{(k)}(0) = \mathbb{E}\left(X^k\right)$, $k \in \mathbb{N}$;

iv. X, Y independent implies that $M_{X+Y}(t) = M_X(t) M_Y(t)$.

Proposition 1.9. (Notable Probabilistic Inequalities)

i. **Markov's Inequality**: $X \geqslant 0 \Rightarrow \mathbb{P}(X \geqslant a) \leqslant \frac{\mathbb{E}(X)}{a}$, $a > 0$

ii. **Chebyshev's Inequality**: $\mathbb{P}\left(|X - \mathbb{E}(X)| \geqslant a\right) \leqslant \frac{\text{Var}(X)}{a^2}$, $a > 0$

iii. **Cauchy - Schwarz Inequality**: $[\mathbb{E}(XY)]^2 \leqslant \mathbb{E}\left(X^2\right) \mathbb{E}\left(Y^2\right)$

iv. **Covariance Inequality**: $[\text{Cov}(X, Y)]^2 \leqslant \text{Var}\,(X)\, \text{Var}\,(Y)$

v. **Jensen's Inequality**: f convex implies that $f\left(\mathbb{E}(X)\right) \leqslant \mathbb{E}\left[f(X)\right]$

Note 1.1. An easy way to remember the direction in Jensen's inequality is through the non-negativity property of the variance of a random variable X. More specifically, we know that:

$$\text{Var}(X) = \mathbb{E}\left(X^2\right) - [\mathbb{E}(X)]^2 \geqslant 0 \quad \Rightarrow \quad [\mathbb{E}(X)]^2 \leqslant \mathbb{E}\left(X^2\right) \quad \Rightarrow$$

$$f\left(\mathbb{E}(X)\right) \leqslant \mathbb{E}\left[f(X)\right],$$

where $f(x) = x^2$ is a convex function in $\mathbb{R}$.

Chapter 2

Exponential Family of Distributions

2.1 Introduction

The exponential family of distributions is a class of distributions which includes many of the most widely used (discrete and continuous) distributions. Its usefulness lies in the fact that the distributions which belong to it have some common properties, which allow us to formulate various propositions that are valid for all them. Many well-known results about these distributions can arise as special cases of these propositions.

Definition 2.1. i. The set Θ which contains all the values that an unknown parameter ϑ can take is called the *parameter space*.

ii. The set $S = \{x \in \mathbb{R} : f(x;\vartheta) > 0\}$ is called the *support* of the distribution with PMF or PDF $f(x;\vartheta)$.

2.2 One-parameter Exponential Family

Definition 2.2. A distribution with unknown parameter $\vartheta \in \Theta \subseteq \mathbb{R}$ and PMF or PDF $f(x;\vartheta)$ for $x \in S \subseteq \mathbb{R}$ belongs to the *one-parameter (full) exponential family* if the support S doesn't depend on the value of ϑ and it holds that:

$$f(x;\vartheta) = h(x)e^{Q(\vartheta)T(x)-A(\vartheta)}.$$

If $Q(\vartheta) = \vartheta$, then we say that the exponential family is in *canonical form.*

Note 2.1. Indicatively, we mention that the following distributions belong to the one-parameter exponential family: Bernoulli, binomial with known number of trials, geometric, negative binomial with known number of trials, Poisson and exponential.

Proposition 2.1. If a random variable X has PMF or PDF $f(x;\vartheta) = h(x)e^{\vartheta T(x) - A(\vartheta)}$ in canonical form, then it holds that:

$$\mathbb{E}[T(X)] = A'(\vartheta), \quad \mathrm{Var}[T(X)] = A''(\vartheta), \quad M_T(t) = \mathbb{E}\left[e^{tT(X)}\right] = e^{A(t+\vartheta) - A(\vartheta)}.$$

Proof. Without loss of generality, suppose that X is a continuous random variable. We observe that:

$$\int_{\mathbb{R}} f(x;\vartheta)dx = 1 \quad \Rightarrow \quad \int_{\mathbb{R}} h(x)e^{\vartheta T(x)}dx = e^{A(\vartheta)}.$$

In differentiating this expression with respect to ϑ, we can interchange the order of differentiation and integration by suitable application of the dominated convergence theorem in order to get that:

$$\frac{\partial}{\partial\vartheta}\int_{\mathbb{R}} h(x)e^{\vartheta T(x)}dx = A'(\vartheta)e^{A(\vartheta)} \quad \Rightarrow \quad \int_{\mathbb{R}} h(x)\frac{\partial}{\partial\vartheta}e^{\vartheta T(x)}dx = A'(\vartheta)e^{A(\vartheta)} \quad \Rightarrow$$

$$\int_{\mathbb{R}} h(x)T(x)e^{\vartheta T(x)}dx = A'(\vartheta)e^{A(\vartheta)} \quad \Rightarrow$$

$$A'(\vartheta) = \int_{\mathbb{R}} T(x)\underbrace{h(x)e^{\vartheta T(x) - A(\vartheta)}}_{f(x;\vartheta)}dx = \mathbb{E}[T(X)].$$

Similarly, we calculate that:

$$\int_{\mathbb{R}} h(x)T^2(x)e^{\vartheta T(x)}dx = \left[A'(\vartheta)\right]^2 e^{A(\vartheta)} + A''(\vartheta)e^{A(\vartheta)} \quad \Rightarrow$$

$$\mathbb{E}\left[T^2(X)\right] = \left[A'(\vartheta)\right]^2 + A''(\vartheta) \quad \Rightarrow$$

$$A''(\vartheta) = \mathbb{E}\left[T^2(X)\right] - \left[\mathbb{E}\left(T(X)\right)\right]^2 = \mathrm{Var}\left[T(X)\right].$$

As far as the moment generating function is concerned, we calculate that:

$$\begin{aligned} M_T(t) &= \int_{\mathbb{R}} e^{tT(x)}f(x;\vartheta)dx = \int_{\mathbb{R}} e^{tT(x)}h(x)e^{\vartheta T(x) - A(\vartheta)}dx \\ &= e^{-A(\vartheta)}\int_{\mathbb{R}} h(x)e^{(t+\vartheta)T(x)}dx \\ &= e^{A(t+\vartheta) - A(\vartheta)}\int_{\mathbb{R}} \underbrace{h(x)e^{(t+\vartheta)T(x) - A(t+\vartheta)}}_{f(x;t+\vartheta)}dx \\ &= e^{A(t+\vartheta) - A(\vartheta)}. \end{aligned}$$

□

Note 2.2. If $Q(\vartheta) \neq \vartheta$, then the exponential family may be converted to canonical form with the reparameterization $\eta = Q(\vartheta)$.

Example 2.1. (Binomial with known number of trials)

$$
\begin{aligned}
f(x;p) &= \binom{N}{x} p^x (1-p)^{N-x} = \binom{N}{x} e^{x \log p + (N-x)\log(1-p)} \\
&= \binom{N}{x} e^{x[\log p - \log(1-p)] + N \log(1-p)} \\
&= \binom{N}{x} \exp\left\{ x \log \frac{p}{1-p} - N \log \frac{1}{1-p} \right\},
\end{aligned}
$$

$$
h(x) = \binom{N}{x}, \quad Q(p) = \log \frac{p}{1-p}, \quad T(x) = x, \quad A(p) = N \log \frac{1}{1-p}.
$$

Consider the following reparameterization:

$$
\eta = \log \frac{p}{1-p} \in \mathbb{R} \quad \Rightarrow \quad (1-p)e^\eta = p \quad \Rightarrow \quad p = \frac{e^\eta}{e^\eta + 1} = \frac{1}{1 + e^{-\eta}},
$$

$$
f(x;\eta) = \binom{N}{x} e^{\eta x - N \log(e^\eta + 1)}, \quad A(\eta) = N \log\left(e^\eta + 1\right).
$$

Then, it follows that:

$$
\mathbb{E}\left[T(X)\right] = \mathbb{E}(X) = A'(\eta) = \frac{N e^\eta}{e^\eta + 1} = Np,
$$

$$
\operatorname{Var}\left[T(X)\right] = \operatorname{Var}(X) = A''(\eta) = \frac{N e^\eta}{\left(e^\eta + 1\right)^2} = \frac{N e^\eta}{e^\eta + 1} \frac{1}{e^\eta + 1} = Np(1-p),
$$

$$
\begin{aligned}
M_T(t) = M_X(t) = \mathbb{E}\left(e^{tX}\right) &= e^{A(t+\eta) - A(\eta)} = e^{N \log\left(e^{t+\eta}+1\right) - N \log(e^\eta + 1)} \\
&= \exp\left\{ N \log \frac{e^{t+\eta} + 1}{e^\eta + 1} \right\} = \left(\frac{e^t e^\eta + 1}{e^\eta + 1} \right)^N \\
&= \left(\frac{e^\eta}{e^\eta + 1} e^t + \frac{1}{e^\eta + 1} \right)^N = \left(pe^t + 1 - p \right)^N, \quad t \in \mathbb{R}. \quad \square
\end{aligned}
$$

2.3 Multiparameter Exponential Family

Definition 2.3. A distribution with unknown parameter vector $\vartheta \in \Theta \subseteq \mathbb{R}^s$ and PMF or PDF $f(x;\vartheta)$ for $x \in S \subseteq \mathbb{R}$ belongs to the *multiparameter exponential family* if the support S doesn't depend on the value of ϑ and it holds that:

$$
f(x;\vartheta) = h(x) e^{\langle Q(\vartheta), T(x) \rangle - A(\vartheta)},
$$

where $Q : \Theta \to \mathbb{R}^d$ and $T : S \to \mathbb{R}^d$ with $d \geqslant s$. If $s = d$, i.e. the dimension of the vector ϑ is equal to the dimension of the range of the functions Q and T, then we say that it constitutes a *full* exponential family. Otherwise, we say that it constitutes a *curved* exponential family. If $Q(\vartheta) = \vartheta$, then we say that the exponential family is in *canonical form*.

Note 2.3. Indicatively, we mention that the following distributions belong to the 2-parameter full exponential family: normal, gamma and beta. In contrast, we can easily see that the continuous uniform distribution on $[\vartheta_1, \vartheta_2]$ does **not** belong to the exponential family, since the support $S = [\vartheta_1, \vartheta_2]$ depends on the value of the parameter vector $\vartheta = (\vartheta_1, \vartheta_2)$.

Example 2.2. (Gamma)

$$f(x; k, \lambda) = \frac{\lambda^k}{\Gamma(k)} x^{k-1} e^{-\lambda x} = \frac{1}{x} \exp\left\{ k \log x - \lambda x - k \log \frac{1}{\lambda} - \log \Gamma(k) \right\},$$

$$h(x) = \frac{1}{x}, \quad Q(k, \lambda) = (k, -\lambda), \quad T(x) = (\log x, x), \quad A(k, \lambda) = k \log \frac{1}{\lambda} + \log \Gamma(k).$$

Hence, the gamma distribution belongs to the 2-parameter full exponential family. □

Example 2.3. (Weibull) For $k > 0$, $\lambda > 0$ and $x > 0$, we calculate that:

$$f(x; k, \lambda) = k\lambda x^{k-1} e^{-\lambda x^k} = k x^{k-1} \exp\left\{ -\lambda x^k - \log \frac{1}{\lambda} \right\}.$$

We observe that there exists no way to write the term λx^k as a product of a function of the parameter vector $\vartheta = (k, \lambda)$ and a function of x. Thus, the Weibull distribution does **not** belong to the two-parameter exponential family. However, if k is a known constant, then we calculate that:

$$f(x; \lambda) = k\lambda x^{k-1} e^{-\lambda x^k} = k x^{k-1} \exp\left\{ -\lambda x^k - \log \frac{1}{\lambda} \right\},$$

$$h(x) = k x^{k-1}, \quad Q(\lambda) = -\lambda, \quad T(x) = x^k, \quad A(\lambda) = \log \frac{1}{\lambda}.$$

Therefore, the Weibull distribution with known k belongs to the one-parameter exponential family. □

Proposition 2.2. If $f(x; \vartheta) = h(x) e^{\langle \vartheta, T(x) \rangle - A(\vartheta)}$ is the PMF or PDF of a random variable X in canonical form, then the following hold for $j, k = 1, 2, \ldots, s$:

$$\mathbb{E}[T_j(X)] = \frac{\partial A}{\partial \vartheta_j}, \quad \mathrm{Var}[T_j(X)] = \frac{\partial^2 A}{\partial \vartheta_j^2}, \quad \mathrm{Cov}[T_j(X), T_k(X)] = \frac{\partial^2 A}{\partial \vartheta_j \partial \vartheta_k},$$

$$M_T(t) = \mathbb{E}\left[e^{\langle t, T(X) \rangle} \right] = e^{A(t+\vartheta) - A(\vartheta)}.$$

Proof. Without loss of generality, suppose that X is a continuous random variable. We observe that:

$$\int_{\mathbb{R}} f(x; \vartheta) dx = 1 \quad \Rightarrow \quad \int_{\mathbb{R}} h(x) e^{\langle \vartheta, T(x) \rangle} dx = e^{A(\vartheta)}.$$

In differentiating this expression with respect to ϑ_j, we can interchange the order of differentiation and integration by suitable application of the dominated convergence theorem in order to get that:

$$\int_{\mathbb{R}} h(x)T_j(x)e^{\langle\vartheta,T(x)\rangle}dx = \frac{\partial A}{\partial\vartheta_j}e^{A(\vartheta)} \quad \Rightarrow$$

$$\frac{\partial A}{\partial\vartheta_j} = \int_{\mathbb{R}} T_j(x)\underbrace{h(x)e^{\langle\vartheta,T(x)\rangle-A(\vartheta)}}_{f(x;\vartheta)}dx = \mathbb{E}[T_j(X)].$$

By differentiating with respect to ϑ_k, we calculate that:

$$\int_{\mathbb{R}} h(x)T_j(x)T_k(x)e^{\langle\vartheta,T(x)\rangle}dx = \left(\frac{\partial A}{\partial\vartheta_j}\frac{\partial A}{\partial\vartheta_k} + \frac{\partial^2 A}{\partial\vartheta_j\partial\vartheta_k}\right)e^{A(\vartheta)} \quad \Rightarrow$$

$$\mathbb{E}\left[T_j(X)T_k(X)\right] = \frac{\partial A}{\partial\vartheta_j}\frac{\partial A}{\partial\vartheta_k} + \frac{\partial^2 A}{\partial\vartheta_j\partial\vartheta_k} \quad \Rightarrow$$

$$\frac{\partial^2 A}{\partial\vartheta_j\partial\vartheta_k} = \mathbb{E}\left[T_j(X)T_k(X)\right] - \mathbb{E}\left[T_j(X)\right]\mathbb{E}\left[T_k(X)\right] = \operatorname{Cov}\left[T_j(X),T_k(X)\right].$$

For $j = k$, we observe that:

$$\operatorname{Cov}\left[T_j(X),T_j(X)\right] = \operatorname{Var}\left[T_j(X)\right] = \frac{\partial^2 A}{\partial\vartheta_j^2}.$$

As far as the moment generating function is concerned, we calculate that:

$$\begin{aligned} M_T(t) &= \int_{\mathbb{R}} e^{\langle t,T(x)\rangle}f(x;\vartheta)dx = \int_{\mathbb{R}} e^{\langle t,T(x)\rangle}h(x)e^{\langle\vartheta,T(x)\rangle-A(\vartheta)}dx \\ &= e^{-A(\vartheta)}\int_{\mathbb{R}} h(x)e^{\langle t+\vartheta,T(x)\rangle}dx \\ &= e^{A(t+\vartheta)-A(\vartheta)}\int_{\mathbb{R}}\underbrace{h(x)e^{\langle t+\vartheta,T(x)\rangle-A(t+\vartheta)}}_{f(x;t+\vartheta)}dx \\ &= e^{A(t+\vartheta)-A(\vartheta)}. \end{aligned}$$

□

Note 2.4. If $Q(\vartheta) \neq \vartheta$, then the exponential family may be converted to canonical form with the reparameterization $\eta = Q(\vartheta)$.

Example 2.4. (Normal with mean ϑ_1 and variance ϑ_2)

$$\begin{aligned} f(x;\vartheta) &= \frac{1}{\sqrt{2\pi\vartheta_2}}\exp\left\{-\frac{1}{2\vartheta_2}(x-\vartheta_1)^2\right\} \\ &= \frac{1}{\sqrt{2\pi}}\exp\left\{\frac{\vartheta_1}{\vartheta_2}x - \frac{1}{2\vartheta_2}x^2 - \frac{\vartheta_1^2}{2\vartheta_2} - \frac{1}{2}\log\vartheta_2\right\}, \end{aligned}$$

$$h(x) = \frac{1}{\sqrt{2\pi}}, \quad Q(\vartheta) = \left(\frac{\vartheta_1}{\vartheta_2}, -\frac{1}{2\vartheta_2}\right), \quad T(x) = \left(x, x^2\right), \quad A(\vartheta) = \frac{\vartheta_1^2}{2\vartheta_2} + \frac{1}{2}\log\vartheta_2.$$

Consider the following reparameterization:

$$\eta = (\eta_1, \eta_2) = \left(\frac{\vartheta_1}{\vartheta_2}, -\frac{1}{2\vartheta_2}\right) \in \mathbb{R} \times (-\infty, 0) \quad \Rightarrow$$

$$\vartheta_2 = -\frac{1}{2\eta_2}, \quad \vartheta_1 = -\frac{\eta_1}{2\eta_2},$$

$$f(x;\eta) = \frac{1}{\sqrt{2\pi}} \exp\left\{\eta_1 x + \eta_2 x^2 + \frac{\eta_1^2}{4\eta_2} + \frac{1}{2}\log\left(-2\eta_2\right)\right\},$$

$$A(\eta) = -\frac{\eta_1^2}{4\eta_2} - \frac{1}{2}\log\left(-2\eta_2\right).$$

Then, it follows that:

$$\mathbb{E}\left[T_1(X)\right] = \mathbb{E}\left(X\right) = \frac{\partial A(\eta)}{\partial \eta_1} = -\frac{\eta_1}{2\eta_2} = \vartheta_1,$$

$$\mathbb{E}\left[T_2(X)\right] = \mathbb{E}\left(X^2\right) = \frac{\partial A(\eta)}{\partial \eta_2} = \frac{\eta_1^2}{4\eta_2^2} - \frac{1}{2\eta_2} = \vartheta_1^2 + \vartheta_2,$$

$$\mathrm{Var}\left[T_1(X)\right] = \mathrm{Var}\left(X\right) = \frac{\partial^2 A(\eta)}{\partial \eta_1^2} = -\frac{1}{2\eta_2} = \vartheta_2,$$

$$\mathrm{Var}\left[T_2(X)\right] = \mathrm{Var}\left(X^2\right) = \frac{\partial^2 A(\eta)}{\partial \eta_2^2} = -\frac{\eta_1^2}{2\eta_2^3} + \frac{1}{2\eta_2^2} = 4\vartheta_1^2\vartheta_2 + 2\vartheta_2^2,$$

$$\mathrm{Cov}\left[T_1(X), T_2(X)\right] = \mathrm{Cov}\left(X, X^2\right) = \frac{\partial^2 A(\eta)}{\partial \eta_1 \partial \eta_2} = \frac{\eta_1}{2\eta_2^2} = 2\vartheta_1\vartheta_2. \quad \square$$

Definition 2.4. A multivariate distribution with unknown parameter $\vartheta \in \Theta \subseteq \mathbb{R}^s$ and joint PMF or PDF $f(x;\vartheta)$ for $x \in S \subseteq \mathbb{R}^n$ belongs to the *multivariate exponential family* if the support S doesn't depend on the value of ϑ and it holds that:

$$f(x;\vartheta) = h(x)e^{\langle Q(\vartheta), T(x)\rangle - A(\vartheta)}.$$

Proposition 2.3. Suppose that $X_1, \ldots, X_n$ are independent and identically distributed (iid) random variables from a distribution which belongs to the univariate exponential family. Then, the joint distribution of the random vector $X = (X_1, \ldots, X_n)$ belongs to the multivariate exponential family with PMF or PDF given by:

$$f(x;\vartheta) = h^*(x)e^{\langle Q(\vartheta), T^*(x)\rangle - A^*(\vartheta)}, \quad x \in S^n,$$

$$h^*(x) = \prod_{i=1}^{n} h(x_i), \quad T^*(x) = \sum_{i=1}^{n} T(x_i), \quad A^*(\vartheta) = nA(\vartheta).$$

Proof. We calculate that:

$$\begin{aligned} f(x;\vartheta) &= \prod_{i=1}^{n} f(x_i;\vartheta) = \prod_{i=1}^{n} h(x_i) e^{\langle Q(\vartheta), T(x_i)\rangle - A(\vartheta)} \\ &= \prod_{i=1}^{n} h(x_i) \exp\left\{ \sum_{i=1}^{n} \langle Q(\vartheta), T(x_i)\rangle - nA(\vartheta) \right\} \\ &= h^*(x) \exp\left\{ \left\langle Q(\vartheta), \sum_{i=1}^{n} T(x_i) \right\rangle - A^*(\vartheta) \right\} \\ &= h^*(x) e^{\langle Q(\vartheta), T^*(x)\rangle - A^*(\vartheta)}. \end{aligned}$$

□

Chapter 3

Point Estimation

3.1 Introduction

Definition 3.1. i. An n-dimensional random vector $X = (X_1, \ldots, X_n)$ is called a *sample* of size n.

ii. An n-dimensional random vector $X = (X_1, \ldots, X_n)$ is called a *random sample* of size n if the random variables $X_1, \ldots, X_n$ are independent and identically distributed (iid).

Definition 3.2. i. A function $T(X) = T(X_1, \ldots, X_n)$ which doesn't depend on the value of the unknown parameter ϑ is called a *statistic*.

ii. A statistic $T(X)$ is called an *estimator* of the parametric function $g(\vartheta)$ if it holds that $T(S) \subseteq g(\Theta)$.

Note 3.1. As can be seen from the previous definition, we could consider any arbitrary function of the sample X as an estimator of ϑ, as long as this function takes values on the parameter space Θ. However, this condition alone is not enough to give us a good estimate of the true value of ϑ in practice. For this reason, various criteria have been developed to judge whether an estimator of ϑ is "good" or not. In this chapter we will study these criteria for "good" estimators, such as unbiasedness, the mean squared error criterion, sufficiency, efficiency and consistency. At the end of the chapter we will study 2 of the most widely used methods of finding estimators - the maximum likelihood method and the method of moments.

3.2 Unbiased Estimators

Definition 3.3. i. A statistic $T(X)$ is called an *unbiased estimator* of the parametric function $g(\vartheta)$ if it holds that $\mathbb{E}_\vartheta\left[T(X)\right] = g(\vartheta)\ \forall \vartheta \in \Theta$.

ii. The function $\text{bias}_{g(\vartheta)}\left[T(X)\right] = \mathbb{E}_\vartheta\left[T(X)\right] - g(\vartheta)$ is called the *bias* of the estimator $T(X)$ with respect to the parametric function $g(\vartheta)$.

Interpretation: The property of unbiasedness ensures that an estimator of ϑ takes values close to the true value of ϑ, but it doesn't provide any information about how tightly concentrated all the most probable values of the estimator are around that value. Therefore, this property doesn't suffice in order to characterize a "good" estimator, since it could potentially take values very far away from the true value of ϑ with high probability. In order to ensure that all the most probable values of the estimator are tightly concentrated around the true value of ϑ, we must also demand that the estimator have as small a variance as possible.

Note 3.2. We observe that a statistic $T(X)$ is an unbiased estimator of $g(\vartheta)$ if and only if $\text{bias}_{g(\vartheta)}\left[T(X)\right] = 0 \ \forall \vartheta \in \Theta$. For a given parametric function $g(\vartheta)$ there may not exist any unbiased estimator, there may exist a unique unbiased estimator, or there may exist multiple unbiased estimators.

Definition 3.4. i. The statistic $\overline{X} = \frac{1}{n}\sum_{i=1}^n X_i$ is called the *sample mean*.

ii. Consider the following statistic:

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

which is called the *sample variance*.

Proposition 3.1. Let $X_1, \dots, X_n$ be a random sample from a distribution with unknown parameter ϑ. Then, it follows that:

i. The sample mean $\overline{X}$ is an unbiased estimator of $g_1(\vartheta) = \mathbb{E}_\vartheta(X_1)$;

ii. $\text{Var}_\vartheta\left(\overline{X}\right) = \frac{1}{n}\text{Var}_\vartheta(X_1)$;

iii. The sample variance S^2 is an unbiased estimator of $g_2(\vartheta) = \text{Var}_\vartheta(X_1)$.

Proof. i. We calculate that:

$$\mathbb{E}_\vartheta\left(\overline{X}\right) = \frac{1}{n}\mathbb{E}_\vartheta\left(\sum_{i=1}^n X_i\right) = \frac{1}{n}\sum_{i=1}^n \mathbb{E}_\vartheta(X_i) = \frac{1}{n}\sum_{i=1}^n \mathbb{E}_\vartheta(X_1) = \mathbb{E}_\vartheta(X_1) = g_1(\vartheta).$$

ii. We calculate that:

$$\text{Var}_\vartheta\left(\overline{X}\right) = \frac{1}{n^2}\text{Var}_\vartheta\left(\sum_{i=1}^n X_i\right) = \frac{1}{n^2}\sum_{i=1}^n \text{Var}_\vartheta(X_i) = \frac{1}{n}\text{Var}_\vartheta(X_1).$$

iii. We calculate that:

$$\mathbb{E}_\vartheta\left(\overline{X}^2\right) = \operatorname{Var}_\vartheta\left(\overline{X}\right) + \left[\mathbb{E}_\vartheta\left(\overline{X}\right)\right]^2 = \frac{1}{n}\operatorname{Var}_\vartheta(X_1) + \left[\mathbb{E}_\vartheta(X_1)\right]^2,$$

$$\begin{aligned}\mathbb{E}_\vartheta\left(S^2\right) &= \frac{1}{n-1}\left[\sum_{i=1}^n \mathbb{E}_\vartheta\left(X_i^2\right) - n\mathbb{E}_\vartheta\left(\overline{X}^2\right)\right] \\ &= \frac{n\operatorname{Var}_\vartheta(X_1) + n\left[\mathbb{E}_\vartheta(X_1)\right]^2 - \operatorname{Var}_\vartheta(X_1) - n\left[\mathbb{E}_\vartheta(X_1)\right]^2}{n-1} \\ &= \operatorname{Var}_\vartheta(X_1) = g_2(\vartheta).\end{aligned}$$

□

Example 3.1. Let $X_1, \ldots, X_n \sim \operatorname{Bin}(N, p)$ be a random sample with known N. We know that $\mathbb{E}(X_1) = Np$ and $\operatorname{Var}(X_1) = Np(1-p)$. According to the previous note, it follows that $\mathbb{E}(\overline{X}) = Np$ and $\mathbb{E}\left(S^2\right) = Np(1-p)$. Furthermore, we observe that:

$$\mathbb{E}\left(\frac{1}{N}\overline{X}\right) = p, \quad \mathbb{E}\left(\frac{1}{N}S^2\right) = p(1-p).$$

Therefore, $T_1(X) = \frac{1}{N}\overline{X}$ is an unbiased estimator of p and $T_2(X) = \frac{1}{N}S^2$ is an unbiased estimator of the parametric function $g(p) = p(1-p)$. □

Example 3.2. Let $X \sim \operatorname{Poisson}(\lambda)$ be a sample of size 1. We want to show that there doesn't exist any unbiased estimator of the parametric function $g(\lambda) = \frac{1}{\lambda}$. Suppose that the statistic $T(X)$ is an unbiased estimator of $g(\lambda)$, i.e. it holds that:

$$\mathbb{E}\left[T(X)\right] = g(\lambda) \quad \Leftrightarrow \quad \sum_{x=0}^\infty T(x)e^{-\lambda}\frac{\lambda^x}{x!} = \frac{1}{\lambda} \quad \Leftrightarrow$$

$$\lambda\sum_{x=0}^\infty T(x)\frac{\lambda^x}{x!} = e^\lambda \quad \Leftrightarrow \quad \sum_{x=0}^\infty \frac{T(x)}{x!}\lambda^{x+1} = \sum_{x=0}^\infty \frac{\lambda^x}{x!} \quad \Leftrightarrow$$

$$\sum_{x=1}^\infty \frac{T(x-1)}{(x-1)!}\lambda^x = \sum_{x=0}^\infty \frac{1}{x!}\lambda^x.$$

Since the left-hand side is a power series without a constant term and the right-hand side is a power series with a constant term equal to 1, it's impossible for them to be equal to each other. Thus, there doesn't exist any unbiased estimator of $g(\lambda) = \frac{1}{\lambda}$. □

3.3 Mean Squared Error

Definition 3.5. The function $\operatorname{MSE}_{g(\vartheta)}\left[T(X)\right] = \mathbb{E}_\vartheta[(T(X) - g(\vartheta))^2]$ is called the *mean squared error* (MSE) of the estimator $T(X)$ with respect to the parametric function $g(\vartheta)$.

Mean Squared Error Criterion: An estimator $T^*(X)$ of $g(\vartheta)$ is considered "better" than some other estimator $T(X)$ of $g(\vartheta)$ according to the MSE criterion if it holds that $\text{MSE}_{g(\vartheta)}\left[T^*(X)\right] \leqslant \text{MSE}_{g(\vartheta)}\left[T(X)\right]\ \forall\vartheta \in \Theta$.

Note 3.3. The mean squared error function can be decomposed as follows:

$$\text{MSE}_{g(\vartheta)}\left[T(X)\right] = \text{Var}_{\vartheta}\left[T(X)\right] + \text{bias}^2_{g(\vartheta)}\left[T(X)\right].$$

If $\text{bias}_{g(\vartheta)}\left[T(X)\right] = 0$, i.e. $T(X)$ is an unbiased estimator of $g(\vartheta)$, then we observe that $\text{MSE}_{g(\vartheta)}\left[T(X)\right] = \text{Var}_{\vartheta}\left[T(X)\right]$. In other words, if we restrict ourselves to considering only unbiased estimators of $g(\vartheta)$, then the "best" among them according to the MSE criterion is the one which achieves the smallest possible variance. This unbiased estimator which achieves the smallest possible variance is called the uniformly minimum-variance unbiased estimator (UMVUE) $g(\vartheta)$, and we will study some of its properties in section 3.7. However, this doesn't exclude the possibility of there existing a biased estimator of $g(\vartheta)$ with smaller MSE than the UMVUE of $g(\vartheta)$, and thus smaller MSE than any other unbiased estimator of $g(\vartheta)$.

3.4 Sufficiency

Definition 3.6. A statistic $T(X)$ is called *sufficient* for the parameter ϑ if the conditional distribution of the sample X given that $T(X) = t$ doesn't depend on the value of ϑ $\forall\vartheta \in \Theta$ and $\forall t \in T(S)$.

Interpretation: A sufficient statistic gathers all the information contained in the sample for the unknown parameter. In other words, it suffices to compute the value of a sufficient statistic from a sample of observations, and we will have all the information we need to estimate the unknown parameter, without further access to the individual observations.

Theorem 3.1. (Fisher - Neyman Factorization Criterion) Let X be a sample with joint PMF or PDF $f(x;\vartheta)$ for $\vartheta \in \Theta$ and $x \in S$. A statistic $T(X)$ is sufficient for the parameter ϑ if and only if there exist non-negative functions g, h such that:

$$f(x;\vartheta) = g(T(x),\vartheta)h(x).$$

Proof. Assume that the distribution of the sample is discrete. First, suppose that $f(x;\vartheta) = g(T(x),\vartheta)h(x)$. We know that:

$$\mathbb{P}(X = x \mid T(X) = t) = \frac{\mathbb{P}(X = x, T(X) = t)}{\mathbb{P}(T(X) = t)}.$$

We discern the following 2 cases:

1. If $T(x) \neq t$, then it follows that:

$$\{X = x\} \cap \{T(X) = t\} = \varnothing \quad \Rightarrow \quad \mathbb{P}(X = x, T(X) = t) = 0 \quad \Rightarrow$$

$$\mathbb{P}(X = x \mid T(x) = t) = 0.$$

2. If $T(x) = t$, then we calculate that:

$$\begin{aligned}\mathbb{P}(X = x \mid T(X) = t) = \frac{\mathbb{P}(X = x)}{\mathbb{P}(T(X) = t)} &= \frac{g(t, \vartheta)h(x)}{\sum_{x:\ T(x)=t} f(x; \vartheta)} \\ &= \frac{g(t, \vartheta)h(x)}{\sum_{x:\ T(x)=t} g(t, \vartheta)h(x)} = \frac{h(x)}{\sum_{x:\ T(x)=t} h(x)},\end{aligned}$$

   which does not depend on the value of ϑ.

In both cases, the conditional probability $\mathbb{P}(X = x \mid T(X) = t)$ does not depend on the value of ϑ, which implies that $T(X)$ is a sufficient statistic for ϑ.

Conversely, suppose that $T(X)$ is a sufficient statistic for ϑ, which implies that there exists a function $\varphi(x, t)$ such that $\mathbb{P}(X = x \mid T(X) = t) = \varphi(x, t)$. Once again, we discern the following 2 cases:

1. If $T(x) \neq t$, then it follows that $\varphi(x, t) = 0$.
2. If $T(x) = t$, then it follows that $\varphi(x, t) = \varphi(x, T(x)) = h(x)$ for some function h. Then, we calculate that:

$$\begin{aligned}f(x; \vartheta) &= \mathbb{P}(X = x \mid T(X) = t)\mathbb{P}(T(X) = t) = \varphi(x, t)\mathbb{P}(T(X) = t) \\ &= h(x) f_T(t; \vartheta) = h(x) g(T(x), \vartheta).\end{aligned}$$

The proof for the general case can be found in Keener, Section 6.4. □

Note 3.4. Suppose that the statistic $T(X)$ is sufficient for ϑ. If $x, y \in S$ with $T(x) = T(y)$, then we observe that:

$$\frac{f(x; \vartheta)}{f(y; \vartheta)} = \frac{h(x)}{h(x)},$$

which doesn't depend on the value of ϑ. Conversely, if that ratio depends on the value of ϑ, then the statistic $T(X)$ isn't sufficient for ϑ.

Corollary 3.1. Suppose that the statistic $T(X)$ is sufficient for ϑ.

i. If it holds that $T = \psi(T^*)$ for some function ψ, then $T^*(X)$ is sufficient for ϑ.

ii. If it holds that $\vartheta = \varphi(\eta)$ for some function φ, then $T(X)$ is sufficient for η too.

Proof. i. According to the Fisher - Neyman factorization criterion, it follows that:

$$f(x;\vartheta) = g(T(x),\vartheta)h(x) = g\left(\psi\left(T^*(x)\right),\vartheta\right)h(x) = g^*\left(T^*(x),\vartheta\right)h(x),$$

where $g^*\left(t,\vartheta\right) = g\left(\psi(t),\vartheta\right)$. Hence, $T^*(X)$ is a sufficient statistic for ϑ.

ii. According to the Fisher - Neyman factorization criterion, it follows that:

$$f(x;\vartheta) = g(T(x),\vartheta)h(x) = g\left(T(x),\varphi(\eta)\right)h(x) = g^*\left(T(x),\eta\right)h(x),$$

where $g^*\left(t,\eta\right) = g\left(t,\varphi(\eta)\right)$. Hence, $T(X)$ is a sufficient statistic for η. □

Example 3.3. Let $X_1,\dots,X_n \sim \text{Exp}(\lambda)$ be a random sample. We know that:

$$f(x;\lambda) = \prod_{i=1}^{n} f(x_i;\lambda) = \prod_{i=1}^{n}\left[\lambda e^{-\lambda x_i}\mathbb{1}_{(0,\infty)}(x_i)\right] = \lambda^n \exp\left\{-\lambda\sum_{i=1}^{n} x_i\right\}\mathbb{1}_{(0,\infty)^n}(x),$$

where $T(x) = \sum_{i=1}^n x_i$, $g(t,\lambda) = \lambda^n e^{-\lambda t}$ and $h(x) = \mathbb{1}_{(0,\infty)^n}(x)$. According to the Fisher - Neyman factorization theorem, it follows that the statistic $T(X) = \sum_{i=1}^n X_i$ is sufficient for λ. □

Example 3.4. Let $X_1,\dots,X_n \sim \text{Laplace}(\mu,\lambda)$ be a random sample with $\mu \in \mathbb{R}$, known $\lambda > 0$ and PDF $f(x;\mu) = \frac{\lambda}{2}e^{-\lambda|x-\mu|}$ for $x \in \mathbb{R}$. We calculate that:

$$f(x;\mu) = \left(\frac{\lambda}{2}\right)^n \exp\left\{-\lambda\sum_{i=1}^{n}|x_i-\mu|\right\},$$

where $T(x) = (x_1,x_2,\dots,x_n)$, $g(t,\mu) = e^{-\lambda\sum_{i=1}^n|t_i-\mu|}$ and $h(x) = \left(\frac{\lambda}{2}\right)^n$. According to the Fisher - Neyman factorization theorem, $T(X) = (X_1,X_2,\dots,X_n)$ is sufficient for μ. We observe that the sufficient statistic we calculated was the entire sample X, and we wouldn't have been able to find any lower-dimensional sufficient statistic than that. The term $\sum_{i=1}^n|X_i-\mu|$ which appears in the joint PDF of the sample doesn't constitute a statistic, since it depends on value of the unknown parameter μ. □

Definition 3.7. We denote the *order statistics* of the sample X by $X_{(1)},\dots,X_{(n)}$. In particular, it holds that $X_{(1)} = \min\{X_1,\dots,X_n\}$ and $X_{(n)} = \max\{X_1,\dots,X_n\}$.

Example 3.5. Let $X_1,\dots,X_n \sim \mathcal{U}(0,\vartheta)$ be a random sample. We know that:

$$\begin{aligned} f(x;\vartheta) &= \frac{1}{\vartheta^n}\prod_{i=1}^{n}\mathbb{1}_{[0,\vartheta]}(x_i) = \vartheta^{-n}\mathbb{1}_{[0,\vartheta]}\left(x_{(1)}\right)\mathbb{1}_{[0,\vartheta]}\left(x_{(n)}\right) \\ &= \vartheta^{-n}\mathbb{1}_{[0,\infty)}\left(x_{(1)}\right)\mathbb{1}_{(-\infty,\vartheta]}\left(x_{(n)}\right), \end{aligned}$$

where $T(x) = x_{(n)}$, $g(t,\vartheta) = \vartheta^{-n}\mathbb{1}_{(-\infty,\vartheta]}(t)$ and $h(x) = \mathbb{1}_{[0,\infty)}\left(x_{(1)}\right)$. According to the Fisher - Neyman factorization theorem, $T(X) = X_{(n)}$ is sufficient for ϑ. □

Example 3.6. Let $X_1, \dots, X_n \sim \mathcal{U}(\vartheta, \vartheta+1)$ be a random sample. We know that:

$$f(x;\vartheta) = \prod_{i=1}^{n} \mathbb{1}_{[\vartheta,\vartheta+1]}(x_i) = \mathbb{1}_{[\vartheta,\vartheta+1]}\left(x_{(1)}\right) \mathbb{1}_{[\vartheta,\vartheta+1]}\left(x_{(n)}\right),$$

where $T(x) = \left(x_{(1)}, x_{(n)}\right)$, $g(t_1, t_2, \vartheta) = \mathbb{1}_{[\vartheta,\vartheta+1]}(t_1)\, \mathbb{1}_{[\vartheta,\vartheta+1]}(t_2)$ and $h(x) = 1$. According to the Fisher - Neyman factorization theorem, it follows that the statistic $T(X) = \left(X_{(1)}, X_{(n)}\right)$ is sufficient for ϑ. □

Example 3.7. Let $X_1, \dots, X_n$ be a random sample with $f(x;\lambda,k) = \lambda e^{-\lambda(x-k)}$ for $\lambda > 0$, $k \in \mathbb{R}$ and $x \geqslant k$. We calculate that:

$$\begin{aligned} f(x;\lambda,k) &= \lambda^n \exp\left\{-\lambda \sum_{i=1}^{n}(x_i - k)\right\} \prod_{i=1}^{n} \mathbb{1}_{[k,\infty)}(x_i) \\ &= \lambda^n \exp\left\{-\lambda \sum_{i=1}^{n} x_i + n\lambda k\right\} \mathbb{1}_{[k,\infty)}\left(x_{(1)}\right), \end{aligned}$$

where $T(x) = \left(\sum_{i=1}^{n} x_i, x_{(1)}\right)$, $g(t_1, t_2, \lambda, k) = \lambda^n e^{-\lambda t_1 + n\lambda k} \mathbb{1}_{[k,\infty)}(t_2)$ and $h(x) = 1$. According to the Fisher - Neyman factorization theorem, it follows that the statistic $T(X) = \left(\sum_{i=1}^{n} X_i, X_{(1)}\right)$ is sufficient for $\vartheta = (\lambda, k)$. □

Proposition 3.2. (Sufficiency in the Exponential Family) Suppose that the distribution of the sample X belongs to the multivariate exponential family with PMF or PDF $f(x;\vartheta) = h(x) e^{\langle Q(\vartheta), T(x)\rangle - A(\vartheta)}$ for $\vartheta \in \Theta$ and $x \in S$. Then, the statistic $T(X)$ is sufficient for ϑ.

Proof. We observe that the PMF or PDF can be written as $f(x;\vartheta) = g(T(x),\vartheta)h(x)$, where $g(t,\vartheta) = e^{\langle Q(\vartheta), t\rangle - A(\vartheta)}$. According to the Fisher - Neyman factorization criterion, it immediately follows that the statistic $T(X)$ is sufficient for ϑ. □

Example 3.8. Let $X_1, \dots, X_n \sim \mathcal{N}\left(\vartheta, \vartheta^2\right)$ be a random sample with $\vartheta \neq 0$. We calculate that:

$$\begin{aligned} f(x;\vartheta) &= \left(\frac{1}{\sqrt{2\pi\vartheta^2}}\right)^n \exp\left\{-\frac{1}{2\vartheta^2} \sum_{i=1}^{n}(x_i - \vartheta)^2\right\} \\ &= (2\pi)^{-n/2} |\vartheta|^{-n} \exp\left\{\frac{1}{\vartheta} \sum_{i=1}^{n} x_i - \frac{1}{2\vartheta^2} \sum_{i=1}^{n} x_i^2 - \frac{n}{2}\right\} \\ &= (2\pi e)^{-n/2} \exp\left\{\frac{1}{\vartheta} \sum_{i=1}^{n} x_i - \frac{1}{2\vartheta^2} \sum_{i=1}^{n} x_i^2 - n \log|\vartheta|\right\}, \end{aligned}$$

where we let $h(x) = (2\pi e)^{-n/2}$, $Q(\vartheta) = \left(\frac{1}{\vartheta}, -\frac{1}{2\vartheta^2}\right)$, $T(x) = \left(\sum_{i=1}^{n} x_i, \sum_{i=1}^{n} x_i^2\right)$ and

$A(\vartheta) = n \log|\vartheta|$. According to the proposition about sufficiency in the exponential family, it follows that the statistic $T(X) = \left(\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2\right)$ is sufficient for ϑ. We observe that the distribution of the sample only has 1 unknown parameter, whereas the sufficient statistic $T(X)$ is 2-dimensional, so it's a curved exponential family. □

Note 3.5. To sum up, we have 3 methods at our disposal for showing that a statistic is sufficient for some unknown parameter: the definition of sufficiency (usually unwieldy in practice), the Fisher - Neyman factorization theorem (more straightforward than the definition) and the proposition about sufficiency in the exponential family (which may be easily combined with proving the completeness of the statistic). In table 3.1, we summarize some notable sufficient statistics for the parameters of some widely used families of distributions.

Bernoulli(p)	$\sum_{i=1}^n X_i$
Bin(N,p) with known N	
Geom(p)	
NegBin(N,p) with known N	
Poisson(λ)	
Exp(λ)	
$\mathcal{N}\left(\mu,\sigma^2\right)$ with known σ^2	
$\mathcal{N}\left(\mu,\sigma^2\right)$ with known μ	$\sum_{i=1}^n (X_i-\mu)^2$
$\mathcal{N}\left(\mu,\sigma^2\right)$	$\left(\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2\right)$
Gamma(k,λ)	$(\sum_{i=1}^n \log X_i, \sum_{i=1}^n X_i)$
Beta$(\vartheta_1,\vartheta_2)$	$(\sum_{i=1}^n \log X_i, \sum_{i=1}^n \log(1-X_i))$
$\mathcal{U}(\vartheta_1,\vartheta_2)$	$\left(X_{(1)}, X_{(n)}\right)$

Table 3.1: Notable Sufficient Statistics

3.5 Completeness

Definition 3.8. A statistic $T(X)$ is called *complete* (for the distribution of the sample) if $\mathbb{E}_\vartheta[\varphi(T)] = 0\ \forall\vartheta\in\Theta$ implies that $\varphi(T) = 0$ with probability 1 (almost surely) for any function φ.

Note 3.6. We observe that any unbiased estimator of 0 which is a function of a complete statistic must be almost identically equal to 0. Therefore, if there exist 2 different functions $\varphi(T)$ and $\psi(T)$ which are both unbiased estimators of $g(\vartheta)$, then the statistic $T(X)$ cannot be complete. Conversely, if $T(X)$ is a complete statistic, then there exists at most one function $\varphi(T)$ which is an unbiased estimator of $g(\vartheta)$.

Definition 3.9. A statistic $A(X)$ whose distribution doesn't depend on any unknown parameter ϑ is called *ancillary*.

Note 3.7. If it holds that $A(X) = \varphi(T)$ for some function φ and $A(X)$ is an ancillary statistic, then $T(X)$ cannot be complete. Conversely, if $T(X)$ is a complete statistic, then any function of it cannot be ancillary. In other words, any function of a complete statistic is informative about the unknown parameter ϑ.

Proposition 3.3. If $T(X)$ is a complete statistic and it holds that $T = \psi(T^*)$ for some injective function ψ, then $T^*(X)$ is also a complete statistic.

Proof. Suppose that $\mathbb{E}_\vartheta[\varphi(T^*)] = 0 \ \forall \vartheta \in \Theta$. Since the statistic $T(X)$ is complete and the function ψ is invertible, it follows that:

$$\mathbb{E}_\vartheta[\varphi(T^*)] = \mathbb{E}_\vartheta[\varphi\left(\psi^{-1}(T)\right)] = \mathbb{E}_\vartheta[(\varphi \circ \psi^{-1})(T)] = 0 \quad \Rightarrow$$

$$(\varphi \circ \psi^{-1})(T) = \varphi\left(\psi^{-1}(T)\right) = \varphi(T^*) \stackrel{\text{a.s.}}{=} 0.$$

Therefore, $T^*(X)$ is also a complete statistic. □

Note 3.8. In practice, first we find a sufficient statistic for ϑ using one of the methods presented in the previous paragraph, and then we check if it's also complete. To check whether the definition of completeness holds, we need to determine the distribution of the sufficient statistic $T(X)$, so that we can calculate the expectation $\mathbb{E}_\vartheta[\varphi(T)]$. There are 2 notable cases to consider:

i. If the statistic $T(X) = \sum_{i=1}^n X_i$ is sufficient for ϑ, the distribution of $T(X)$ follows directly from the properties of MGFs.

ii. If $X_1, \dots, X_n$ is a random sample with PDF $f(x;\vartheta)$, CDF $F(x;\vartheta)$ and sufficient statistic $T(X) = X_{(n)}$ or $T(X) = X_{(1)}$ for ϑ, then the PDF of $T(X)$ can be calculated as follows:

$$\begin{aligned} F_{X_{(n)}}(x) &= \mathbb{P}\left(\max\{X_1, \dots, X_n\} \leqslant x\right) = \mathbb{P}\left(X_1 \leqslant x, \dots, X_n \leqslant x\right) \\ &= \mathbb{P}(X_1 \leqslant x) \cdots \mathbb{P}(X_n \leqslant x) = \left[F(x;\vartheta)\right]^n, \end{aligned}$$

$$f_{X_{(n)}}(x) = n f(x;\vartheta) \left[F(x;\vartheta)\right]^{n-1},$$

$$\begin{aligned} F_{X_{(1)}}(x) &= 1 - \mathbb{P}\left(\min\{X_1, \dots, X_n\} > x\right) = 1 - \mathbb{P}(X_1 > x, \dots, X_n > x) \\ &= 1 - \mathbb{P}(X_1 > x) \cdots \mathbb{P}(X_n > x) = 1 - \left[1 - F(x;\vartheta)\right]^n, \end{aligned}$$

$$f_{X_{(1)}}(x) = n f(x;\vartheta) \left[1 - F(x;\vartheta)\right]^{n-1}.$$

Note 3.9. To compute the expectation $\mathbb{E}_\vartheta[\varphi(T)]$, we distinguish the following cases:

i. The distribution of T is discrete: The expectation takes the form of a series (or a sum). Specifically, if it takes the following form:

$$\mathbb{E}_\vartheta[\varphi(T)] = \sum_{t=0}^{\infty} \varphi(t)\psi(t)\,[u(\vartheta)]^t = 0, \quad \forall \vartheta \in \Theta,$$

then we infer that $\varphi(t)\psi(t) = 0\ \forall t \in T\,(S)$. Furthermore, if $\psi(t) \neq 0\ \forall t \in T\,(S)$, then we conclude that $\varphi(t) = 0\ \forall t \in T\,(S)$.

ii. The distribution of T is continuous and its support is $(0, \infty)$: Suppose that the integral takes the following form:

$$\mathbb{E}_\vartheta[\varphi(T)] = \int_0^\infty \varphi(t)\psi(t)w(\vartheta)e^{-u(\vartheta)t}dt = w(\vartheta)\int_0^\infty \varphi(t)\psi(t)e^{-u(\vartheta)t}dt = 0,$$

$\forall \vartheta \in \Theta$. If $w(\vartheta) \neq 0\ \forall \vartheta \in \Theta$, then if follows that:

$$\int_0^\infty \varphi(t)\psi(t)e^{-u(\vartheta)t}dt = 0, \quad \forall \vartheta \in \Theta.$$

The last integral is the Laplace transform of the function $\varphi(t)\psi(t)$ evaluated at $u(\vartheta)$. We know that the Laplace transform is injective on classes of almost surely equal functions. Additionally, the Laplace transform of the zero function is equal to 0, so we infer that $\varphi(t)\psi(t) = 0$ almost surely. Furthermore, if $\psi(t) \neq 0$, then we conclude that $\varphi(t) = 0$ almost surely.

iii. The distribution of T is continuous and its support is the real line: Similarly, suppose that the integral takes the following form:

$$\mathbb{E}_\vartheta[\varphi(T)] = \int_{-\infty}^\infty \varphi(t)\psi(t)w(\vartheta)e^{-u(\vartheta)t}dt = w(\vartheta)\int_{-\infty}^\infty \varphi(t)\psi(t)e^{-u(\vartheta)t}dt = 0,$$

$\forall \vartheta \in \Theta$. If $w(\vartheta) \neq 0\ \forall \vartheta \in \Theta$, then it follows that:

$$\int_{-\infty}^\infty \varphi(t)\psi(t)e^{-u(\vartheta)t}dt = 0, \quad \forall \vartheta \in \Theta.$$

The last integral is the two-sided Laplace transform of the function $\varphi(t)\psi(t)$ evaluated at $u(\vartheta)$, which is also injective on classes of almost surely equal functions. The two-sided Laplace transform of the zero function is also equal to 0. If $\varphi(t) \neq 0$, then we arrive at the desired result in the same manner as before.

iv. The distribution of T is continuous and its support depends on ϑ: The expectation takes the form of a Riemann integral with at least one integration limit which is a function of ϑ:

$$\mathbb{E}_\vartheta[\varphi(T)] = \int_a^{u(\vartheta)} \varphi(t)\psi(t)w(\vartheta)dt = w(\vartheta)\int_a^{u(\vartheta)} \varphi(t)\psi(t)dt = 0, \quad \forall \vartheta \in \Theta.$$

If $w(\vartheta) \neq 0 \ \forall \vartheta \in \Theta$, then we infer that:

$$\int_a^{u(\vartheta)} \varphi(t)\psi(t)dt = 0, \quad \forall \vartheta \in \Theta.$$

According to the fundamental theorem of calculus, it follows that:

$$u'(\vartheta)\varphi\left(u(\vartheta)\right)\psi\left(u(\vartheta)\right) = 0, \quad \forall \vartheta \in \Theta.$$

If $u'(\vartheta)\psi\left(u(\vartheta)\right) \neq 0 \ \forall \vartheta \in \Theta$, then we conclude that $\varphi\left(u(\vartheta)\right) = 0 \ \forall \vartheta \in \Theta$, i.e. $\varphi(t) = 0 \ \forall t \in u(\Theta)$. If $T(S) \subseteq u(\Theta)$, then the desired result follows.

Example 3.9. Let $X_1, \ldots, X_n \sim \text{Poisson}(\lambda)$ be a random sample. We know that $T(X) = \sum_{i=1}^n X_i \sim \text{Poisson}(n\lambda)$ is sufficient for λ. Suppose that $\mathbb{E}_\lambda[\varphi(T)] = 0 \ \forall \lambda > 0$. Then, we calculate that:

$$\mathbb{E}_\lambda[\varphi(T)] = \sum_{t=0}^{\infty} \varphi(t)\mathbb{P}_\lambda(T = t) = \sum_{t=0}^{\infty} \varphi(t) e^{-n\lambda t}\frac{(n\lambda)^t}{t!} = \sum_{t=0}^{\infty} \frac{\varphi(t)}{t!}\left(n\lambda e^{-n\lambda}\right)^t = 0,$$

$\forall \lambda > 0$. It follows that $\frac{\varphi(t)}{t!} = 0$ for $t = 0, 1, \ldots$, which implies that $\varphi(t) = 0$ for $t = 0, 1, \ldots$. Therefore, the statistic $T(X) = \sum_{i=1}^n X_i$ is complete. □

Example 3.10. Let $X_1, \ldots, X_n \sim \text{Exp}(\lambda)$ be a random sample. We know that the statistic $T(X) = \sum_{i=1}^n X_i \sim \text{Gamma}(n, \lambda)$ is sufficient for λ. Suppose that $\mathbb{E}_\lambda[\varphi(T)] = 0 \ \forall \lambda > 0$. Then, we calculate that:

$$\mathbb{E}_\lambda[\varphi(T)] = \int_0^\infty f_T(t)\varphi(t)dt = \frac{\lambda^n}{(n-1)!}\int_0^\infty t^{n-1}e^{-\lambda t}\varphi(t)dt = 0, \quad \forall \lambda > 0 \quad \Rightarrow$$

$$\int_0^\infty \varphi(t)t^{n-1}e^{-\lambda t}dt = 0, \quad \forall \lambda > 0.$$

The last integral is the Laplace transform of the function $\varphi(t)t^{n-1}$ evaluated at λ. According to note 3.9, we infer that $\varphi(t)t^{n-1} = 0 \ \forall t > 0$, which implies that $\varphi(t) = 0$ $\forall t > 0$. Therefore, the statistic $T(X) = \sum_{i=1}^n X_i$ is complete. □

Example 3.11. Let $X_1, \ldots, X_n \sim \mathcal{U}\left(\vartheta^2, 1\right)$ be a random sample with $\vartheta \in (0, 1)$. We can easily show that the statistic $T(X) = X_{(1)}$ is sufficient for ϑ. According to note 3.8, we calculate that:

$$f_{X_{(1)}}(t) = \frac{n}{\left(1-\vartheta^2\right)^n}(1-t)^{n-1}, \quad t \in \left(\vartheta^2, 1\right).$$

Suppose that $\mathbb{E}_\vartheta[\varphi(T)] = 0 \ \forall \vartheta \in (0, 1)$. Then, we calculate that:

$$\mathbb{E}_\vartheta[\varphi(T)] = \int_{\vartheta^2}^1 f_{X_{(1)}}(t)\varphi(t)dt = \frac{n}{\left(1-\vartheta^2\right)^n}\int_{\vartheta^2}^1 (1-t)^{n-1}\varphi(t)dt = 0, \quad \forall \vartheta \in (0, 1) \quad \Rightarrow$$

$$\int_{\vartheta^2}^{1}(1-t)^{n-1}\varphi(t)dt = 0, \quad \forall\vartheta \in (0,1).$$

According to the fundamental theorem of calculus, we infer that:

$$-2\vartheta\left(1-\vartheta^2\right)^{n-1}\varphi\left(\vartheta^2\right) = 0, \quad \forall\vartheta \in (0,1) \quad \Rightarrow \quad \varphi(t) = 0, \quad \forall t \in (0,1) \supseteq \left(\vartheta^2, 1\right).$$

Therefore, the statistic $T(X) = X_{(1)}$ is complete. □

Example 3.12. Let $X_1, \ldots, X_n \sim \mathcal{U}(-\vartheta, \vartheta)$ be a random sample with $\vartheta > 0$. We can easily show that the statistic $T(X) = \left(X_{(1)}, X_{(n)}\right)$ is sufficient for ϑ. According to note 3.8, we calculate that:

$$f_{X_{(n)}}(t) = \frac{n}{(2\vartheta)^n}(t+\vartheta)^{n-1}, \quad f_{X_{(1)}}(t) = \frac{n}{(2\vartheta)^n}(\vartheta - t)^{n-1},$$

$$\begin{aligned}
\mathbb{E}_\vartheta\left[X_{(n)}\right] &= \int_{-\vartheta}^{\vartheta} n(t+\vartheta)^{n-1}\frac{t}{(2\vartheta)^n}dt = \left[(t+\vartheta)^n\frac{t}{(2\vartheta)^n}\right]_{-\vartheta}^{\vartheta} - \frac{1}{(2\vartheta)^n}\int_{-\vartheta}^{\vartheta}(t+\vartheta)^n dt \\
&= \vartheta - \frac{1}{(2\vartheta)^n}\left[\frac{1}{(n+1)}(t+\vartheta)^{n+1}\right]_{-\vartheta}^{\vartheta} = \vartheta - \frac{2\vartheta}{n+1}, \\
\mathbb{E}_\vartheta\left[X_{(1)}\right] &= \int_{-\vartheta}^{\vartheta} n(\vartheta - t)^{n-1}\frac{t}{(2\vartheta)^n}dt = -\left[(\vartheta - t)^n\frac{t}{(2\vartheta)^n}\right]_{-\vartheta}^{\vartheta} + \frac{1}{(2\vartheta)^n}\int_{-\vartheta}^{\vartheta}(\vartheta - t)^n dt \\
&= -\vartheta - \frac{1}{(2\vartheta)^n}\left[\frac{1}{(n+1)}(\vartheta - t)^{n+1}\right]_{-\vartheta}^{\vartheta} = -\vartheta + \frac{2\vartheta}{n+1}.
\end{aligned}$$

We observe that $\mathbb{E}_\vartheta\left[X_{(1)} + X_{(n)}\right] = 0 \ \forall\vartheta > 0$, i.e. the statistic $X_{(1)} + X_{(n)}$ is an unbiased estimator of 0 which is a function of $T(X)$. According to note 3.6, the statistic $T(X) = \left(X_{(1)}, X_{(n)}\right)$ is **not** complete. Alternatively, we let $Y_i = |X_i|$ for $i = 1, 2, \ldots, n$ and calculate that:

$$\begin{aligned}
F_{Y_1}(y) &= \mathbb{P}\left(|X_1| \leqslant y\right) = \mathbb{P}(-y \leqslant X_1 \leqslant y) = \mathbb{P}(X_1 \leqslant y) - \mathbb{P}(X_1 < -y) \\
&= F(y;\vartheta) - F(-y;\vartheta) = \frac{y+\vartheta}{2\vartheta} - \frac{-y+\vartheta}{2\vartheta} = \frac{y}{\vartheta}, \quad y \in (0,\vartheta),
\end{aligned}$$

i.e. $Y_i = |X_i| \sim \mathcal{U}(0,\vartheta)$ for $i = 1, 2, \ldots, n$. According to example 3.5 (page 30), the statistic $T^*(X) = \max\{|X_1|, \ldots, |X_n|\}$ is also sufficient for ϑ. In the same manner as in the previous example, we can show that the statistic $T^*(X)$ is complete. □

Theorem 3.2. (Complete Sufficiency in the Exponential Family) Suppose that the distribution of the sample X belongs to the multivariate **full** exponential family with $f(x;\vartheta) = h(x)e^{\langle Q(\vartheta), T(x)\rangle - A(\vartheta)}$ for $\vartheta \in \Theta \subseteq \mathbb{R}^s$ and $x \in S$. Additionally, if the set $Q(\Theta) = \{Q(\vartheta) : \vartheta \in \Theta\} \subseteq \mathbb{R}^s$ contains a non-empty, open subset of $\mathbb{R}^s$, then the statistic $T(X)$ is sufficient for ϑ and complete.

Proof. According to the proposition on sufficiency in the exponential family, we already know that $T(X)$ is a sufficient statistic for ϑ. Without loss of generality, assume that the distribution of the sample is continuous. According to the change of variables formula, the PDF of $T(X)$ is given by $f_T(t) = r(t)e^{\langle Q(\vartheta),t\rangle - nA(\vartheta)}$ for some function r. Suppose that $\mathbb{E}_\vartheta[\varphi(T)] = 0 \ \forall \vartheta \in \Theta$. Then, we calculate that:

$$\mathbb{E}_\vartheta\left[\varphi(T)\right] = \int_{\mathbb{R}^s} \varphi(t)r(t)e^{\langle Q(\vartheta),t\rangle - nA(\vartheta)}dt = 0 \quad \Rightarrow \quad \int_{\mathbb{R}^s} \varphi(t)r(t)e^{-\langle -Q(\vartheta),t\rangle}dt = 0.$$

The last integral is the s-dimensional two-sided Laplace transform of the function $\varphi(t)r(t)$ evaluated at $-Q(\vartheta)$, which is injective on classes of almost surely equal functions. The multidimensional two-sided Laplace transform of the zero function is equal to 0. Since $r(t) \neq 0$ by the definition of the PDF of $T(X)$, it follows that $\varphi(t) = 0 \ \forall t \in \mathbb{R}^s$. Therefore, the statistic $T(X)$ is complete. □

Example 3.13. Let $X_1, \ldots, X_n \sim \mathcal{N}(\vartheta_1, \vartheta_2)$ be a random sample. According to example 2.4 (page 21), the distribution of the random variables $X_1, \ldots, X_n$ belongs to the univariate exponential family. According to proposition 2.3 (page 22), the joint distribution of the sample X belongs to the multivariate exponential family with the following PDF:

$$f(x;\vartheta) = (2\pi)^{-n/2} \exp\left\{\frac{\vartheta_1}{\vartheta_2}\sum_{i=1}^n x_i - \frac{1}{2\vartheta_2}\sum_{i=1}^n x_i^2 - \frac{n\vartheta_1^2}{2\vartheta_2} - \frac{n}{2}\log\vartheta_2\right\},$$

$$Q(\vartheta) = \left(\frac{\vartheta_1}{\vartheta_2}, -\frac{1}{2\vartheta_2}\right), \quad T(x) = \left(\sum_{i=1}^n x_i, \sum_{i=1}^n x_i^2\right).$$

The dimension of the function $T(x)$ is equal to the dimension of the parameter $(\vartheta_1, \vartheta_2)$, and the set $Q(\Theta) = \left\{\left(\frac{\vartheta_1}{\vartheta_2}, -\frac{1}{2\vartheta_2}\right) : (\vartheta_1, \vartheta_2) \in \mathbb{R} \times (0, \infty)\right\} = \mathbb{R} \times (-\infty, 0)$ contains a non-empty, open subset of $\mathbb{R}^2$. According to the complete sufficiency theorem in the exponential family, the statistic $T(X) = \left(\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2\right)$ is sufficient for $(\vartheta_1, \vartheta_2)$ and complete. □

Example 3.14. Let $X_1, \ldots, X_n \sim \mathcal{N}\left(\vartheta, \vartheta^2\right)$ be a random sample with $\vartheta \neq 0$. According to example 3.8 (page 31), the statistic $T(X) = \left(\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2\right)$ is sufficient for ϑ. Furthermore, we observe that $T(X) = \left(n\overline{X}, (n-1)S^2 + n\overline{X}^2\right) = \psi\left(\overline{X}, S^2\right)$. According to corollary 3.1 (page 29), we infer that the statistic $T^*(X) = \left(\overline{X}, S^2\right)$ is also sufficient for ϑ. According to proposition 3.1 (page 26), we know that $\mathbb{E}_\vartheta\left(S^2\right) = \vartheta^2$. Additionally, we calculate that:

$$\mathbb{E}_\vartheta\left(\overline{X}^2\right) = \frac{1}{n}\vartheta^2 + \vartheta^2 = \frac{n+1}{n}\vartheta^2 \quad \Rightarrow \quad \mathbb{E}_\vartheta\left(S^2 - \frac{n}{n+1}\overline{X}^2\right) = 0, \quad \forall \vartheta \neq 0,$$

i.e. there exist 2 unbiased estimators of the parametric function $g(\vartheta) = \vartheta^2$ which are both a function of $T^*(X)$. According to note 3.6, the statistic $T^*(X) = \left(\overline{X}, S^2\right)$ is **not** complete. We observe that the complete sufficiency theorem in the exponential family doesn't apply in this particular case, since this is a curved exponential family. □

Note 3.10. To sum up, we have 2 methods at our disposal for checking whether a statistic is complete: the definition (which requires knowledge of the distribution of the statistic) and the complete sufficiency theorem in the exponential family (easy to check whether its conditions hold). In table 3.2, we summarize the distributions of some notable complete sufficient statistics.

<table>
<tr><td>Bernoulli(p)</td><td rowspan="8">$\sum_{i=1}^n X_i$</td><td>Bin(n, p)</td></tr>
<tr><td>Bin(N, p) with known N</td><td>Bin(nN, p)</td></tr>
<tr><td>Geom(p)</td><td>NegBin(n, p)</td></tr>
<tr><td>NegBin(N, p) with known N</td><td>NegBin(nN, p)</td></tr>
<tr><td>Poisson(λ)</td><td>Poisson($n\lambda$)</td></tr>
<tr><td>Gamma(k, λ) with known k</td><td>Gamma(nk, λ)</td></tr>
<tr><td>$\mathcal{N}\left(\mu, \sigma^2\right)$ with known σ^2</td><td>$\mathcal{N}(n\mu, n\sigma^2)$</td></tr>
<tr><td>Exp(ϑ)</td><td rowspan="3">Gamma(n, ϑ)</td></tr>
<tr><td>Beta($\vartheta, 1$)</td><td>$-\sum_{i=1}^n \log X_i$</td></tr>
<tr><td>Beta($1, \vartheta$)</td><td>$-\sum_{i=1}^n \log(1 - X_i)$</td></tr>
</table>

TABLE 3.2: Distributions of Notable Complete Sufficient Statistics

Note 3.11. (χ^2 distribution with ν degrees of freedom)

i. If $X \sim \chi^2_\nu \equiv \text{Gamma}\left(\frac{\nu}{2}, \frac{1}{2}\right)$, then $\mathbb{E}(X) = \nu$ and $\text{Var}(X) = 2\nu$.

ii. If $X \sim \text{Gamma}(k, \vartheta)$, then $2\vartheta X \sim \text{Gamma}\left(k, \frac{1}{2}\right) \equiv \chi^2_{2k}$.

iii. If $X \sim \mathcal{N}(\mu, \sigma^2)$, then $\frac{X-\mu}{\sigma} \sim \mathcal{N}(0, 1)$ and $\left(\frac{X-\mu}{\sigma}\right)^2 \sim \chi^2_1$.

iv. If $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ are iid, then $\sum_{i=1}^n \left(\frac{X_i-\mu}{\sigma}\right)^2 \sim \chi^2_n$.

v. If $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ are iid, then $\frac{n-1}{\sigma^2} S^2 = \sum_{i=1}^n \left(\frac{X_i-\overline{X}}{\sigma}\right)^2 \sim \chi^2_{n-1}$.

Note 3.12. If $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ are iid random variables, then it follows that:

$$\mathbb{E}\left(\frac{n-1}{\sigma^2} S^2\right) = n - 1 \quad \Rightarrow \quad \mathbb{E}\left(S^2\right) = \sigma^2,$$

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

Theorem 3.3. (Basu) Suppose that the statistic $T(X)$ is sufficient for ϑ and complete. If $A(X)$ is an ancillary statistic, then $T(X)$ and $A(X)$ are independent.

Proof. According to the law of iterated expectations, it follows that:

$$\mathbb{E}\left[h(A(X))\right] = \mathbb{E}_\vartheta\left[\mathbb{E}\left(h(A(X)) \mid T(X)\right)\right].$$

Note that $\mathbb{E}\left[h(A(X))\right]$ does not depend on the value of ϑ because the statistic $A(X)$ is ancillary, while $\mathbb{E}\left[h(A(X)) \mid T(X)\right]$ does not depend on the value of ϑ because we are conditioning on the sufficient statistic $T(X)$ for ϑ. Now, consider the following function:

$$g(t) = \mathbb{E}\left[h(A(X)) \mid T(X) = t\right] - \mathbb{E}\left[h(A(X))\right].$$

Based on our previous argument, we know that $\mathbb{E}_\vartheta\left[g(T)\right] = 0 \ \forall \vartheta \in \Theta$. Since the statistic $T(X)$ is complete, it follows that:

$$g(T) \overset{\text{a.s.}}{=} 0 \quad \Rightarrow \quad \mathbb{E}\left[h(A(X)) \mid T(X)\right] \overset{\text{a.s.}}{=} \mathbb{E}\left[h(A(X))\right].$$

Since the function h was arbitrary, we deduce that $T(X)$ and $A(X)$ are mutually independent. □

Note 3.13. A well-known application of Basu's theorem lies in proving the independence of the statistics $\overline{X}$ and S^2 if the random variables $X_1, \ldots, X_n \sim \mathcal{N}\left(\mu, \sigma^2\right)$ are iid. In fact, the independence of the sample mean and the sample variance characterizes the normal distribution - no other distribution has this property. We fix σ^2. Then, we know that the statistic $\overline{X}$ is sufficient for μ and complete. We also know that $\frac{n-1}{\sigma^2}S^2 \sim \chi^2_{n-1}$, i.e. S^2 is an ancillary statistic. According to Basu's theorem, it follows that the statistics $\overline{X}$ and S^2 are independent. □

Definition 3.10. i. The statistic $R(X) = X_{(n)} - X_{(1)}$ is called the *sample range.*

ii. The statistic $M(X) = \frac{X_{(1)} + X_{(n)}}{2}$ is called the *sample midpoint.*

Example 3.15. Let $X_1, \ldots, X_n \sim \mathcal{N}\left(\mu, \sigma^2\right)$ be a random sample. We want to show that the statistics $\left(\overline{X}, S^2\right)$ and $A(X) = \frac{R(X)}{S(X)}$ are independent. We know that the statistic $T(X) = \left(\overline{X}, S^2\right)$ is sufficient for $\vartheta = \left(\mu, \sigma^2\right)$ and complete. Furthermore, we let $Z_i = \frac{X_i - \mu}{\sigma} \sim \mathcal{N}(0,1)$ for $i = 1, 2, \ldots, n$, so it follows that $\overline{X} = \sigma\overline{Z} + \mu$ and $X_{(i)} = \sigma Z_{(i)} + \mu$ for $i = 1, 2, \ldots, n$. We calculate that:

$$R(X) = \sigma Z_{(n)} + \mu - \left[\sigma Z_{(1)} + \mu\right] = \sigma\left[Z_{(n)} - Z_{(1)}\right],$$

$$S^2(X) = \frac{1}{n-1}\sum_{i=1}^{n}\left[\sigma Z_i + \mu - \left(\sigma\overline{Z} + \mu\right)\right]^2 = \frac{\sigma^2}{n-1}\sum_{i=1}^{n}\left(Z_i - \overline{Z}\right)^2,$$

$$A(X) = \frac{Z_{(n)} - Z_{(1)}}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(Z_i - \overline{Z}\right)^2}},$$

i.e. the statistic $A(X)$ is ancillary. According to Basu's theorem, it follows that the statistics $\left(\overline{X}, S^2\right)$ and $A(X)$ are independent. □

Example 3.16. Let $X_1, \ldots, X_n \sim \mathcal{U}(\vartheta, \vartheta+1)$ be a random sample with $\vartheta \in \mathbb{R}$. We want to show that the statistic $T(X) = (R, M)$ is sufficient for ϑ but not complete. According to example 3.6 (page 31), the statistic $T^*(X) = \left(X_{(1)}, X_{(n)}\right)$ is sufficient for ϑ. Additionally, we observe that $T^*(X) = \left(\frac{2M+R}{2}, \frac{2M-R}{2}\right) = \psi\left(R, M\right)$. According to corollary 3.1 (page 29), we infer that the statistic $T(X) = (R, M)$ is also sufficient for ϑ. Furthermore, we let $U_i = X_i - \vartheta \sim \mathcal{U}(0,1)$ for $i = 1, 2, \ldots, n$, so it follows that $X_{(i)} = U_{(i)} + \vartheta$ for $i = 1, 2, \ldots, n$. We calculate that:

$$R(X) = X_{(n)} - X_{(1)} = U_{(n)} + \vartheta - \left[U_{(1)} + \vartheta\right] = U_{(n)} - U_{(1)},$$

i.e. the statistic $R(X)$ is ancillary. According to Basu's theorem, it follows that the statistic $T(X) = (R, M)$ is **not** complete, since it's not independent of the ancillary statistic $R(X)$. □

Note 3.14. While the statistic $R(X)$ is ancillary, it's sufficient for ϑ in conjunction with the statistic $M(X)$. In other words, while it doesn't by itself contain any information for the value of ϑ, in conjunction with some other statistic it provides information about the precision with which we can estimate ϑ. For example, if we observe the value $m = 2$ for the statistic $M(X)$, then it follows that ϑ must lie on $[1, 2]$. If we also observe the value $r = 1$ for the statistic $R(X)$, then we calculate that $x_{(1)} = 1.5$ and $x_{(n)} = 2.5$, which implies that ϑ must be equal to $1.5 \in [1, 2]$. If we instead observe $r = 0.5$, then $x_{(1)} = 1.75$ and $x_{(n)} = 2.25$, so ϑ must lie on $[1.25, 1.75] \subset [1, 2]$. However, if we only observe some value r, we obviously cannot draw any conclusion about the value of ϑ.

3.6* Minimal Sufficiency

Definition 3.11. A statistic $T(X)$ is called *minimal sufficient* for the unknown parameter ϑ if it's sufficient for ϑ and for every other sufficient statistic $T^*(X)$ of ϑ there exists a function ψ such that $T(X) = \psi(T^*)$.

Interpretation: A minimal sufficient statistic for ϑ is a function of every other sufficient statistic of ϑ, so it concentrates all the information that a sample holds about ϑ as efficiently as possible. For example, the sample itself is always a sufficient statistic for any unknown parameter ϑ, but it's usually possible to summarize the information that the sample contains about ϑ much more efficiently than that.

Proposition 3.4. If the statistic $T(X)$ is minimal sufficient for ϑ and $T(X) = \psi\left(T^*\right)$ for some injective function ψ, then $T^*(X)$ is also minimal sufficient for ϑ.

Proof. First, note that the statistic $T^*(X)$ is sufficient for ϑ according to corollary 3.1 (page 29). Suppose that $V(X)$ is another sufficient statistic for ϑ. According to the definition of minimal sufficiency, there exists a function φ such that $T = \varphi(V)$. Since the function ψ is invertible, we get that $T^* = \left(\psi^{-1} \circ \varphi\right)(V)$. Since the sufficient statistic $T^*(X)$ for ϑ is a function of any other arbitrary sufficient statistic $V(X)$ for ϑ, we conclude that it is a minimal sufficient statistic for ϑ. $\square$

Theorem 3.4. If a minimal sufficient statistic for ϑ exists, then any statistic which is sufficient for ϑ and complete is minimal sufficient for ϑ.

Proof. Consider a statistic $T(X)$ which is sufficient for ϑ and complete, a statistic $M(X)$ which is minimal sufficient for ϑ and let $g\left(T(X)\right) = T(X) - E\left[T(X) \mid M(X)\right]$. Note that $E\left[T(X) \mid M(X)\right]$ does not depend on the value of ϑ since we are conditioning on the sufficient statistic $M(X)$ for ϑ. Suppose that $E_\vartheta\left[g(T)\right] = 0 \ \forall \vartheta \in \Theta$. Since the statistic $T(X)$ is complete, it follows that:

$$g(T) \overset{\text{a.s.}}{=} 0 \quad \Rightarrow \quad T(X) = \underbrace{E\left[T(X) \mid M(X)\right]}_{\psi(M)}.$$

Since $M(X)$ is a minimal sufficient statistic for ϑ, we conclude that $T(X)$ is also minimal sufficient statistic for ϑ according to the previous proposition. $\square$

Note 3.15. The converse is generally not true, i.e. a minimal sufficient statistic for ϑ isn't necessarily complete.

Theorem 3.5. Let X be a sample with joint PMF or PDF $f(x;\vartheta)$. For a given $x \in \mathbb{R}^n$, we let $\Theta_x = \{\vartheta \in \Theta : f(x;\vartheta)\} \subseteq \Theta$ be the subset of the parameter space under which it's possible to observe the sample x. Suppose that there exists some statistic $T(X)$ such that $\forall x, y \in S$ the following equivalency holds:

$$T(x) = T(y) \quad \Leftrightarrow \quad \Theta_x = \Theta_y \quad \text{and} \quad \frac{f(x;\vartheta)}{f(y;\vartheta)} = h(x,y),$$

where h is a non-negative function which doesn't depend on the value of $\vartheta \in \Theta_x$. Then, the statistic $T(X)$ is minimal sufficient for ϑ.

Proof. Let $A_t = \{x \in \mathbb{R}^n : T(x) = t\}$ and $x_t \in A_t$. For any sample point $x \in \mathbb{R}^n$, we notice that $T\left(x_{T(x)}\right) = T(x)$. By assumption, it follows that $\Theta_x = \Theta_{x_{T(x)}}$ and the ratio $\frac{f(x;\vartheta)}{f\left(x_{T(x)};\vartheta\right)} = h(x)$ does not depend on the value of $\vartheta \in \Theta_x$. Then, we observe that:

$$f(x;\vartheta) = \frac{f(x;\vartheta)}{f\left(x_{T(x)};\vartheta\right)} \underbrace{f\left(x_{T(x)};\vartheta\right)}_{g(T(x),\vartheta)} = h(x) g\left(T(x),\vartheta\right).$$

According to the Fisher - Neyman factorization criterion, the statistic $T(X)$ is sufficient for ϑ. Now, suppose that $T^*(X)$ is another sufficient statistic for ϑ, which implies that $f(x;\vartheta) = g^*\left(T^*(x),\vartheta\right) h^*(x)$ for some functions g^*, h^*. If $T^*(x) = T^*(y)$, then it follows that:

$$\frac{f(x;\vartheta)}{f(y;\vartheta)} = \frac{g^*\left(T^*(x),\vartheta\right) h^*(x)}{g^*\left(T^*(y),\vartheta\right) h^*(y)} = \frac{h^*(x)}{h^*(y)},$$

which does not depend on the value of ϑ. Furthermore, we notice that:

$$\begin{aligned}\Theta_x &= \{\vartheta \in \Theta : f(x;\vartheta) > 0\} = \{\vartheta \in \Theta : g^*\left(T^*(x),\vartheta\right) > 0\} \\ &= \{\vartheta \in \Theta : g^*\left(T^*(y),\vartheta\right) > 0\} = \{\vartheta \in \Theta : f(y;\vartheta) > 0\} = \Theta_y.\end{aligned}$$

By assumption, it follows that $T(x) = T(y)$. Since $T^*(x) = T^*(y)$ implies that $T(x) = T(y)$ for any arbitrary sample points $x, y \in \mathbb{R}^n$, we deduce that $T(X)$ is a function of $T^*(X)$. Since the sufficient statistic $T(X)$ for ϑ is a function of any other arbitrary sufficient statistic $T^*(X)$ for ϑ, we conclude that it is a minimal sufficient statistic for ϑ. □

Example 3.17. Let $X_1, \ldots, X_n \sim \mathcal{N}(\vartheta, \vartheta)$ be a random sample with $\vartheta > 0$. We observe that $\Theta_x = (0, \infty)$ doesn't depend on the observed sample $x \in \mathbb{R}^n$. We calculate that:

$$\begin{aligned}f(x;\vartheta) &= \left(\frac{1}{\sqrt{2\pi\vartheta}}\right)^n \exp\left\{-\frac{1}{2\vartheta}\sum_{i=1}^n (x_i - \vartheta)^2\right\} \\ &= (2\pi)^{-n/2}\vartheta^{-n/2} \exp\left\{\sum_{i=1}^n x_i - \frac{1}{2\vartheta}\sum_{i=1}^n x_i^2 - \frac{n\vartheta}{2}\right\} \\ &= (2\pi)^{-n/2}\vartheta^{-n/2} e^{-n\vartheta/2} e^{n\overline{x}} \exp\left\{-\frac{1}{2\vartheta}\sum_{i=1}^n x_i^2\right\},\end{aligned}$$

Let $T(X) = \sum_{i=1}^n X_i^2$. For $x, y \in \mathbb{R}^n$, we observe that:

$$T(x) = T(y) \quad \Leftrightarrow \quad \frac{f(x;\vartheta)}{f(y;\vartheta)} = e^{n(\overline{x} - \overline{y})},$$

which is constant with respect to ϑ. Therefore, $T(X) = \sum_{i=1}^n X_i^2$ is a minimal sufficient statistic for ϑ. □

Example 3.18. Let $X_1, \ldots, X_n \sim \mathcal{N}\left(\vartheta, \vartheta^2\right)$ be a random sample with $\vartheta > 0$. We observe that $\Theta_x = (0, \infty)$ doesn't depend on the observed sample $x \in \mathbb{R}^n$. According to example 3.8 (page 31), we know that:

$$f(x;\vartheta) = (2\pi e)^{-n/2}\vartheta^{-n} \exp\left\{\frac{1}{\vartheta}\sum_{i=1}^n x_i - \frac{1}{2\vartheta^2}\sum_{i=1}^n x_i^2\right\},$$

Let $T(X) = \left(\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2\right)$. For $x, y \in \mathbb{R}^n$, we observe that:

$$T(x) = T(y) \quad \Leftrightarrow \quad \frac{f(x;\vartheta)}{f(y;\vartheta)} = 1,$$

which is constant with respect to ϑ. Therefore, $T(X) = \left(\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2\right)$ is a minimal sufficient statistic for ϑ. □

Example 3.19. Let $X_1, \ldots, X_n \sim \text{Laplace}(\mu, \lambda)$ be a random sample with $\mu \in \mathbb{R}$, known $\lambda > 0$ and $f(x;\mu) = \frac{\lambda}{2} e^{-\lambda|x-\mu|}$ for $x \in \mathbb{R}$. We observe that $\Theta_x = \mathbb{R}$ doesn't depend on the observed sample $x \in \mathbb{R}^n$. We calculate that:

$$f(x;\mu) = \left(\frac{\lambda}{2}\right)^n \exp\left\{-\lambda \sum_{i=1}^n |x_i - \mu|\right\} = \left(\frac{\lambda}{2}\right)^n \exp\left\{-\lambda \sum_{i=1}^n \left|x_{(i)} - \mu\right|\right\}.$$

Let $T(X) = \left(X_{(1)}, X_{(2)}, \ldots, X_{(n)}\right)$. For $x, y \in \mathbb{R}^n$, we observe that:

$$T(x) = T(y) \quad \Leftrightarrow \quad \frac{f(x;\mu)}{f(y;\mu)} = 1,$$

which is constant with respect to μ. Therefore, $T(X) = \left(X_{(1)}, X_{(2)}, \ldots, X_{(n)}\right)$ is a minimal sufficient statistic for μ. □

Example 3.20. Let $X_1, \ldots, X_n \sim \text{Pareto}(k, \lambda)$ be a random sample with $k > 0$, known $\lambda > 0$ and $f(x;k) = \frac{\lambda k^\lambda}{x^{\lambda+1}}$ for $x \geqslant k$. We observe that $\Theta_x = \left(0, x_{(1)}\right]$ depends on the observed sample $x \in (0, \infty)^n$. We calculate that:

$$f(x;k) = \lambda^n k^{n\lambda} \prod_{i=1}^n x_i^{-\lambda-1} \mathbb{1}_{[k,\infty)}\left(x_{(1)}\right).$$

Let $T(X) = X_{(1)}$. For $x, y \in (0, \infty)^n$, we observe that:

$$T(x) = T(y) \quad \Leftrightarrow \quad \Theta_x = \Theta_y \quad \text{and} \quad \frac{f(x;k)}{f(y;k)} = 1,$$

which is constant with respect to $k \in \left(0, x_{(1)}\right]$. Therefore, $T(X) = X_{(1)}$ is a minimal sufficient statistic for k. □

Example 3.21. Let $X_1, \ldots, X_n \sim \text{Pareto}(k, \lambda)$ be a random sample with $k > 0$, $\lambda > 0$ and $f(x;k,\lambda) = \frac{\lambda k^\lambda}{x^{\lambda+1}}$ for $x \geqslant k$. We observe that $\Theta_x = \left(0, x_{(1)}\right] \times (0, \infty)$ depends on the observed sample $x \in (0, \infty)^n$. We calculate that:

$$f(x;\vartheta) = \lambda^n k^{n\lambda} \exp\left\{-(\lambda+1) \sum_{i=1}^n \log x_i\right\} \mathbb{1}_{[k,\infty)}\left(x_{(1)}\right).$$

Let $T(X) = \left(\sum_{i=1}^n \log X_i, X_{(1)}\right)$. For $x, y \in (0,\infty)^n$, we observe that:

$$T(x) = T(y) \quad \Leftrightarrow \quad \Theta_x = \Theta_y \quad \text{and} \quad \frac{f(x;\vartheta)}{f(y;\vartheta)} = 1,$$

which is constant with respect to $(k,\lambda) \in \left(0, x_{(1)}\right] \times (0,\infty)$. Therefore, the statistic $T(X) = \left(\sum_{i=1}^n \log X_i, X_{(1)}\right)$ is minimal sufficient for $\vartheta = (k,\lambda)$. □

Example 3.22. Let $X_1, \ldots, X_n \sim \mathcal{U}(\vartheta, \vartheta+1)$ be a random sample with $\vartheta \in \mathbb{R}$. We observe that $\Theta_x = \left[x_{(n)} - 1, x_{(1)}\right]$ depends on the observed sample $x \in \mathbb{R}^n$. According to example 3.6 (page 31), we know that:

$$f(x;\vartheta) = \mathbb{1}_{[\vartheta,\vartheta+1]}\left(x_{(1)}\right) \mathbb{1}_{[\vartheta,\vartheta+1]}\left(x_{(n)}\right).$$

Let $T(X) = \left(X_{(1)}, X_{(n)}\right)$. For $x, y \in \mathbb{R}^n$, we observe that:

$$T(x) = T(y) \quad \Leftrightarrow \quad \Theta_x = \Theta_y \quad \text{and} \quad \frac{f(x;\vartheta)}{f(y;\vartheta)} = 1,$$

which is constant with respect to $\vartheta \in \left[x_{(n)} - 1, x_{(1)}\right]$. Therefore, $T(X) = \left(X_{(1)}, X_{(n)}\right)$ is a minimal sufficient statistic for ϑ. □

3.7 Uniformly Minimum-Variance Unbiased Estimators

Definition 3.12. A statistic $\delta(X)$ is called a *uniformly minimum-variance unbiased estimator* (UMVUE) of the parametric function $g(\vartheta)$ if it's an unbiased estimator of $g(\vartheta)$ with finite variance and for every other unbiased estimator $V(X)$ of $g(\vartheta)$ it holds that $\mathrm{Var}_\vartheta\left[\delta(X)\right] \leqslant \mathrm{Var}_\vartheta\left[V(X)\right]\ \forall \vartheta \in \Theta$.

Theorem 3.6. Let $\mathcal{U}_0 = \left\{U(X) : \mathbb{E}_\vartheta\left[U(X)\right] = 0 \text{ and } \mathbb{E}_\vartheta\left[U^2(X)\right] < \infty\ \forall \vartheta \in \Theta\right\}$ be the class of unbiased estimators of 0 with finite variance and $\delta(X)$ be an unbiased estimator of $g(\vartheta)$ with finite variance. Then, the statistic $\delta(X)$ is a UMVUE of $g(\vartheta)$ if and only if $\mathrm{Cov}_\vartheta\left[\delta(X), U(X)\right] = 0\ \forall \vartheta \in \Theta$ and $\forall U(X) \in \mathcal{U}_0$.

Proof. First, suppose that $\delta(X)$ is a UMVUE of $g(\vartheta)$. Let $U(X) \in \mathcal{U}_0$ and consider the estimator $\delta^*(X) = \delta(X) + \lambda U(X)$ of $g(\vartheta)$ for $\lambda \in \mathbb{R}$. Then, we observe that:

$$E_\vartheta\left[\delta^*(X)\right] = E_\vartheta\left[\delta(X)\right] + \lambda E_\vartheta\left[U(X)\right] = g(\vartheta),$$

so $\delta^*(X)$ is an unbiased estimator of $g(\vartheta)$. Since $\delta(X)$ is a UMVUE of $g(\vartheta)$, it follows that:

$$\mathrm{Var}_\vartheta\left[\delta(X)\right] \leqslant \mathrm{Var}_\vartheta\left[\delta^*(X)\right] = \mathrm{Var}_\vartheta\left[\delta(X)\right] + \lambda^2 \mathrm{Var}_\vartheta\left[U(X)\right] + 2\lambda \mathrm{Cov}_\vartheta\left[\delta(X), U(X)\right],$$

which implies that $\lambda^2\mathrm{Var}_\vartheta\left[U(X)\right] + 2\lambda\mathrm{Cov}_\vartheta\left[\delta(X), U(X)\right] \geqslant 0\ \forall\lambda \in \mathbb{R}$. For this condition to be satisfied, the determinant of this quadratic equation with respect to λ must be non-positive, i.e. $\left[2\mathrm{Cov}_\vartheta\left(\delta(X), U(X)\right)\right]^2 \leqslant 0\ \forall\vartheta \in \Theta$, which implies that $\mathrm{Cov}_\vartheta[\delta(X), U(X)] = 0\ \forall\vartheta \in \Theta$.

Conversely, suppose that $\mathrm{Cov}_\vartheta\left[\delta(X), U(X)\right] = 0\ \forall\vartheta \in \Theta$ and $\forall U(X) \in \mathcal{U}_0$. Consider an unbiased estimator $\delta^*(X)$ of $g(\vartheta)$. Then, we note that $E_\vartheta\left[\delta(X) - \delta^*(X)\right] = 0$, which implies that $\delta(X) - \delta^*(X) \in \mathcal{U}_0$. By assumption and use of the covariance inequality, we deduce that:

$$\mathrm{Cov}_\vartheta\left[\delta(X), \delta(X) - \delta^*(X)\right] = 0 \quad \Rightarrow \quad \mathrm{Var}_\vartheta\left[\delta(X)\right] = \mathrm{Cov}_\vartheta\left[\delta(X), \delta^*(X)\right] \quad \Rightarrow$$

$$\left[\mathrm{Var}_\vartheta\left(\delta(X)\right)\right]^2 = \left[\mathrm{Cov}_\vartheta\left(\delta(X), \delta^*(X)\right)\right]^2 \leqslant \mathrm{Var}_\vartheta\left[\delta(X)\right]\mathrm{Var}_\vartheta\left[\delta^*(X)\right] \quad \Rightarrow$$

$$\mathrm{Var}_\vartheta\left[\delta(X)\right] \leqslant \mathrm{Var}_\vartheta\left[\delta^*(X)\right].$$

Since δ^* was an arbitrary unbiased estimator of $g(\vartheta)$, we conclude that the statistic $\delta(X)$ is a UMVUE of $g(\vartheta)$. □

Corollary 3.2. Let $U(X) \in \mathcal{U}_0$ and $V(X)$ be an unbiased estimator of the parametric function $g(\vartheta)$ with finite variance. If the constant $c = \frac{\mathrm{Cov}_\vartheta[V(X), U(X)]}{\mathrm{Var}_\vartheta[U(X)]} \neq 0$ doesn't depend on the value of ϑ, then $V^*(X) = V(X) - cU(X)$ is also an unbiased estimator of $g(\vartheta)$ and it holds that $\mathrm{Var}_\vartheta\left[V^*(X)\right] \leqslant \mathrm{Var}_\vartheta\left[V(X)\right]\ \forall\vartheta \in \Theta$.

Proof. First, we observe that:

$$\mathbb{E}_\vartheta[V^*(X)] = \mathbb{E}_\vartheta[V(X)] - c\mathbb{E}_\vartheta[U(X)] = g(\vartheta),$$

which implies that $V^*(X)$ is an unbiased estimator of $g(\vartheta)$. Then, we calculate that:

$$\begin{aligned}
\mathrm{Var}_\vartheta\left[V^*(X)\right] &= \mathrm{Var}_\vartheta\left[V(X) - cU(X)\right] \\
&= \mathrm{Var}_\vartheta\left[V(X)\right] + c^2\mathrm{Var}_\vartheta\left[U(X)\right] - 2c\mathrm{Cov}_\vartheta\left[V(X), U(X)\right] \\
&= \mathrm{Var}_\vartheta\left[V(X)\right] + \frac{\mathrm{Cov}^2_\vartheta\left[V(X), U(X)\right]}{\mathrm{Var}_\vartheta\left[U(X)\right]} - 2\frac{\mathrm{Cov}^2_\vartheta\left[V(X), U(X)\right]}{\mathrm{Var}_\vartheta\left[U(X)\right]} \\
&= \mathrm{Var}_\vartheta\left[V(X)\right] - \frac{\mathrm{Cov}^2_\vartheta\left[V(X), U(X)\right]}{\mathrm{Var}_\vartheta\left[U(X)\right]} \leqslant \mathrm{Var}_\vartheta\left[V(X)\right].
\end{aligned}$$

□

Note 3.16. This result is often used in Monte Carlo simulation as a variance reduction technique and referred to as the control variates method.

Corollary 3.3. If there exists a UMVUE for a parametric function $g(\vartheta)$, then it is unique.

Proof. Suppose $\delta_1(X)$ and $\delta_2(X)$ are 2 different UMVUEs of $g(\vartheta)$. Then, it follows that $\delta_1(X) - \delta_2(X) \in \mathcal{U}_0$. According to the previous theorem, we infer that:

$$\mathrm{Cov}_\vartheta\left[\delta_1(X), \delta_1(X) - \delta_2(X)\right] = \mathrm{Cov}_\vartheta\left[\delta_2(X), \delta_1(X) - \delta_2(X)\right] = 0 \quad \Rightarrow$$

$$\mathrm{Var}_\vartheta\left[\delta_1(X)\right] = \mathrm{Var}_\vartheta\left[\delta_2(X)\right] = \mathrm{Cov}_\vartheta\left[\delta_1(X), \delta_2(X)\right] \quad \Rightarrow$$

$$\mathrm{Corr}_\vartheta\left[\delta_1(X), \delta_2(X)\right] = \frac{\mathrm{Cov}_\vartheta\left[\delta_1(X), \delta_2(X)\right]}{\sqrt{\mathrm{Var}_\vartheta\left[\delta_1(X)\right] \mathrm{Var}_\vartheta\left[\delta_2(X)\right]}} = 1.$$

Hence, there exist constants $c_0 \in \mathbb{R}$ and $c_1 > 0$ such that $\delta_1(X) \overset{\text{a.s.}}{=} c_0 + c_1\delta_2(X)$. Since it holds that $\mathrm{Var}_\vartheta\left[\delta_1(X)\right] = \mathrm{Var}_\vartheta\left[\delta_2(X)\right]$, we infer that $c_1^2 = 1$, which implies that $c_1 = 1$. Furthermore, since it holds that $\mathbb{E}_\vartheta\left[\delta_1(X)\right] = \mathbb{E}_\vartheta\left[\delta_2(X)\right]$, it follows that $c_0 = 0$. Therefore, we conclude that $\delta_1(X) \overset{\text{a.s.}}{=} \delta_2(X)$. Since any 2 arbitrary UMVUEs of $g(\vartheta)$ must be almost surely equal to each other, we conclude that the UMVUE of $g(\vartheta)$ is unique, provided that it exists. □

Corollary 3.4. If the statistics $\delta_1(X), \ldots, \delta_d(X)$ are the UMVUEs of the parametric functions $g_1(\vartheta), \ldots, g_d(\vartheta)$ respectively, then the statistic $\delta(X) = \sum_{j=1}^d c_j\delta_j(X)$ is the UMVUE of the parametric function $g(\vartheta) = \sum_{j=1}^d c_j g_j(\vartheta)$.

Proof. First, we observe that:

$$\mathbb{E}_\vartheta\left[\delta(X)\right] = \sum_{j=1}^d c_j \mathbb{E}_\vartheta\left[\delta_j(X)\right] = \sum_{j=1}^d c_j g_j(\vartheta) = g(\vartheta),$$

so $\delta(X)$ is an unbiased estimator of $g(\vartheta)$. Now, let $U(X) \in \mathcal{U}_0$. According to the previous theorem and the linearity of the covariance operator, we get that:

$$\mathrm{Cov}_\vartheta\left[\delta(X), U(X)\right] = \sum_{j=1}^d c_j \mathrm{Cov}_\vartheta\left[\delta_j(X), U(X)\right] = \sum_{j=1}^d c_j \cdot 0 = 0.$$

Since the unbiased estimator $\delta(X)$ of $g(\vartheta) = \sum_{j=1}^d c_j g_j(\vartheta)$ is uncorrelated with any arbitrary unbiased estimator 0, we conclude that it is the UMVUE of the parametric function $g(\vartheta)$. □

Theorem 3.7. (Rao - Blackwell) Let $V(X)$ be an unbiased estimator of the parametric function $g(\vartheta)$ with finite variance and $T(X)$ be a sufficient statistic for ϑ. Then, the statistic $V^*(X) = \mathbb{E}\left[V(X) \mid T(X)\right]$ is also an unbiased estimator of $g(\vartheta)$ and it holds that $\mathrm{Var}_\vartheta\left[V^*(X)\right] \leqslant \mathrm{Var}_\vartheta\left[V(X)\right]$ $\forall \vartheta \in \Theta$.

Proof. According to the law of iterated expectations, we calculate that:

$$\mathbb{E}_\vartheta\left[V^*(X)\right] = \mathbb{E}_\vartheta\left[\mathbb{E}\left(V(X) \mid T(X)\right)\right] = \mathbb{E}_\vartheta\left[V(X)\right] = g(\vartheta).$$

According to Jensen's inequality and the law of iterated expectations, we calculate that:

$$\begin{aligned}\mathrm{Var}_\vartheta\left[V^*(X)\right] &= \mathbb{E}_\vartheta\left[\left(V^*(X)-\vartheta\right)^2\right] = \mathbb{E}_\vartheta\left[\left(\mathbb{E}\left(V(X)\mid T(X)\right)-\vartheta\right)^2\right] \\ &= \mathbb{E}_\vartheta\left[\left(\mathbb{E}\left(V(X)-\vartheta\mid T(X)\right)\right)^2\right] \leqslant \mathbb{E}_\vartheta\left[\mathbb{E}\left(\left(V(X)-\vartheta\right)^2\mid T(X)\right)\right] \\ &= \mathbb{E}_\vartheta\left[\left(V(X)-\vartheta\right)^2\right] = \mathrm{Var}_\vartheta\left[V(X)\right].\end{aligned}$$

□

Theorem 3.8. (Lehmann - Scheffé) Let $V(X)$ be an unbiased estimator of the parametric function $g(\vartheta)$ with finite variance. If the statistic $T(X)$ is sufficient for ϑ and complete, then the statistic $\delta(X)=\mathbb{E}\left[V(X)\mid T(X)\right]$ is the UMVUE of $g(\vartheta)$.

Proof. Suppose that there exists some unbiased estimator $\delta^*(X)$ of $g(\vartheta)$ such that $\mathrm{Var}_\vartheta\left[\delta^*(X)\right] < \mathrm{Var}_\vartheta\left[\delta(X)\right]$ for some $\vartheta\in\Theta$. Let $V^*(X)=\mathbb{E}\left[\delta^*(X)\mid T(X)\right]$. According to the Rao - Blackwell theorem, the statistic $V^*(X)$ is an unbiased estimator of $g(\vartheta)$ with $\mathrm{Var}_\vartheta\left[V^*(X)\right]\leqslant \mathrm{Var}_\vartheta\left[\delta^*(X)\right]$. Then, we observe that:

$$0=\mathbb{E}_\vartheta\left[V^*(X)-\delta(X)\right]=\mathbb{E}_\vartheta\left[\mathbb{E}\left(\delta^*(X)-V(X)\mid T(X)\right)\right].$$

Since the statistic $T(X)$ is complete, it follows that $\mathbb{E}\left[\delta^*(X)-V(X)\mid T(X)\right]\overset{\text{a.s.}}{=}0$, which implies that $V^*(X)\overset{\text{a.s.}}{=}\delta(X)$. Hence, we deduce that:

$$\mathrm{Var}_\vartheta\left[\delta(X)\right]=\mathrm{Var}_\vartheta\left[V^*(X)\right]\leqslant \mathrm{Var}_\vartheta\left[\delta^*(X)\right]<\mathrm{Var}_\vartheta\left[\delta(X)\right],$$

which is a contradiction. Since no other unbiased estimator of $g(\vartheta)$ can achieve smaller variance than that of $\delta(X)$ at any arbitrary parameter value $\vartheta\in\Theta$, we conclude that $\delta(X)=\mathbb{E}\left[V(X)\mid T(X)\right]$ is the UMVUE of $g(\vartheta)$. □

Corollary 3.5. Suppose that the statistic $T(X)$ is sufficient for ϑ and complete. If it holds that $\mathbb{E}_\vartheta\left[\psi(T)\right]=g(\vartheta)$ for some function ψ, then the statistic $\delta(X)=\psi(T)$ is the UMVUE of the parametric function $g(\vartheta)$.

Proof. According to the Lehmann - Scheffé theorem, we know that the statistic $\delta(X)=\mathbb{E}\left[\psi(T)\mid T(X)\right]$ is the UMVUE of $g(\vartheta)$. By the properties of conditional expectation, we also know that $\mathbb{E}\left[\psi(T)\mid T(X)\right]=\psi(T)$. Therefore, we conclude that $\delta(X)=\psi(T)$ is the UMVUE of $g(\vartheta)$. □

Note 3.17. To sum up, in order to calculate the UMVUE of a parametric function $g(\vartheta)$, we first need to find a statistic $T(X)$ which is sufficient for ϑ and complete. Then, we have 2 methods at our disposal:

i. If we can determine any unbiased estimator $V(X)$ of $g(\vartheta)$ and it's easy to calculate the conditional expectation $\psi(t) = \mathbb{E}_\vartheta [V(X) \mid T = t]$, then the statistic $\psi(T)$ is the UMVUE of $g(\vartheta)$ according to the Lehmann - Scheffé theorem. However, finding any unbiased estimator of $g(\vartheta)$ is not always a trivial task.

ii. If we can determine a function ψ of $T(X)$ such that $\mathbb{E}_\vartheta [\psi(T)] = c_0 + c_1 g(\vartheta)$, then $\delta(X) = \frac{\psi(T)-c_0}{c_1}$ is the UMVUE of $g(\vartheta)$ according to the previous corollary. However, finding such a function ψ is also not always a trivial task.

Example 3.23. Let $X_1, \ldots, X_n \sim \text{Bernoulli}(p)$ be a random sample. We want to find the UMVUEs of the parametric functions $g_1(p) = p^2$ and $g_2(p) = p(1-p)$. We know that the statistic $T(X) = \sum_{i=1}^n X_i \sim \text{Bin}(n, p)$ is sufficient for p and complete. We calculate that:

$$\begin{aligned} \mathbb{E}\left(T^2\right) &= \text{Var}(T) + [\mathbb{E}(T)]^2 = np(1-p) + (np)^2 = np - np^2 + n^2p^2 \\ &= \mathbb{E}(T) + n(n-1)p^2 \quad \Rightarrow \quad \mathbb{E}\left[\frac{T(T-1)}{n(n-1)}\right] = p^2 \end{aligned}$$

According to corollary 3.5, $\psi_1(T) = \frac{T(T-1)}{n(n-1)}$ is the UMVUE of $g_1(p)$. We observe that:

$$\mathbb{E}\left(T^2\right) = np(1-p) + n^2p^2 - n^2p + n^2p = np(1-p) - n^2p(1-p) + n\mathbb{E}(T) \quad \Rightarrow$$

$$\mathbb{E}\left(nT - T^2\right) = n(n-1)p(1-p) \quad \Rightarrow \quad \mathbb{E}\left[\frac{T(n-T)}{n(n-1)}\right] = p(1-p).$$

According to corollary 3.5, the statistic $\psi_2(T) = \frac{T(n-T)}{n(n-1)}$ is the UMVUE of $g_2(p)$. □

Example 3.24. Let $X_1, \ldots, X_n \sim \text{Exp}(\lambda)$ be a random sample. We want to find the UMVUEs of the parametric function $g(\lambda) = \frac{1}{\lambda^2}$ and λ. According to example 3.10 (page 35), we know that the statistic $T(X) = \sum_{i=1}^n X_i \sim \text{Gamma}(n, \lambda)$ is sufficient for λ and complete. We calculate that:

$$\mathbb{E}\left(T^2\right) = \text{Var}(T) + [\mathbb{E}(T)]^2 = \frac{n}{\lambda^2} + \frac{n^2}{\lambda^2} \quad \Rightarrow \quad \mathbb{E}\left[\frac{T^2}{n(n+1)}\right] = \frac{1}{\lambda^2}.$$

According to corollary 3.5, $\psi_1(T) = \frac{T^2}{n(n+1)}$ is the UMVUE of $g(\lambda)$. Next, we calculate that:

$$\begin{aligned} \mathbb{E}\left(\frac{1}{T}\right) &= \int_0^\infty \frac{1}{x} \frac{\lambda^n}{(n-1)!} x^{n-1} e^{-\lambda x} dx = \frac{\lambda^n}{(n-1)!} \int_0^\infty x^{n-2} e^{-\lambda x} dx \\ &= \frac{\lambda^n}{(n-1)!} \frac{(n-2)!}{\lambda^{n-1}} = \frac{\lambda}{n-1} \quad \Rightarrow \quad \mathbb{E}\left(\frac{n-1}{T}\right) = \lambda. \end{aligned}$$

According to corollary 3.5, the statistic $\psi_2(T) = \frac{n-1}{T}$ is the UMVUE of λ. □

Example 3.25. Let $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ be a random sample with known σ^2. We want to find the UMVUE of the parametric function $g(\mu) = e^{\mu t}$ for $t \in \mathbb{R}$. We know that the statistic $T(X) = \overline{X}$ is sufficient for μ and complete. We also know that:

$$M_{X_1}(t) = \mathbb{E}\left(e^{tX_1}\right) = \exp\left\{\mu t + \frac{1}{2}\sigma^2 t^2\right\},$$

$$M_T(t) = \mathbb{E}\left(e^{t\overline{X}}\right) = \prod_{i=1}^{n} M_{X_i}(t/n) = [M_{X_1}(t/n)]^n = \exp\left\{\mu t + \frac{1}{2n}\sigma^2 t^2\right\},$$

$$\mathbb{E}\left(\exp\left\{t\overline{X} - \frac{1}{2n}\sigma^2 t^2\right\}\right) = e^{\mu t}.$$

According to corollary 3.5, $\psi(\overline{X}) = \exp\left\{t\overline{X} - \frac{1}{2n}\sigma^2 t^2\right\}$ is the UMVUE of $g(\mu)$. □

Example 3.26. Let $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ be a random sample. We want to find the UMVUEs of the parametric functions $g_1(\mu, \sigma^2) = \sigma^2$ and $g_2(\mu, \sigma^2) = \mu^2$. According to example 3.13 (page 37), we know that the statistic $T(X) = \left(\overline{X}, S^2\right)$ is sufficient for $\vartheta = (\mu, \sigma^2)$ and complete. According to proposition 3.1 (page 26), we also know that $\mathbb{E}(S^2) = \sigma^2$. Hence, the statistic $\psi_1\left(\overline{X}, S^2\right) = S^2$ is the UMVUE of $g_1(\vartheta)$ according to corollary 3.5. Next, we calculate that:

$$\mathbb{E}\left(\overline{X}^2\right) = \mathrm{Var}(\overline{X}) + \left[\mathbb{E}(\overline{X})\right]^2 = \frac{1}{n}\sigma^2 + \mu^2 \quad \Rightarrow \quad \mathbb{E}\left(\overline{X}^2 - \frac{1}{n}S^2\right) = \mu^2.$$

According to corollary 3.5, $\psi_2\left(\overline{X}, S^2\right) = \overline{X}^2 - \frac{1}{n}S^2$ is the UMVUE of $g_2(\vartheta)$. □

Example 3.27. Let $X_1, \ldots, X_n \sim \mathcal{U}(0, \vartheta)$ be a random sample. If the function $g : (0, \infty) \to \mathbb{R}$ is differentiable, we want to find the UMVUE of the parametric function $g(\vartheta)$. We know that $T(X) = X_{(n)}$ is sufficient for ϑ and complete. For $t \in (0, \vartheta)$, we calculate that:

$$f_{X_{(n)}}(t) = \frac{n}{\vartheta^n} t^{n-1}.$$

Suppose that $\mathbb{E}_\vartheta[\psi(T)] = g(\vartheta)\ \forall \vartheta > 0$. Then, we calculate that:

$$\int_0^\vartheta f_{X_{(n)}}(t)\psi(t)dt = \frac{n}{\vartheta^n}\int_0^\vartheta t^{n-1}\psi(t)dt = g(\vartheta) \quad \Rightarrow$$

$$n\int_0^\vartheta t^{n-1}\psi(t)dt = \vartheta^n g(\vartheta) \quad \Rightarrow \quad n\vartheta^{n-1}\psi(\vartheta) = n\vartheta^{n-1}g(\vartheta) + \vartheta^n g'(\vartheta) \quad \Rightarrow$$

$$\psi(\vartheta) = g(\vartheta) + \frac{\vartheta}{n}g'(\vartheta), \quad \forall \vartheta \in (0, \infty) \supseteq (0, \vartheta).$$

According to corollary 3.5, $\psi(T) = g(T) + \frac{T}{n}g'(T)$ is the UMVUE of $g(\vartheta)$. □

Example 3.28. Let $X_1, \ldots, X_n \sim \text{Poisson}(\lambda)$ be a random sample. We want to find the UMVUEs of $g_1(\lambda) = \lambda^k e^{-\lambda}$, $g_2(\lambda) = e^{-k\lambda}$ and $g_3(\lambda) = \lambda^k$ for $k \leqslant n$. We know that $T(X) = \sum_{i=1}^n X_i \sim \text{Poisson}(\lambda)$ is sufficient for λ and complete. Then, we observe that:

$$\mathbb{E}\left[\mathbb{1}_{\{k\}}(X_1)\right] = \mathbb{P}(X_1 = k) = e^{-\lambda}\frac{\lambda^k}{k!} \quad \Rightarrow \quad \mathbb{E}\left[k!\mathbb{1}_{\{k\}}(X_1)\right] = e^{-\lambda}\lambda^k.$$

Thus, $V_1(X) = k!\mathbb{1}_{\{k\}}(X_1)$ is an unbiased estimator of $g_1(\lambda)$. For $t = k, k+1, \ldots$, we calculate that:

$$\begin{aligned}
\mathbb{E}\left(V_1 \mid T = t\right) &= \mathbb{E}\left(k!\mathbb{1}_{\{k\}}(X_1) \mid T = t\right) = k!\mathbb{P}\left(X_1 = k \mid T = t\right) \\
&= \frac{k!\mathbb{P}\left(X_1 = k, \sum_{i=1}^n X_i = t\right)}{\mathbb{P}\left(\sum_{i=1}^n X_i = t\right)} = \frac{k!\mathbb{P}\left(X_1 = k, \sum_{i=2}^n X_i = t-k\right)}{\mathbb{P}\left(\sum_{i=1}^n X_i = t\right)} \\
&= \frac{k!\mathbb{P}(X_1 = k)\mathbb{P}\left(\sum_{i=2}^n X_i = t-k\right)}{\mathbb{P}\left(\sum_{i=1}^n X_i = t\right)} \\
&= \frac{k!e^{-\lambda}\lambda^k/k! \cdot e^{-(n-1)\lambda}\left[(n-1)\lambda\right]^{t-k}/(t-k)!}{e^{-n\lambda}(n\lambda)^t/t!} \\
&= \frac{t!}{(t-k)!}\left(\frac{1}{n}\right)^k\left(1 - \frac{1}{n}\right)^{t-k}.
\end{aligned}$$

Therefore, the statistic $\psi_1(T) = \frac{T!}{(T-k)!}\left(\frac{1}{n}\right)^k\left(1-\frac{1}{n}\right)^{T-k}\mathbb{1}_{\{k,k+1,\ldots\}}(T)$ is the UMVUE of $g_1(\lambda)$ according to the Lehmann - Scheffé theorem. Next, we observe that:

$$\mathbb{E}\left[\mathbb{1}_{\{0\}}(X_1)\cdots\mathbb{1}_{\{0\}}(X_k)\right] = \mathbb{P}(X_1 = 0, \ldots, X_k = 0) \overset{\text{iid}}{=} \left[\mathbb{P}(X_1 = 0)\right]^k = e^{-k\lambda}.$$

Hence, the statistic $V_2(X) = \mathbb{1}_{\{0\}}(X_1)\cdots\mathbb{1}_{\{0\}}(X_k)$ is an unbiased estimator of $g_2(\lambda)$. For $k \leqslant n$, we calculate that:

$$\begin{aligned}
\mathbb{E}\left(V_2 \mid T = t\right) &= \mathbb{P}\left(X_1 = 0, \ldots, X_k = 0 \mid T = t\right) \\
&= \frac{\mathbb{P}\left(X_1 = 0, \ldots, X_k = 0, \sum_{i=1}^n X_i = t\right)}{\mathbb{P}\left(\sum_{i=1}^n X_i = t\right)} \\
&= \frac{\mathbb{P}(X_1 = 0)\cdots\mathbb{P}(X_k = 0)\mathbb{P}\left(\sum_{i=k+1}^n X_i = t\right)}{\mathbb{P}\left(\sum_{i=1}^n X_i = t\right)} \\
&= \frac{e^{-k\lambda}e^{-(n-k)\lambda}\left[(n-k)\lambda\right]^t/t!}{e^{-n\lambda}(n\lambda)^t/t!} = \left(1 - \frac{k}{n}\right)^t.
\end{aligned}$$

According to the Lehmann - Scheffé theorem, the statistic $\psi_2(T) = \left(1 - \frac{k}{n}\right)^T$ is the UMVUE of $g_2(\lambda)$. Now, suppose that $\mathbb{E}_\lambda\left[\psi_3(T)\right] = g_3(\lambda)\ \forall \lambda > 0$. Then, we calculate that:

$$\sum_{t=0}^{\infty}\psi_3(t)e^{-n\lambda}\frac{(n\lambda)^t}{t!} = \lambda^k \quad \Rightarrow \quad \sum_{t=0}^{\infty}\frac{n^t\psi_3(t)}{t!}\lambda^t = \lambda^k e^{n\lambda} \quad \Rightarrow$$

$$\sum_{t=0}^{\infty}\frac{n^t\psi_3(t)}{t!}\lambda^t = \lambda^k\sum_{t=0}^{\infty}\frac{(n\lambda)^t}{t!} \quad \Rightarrow \quad \sum_{t=0}^{\infty}\frac{n^t\psi_3(t)}{t!}\lambda^t = \sum_{t=0}^{\infty}\frac{n^t}{t!}\lambda^{t+k} \quad \Rightarrow$$

$$\sum_{t=0}^{\infty}\frac{n^t\psi_3(t)}{t!}\lambda^t = \sum_{t=k}^{\infty}\frac{n^{t-k}}{(t-k)!}\lambda^t \quad \Rightarrow$$

$$\frac{n^t \psi_3(t)}{t!} = \begin{cases} 0, & t = 0, 1, \dots, k-1 \\ \frac{n^{t-k}}{(t-k)!}, & t = k, k+1, \dots \end{cases} \quad \Rightarrow \quad \psi_3(t) = \begin{cases} 0, & t = 0, 1, \dots, k-1 \\ \binom{t}{k} \frac{k!}{n^k}, & t = k, k+1, \dots \end{cases}.$$

According to corollary 3.5, $\psi_3(T) = \binom{T}{k} \frac{k!}{n^k} \mathbb{1}_{\{k,k+1,\dots\}}(T)$ is the UMVUE of $g_3(\lambda)$. □

Note 3.18. If $X_1, \dots, X_n \sim \text{Poisson}(\lambda)$ are iid, then $(X_1 \mid \sum_{i=1}^n X_i = t) \sim \text{Bin}\left(t, \frac{1}{n}\right)$ independently of the value λ.

3.8 Cramér - Rao Inequality

Definition 3.13. i. The function $\mathcal{S}_X(\vartheta) = \frac{\partial}{\partial \vartheta} \log f(X; \vartheta)$ is called the *score function* of the sample X for the parameter ϑ.

ii. The parametric function $\mathcal{I}_X(\vartheta) = \mathbb{E}_\vartheta \left[\mathcal{S}_X^2(\vartheta)\right]$ is called the *Fisher information* of the sample X for the parameter ϑ.

Proposition 3.5. i. If $X_1, \dots, X_n$ are independent, then $\mathcal{S}_X(\vartheta) = \sum_{i=1}^n \mathcal{S}_{X_i}(\vartheta)$.

ii. If $g(\eta) = \vartheta$ is a continuously differentiable function, $\mathcal{I}_X(\eta) = \mathcal{I}_X(g(\eta)) \left[g'(\eta)\right]^2$.

Proof. i. We observe that:

$$\begin{aligned} \mathcal{S}_X(\vartheta) &= \frac{\partial}{\partial \vartheta} \log \prod_{i=1}^n f(X_i; \vartheta) = \frac{\partial}{\partial \vartheta} \sum_{i=1}^n \log f(X_i; \vartheta) \\ &= \sum_{i=1}^n \frac{\partial}{\partial \vartheta} \log f(X_i; \vartheta) = \sum_{i=1}^n \mathcal{S}_{X_i}(\vartheta). \end{aligned}$$

ii. According to the chain rule, we calculate that:

$$\begin{aligned} \mathcal{I}_X(\eta) = \mathbb{E}_\eta \left[\mathcal{S}_X^2(\eta)\right] &= \mathbb{E}_\eta \left[\left(\frac{\partial}{\partial \eta} \log f(X; \eta)\right)^2\right] = \mathbb{E}_\vartheta \left[\left(\frac{\partial}{\partial \vartheta} \log f(X; \vartheta) \frac{\partial \vartheta}{\partial \eta}\right)^2\right] \\ &= \mathbb{E}_\vartheta \left[\left(\frac{\partial}{\partial \vartheta} \log f(X; \vartheta)\right)^2\right] \left(\frac{\partial \vartheta}{\partial \eta}\right)^2 = \mathcal{I}_X(g(\eta)) \left[g'(\eta)\right]^2. \end{aligned}$$

□

Regularity Conditions: Without loss of generality, assume that the distribution of the sample is continuous with joint PDF $f(x; \vartheta)$ for $\vartheta \in \Theta \subseteq \mathbb{R}$ and $x \in S$. We define the following regularity conditions:

I. The parameter space Θ is an open subset of $\mathbb{R}$.

II. The support $S = \{x \in \mathbb{R}^n : f(x; \vartheta) > 0\}$ doesn't depend on the value of ϑ.

III. $\frac{\partial}{\partial \vartheta} f(x; \vartheta) < \infty \ \forall x \in S$ and $\forall \vartheta \in \Theta$.

IV. $\int_S \frac{\partial}{\partial\vartheta} f(x;\vartheta)dx = \frac{\partial}{\partial\vartheta}\int_S f(x;\vartheta)dx = 0\ \forall\vartheta\in\Theta$.

V. $\mathcal{I}_X(\vartheta)\in(0,\infty)\ \forall\vartheta\in\Theta$.

Proposition 3.6. Suppose that the following regularity conditions are satisfied:

VI. $\frac{\partial^2}{\partial\vartheta^2} f(x;\vartheta) < \infty\ \forall x\in S$ and $\forall\vartheta\in\Theta$.

VII. $\int_S \frac{\partial^2}{\partial\vartheta^2} f(x;\vartheta)dx = \frac{\partial^2}{\partial\vartheta^2}\int_S f(x;\vartheta)dx = 0\ \forall\vartheta\in\Theta$.

Then, it follows that:

$$\mathcal{I}_X(\vartheta) = -\mathbb{E}_\vartheta\left[\frac{\partial}{\partial\vartheta}\mathcal{S}_X(\vartheta)\right] = -\mathbb{E}_\vartheta\left[\frac{\partial^2}{\partial\vartheta^2}\log f(X;\vartheta)\right].$$

Proof. According to integration by parts, we calculate that:

$$\begin{aligned}\mathbb{E}_\vartheta\left[\frac{\partial^2}{\partial\vartheta^2}\log f(X;\vartheta)\right] &= \int_S f(x;\vartheta)\frac{\partial^2}{\partial\vartheta^2}\log f(x;\vartheta)dx = \int_S f(x;\vartheta)\frac{\partial}{\partial\vartheta}\mathcal{S}_x(\vartheta)dx \\ &= \int_S\left[\frac{\partial}{\partial\vartheta}\left[f(x;\vartheta)\mathcal{S}_x(\vartheta)\right] - \mathcal{S}_x(\vartheta)\frac{\partial}{\partial\vartheta}f(x;\vartheta)\right]dx,\end{aligned}$$

where $\mathcal{S}_x(\vartheta) = \frac{\partial}{\partial\vartheta}\log f(x;\vartheta) = \frac{1}{f(x;\vartheta)}\frac{\partial}{\partial\vartheta}f(x;\vartheta)$ is the score function. According to regularity condition VII, we calculate that:

$$\begin{aligned}\int_S \frac{\partial}{\partial\vartheta}\left[f(x;\vartheta)\mathcal{S}_x(\vartheta)\right]dx &= \int_S\frac{\partial}{\partial\vartheta}\left[f(x;\vartheta)\frac{1}{f(x;\vartheta)}\frac{\partial}{\partial\vartheta}f(x;\vartheta)\right]dx \\ &= \int_S\frac{\partial^2}{\partial\vartheta^2}f(x;\vartheta)dx = \frac{\partial^2}{\partial\vartheta^2}\int_S f(x;\vartheta)dx = 0.\end{aligned}$$

Finally, we conclude that:

$$\begin{aligned}-\mathbb{E}_\vartheta\left[\frac{\partial^2}{\partial\vartheta^2}\log f(X;\vartheta)\right] &= \int_S\mathcal{S}_x(\vartheta)\frac{\partial}{\partial\vartheta}f(x;\vartheta)dx \\ &= \int_S\mathcal{S}_x(\vartheta)f(x;\vartheta)\frac{1}{f(x;\vartheta)}\frac{\partial}{\partial\vartheta}f(x;\vartheta)dx \\ &= \int_S\mathcal{S}_x(\vartheta)f(x;\vartheta)\frac{\partial}{\partial\vartheta}\log f(x;\vartheta)dx \\ &= \int_S f(x;\vartheta)\left[\frac{\partial}{\partial\vartheta}\log f(x;\vartheta)\right]^2dx \\ &= \mathbb{E}_\vartheta\left[\left(\frac{\partial}{\partial\vartheta}\log f(X;\vartheta)\right)^2\right] = \mathbb{E}_\vartheta\left[\mathcal{S}_X^2(\vartheta)\right] = \mathcal{I}_X(\vartheta).\end{aligned}$$

□

Proposition 3.7. Let X be a sample which satisfies the regularity conditions I-V.

i. It holds that $\mathbb{E}\left[\mathcal{S}_X(\vartheta)\right] = 0$ and $\mathrm{Var}_\vartheta\left[\mathcal{S}_X(\vartheta)\right] = \mathbb{E}_\vartheta\left[\mathcal{S}_X^2(\vartheta)\right] = \mathcal{I}_X(\vartheta)\ \forall\vartheta\in\Theta$.

ii. If $X_1,\ldots,X_n$ are independent, then $\mathcal{I}_X(\vartheta) = \sum_{i=1}^n \mathcal{I}_{X_i}(\vartheta)$.

iii. If $X_1, \ldots, X_n$ are iid, then $\mathcal{I}_X(\vartheta) = n\mathcal{I}_{X_1}(\vartheta)$.

Proof. i. According to regularity condition IV, we calculate that:

$$\begin{aligned}\mathbb{E}\left[\mathcal{S}_X(\vartheta)\right] &= \int_S f(x;\vartheta)\mathcal{S}_x(\vartheta)dx = \int_S f(x;\vartheta)\frac{\partial}{\partial\vartheta}\log f(x;\vartheta)dx \\ &= \int_S f(x;\vartheta)\frac{1}{f(x;\vartheta)}\frac{\partial}{\partial\vartheta}f(x;\vartheta)dx = \int_S \frac{\partial}{\partial\vartheta}f(x;\vartheta)dx = \frac{\partial}{\partial\vartheta}\int_S f(x;\vartheta)dx = 0.\end{aligned}$$

Hence, we infer that:

$$\mathrm{Var}_\vartheta\left[\mathcal{S}_X(\vartheta)\right] = \mathbb{E}_\vartheta\left[\mathcal{S}_X^2(\vartheta)\right] - \left[\mathbb{E}_\vartheta\left(\mathcal{S}_X(\vartheta)\right)\right]^2 = \mathbb{E}_\vartheta\left[\mathcal{S}_X^2(\vartheta)\right] = \mathcal{I}_X(\vartheta).$$

ii. Since the random variables $X_1, X_2, \ldots, X_n$ are independent, it follows that:

$$\mathcal{I}_X(\vartheta) = \mathrm{Var}_\vartheta\left[\mathcal{S}_X(\vartheta)\right] = \mathrm{Var}_\vartheta\left[\sum_{i=1}^n \mathcal{S}_{X_i}(\vartheta)\right] = \sum_{i=1}^n \mathrm{Var}_\vartheta\left[\mathcal{S}_{X_i}(\vartheta)\right] = \sum_{i=1}^n \mathcal{I}_{X_i}(\vartheta).$$

iii. Since the random variables $X_1, X_2, \ldots, X_n$ are iid, it follows that:

$$\mathcal{I}_{X_1}(\vartheta) = \mathcal{I}_{X_2}(\vartheta) = \cdots = \mathcal{I}_{X_n}(\vartheta) \quad \Rightarrow \quad \mathcal{I}_X(\vartheta) = \sum_{i=1}^n \mathcal{I}_{X_i}(\vartheta) = n\mathcal{I}_{X_1}(\vartheta).$$

□

Theorem 3.9. (Cramér - Rao Inequality) Let X be a sample with joint PDF $f(x;\vartheta)$ for $\vartheta \in \Theta \subseteq \mathbb{R}$ and $x \in S$ which satisfies the regularity conditions I-V. Suppose that the statistic $T(X)$ is an estimator of $g(\vartheta)$ with finite variance which satisfies the following regularity condition:

VIII. $\int_S T(x)\frac{\partial}{\partial\vartheta}f(x;\vartheta)dx = \frac{\partial}{\partial\vartheta}\int_S T(x)f(x;\vartheta)dx = \frac{\partial}{\partial\vartheta}\mathbb{E}_\vartheta\left[T(X)\right]\ \forall\vartheta \in \Theta,$

where $\mathbb{E}_\vartheta\left[T(X)\right] = g(\vartheta) + \mathrm{bias}_{g(\vartheta)}\left[T(X)\right]$. Then, it follows that:

$$\mathrm{Var}_\vartheta\left[T(X)\right] \geqslant \frac{1}{\mathcal{I}_X(\vartheta)}\left[\frac{\partial}{\partial\vartheta}\mathbb{E}_\vartheta\left[T(X)\right]\right]^2, \quad \forall\vartheta \in \Theta.$$

Proof. By making use of the regularity conditions VII and IV, we calculate that:

$$\begin{aligned}\frac{\partial}{\partial\vartheta}\mathbb{E}_\vartheta\left[T(X)\right] &= \frac{\partial}{\partial\vartheta}\int_S T(x)f(x;\vartheta)dx = \int_S T(x)\frac{\partial}{\partial\vartheta}f(x;\vartheta)dx \\ &= \int_S T(x)f(x;\vartheta)\frac{1}{f(x;\vartheta)}\frac{\partial}{\partial\vartheta}f(x;\vartheta)dx \\ &= \int_S f(x;\vartheta)T(x)\frac{\partial}{\partial\vartheta}\log f(x;\vartheta)dx = \mathbb{E}_\vartheta\left[T(X)\mathcal{S}_X(\vartheta)\right] \\ &= \mathrm{Cov}_\vartheta\left[T(X), \mathcal{S}_X(\vartheta)\right] + \mathbb{E}_\vartheta\left[T(X)\right]\mathbb{E}_\vartheta\left[\mathcal{S}_X(\vartheta)\right] = \mathrm{Cov}_\vartheta\left[T(X), \mathcal{S}_X(\vartheta)\right],\end{aligned}$$

where $\mathbb{E}_\vartheta\left[\mathcal{S}_X(\vartheta)\right] = 0$ according to the previous proposition. By making use of the covariance inequality, we deduce that:

$$\begin{aligned}\left[\frac{\partial}{\partial\vartheta}\mathbb{E}_\vartheta\left[T(X)\right]\right]^2 &= \left[\mathrm{Cov}_\vartheta\left(T(X), \mathcal{S}_X(\vartheta)\right)\right]^2 \\ &\leqslant \mathrm{Var}_\vartheta\left[T(X)\right]\mathrm{Var}_\vartheta\left[\mathcal{S}_X(\vartheta)\right] = \mathrm{Var}_\vartheta\left[T(X)\right]\mathcal{I}_X(\vartheta),\end{aligned}$$

where $\mathrm{Var}_\vartheta\left[\mathcal{S}_X(\vartheta)\right] = \mathcal{I}_X(\vartheta)$ according to the previous proposition. Therefore, we conclude that:

$$\mathrm{Var}_\vartheta\left[T(X)\right] \geqslant \frac{1}{\mathcal{I}_X(\vartheta)}\left[\frac{\partial}{\partial\vartheta}\mathbb{E}_\vartheta\left[T(X)\right]\right]^2.$$

□

Corollary 3.6. If the statistic $T(X)$ is an unbiased estimator of the parametric function $g(\vartheta)$, then it follows that:

$$\mathrm{Var}_\vartheta\left[T(X)\right] \geqslant \frac{\left[g'(\vartheta)\right]^2}{\mathcal{I}_X(\vartheta)}, \quad \forall\vartheta\in\Theta.$$

Proof. If $T(X)$ is an unbiased estimator of $g(\vartheta)$, then $\frac{\partial}{\partial\vartheta}\mathbb{E}_\vartheta\left[T(X)\right] = g'(\vartheta)$, so the desired result follows immediately from the Cramér - Rao inequality. □

Definition 3.14. i. An unbiased estimator $T(X)$ of the parametric function $g(\vartheta)$ which achieves the Cramér - Rao lower bound, i.e. for which it holds that:

$$\mathrm{Var}_\vartheta\left[T(X)\right] = \frac{\left[g'(\vartheta)\right]^2}{\mathcal{I}_X(\vartheta)}, \quad \forall\vartheta\in\Theta,$$

is called an *efficient* estimator of $g(\vartheta)$.

ii. Let $T(X)$ be an unbiased estimator of the parametric function $g(\vartheta)$. The following ratio:

$$e_{g(\vartheta)}\left[T(X)\right] = \frac{\left[g'(\vartheta)\right]^2/\mathcal{I}_X(\vartheta)}{\mathrm{Var}_\vartheta\left[T(X)\right]} \in [0,1].$$

is called the *efficiency* of $T(X)$ with respect to $g(\vartheta)$.

Note 3.19. We observe that the statistic $T(X)$ is an efficient estimator of $g(\vartheta)$ if and only if $e_{g(\vartheta)}\left[T(X)\right] = 1$ $\forall\vartheta\in\Theta$. If $T(X)$ is an efficient estimator of $g(\vartheta)$, then it's also the unique UMVUE of $g(\vartheta)$. The converse is generally not true. If $T(X)$ is the UMVUE of $g(\vartheta)$, it's not necessarily an efficient estimator of $g(\vartheta)$, i.e. it doesn't necessarily achieve the Cramér - Rao lower bound. In this case, it follows that there doesn't exist any efficient estimator of $g(\vartheta)$.

Proposition 3.8. A statistic $T(X)$ is an efficient estimator of $g(\vartheta)$ if and only if there exists a function $k(\vartheta) \neq 0$ such that $\mathcal{S}_X(\vartheta) = k(\vartheta)\left[T(X) - g(\vartheta)\right]$ $\forall\vartheta\in\Theta$. Then, it holds that $k(\vartheta) = \frac{\mathcal{I}_X(\vartheta)}{g'(\vartheta)}$.

Proof. We know that $T(X)$ is an efficient estimator of $g(\vartheta)$ if and only if it's an unbiased estimator of $g(\vartheta)$ and it achieves the Cramér - Rao lower bound. Furthermore, we know that $T(X)$ achieves the Cramér - Rao lower bound if and only if the covariance inequality invoked in the proof of the Cramér - Rao inequality holds as an equality $\forall \vartheta \in \Theta$. We also know that the covariance inequality holds as an equality if and only if there exist functions $c_0(\vartheta)$ and $c_1(\vartheta) \neq 0$ such that $T(X) = c_0(\vartheta) + c_1(\vartheta)\mathcal{S}_X(\vartheta)$ $\forall \vartheta \in \Theta$. Then, we calculate that:

$$g(\vartheta) = \mathbb{E}\left[T(X)\right] = c_0(\vartheta) + c_1(\vartheta)\mathbb{E}_\vartheta\left[\mathcal{S}_X(\vartheta)\right] = c_0(\vartheta),$$

$$0 = c_0'(\vartheta) + c_1'(\vartheta)\mathcal{S}_X(\vartheta) + c_1(\vartheta)\frac{\partial}{\partial \vartheta}\mathcal{S}_X(\vartheta) \quad \Rightarrow$$

$$g'(\vartheta) + c_1'(\vartheta)\mathbb{E}_\vartheta\left[\mathcal{S}_X(\vartheta)\right] + c_1(\vartheta)\mathbb{E}_\vartheta\left[\frac{\partial}{\partial \vartheta}\mathcal{S}_X(\vartheta)\right] = 0 \quad \Rightarrow \quad c_1(\vartheta) = \frac{g'(\vartheta)}{\mathcal{I}_X(\vartheta)}.$$

Therefore, we conclude that:

$$T(X) = g(\vartheta) + \frac{g'(\vartheta)}{\mathcal{I}_X(\vartheta)}\mathcal{S}_X(\vartheta) \quad \Leftrightarrow \quad \mathcal{S}_X(\vartheta) = \underbrace{\frac{\mathcal{I}_X(\vartheta)}{g'(\vartheta)}}_{k(\vartheta)}\left[T(X) - g(\vartheta)\right].$$

□

Proposition 3.9. Suppose that the distribution of the sample X belongs to the one-parameter multivariate exponential family with $f(x;\vartheta) = h(x)e^{Q(\vartheta)T(x) - A(\vartheta)}$. If the parameter space Θ is an open subset of $\mathbb{R}$ and the function $Q : \Theta \to \mathbb{R}$ is continuously differentiable with $Q'(\vartheta) \neq 0 \ \forall \vartheta \in \Theta$, then all of the regularity conditions are satisfied. Additionally, the statistic $T(X)$ is an efficient estimator of the parametric function $g(\vartheta) = \frac{A'(\vartheta)}{Q'(\vartheta)}$. In fact, an efficient estimator of $h(\vartheta)$ exists if and only if the distribution of the sample belongs to the exponential family and the parametric function $h(\vartheta)$ is of the form $h(\vartheta) = c_0 + c_1 g(\vartheta)$ for some constants $c_0 \in \mathbb{R}$, $c_1 \neq 0$.

Proof. Regularity conditions I and II are satisfied by assumption. The rest of the regularity conditions can be shown to be satisfied by suitable application of the dominated convergence theorem. Then, we calculate that:

$$\mathcal{S}_X(\vartheta) = \frac{\partial}{\partial \vartheta}\log f(X;\vartheta) = Q'(\vartheta)T(X) - A'(\vartheta) = Q'(\vartheta)\left[T(X) - \frac{A'(\vartheta)}{Q'(\vartheta)}\right].$$

According to the previous proposition, it follows that the statistic $T(X)$ is an efficient estimator of $g(\vartheta) = \frac{A'(\vartheta)}{Q'(\vartheta)}$. According to the proof of the previous proposition, we know that $T(X)$ is an efficient estimator of $g(\vartheta)$ if and only if there exist functions

$c_0(\vartheta)$ and $c_1(\vartheta) \neq 0$ such that $\mathcal{S}_X(\vartheta) = c_0(\vartheta) + c_1(\vartheta)T(X)\ \forall \vartheta \in \Theta$. Then, we calculate that:

$$\log f(X;\vartheta) = T(X) \underbrace{\int c_1(\vartheta)d\vartheta}_{C_1(\vartheta)+A_1(X)} + \underbrace{\int c_0(\vartheta)d\vartheta}_{C_0(\vartheta)+A_0(X)} \quad \Leftrightarrow$$

$$f(x;\vartheta) = \underbrace{e^{A_0(x)+A_1(x)T(x)}}_{h(x)} e^{C_1(\vartheta)T(x)+C_2(\vartheta)}.$$

Therefore, $T(X)$ is an efficient estimator of $g(\vartheta)$ if and only if the distribution of the sample belongs to the exponential family of distributions with $Q(\vartheta) = C_1(\vartheta)$ and $A(\vartheta) = -C_2(\vartheta)$. □

Example 3.29. Let $X_1, \ldots, X_n \sim \text{Bernoulli}(p)$ be a random sample. We calculate that:

$$\log f(x;p) = \log p \sum_{i=1}^n x_i + \log(1-p)\left(n - \sum_{i=1}^n x_i\right),$$

$$\mathcal{S}_X(p) = \frac{\partial}{\partial p}\log f(X;p) = \frac{1}{p}\sum_{i=1}^n X_i - \frac{1}{1-p}\left(n - \sum_{i=1}^n X_i\right)$$

$$= \frac{1}{p(1-p)}\left(\sum_{i=1}^n X_i - np\right) = \frac{n}{p(1-p)}\left(\overline{X} - p\right),$$

where $k(p) = \frac{n}{p(1-p)} \neq 0\ \forall p \in (0,1)$. According to proposition 3.8, the statistic $T(X) = \overline{X}$ is an efficient estimator of the parametric function $g(p) = p$. Alternatively, we observe that the parameter space $\Theta = (0,1)$ is an open subset of $\mathbb{R}$ and the distribution of the sample X belongs to the exponential family with joint PMF:

$$f(x;p) = \exp\left\{n\left[\log p - \log(1-p)\right]\overline{x} - n\log\frac{1}{1-p}\right\},$$

where $T(x) = \overline{x}$, $Q(p) = n\left[\log p - \log(1-p)\right]$ and $A(p) = n\log\frac{1}{1-p}$. We calculate that:

$$Q'(p) = \frac{n}{p} + \frac{n}{1-p} = \frac{n}{p(1-p)} \neq 0, \quad A'(p) = \frac{n}{1-p},$$

so all regularity conditions are satisfied. According to proposition 3.9, $T(X) = \overline{X}$ is an efficient estimator of $g(p) = \frac{A'(p)}{Q'(p)} = p$. Alternatively, we calculate that:

$$\frac{\partial^2}{\partial p^2}\log f(X;p) = -\frac{1}{p^2}\sum_{i=1}^n X_i - \frac{1}{(1-p)^2}\left(n - \sum_{i=1}^n X_i\right),$$

$$\mathcal{I}_X(p) = -\mathbb{E}\left[\frac{\partial^2}{\partial p^2}\log f(X;p)\right] = \frac{1}{p^2}\sum_{i=1}^n \mathbb{E}(X_i) - \frac{1}{(1-p)^2}\left[n - \sum_{i=1}^n \mathbb{E}(X_i)\right]$$

$$= \frac{np}{p^2} + \frac{n-np}{(1-p)^2} = \frac{n}{p(1-p)} \in (0,\infty).$$

We also know that:

$$\mathbb{E}(\overline{X}) = \mathbb{E}(X_1) = p = g(p),$$

$$\mathrm{Var}(\overline{X}) = \frac{\mathrm{Var}(X_1)}{n} = \frac{1}{n}p(1-p) = \frac{[g'(p)]^2}{\mathcal{I}_X(p)}.$$

Therefore, we conclude that $T(X) = \overline{X}$ is an efficient estimator of p. □

Example 3.30. Let $X_1, \ldots, X_n$ be a random sample with $f(x;\vartheta) = \frac{\log\vartheta}{\vartheta-1}\vartheta^x$ for $\vartheta > 1$ and $x \in (0,1)$. We calculate that:

$$\log f(x;\vartheta) = n\log\log\vartheta - n\log(\vartheta-1) + \log\vartheta\sum_{i=1}^n x_i,$$

$$\begin{aligned}\mathcal{S}_X(\vartheta) = \frac{\partial}{\partial\vartheta}\log f(X;\vartheta) &= \frac{n}{\vartheta\log\vartheta} - \frac{n}{\vartheta-1} + \frac{1}{\vartheta}\sum_{i=1}^n X_i \\ &= \frac{1}{\vartheta}\left(\sum_{i=1}^n X_i - \frac{n\vartheta}{\vartheta-1} + \frac{n}{\log\vartheta}\right) = \frac{n}{\vartheta}\left[\overline{X} - \left(\frac{\vartheta}{\vartheta-1} - \frac{1}{\log\vartheta}\right)\right],\end{aligned}$$

where $k(\vartheta) = \frac{n}{\vartheta} \neq 0\ \forall\vartheta > 1$. According to proposition 3.8, the statistic $T(X) = \overline{X}$ is an efficient estimator of the parametric function $g(\vartheta) = \frac{\vartheta}{\vartheta-1} - \frac{1}{\log\vartheta}$. Alternatively, we observe that the parameter space $\Theta = (1,\infty)$ is an open subset of $\mathbb{R}$ and the distribution of the sample X belongs to the exponential family with the following joint PDF:

$$f(x;\vartheta) = \exp\left\{n\overline{x}\log\vartheta - n\left[\log(\vartheta-1) + \log\frac{1}{\log\vartheta}\right]\right\},$$

where $T(x) = \overline{x}$, $Q(\vartheta) = n\log\vartheta$ and $A(\vartheta) = n\left[\log(\vartheta-1) + \log\frac{1}{\log\vartheta}\right]$. We calculate that:

$$Q'(\vartheta) = \frac{n}{\vartheta} \neq 0, \quad A'(\vartheta) = \frac{n}{\vartheta-1} - \frac{n}{\vartheta\log\vartheta},$$

so all of the regularity conditions are satisfied. According to proposition 3.9, the statistic $T(X) = \overline{X}$ is an efficient estimator of $g(\vartheta) = \frac{A'(\vartheta)}{Q'(\vartheta)} = \frac{\vartheta}{\vartheta-1} - \frac{1}{\log\vartheta}$. We note that it would have been exceptionally arduous to calculate the variance of $T(X)$ to compare it against the Cramér - Rao lower bound. □

Note 3.20. If we know of an unbiased estimator of $g(\vartheta)$, it suffices to calculate its variance and compare it against the Cramér - Rao lower bound to check whether it's efficient. Otherwise, we can apply proposition 3.8 or proposition 3.9 to check whether an efficient estimator of $g(\vartheta)$ exists or not. Indicatively, in table 3.3 we summarize the Fisher information of 1 observation for the parameters of some widely used distributions.

Bernoulli(p)	$1/p(1-p)$
Bin(N, p) with known N	$N/p(1-p)$
Poisson(λ)	$1/\lambda$
Exp(ϑ)	$1/\vartheta^2$
Beta($\vartheta, 1$)	
Beta($1, \vartheta$)	
Gamma(k, λ) with known k	k/λ^2
$\mathcal{N}(\mu, \sigma^2)$ with known σ^2	$1/\sigma^2$
$\mathcal{N}(\mu, \sigma^2)$ with known μ	$1/2\sigma^4$

TABLE 3.3: Fisher Information of Notable Distributions

3.9* Multivariate Cramér - Rao Inequality

Definition 3.15. Let $X = (X_1, X_2, \dots, X_r) \in \mathbb{R}^r$ and $Y = (Y_1, Y_2, \dots, Y_s) \in \mathbb{R}^s$ be 2 random vectors. Then, we define:

- $E(X) = [E(X_1), E(X_2), \dots, E(X_r)]^{\mathrm{T}} \in \mathbb{R}^r$ the mean vector of X;
- $\mathrm{Var}(X) = E\left(XX^{\mathrm{T}}\right) - E(X)\left[E(X)\right]^{\mathrm{T}} \in \mathbb{R}^{r\times r}$ the covariance matrix of X;
- $\mathrm{Cov}(X, Y) = E\left(XY^{\mathrm{T}}\right) - E(X)\left[E(Y)\right]^{\mathrm{T}} \in \mathbb{R}^{r\times s}$ the covariance matrix between X and Y.

Proposition 3.10. Let $X \in \mathbb{R}^r$, $Y \in \mathbb{R}^s$ be random vectors, $A \in \mathbb{R}^{m\times r}$, $B \in \mathbb{R}^{m\times s}$ be constant matrices and $c \in \mathbb{R}^m$, $d \in \mathbb{R}^m$ be constant vectors. Then, we know that:

i. $E(AX + c) = AE(X) + c$,

ii. $E(AX + BY) = AE(X) + BE(Y)$,

iii. $\mathrm{Var}(AX + c) = A\mathrm{Var}(X)A^{\mathrm{T}}$,

iv. $\mathrm{Var}(X)$ is symmetric and positive semi-definite, i.e. $u^{\mathrm{T}}\mathrm{Var}(X)u \geqslant 0 \ \forall u \in \mathbb{R}^r$.

v. $\mathrm{Cov}(X, c) = \mathbf{0}_{r\times m}$,

vi. $\mathrm{Cov}(X, X) = \mathrm{Var}(X)$,

vii. $\mathrm{Cov}(Y, X) = \left[\mathrm{Cov}(X, Y)\right]^{\mathrm{T}}$,

viii. $\mathrm{Cov}(AX + c, BY + d) = A\mathrm{Cov}(X, Y)B^{\mathrm{T}}$,

ix. $\mathrm{Var}(AX+BY) = A\mathrm{Var}(X)A^{\mathrm{T}} + B\mathrm{Var}(Y)B^{\mathrm{T}} + A\mathrm{Cov}(X, Y)B^{\mathrm{T}} + B\mathrm{Cov}(Y, X)A^{\mathrm{T}}$,

x. X, Y independent $\Rightarrow \mathrm{Cov}(X, Y) = \mathbf{0}_{r\times s} \Rightarrow E\left(XY^{\mathrm{T}}\right) = E(X)\left[E(Y)\right]^{\mathrm{T}}$.

Definition 3.16. i. The function $\mathcal{S}_X(\vartheta) = \nabla_\vartheta \log f(X; \vartheta) \in \mathbb{R}^s$ is called the *score function* of the sample X for the parameter $\vartheta \in \mathbb{R}^s$.

ii. The parametric function $\mathcal{I}_X(\vartheta) = \mathbb{E}\left[\mathcal{S}_X(\vartheta)\mathcal{S}_X^{\mathrm{T}}(\vartheta)\right] \in \mathbb{R}^{s\times s}$ is called the *Fisher information matrix* of the sample X for the parameter ϑ.

Proposition 3.11. If $g(\eta) = \vartheta$ is a continuously differentiable function with Jacobian matrix $\mathcal{J}_g(\eta) \in \mathbb{R}^{s\times d}$, then $\mathcal{I}_X(\eta) = \mathcal{J}_g^{\mathrm{T}}(\eta)\mathcal{I}_X\left(g(\eta)\right)\mathcal{J}_g(\eta) \in \mathbb{R}^{d\times d}$.

Proof. According to the chain rule, we calculate that:

$$\begin{aligned}
\mathcal{I}_X(\eta) &= \mathbb{E}_\eta\left[\mathcal{S}_X(\eta)\mathcal{S}_X^{\mathrm{T}}(\eta)\right] = \mathbb{E}_\eta\left[\nabla_\eta \log f(X;\eta)\nabla_\eta^{\mathrm{T}} \log f(X;\eta)\right] \\
&= \mathbb{E}_\vartheta\left[\mathcal{J}_g^{\mathrm{T}}(\eta)\nabla_\vartheta \log f(X;\vartheta)\nabla_\vartheta^{\mathrm{T}} \log f(X;\vartheta)\mathcal{J}_g(\eta)\right] \\
&= \mathcal{J}_g^{\mathrm{T}}(\eta)\mathbb{E}_\vartheta\left[\nabla_\vartheta \log f(X;\vartheta)\nabla_\vartheta^{\mathrm{T}} \log f(X;\vartheta)\right]\mathcal{J}_g(\eta) \\
&= \mathcal{J}_g^{\mathrm{T}}(\eta)\mathcal{I}_X\left(g(\eta)\right)\mathcal{J}_g(\eta).
\end{aligned}$$

□

Regularity Conditions: Without loss of generality, assume that the distribution of the sample is continuous with joint PDF $f(x;\vartheta)$ for $\vartheta \in \Theta \subseteq \mathbb{R}^s$ and $x \in S$. We define the following regularity conditions:

I. The parameter space Θ is an open subset of $\mathbb{R}^s$.

II. The support $S = \{x \in \mathbb{R}^n : f(x;\vartheta) > 0\}$ doesn't depend on the value of ϑ.

III. $\frac{\partial}{\partial\vartheta_j} f(x;\vartheta) < \infty$ $\forall x \in S$ and $\forall\vartheta \in \Theta$ for $j = 1, 2, \ldots, s$.

IV. $\int_S \frac{\partial}{\partial\vartheta_j} f(x;\vartheta)dx = \frac{\partial}{\partial\vartheta_j}\int_S f(x;\vartheta)dx = 0$ $\forall\vartheta \in \Theta$ for $j = 1, 2, \ldots, s$.

V. The matrix $\mathcal{I}_X(\vartheta) \in \mathbb{R}^{s\times s}$ is positive definite $\forall\vartheta \in \Theta$.

Proposition 3.12. Suppose that the following regularity conditions are satisfied:

VI. $\frac{\partial^2}{\partial\vartheta_j\partial\vartheta_k} f(x;\vartheta) < \infty$ $\forall x \in S$ and $\forall\vartheta \in \Theta$ for $j, k = 1, 2, \ldots, s$.

VII. $\frac{\partial^2}{\partial\vartheta_j\partial\vartheta_k} f(x;\vartheta)dx = \frac{\partial^2}{\partial\vartheta_j\partial\vartheta_k}\int_S f(x;\vartheta)dx = 0$ $\forall\vartheta \in \Theta$ for $j, k = 1, 2, \ldots, s$.

Then, it follows that $\mathcal{I}_X(\vartheta) = -\mathbb{E}_\vartheta\left[\mathcal{H}_X(\vartheta)\right]$, where $\mathcal{H}_X(\vartheta)$ is the Hessian matrix of $\log f(X;\vartheta)$, i.e. the Jacobian matrix of the score function $\mathcal{S}_X(\vartheta)$.

Proof. We follow the same steps as the proof of proposition 3.6 (page 52). □

Proposition 3.13. Let X be a sample which satisfies the regularity conditions I-V. Then, $\mathbb{E}\left[\mathcal{S}_X(\vartheta)\right] = 0$ and $\mathrm{Var}_\vartheta\left[\mathcal{S}_X(\vartheta)\right] = \mathbb{E}_\vartheta\left[\mathcal{S}_X(\vartheta)\mathcal{S}_X^{\mathrm{T}}(\vartheta)\right] = \mathcal{I}_X(\vartheta)$ $\forall\vartheta \in \Theta$.

Proof. We follow the same steps as the proof of proposition 3.7 (page 52). □

Lemma 3.1. (Multivariate Covariance Inequality) Let $X \in \mathbb{R}^n$ and $Y \in \mathbb{R}^m$ be 2 random vectors with positive definite covariance matrices. Then, it follows that the matrix difference $\text{Var}(X) - \text{Cov}(X, Y)\,[\text{Var}(Y)]^{-1}\,\text{Cov}(Y, X)$ is positive semi-definite.

Proof. Without loss of generality, assume that $\mathbb{E}(X) = \mathbb{E}(Y) = 0$. For any constant matrix $A \in \mathbb{R}^{n\times m}$, we know that the matrix $(X + AY)(X + AY)^{\text{T}}$ is positive semi-definite. Then, we observe that:

$$(X + AY)(X + AY)^{\text{T}} = XX^{\text{T}} + AYX^{\text{T}} + XY^{\text{T}}A^{\text{T}} + AYY^{\text{T}}A^{\text{T}} \quad \Rightarrow$$

$$\mathbb{E}\left[(X + AY)(X + AY)^{\text{T}}\right] = \text{Var}(X) + A\text{Cov}(Y, X) + \text{Cov}(X, Y)A^{\text{T}} + A\text{Var}(Y)A^{\text{T}}.$$

Let $A = -\text{Cov}(X, Y)\,[\text{Var}(Y)]^{-1}$. Since $\text{Cov}(Y, X) = [\text{Cov}(X, Y)]^{\text{T}}$, it follows that:

$$\mathbb{E}\left[(X + AY)(X + AY)^{\text{T}}\right] = \text{Var}(X) - \text{Cov}(X, Y)\,[\text{Var}(Y)]^{-1}\,\text{Cov}(Y, X).$$

Therefore, we conclude that the matrix $\text{Var}(X) - \text{Cov}(X, Y)\,[\text{Var}(Y)]^{-1}\,\text{Cov}(Y, X)$ is positive semi-definite. □

Theorem 3.10. (Multivariate Cramér - Rao Inequality) Let X be a sample with joint PDF $f(x; \vartheta)$ for $\vartheta \in \Theta \subseteq \mathbb{R}^s$ and $x \in S$ which satisfies the regularity conditions I-V. Suppose that the statistic $T(X)$ is an estimator of the parametric function $g(\vartheta) \in \mathbb{R}^d$ with finite variance which satisfies the following regularity condition:

VIII. $\int_S T_h(x)\frac{\partial}{\partial\vartheta_j}f(x; \vartheta)dx = \frac{\partial}{\partial\vartheta_j}\int_S T_h(x)f(x; \vartheta)dx = \frac{\partial}{\partial\vartheta_j}\mathbb{E}_\vartheta\left[T_h(X)\right]\ \forall\vartheta \in \Theta$,

where $\mathbb{E}_\vartheta\left[T_h(X)\right] = g_h(\vartheta) + \text{bias}_{g_h(\vartheta)}\left[T_h(X)\right]$ for $h = 1, 2, \ldots, d$. Then, the matrix difference $\text{Var}_\vartheta\left[T(X)\right] - \mathcal{J}_m(\vartheta)\mathcal{I}_X^{-1}(\vartheta)\mathcal{J}_m^{\text{T}}(\vartheta) \in \mathbb{R}^{d\times d}$ is positive semi-definite $\forall\vartheta \in \Theta$, where $\mathcal{J}_m \in \mathbb{R}^{d\times s}$ is the Jacobian matrix of $m(\vartheta) = \mathbb{E}_\vartheta\left[T(X)\right]$.

Proof. By making use of the regularity conditions VII and IV, we calculate that:

$$\begin{aligned}
\nabla_\vartheta m_h(\vartheta) &= \nabla_\vartheta\mathbb{E}_\vartheta\left[T_h(X)\right] = \nabla_\vartheta\int_S T_h(x)f(x; \vartheta)dx \\
&= \int_S T_h(x)\nabla_\vartheta f(x; \vartheta)dx = \int_S T_h(x)f(x; \vartheta)\frac{1}{f(x; \vartheta)}\nabla_\vartheta f(x; \vartheta)dx \\
&= \int_S f(x; \vartheta)T_h(x)\nabla_\vartheta \log f(x; \vartheta)dx = \mathbb{E}_\vartheta\left[T_h(X)\mathcal{S}_X(\vartheta)\right] \\
&= \text{Cov}_\vartheta\left[T_h(X), \mathcal{S}_X(\vartheta)\right] + \mathbb{E}_\vartheta\left[T_h(X)\right]\mathbb{E}_\vartheta\left[\mathcal{S}_X(\vartheta)\right] \\
&= \text{Cov}_\vartheta\left[T_h(X), \mathcal{S}_X(\vartheta)\right],
\end{aligned}$$

where $\mathbb{E}_\vartheta\left[\mathcal{S}_X(\vartheta)\right] = 0$ according to the previous proposition. Then, we infer that:

$$\text{Cov}_\vartheta\left[T(X), \mathcal{S}_X(\vartheta)\right]\left[\text{Var}\left(\mathcal{S}_X(\vartheta)\right)\right]^{-1}\text{Cov}_\vartheta\left[\mathcal{S}_X(\vartheta), T(X)\right] = \mathcal{J}_m(\vartheta)\mathcal{I}_X^{-1}(\vartheta)\mathcal{J}_m^{\text{T}}(\vartheta),$$

where $\mathrm{Var}\left[\mathcal{S}_X(\vartheta)\right] = \mathcal{I}_X(\vartheta)$ according to the previous proposition. By making use of the multivariate covariance inequality, we conclude that the matrix difference $\mathrm{Var}_\vartheta\left[T(X)\right] - \mathcal{J}_m(\vartheta)\mathcal{I}_X^{-1}(\vartheta)\mathcal{J}_m^{\mathrm{T}}(\vartheta)$ is positive semi-definite. □

Corollary 3.7. If the statistic $T(X)$ is an unbiased estimator of the parametric function $g(\vartheta)$, then the matrix difference $\mathrm{Var}_\vartheta\left[T(X)\right] - \mathcal{J}_g(\vartheta)\mathcal{I}_X^{-1}(\vartheta)\mathcal{J}_g^{\mathrm{T}}(\vartheta)$ is positive semi-definite $\forall\vartheta \in \Theta$.

Proof. If $T(X)$ is an unbiased estimator of $g(\vartheta)$, then $m(\vartheta) = \mathbb{E}_\vartheta\left[T(X)\right] = g(\vartheta)$, so the result follows immediately from the multivariate Cramér - Rao inequality. □

Definition 3.17. An unbiased estimator $T(X)$ of $g(\vartheta)$ which achieves the Cramér - Rao lower bound, i.e. for which it holds that $\mathrm{Var}_\vartheta\left[T(X)\right] = \mathcal{J}_g(\vartheta)\mathcal{I}_X^{-1}(\vartheta)\mathcal{J}_g^{\mathrm{T}}(\vartheta)$ $\forall\vartheta \in \Theta$, is called an *efficient* estimator of $g(\vartheta)$.

Proposition 3.14. A statistic $T(X)$ is an efficient estimator of $g(\vartheta)$ if and only if there exists a function $K(\vartheta) \in \mathbb{R}^{d\times s}$ such that $K(\vartheta)\mathcal{S}_X(\vartheta) = T(X) - g(\vartheta)$ $\forall\vartheta \in \Theta$. Then, it holds that $K(\vartheta) = \mathcal{J}_g(\vartheta)\mathcal{I}_X^{-1}(\vartheta)$.

Proof. We know that $T(X)$ is an efficient estimator of $g(\vartheta)$ if and only if it's an unbiased estimator of $g(\vartheta)$ and it achieves the Cramér - Rao lower bound. Furthermore, we know that $T(X)$ achieves the Cramér - Rao lower bound if and only if the covariance inequality invoked in the proof of the Cramér - Rao inequality holds as an equality $\forall\vartheta \in \Theta$. We also know that the covariance inequality holds as an equality if and only if there exist functions $c_0(\vartheta) \in \mathbb{R}^d$ and $C_1(\vartheta) \in \mathbb{R}^{d\times s}$ such that $T(X) = c_0(\vartheta) + C_1(\vartheta)\mathcal{S}_X(\vartheta)$ $\forall\vartheta \in \Theta$. Then, we calculate that:

$$g(\vartheta) = \mathbb{E}\left[T(X)\right] = c_0(\vartheta) + C_1(\vartheta)\mathbb{E}_\vartheta\left[\mathcal{S}_X(\vartheta)\right] = c_0(\vartheta),$$

$$0 = \mathcal{J}_{c_0}(\vartheta) + \frac{\partial C_1}{\partial\vartheta}\mathcal{S}_X(\vartheta) + C_1(\vartheta)\mathcal{H}_X(\vartheta) \quad \Rightarrow$$

$$\mathcal{J}_g(\vartheta) + \frac{\partial C_1}{\partial\vartheta}\mathbb{E}_\vartheta\left[\mathcal{S}_X(\vartheta)\right] + C_1(\vartheta)\mathbb{E}_\vartheta\left[\mathcal{H}_X(\vartheta)\right] = 0 \quad \Rightarrow \quad C_1(\vartheta) = \mathcal{J}_g(\vartheta)\mathcal{I}_X^{-1}(\vartheta).$$

Therefore, we conclude that:

$$T(X) = g(\vartheta) + \underbrace{\mathcal{J}_g(\vartheta)\mathcal{I}_X^{-1}(\vartheta)}_{K(\vartheta)}\mathcal{S}_X(\vartheta).$$

□

Proposition 3.15. Suppose that the distribution of the sample X belongs to the multiparameter multivariate full exponential family with $f(x;\vartheta) = h(x)e^{\langle Q(\vartheta),T(x)\rangle - A(\vartheta)}$.

If the parameter space Θ is an open subset of $\mathbb{R}^s$ and the function $Q : \Theta \to \mathbb{R}^s$ is continuously differentiable with invertible Jacobian matrix $\mathcal{J}_Q$, then all of the regularity conditions are satisfied. Furthermore, the statistic $T(X)$ is an efficient estimator of the parametric function $g(\vartheta) = \mathcal{J}_Q^{-1}(\vartheta)\nabla_\vartheta A(\vartheta) \in \mathbb{R}^s$.

Proof. Regularity conditions I and II are satisfied by assumption. The rest of the regularity conditions can be shown to be satisfied by suitable application of the dominated convergence theorem. Then, we calculate that:

$$\begin{aligned}\mathcal{S}_X(\vartheta) = \nabla_\vartheta \log f(X;\vartheta) &= \mathcal{J}_Q(\vartheta)T(X) - \nabla_\vartheta A(\vartheta)\\ &= \mathcal{J}_Q(\vartheta)\left[T(X) - \mathcal{J}_Q^{-1}(\vartheta)\nabla_\vartheta A(\vartheta)\right].\end{aligned}$$

According to the previous proposition, it follows that the statistic $T(X)$ is an efficient estimator of $g(\vartheta) = \mathcal{J}_Q^{-1}(\vartheta)\nabla_\vartheta A(\vartheta)$. □

Example 3.31. Let $X_1, \ldots, X_n \sim \mathcal{N}(\vartheta_1, \vartheta_2)$ be a random sample. We want to show that the statistic $T(X) = \left(\overline{X}, \frac{1}{n}\sum_{i=1}^n X_i^2\right)$ is an efficient estimator of the parametric function $g(\vartheta) = \left(\vartheta_1, \vartheta_1^2 + \vartheta_2\right)$, whereas the sample variance S^2 isn't an efficient estimator of ϑ_2. We observe that the parameter space $\Theta = \mathbb{R} \times (0, \infty)$ is an open subset of $\mathbb{R}^2$. According to example 3.13 (page 37), the distribution of the sample belongs to the exponential family with:

$$T(x) = \left(\overline{x}, \frac{1}{n}\sum_{i=1}^n x_i^2\right), \quad Q(\vartheta) = \left(\frac{n\vartheta_1}{\vartheta_2}, -\frac{n}{2\vartheta_2}\right), \quad A(\vartheta) = \frac{n\vartheta_1^2}{2\vartheta_2} + \frac{n\log\vartheta_2}{2}.$$

We calculate that:

$$\nabla_\vartheta A(\vartheta) = \begin{bmatrix} \frac{n\vartheta_1}{\vartheta_2} \\ -\frac{n\vartheta_1^2}{2\vartheta_2^2} + \frac{n}{2\vartheta_2} \end{bmatrix}, \quad \mathcal{J}_Q(\vartheta) = \begin{bmatrix} \frac{n}{\vartheta_2} & -\frac{n\vartheta_1}{\vartheta_2^2} \\ 0 & \frac{n}{2\vartheta_2^2} \end{bmatrix}, \quad \mathcal{J}_Q^{-1}(\vartheta) = \begin{bmatrix} \frac{\vartheta_2}{n} & \frac{2\vartheta_1\vartheta_2}{n} \\ 0 & \frac{2\vartheta_2^2}{n} \end{bmatrix},$$

so all of the regularity conditions are satisfied. According to proposition 3.15, the statistic $T(X) = \left(\overline{X}, \frac{1}{n}\sum_{i=1}^n X_i^2\right)$ is an efficient estimator of the parametric function $g(\vartheta) = \mathcal{J}_Q^{-1}(\vartheta)\nabla_\vartheta A(\vartheta) = \left(\vartheta_1, \vartheta_1^2 + \vartheta_2\right)$. Alternatively, we calculate that:

$$\log f(x;\vartheta) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\vartheta_2 - \frac{1}{2\vartheta_2}\sum_{i=1}^n (x_i - \vartheta_1)^2,$$

$$\mathcal{S}_X(\vartheta) = \nabla_\vartheta \log f(X;\vartheta) = \begin{bmatrix} \frac{1}{\vartheta_2}\sum_{i=1}^n (X_i - \vartheta_1) \\ -\frac{n}{2\vartheta_2} + \frac{1}{2\vartheta_2^2}\sum_{i=1}^n (X_i - \vartheta_1)^2 \end{bmatrix},$$

$$\mathcal{H}_X(\vartheta) = \mathcal{J}_{\mathcal{S}_X}(\vartheta) = \begin{bmatrix} -\frac{n}{\vartheta_2} & -\frac{1}{\vartheta_2^2}\sum_{i=1}^n (X_i - \vartheta_1) \\ -\frac{1}{\vartheta_2^2}\sum_{i=1}^n (X_i - \vartheta_1) & \frac{n}{2\vartheta_2^2} - \frac{1}{\vartheta_2^3}\sum_{i=1}^n (X_i - \vartheta_1)^2 \end{bmatrix},$$

$$\mathcal{I}_X(\vartheta) = -\mathbb{E}\left[\mathcal{H}_X(\vartheta)\right] = \begin{bmatrix} \frac{n}{\vartheta_2} & 0 \\ 0 & \frac{n}{2\vartheta_2^2} \end{bmatrix}, \quad \mathcal{J}_g(\vartheta) = \begin{bmatrix} 1 & 0 \\ 2\vartheta_1 & 1 \end{bmatrix}.$$

According to example 2.4 (page 21), we know that:

$$\mathbb{E}\left[T(X)\right] = \begin{bmatrix} \vartheta_1 \\ \vartheta_1^2 + \vartheta_2 \end{bmatrix}, \quad \operatorname{Var}\left[T(X)\right] = \begin{bmatrix} \frac{\vartheta_2}{n} & \frac{2\vartheta_1\vartheta_2}{n} \\ \frac{2\vartheta_1\vartheta_1}{n} & \frac{4\vartheta_1^2\vartheta_2 + 2\vartheta_2^2}{n} \end{bmatrix} = \mathcal{J}_g(\vartheta)\mathcal{I}_X^{-1}(\vartheta)\mathcal{J}_g^{\mathrm{T}}(\vartheta).$$

According to proposition 3.14, the statistic $T(X) = \left(\overline{X}, \frac{1}{n}\sum_{i=1}^n X_i^2\right)$ is an efficient estimator of the parametric function $g(\vartheta)$. Alternatively, we observe that:

$$\begin{aligned} \mathcal{J}_g(\vartheta)\mathcal{I}_X^{-1}(\vartheta)\mathcal{S}_X(\vartheta) &= \begin{bmatrix} \frac{\vartheta_2}{n} & 0 \\ \frac{2\vartheta_1\vartheta_2}{n} & \frac{2\vartheta_2^2}{n} \end{bmatrix} \begin{bmatrix} \frac{1}{\vartheta_2}\sum_{i=1}^n (X_i - \vartheta_1) \\ -\frac{n}{2\vartheta_2} + \frac{1}{2\vartheta_2^2}\sum_{i=1}^n (X_i - \vartheta_1)^2 \end{bmatrix} \\ &= \begin{bmatrix} \overline{X} - \vartheta_1 \\ \frac{1}{n}\sum_{i=1}^n X_i^2 - \vartheta_1^2 - \vartheta_2 \end{bmatrix} = T(X) - g(\vartheta). \end{aligned}$$

Therefore, the statistic $T(X) = \left(\overline{X}, \frac{1}{n}\sum_{i=1}^n X_i^2\right)$ is an efficient estimator of the parametric function $g(\vartheta)$. Now, we let $h(\vartheta) = \vartheta_2$ and calculate that $\mathcal{J}_h(\vartheta) = (0, 1)$. According to note 3.12 (page 38), we know that:

$$\operatorname{Var}\left(S^2\right) = \frac{2}{n-1}\vartheta_2^2 > \frac{2}{n}\vartheta_2^2 = \mathcal{J}_h(\vartheta)\mathcal{I}_X^{-1}(\vartheta)\mathcal{J}_h^{\mathrm{T}}(\vartheta),$$

which implies that the sample variance S^2 isn't an efficient estimator of ϑ_2. Since S^2 is the UMVUE of ϑ_2, according to example 3.26 (page 49), it follows that there doesn't exist any efficient estimator of ϑ_2. □

3.10 Asymptotic Distribution of Estimators

Definition 3.18. (Convergence of Random Variables)

i. **Almost sure convergence**: $X_n \xrightarrow{\text{a.s.}} X \;\Leftrightarrow\; \mathbb{P}\left(\lim_{n\to\infty} X_n = X\right) = 1$

ii. **Convergence in probability**: $X_n \xrightarrow{p} X \;\Leftrightarrow\; \lim_{n\to\infty} \mathbb{P}\left(|X_n - X| < \varepsilon\right) = 1$ for all $\varepsilon > 0$

iii. **Convergence in distribution**: $X_n \xrightarrow{d} X \;\Leftrightarrow\; \lim_{n\to\infty} F_{X_n}(x) = F_X(x)$ for every continuity point x of the CDF F_X

Definition 3.19. i. A statistic $T_n(X)$ is called a *strongly consistent* estimator of $g(\vartheta)$ if $T_n(X) \xrightarrow{\text{a.s.}} g(\vartheta)$ $\forall \vartheta \in \Theta$.

ii. A statistic $T_n(X)$ is called a (weakly) *consistent* estimator of $g(\vartheta)$ if $T_n(X) \xrightarrow{p} g(\vartheta)$ $\forall \vartheta \in \Theta$.

iii. A statistic $T_n(X)$ has an *asymptotic distribution* if there exists a sequence $(r_n)_{n\geqslant 1}$ of real numbers with $\lim_{n\to\infty} r_n = \infty$ such that $r_n \left[T_n(X) - g(\vartheta)\right] \xrightarrow{d} Y$ for some random variable Y.

Interpretation: The consistency property ensures that all the most probable values of an estimator of ϑ are concentrated more and more tightly around the true value of ϑ, as we're collecting more and more data. Therefore, it doesn't provide any information about the properties of an estimator based on a sample of a given size, but rather only about its asymptotic behavior.

All results concerning the asymptotic distribution of estimators are consequences of well-known results in the field of measure-theoretic probability theory. We are going to mention those useful probabilistic results without proof and build open them to prove all the needed results in asymptotic statistics. The proofs of those probabilistic results can be found in any standard probability theory textbook such as "Probability and Measure" by Patrick Billingsley and "Probability Theory and Examples" by Rick Durrett.

Proposition 3.16. i. $X_n \xrightarrow{\text{a.s.}} X \Rightarrow X_n \xrightarrow{p} X \;\Rightarrow\; X_n \xrightarrow{d} X$.

ii. If $X = c$ is a degenerate random variable, i.e. it holds that $\mathbb{P}(X = c) = 1$, then $X_n \xrightarrow{p} c \;\Leftrightarrow\; X_n \xrightarrow{d} c$.

iii. If $X_n \xrightarrow{\text{a.s.}/p} X$ and $Y_n \xrightarrow{\text{a.s.}/p} Y$, then $X_n + Y_n \xrightarrow{\text{a.s.}/p} X + Y$ and $X_nY_n \xrightarrow{\text{a.s.}/p} XY$. If $Y_n \neq 0$ and $Y \neq 0$, then $\frac{X_n}{Y_n} \xrightarrow{\text{a.s.}/p} \frac{X}{Y}$.

Corollary 3.8. If the statistic $T_n(X)$ is a strongly consistent estimator of $g(\vartheta)$, then it's also a consistent estimator of $g(\vartheta)$.

Proof. According to the previous proposition, we know that $T_n(X) \xrightarrow{\text{a.s.}} g(\vartheta)$ implies that $T_n(X) \xrightarrow{p} g(\vartheta)$. □

Theorem 3.11. (Slutsky) If $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{d} c$, then $X_n + Y_n \xrightarrow{d} X + c$ and $X_nY_n \xrightarrow{d} cX$. If $Y_n \neq 0$ and $c \neq 0$, then $\frac{X_n}{Y_n} \xrightarrow{d} \frac{X}{c}$.

Definition 3.20. i. A statistic $T_n(X)$ is called an *asymptotically unbiased* estimator of $g(\vartheta)$ if it holds that $\lim_{n\to\infty} \mathbb{E}_\vartheta \left[T_n(X)\right] = g(\vartheta)$.

ii. A statistic $T_n(X)$ is called an *asymptotically efficient* estimator of ϑ if it holds that

$\sqrt{n}\left[T_n(X)-\vartheta\right] \xrightarrow{d} \mathcal{N}\left(0, \mathcal{I}_{X_1}^{-1}(\vartheta)\right)$, i.e. it's asymptotically normal with asymptotic variance which is equal to the Cramér - Rao lower bound.

Note 3.21. The property of asymptotic efficiency ensures that the variance of an estimator becomes as small as possible, as we're collecting more and more data, even if it doesn't achieve the smallest possible variance based on a sample of a given size.

Proposition 3.17. (Sufficient Conditions for Consistency)

i. If the statistic $T_n(X)$ is an unbiased estimator of the parametric function $g(\vartheta)$ with $\lim_{n\to\infty} \operatorname{Var}_\vartheta\left[T_n(X)\right]=0$, then it's a consistent estimator of $g(\vartheta)$.

ii. If the statistic $T_n(X)$ is an asymptotically unbiased estimator of $g(\vartheta)$ and it holds that $\lim_{n\to\infty} \operatorname{Var}_\vartheta\left[T_n(X)\right]=0$, then it's a consistent estimator of $g(\vartheta)$.

iii. If $r_n\left[T_n(X)-g(\vartheta)\right] \xrightarrow{d} Y$ for some sequence $(r_n)_{n\geqslant 1}$ with $\lim_{n\to\infty} r_n=\infty$, then the statistic $T_n(X)$ is a consistent estimator $g(\vartheta)$.

Proof. i. According to Chebyshev's inequality, we know that:

$$\mathbb{P}_\vartheta\left[|T_n(X)-g(\vartheta)| \geqslant \varepsilon\right] \leqslant \frac{\operatorname{Var}_\vartheta\left[T_n(X)\right]}{\varepsilon^2} \xrightarrow{n\to\infty} 0 \quad \Rightarrow$$

$$\lim_{n\to\infty} \mathbb{P}_\vartheta\left[|T_n(X)-g(\vartheta)| \geqslant \varepsilon\right]=0 \quad \Rightarrow \quad \lim_{n\to\infty} \mathbb{P}_\vartheta\left[|T_n(X)-g(\vartheta)|<\varepsilon\right]=1.$$

ii. According to Markov's inequality, we know that:

$$\begin{aligned}\mathbb{P}_\vartheta\left[|T_n(X)-g(\vartheta)| \geqslant \varepsilon\right] &= \mathbb{P}_\vartheta\left[(T_n(X)-g(\vartheta))^2 \geqslant \varepsilon^2\right] \leqslant \frac{1}{\varepsilon^2}\mathbb{E}_\vartheta\left[(T_n(X)-g(\vartheta))^2\right] \\ &= \frac{\operatorname{Var}_\vartheta\left[T_n(X)\right]+\left[\mathbb{E}_\vartheta\left(T_n(X)\right)-g(\vartheta)\right]^2}{\varepsilon^2} \xrightarrow{n\to\infty} 0 \quad \Rightarrow\end{aligned}$$

$$\lim_{n\to\infty} \mathbb{P}_\vartheta\left[|T_n(X)-g(\vartheta)| \geqslant \varepsilon\right]=0 \quad \Rightarrow \quad \lim_{n\to\infty} \mathbb{P}_\vartheta\left[|T_n(X)-g(\vartheta)|<\varepsilon\right]=1.$$

iii. According to Slutsky's theorem, we infer that:

$$T_n(X)=\frac{1}{r_n}\cdot r_n\left[T_n(X)-g(\vartheta)\right]+g(\vartheta) \xrightarrow{d} 0\cdot Y+g(\vartheta)=g(\vartheta).$$

According to proposition 3.16, we conclude that $T_n(X) \xrightarrow{p} g(\vartheta)$. □

Theorem 3.12. (Continuous Mapping Theorem) If $X_n \xrightarrow{\text{a.s.}/p/d} X$ and the function $g:\mathbb{R}\to\mathbb{R}$ is continuous, then $g\left(X_n\right) \xrightarrow{\text{a.s.}/p/d} g(X)$.

Corollary 3.9. If $T_n(X)$ is a (strongly) consistent estimator of ϑ and the function $g:\Theta\to\mathbb{R}$ is continuous, then $g\left(T_n(X)\right)$ is a (strongly) consistent estimator of the parametric function $g(\vartheta)$.

Proof. According to the continuous mapping theorem, we know that $T_n(X) \overset{\text{a.s.}/p}{\to} \vartheta$ implies that $g(T_n(X)) \overset{\text{a.s.}/p}{\to} g(\vartheta)$. □

Theorem 3.13. (Delta Method) If $r_n(X_n - \vartheta) \overset{d}{\to} Y$, where $\lim_{n\to\infty} r_n = \infty$, and the function $g : \Theta \to \mathbb{R}$ is continuously differentiable with $g'(\vartheta) \neq 0$, then it follows that $r_n[g(X_n) - g(\vartheta)] \overset{d}{\to} g'(\vartheta)Y$.

Theorem 3.14.* (Second-Order Delta Method) Suppose that $r_n(X_n - \vartheta) \overset{d}{\to} Y$, where $\lim_{n\to\infty} r_n = \infty$. If the function $g : \Theta \to \mathbb{R}$ is 2 times continuously differentiable with $g'(\vartheta) = 0$ and $g''(\vartheta) \neq 0$, then $r_n^2[g(X_n) - g(\vartheta)] \overset{d}{\to} \frac{1}{2}g''(\vartheta)Y^2$.

Theorem 3.15. (Weak Law of Large Numbers) If $(X_n)_{n\geqslant 1}$ is a sequence of iid random variables with $\mathbb{E}(X_1) = \mu \in \mathbb{R}$, then $\overline{X}_n \overset{p}{\to} \mu$.

Theorem 3.16. (Strong Law of Large Numbers) If $(X_n)_{n\geqslant 1}$ is a sequence of iid random variables with $\mathbb{E}(X_1) = \mu \in \mathbb{R}$, then $\overline{X}_n \overset{\text{a.s.}}{\to} \mu$.

Corollary 3.10. The statistic $T_n(X) = \overline{X}_n$ is a strongly consistent estimator of the parametric function $g(\vartheta) = \mathbb{E}_\vartheta(X_1)$.

Proof. The result follows directly from the strong law of large numbers. □

Theorem 3.17. (Central Limit Theorem) If $(X_n)_{n\geqslant 1}$ is a sequence of iid random variables with $\mathbb{E}(X_1) = \mu \in \mathbb{R}$ and $\mathrm{Var}(X_1) = \sigma^2 \in (0,\infty)$, then it follows that $\sqrt{n}\left(\overline{X}_n - \mu\right) \overset{d}{\to} Y \sim \mathcal{N}\left(0, \sigma^2\right)$.

Lemma 3.2. If the random variable X has continuous CDF $F(x;\vartheta)$, then it holds that $U = F(X;\vartheta) \sim \mathcal{U}(0,1)$.

Proof. Since the CDF $F(x;\vartheta)$ is continuous, we deduce that it's also invertible. For $u \in (0,1)$, we calculate that:

$$F_U(u) = \mathbb{P}[F(X;\vartheta) \leqslant u] = \mathbb{P}\left[X \leqslant F^{-1}(u;\vartheta)\right] = F\left(F^{-1}(u;\vartheta);\vartheta\right) = u.$$

□

Theorem 3.18. If the random variables $U_1, \ldots, U_n \sim \mathcal{U}(0,1)$ are iid, then it follows that $nU_{(1)} \overset{d}{\to} Y$ and $n\left[1 - U_{(n)}\right] \overset{d}{\to} V$, where $Y, V \sim \mathrm{Exp}(1)$ are independent random variables.

Corollary 3.11. If the random sample $X_1, \ldots, X_n$ has continuous CDF $F(x;\vartheta)$, then it follows that $nF\left[X_{(1)};\vartheta\right] \overset{d}{\to} Y$ and $n\left[1 - F\left(X_{(n)};\vartheta\right)\right] \overset{d}{\to} V$, where $Y, V \sim \mathrm{Exp}(1)$ are independent random variables.

Proof. According to the previous lemma, we infer that $U_i = F(X_i; \vartheta) \sim \mathcal{U}(0,1)$ for $i = 1, 2, \ldots, n$, which implies that $U_{(1)} = F\left[X_{(1)}; \vartheta\right]$ and $U_{(n)} = F\left[X_{(n)}; \vartheta\right]$. Hence, the desired result follows from direct application of the previous theorem. □

Lemma 3.3. Let $U \sim \mathcal{U}(0,1)$. Then, the random variable $X = F^{-1}(U; \vartheta)$ has continuous CDF $F(x; \vartheta)$.

Proof. We know that $F_U(u) = \mathbb{P}(U \leqslant u) = u$ for $u \in (0,1)$. For $x \in S$, we calculate that:

$$\begin{aligned} F_X(x) &= \mathbb{P}_{\underline{\vartheta}}(X \leqslant x) = \mathbb{P}_{\underline{\vartheta}}\left[F^{-1}(U; \vartheta) \leqslant x\right] \\ &= \mathbb{P}_{\underline{\vartheta}}\left[F(x; \vartheta) \geqslant U\right] = F_U\left(F(x; \vartheta); \vartheta\right) = F(x; \vartheta). \end{aligned}$$

□

Theorem 3.19.* If the random variables $U_1, \ldots, U_n \sim \mathcal{U}(0,1)$ are iid, then it follows that:

$$\sqrt{n}\left[\text{median}(U) - \frac{1}{2}\right] \xrightarrow{d} Y \sim \mathcal{N}(0, 1/4).$$

Corollary 3.12.* If the random sample $X_1, \ldots, X_n$ has PDF $f(x; \vartheta)$, CDF $F(x; \vartheta)$ and $m = F^{-1}(1/2; \vartheta)$ is the theoretical median of the distribution, then it follows that:

$$\sqrt{n}\left[\text{median}(X) - m\right] \xrightarrow{d} V \sim \mathcal{N}\left(0, \frac{1}{4f^2(m; \vartheta)}\right).$$

Proof. Let $g(x) = F^{-1}(x; \vartheta)$. Since the CDF $F(x; \vartheta)$ is continuous, we infer that g is continuously differentiable. First, we notice that $g(1/2) = m$. Then, we calculate that:

$$\begin{aligned} x = F(g(x); \vartheta) \quad &\Rightarrow \quad g'(x) F'(g(x); \vartheta)) = 1 \quad \Rightarrow \\ g'(x) = \frac{1}{f(g(x); \vartheta))} \quad &\Rightarrow \quad g'(1/2) = \frac{1}{f(m; \vartheta)} \neq 0. \end{aligned}$$

According to the previous lemma, it follows that the random variables $X_1, X_2, \ldots, X_n$ and $g(U_1), g(U_2), \ldots, g(U_n)$ have the same distribution, which implies that the random variables median(X) and $g(\text{median}(U))$ also have the same distribution. By applying the delta method on the asymptotic distribution of the previous theorem, we conclude that:

$$\begin{aligned} \sqrt{n}\left[\text{median}(X) - m\right] &\overset{d}{=} \sqrt{n}\left[g(\text{median}(U)) - g(1/2)\right] \\ &\xrightarrow{d} g'(1/2)\, Y = \frac{1}{f(m; \vartheta)} Y \sim \mathcal{N}\left(0, \frac{1}{4f^2(m; \vartheta)}\right). \end{aligned}$$

□

Proposition 3.18*. If $X_n = (X_{n1}, \ldots, X_{ns})$ and $X = (X_1, \ldots, X_s)$, then it follows that $X_n \overset{\text{a.s.}/p}{\to} X \Leftrightarrow X_{nj} \overset{\text{a.s.}/p}{\to} X_j$ for $j = 1, 2, \ldots, s$.

Theorem 3.20*. (Cramér - Wold) If $X_n = (X_{n1}, \ldots, X_{ns})$ and $X = (X_1, \ldots, X_s)$, then it holds that $X_n \xrightarrow{d} X \Leftrightarrow \sum_{j=1}^{s} c_j X_{nj} \xrightarrow{d} \sum_{j=1}^{s} c_j X_j \ \forall c = (c_1, \ldots, c_s) \in \mathbb{R}^s$.

Definition 3.21*. A random vector $X \in \mathbb{R}^s$ follows the (non-degenerate) *multivariate normal* distribution with mean vector $\mu \in \mathbb{R}^s$ and positive definite covariance matrix $\Sigma \in \mathbb{R}^{s \times s}$, i.e. $X \sim \mathcal{N}_s(\mu, \Sigma)$, if it has the following PDF:

$$f_X(x; \mu, \Sigma) = (2\pi)^{-s/2} |\Sigma|^{-1/2} \exp\left\{-\frac{1}{2}(x-\mu)^{\mathrm{T}} \Sigma^{-1} (x-\mu)\right\}, \quad x \in \mathbb{R}^s.$$

Proposition 3.19*. If $X \sim \mathcal{N}_s(\mu, \Sigma)$, $A \in \mathbb{R}^{d \times s}$ and $c \in \mathbb{R}^d$, then it follows that $AX + c \sim \mathcal{N}_d\left(A\mu + c, A\Sigma A^{\mathrm{T}}\right)$.

Theorem 3.21*. (Multivariate Central Limit Theorem) If $(X_n)_{n \geqslant 1}$ is a sequence of iid random vectors with mean vector $\mathbb{E}(X_1) = \mu \in \mathbb{R}^s$ and positive definite covariance matrix $\mathrm{Var}(X_1) = \Sigma \in \mathbb{R}^{s \times s}$, then it holds that $\sqrt{n}\left(\overline{X}_n - \mu\right) \xrightarrow{d} Y \sim \mathcal{N}_s(0, \Sigma)$.

Theorem 3.22*. (Multivariate Delta Method) Suppose that $r_n(X_n - \vartheta) \xrightarrow{d} Y \in \mathbb{R}^s$, where $\lim_{n\to\infty} r_n = \infty$. If the function $g : \Theta \to \mathbb{R}^d$ is continuously differentiable with Jacobian matrix $\mathcal{J}_g \in \mathbb{R}^{d \times s}$ and the matrix $\mathcal{J}_g(\vartheta)\mathrm{Var}_\vartheta(Y)\mathcal{J}_g^{\mathrm{T}}(\vartheta)$ is positive definite, then it follows that $r_n[g(X_n) - g(\vartheta)] \xrightarrow{d} \mathcal{J}_g(\vartheta)Y$.

Theorem 3.23*. (Multivariate Second-Order Delta Method) Let $r_n(X_n - \vartheta) \xrightarrow{d} Y$, where $\lim_{n\to\infty} r_n = \infty$. If the function $g : \Theta \to \mathbb{R}$ is 2 times continuously differentiable with $\nabla_\vartheta^{\mathrm{T}} g(\vartheta)\mathrm{Var}_\vartheta(Y)\nabla_\vartheta g(\vartheta) = 0$ and Hessian matrix $\mathcal{H}_g(\vartheta) \in \mathbb{R}^{s \times s}$, then it follows that $r_n^2[g(X_n) - g(\vartheta)] \xrightarrow{d} \frac{1}{2} Y^{\mathrm{T}} \mathcal{H}_g(\vartheta) Y$.

Note 3.22. To sum up, there is a multitude of available methods to show that a statistic $T_n(X)$ is a (strongly) consistent estimator of a parametric function $g(\vartheta)$:

i. The definitions of almost sure convergence and convergence in probability;

ii. Showing that $T_n(X)$ is an (asymptotically) unbiased estimator of $g(\vartheta)$ with $\lim_{n\to\infty} \mathrm{Var}_\vartheta[T_n(X)] = 0$;

iii. Combining the laws of large numbers with the continuous mapping theorem and proposition 3.16;

iv. Showing that $T_n(X)$ has an asymptotic distribution via a combination of the definition of convergence in distribution, Slutsky's theorem, the continuous mapping theorem, the delta method, the central limit theorem and corollary 3.11.

Example 3.32. Let $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ be a random sample. We want to find (strongly) consistent estimators of σ^2 and $g(\mu, \sigma^2) = \frac{\mu}{\sigma}$. We also want to show that

$\frac{\overline{X}_n-\mu}{S_n/\sqrt{n}} \xrightarrow{d} Z \sim \mathcal{N}(0,1)$. According to note 3.12 (page 38), we know that the sample variance S_n^2 is an unbiased estimator of σ^2, and it holds that $\operatorname{Var}\left(S_n^2\right) = \frac{2\sigma^4}{n-1} \to 0$ as $n \to \infty$. According to proposition 3.17, it follows that S_n^2 is a consistent estimator of σ^2. Alternatively, we know that:

$$S_n^2 = \frac{1}{n-1}\left(\sum_{i=1}^n X_i^2 - n\overline{X}_n^2\right) = \frac{n}{n-1}\left(\frac{1}{n}\sum_{i=1}^n X_i^2 - \overline{X}_n^2\right).$$

According to the strong law of large numbers, we also know that:

$$\overline{X}_n \xrightarrow{\text{a.s.}} \mu, \quad \frac{1}{n}\sum_{i=1}^n X_i^2 \xrightarrow{\text{a.s.}} \mathbb{E}\left(X_1^2\right) = \operatorname{Var}(X_1) + [\mathbb{E}(X_1)]^2 = \sigma^2 + \mu^2.$$

According to proposition 3.16, we infer that $S_n^2 \xrightarrow{\text{a.s.}} 1 \cdot \left[\left(\sigma^2+\mu^2\right) - \mu^2\right] = \sigma^2$, i.e. S_n^2 is a strongly consistent estimator of σ^2. We also infer that $\frac{\overline{X}_n}{S_n}$ is a strongly consistent estimator of the parametric function $g(\mu,\sigma^2) = \frac{\mu}{\sigma}$. According to the central limit theorem, we know that $\sqrt{n}\left(\overline{X}_n - \mu\right) \xrightarrow{d} Y \sim \mathcal{N}\left(0,\sigma^2\right)$. According to Slutsky's theorem, we conclude that:

$$\frac{\overline{X}_n - \mu}{S_n/\sqrt{n}} \xrightarrow{d} \frac{1}{\sigma}Y = Z \sim \mathcal{N}(0,1). \quad \square$$

Example 3.33. Let $X_1,\dots,X_n \sim \operatorname{Exp}(\lambda)$ be a random sample. We want to examine whether the statistic $T_n(X) = \frac{1}{\overline{X}_n}$ is a (strongly) consistent estimator of λ and calculate its asymptotic distribution. We know that $\sum_{i=1}^n X_i \sim \operatorname{Gamma}(n,\lambda)$. According to example 3.24 (page 48), we also know that $\mathbb{E}\left[T_n(X)\right] = \frac{n\lambda}{n-1} \to \lambda$ as $n \to \infty$, i.e. $T_n(X)$ is an asymptotically unbiased estimator of λ. Additionally, we calculate that:

$$\begin{aligned}\mathbb{E}\left[T_n^2(X)\right] &= n^2\int_0^\infty \frac{1}{x^2}\frac{\lambda^n}{(n-1)!}x^{n-1}e^{-\lambda x}dx = \frac{n^2\lambda^n}{(n-1)!}\int_0^\infty x^{n-3}e^{-\lambda x}dx \\ &= \frac{n^2\lambda^n}{(n-1)!}\frac{(n-3)!}{\lambda^{n-2}} = \frac{n^2\lambda^2}{(n-1)(n-2)},\end{aligned}$$

$$\operatorname{Var}\left[T_n(X)\right] = \frac{n^2\lambda^2}{(n-1)(n-2)} - \frac{n^2\lambda^2}{(n-1)^2} = \frac{n^2\lambda^2}{(n-2)(n-1)^2} \xrightarrow{n\to\infty} 0.$$

According to proposition 3.17, the statistic $T_n(X)$ is a consistent estimator of λ. According to the strong law of large numbers, we know that $\overline{X}_n \xrightarrow{\text{a.s.}} \mathbb{E}(X_1) = \frac{1}{\lambda}$. Hence, we infer that $T_n(X)$ is a strongly consistent estimator of λ, according to proposition 3.16. Furthermore, we know that $\sqrt{n}\left(\overline{X}_n - \frac{1}{\lambda}\right) \xrightarrow{d} Y \sim \mathcal{N}\left(0,\frac{1}{\lambda^2}\right)$, according to the central limit theorem. Since the function $g(x) = \frac{1}{x}$ is continuously differentiable on $\Theta = (0,\infty)$ with $g'(1/\lambda) = -\lambda^2 \neq 0$, we conclude that:

$$\sqrt{n}\left[T_n(X) - \lambda\right] \xrightarrow{d} g'(1/\lambda)\,Y = -\lambda^2 Y \sim \mathcal{N}\left(0,\lambda^2\right),$$

according to the delta method. □

Example 3.34. Let $X_1, \ldots, X_n \sim \text{Pareto}(k, \lambda)$ be a random sample with $k > 0$, known $\lambda > 2$, PDF $f(x; k) = \frac{\lambda k^\lambda}{x^{\lambda+1}}$ and CDF $F(x; k) = 1 - \left(\frac{k}{x}\right)^\lambda$ for $x \geqslant k$. We want to examine whether the statistic $X_{(1)}$ is a consistent estimator of k and calculate its asymptotic distribution. First, we calculate that:

$$F_{X_{(1)}}(x) = 1 - \left(\frac{k}{x}\right)^{n\lambda}, \quad f_{X_{(1)}}(x) = \frac{n\lambda k^{n\lambda}}{x^{n\lambda+1}},$$

i.e. $X_{(1)} \sim \text{Pareto}(k, n\lambda)$. Then, we calculate that:

$$\mathbb{E}\left[X_{(1)}\right] = n\lambda k^{n\lambda} \int_k^\infty \frac{1}{x^{n\lambda}} dx = \frac{n\lambda k^{n\lambda}}{n\lambda - 1} \frac{1}{k^{n\lambda-1}} = \frac{n\lambda k}{n\lambda - 1} \overset{n\to\infty}{\to} k,$$

i.e. $X_{(1)}$ is an asymptotically unbiased estimator of k. Additionally, we calculate that:

$$\mathbb{E}\left[X_{(1)}^2\right] = n\lambda k^{n\lambda} \int_k^\infty \frac{1}{x^{n\lambda-1}} dx = \frac{n\lambda k^{n\lambda}}{n\lambda - 2} \frac{1}{k^{n\lambda-2}} = \frac{n\lambda k^2}{n\lambda - 2},$$

$$\text{Var}\left[X_{(1)}\right] = \frac{n\lambda k^2}{n\lambda - 2} - \frac{n^2\lambda^2 k^2}{(n\lambda - 1)^2} = \frac{n\lambda k^2}{(n\lambda - 1)^2(n\lambda - 2)} \overset{n\to\infty}{\to} 0.$$

According to proposition 3.17, the statistic $X_{(1)}$ is a consistent estimator of k. According to corollary 3.11, we know that:

$$nF\left[X_{(1)}; k\right] = n\left[1 - \frac{k^\lambda}{X_{(1)}^\lambda}\right] = -nk^\lambda \left[\frac{1}{X_{(1)}^\lambda} - \frac{1}{k^\lambda}\right] \overset{d}{\to} Y \sim \text{Exp}(1).$$

Since the function $g(x) = x^{-1/\lambda}$ is continuously differentiable on $\Theta = (0, \infty)$ and it holds that $g'\left(k^{-\lambda}\right) = -\frac{1}{\lambda} k^{\lambda+1} \neq 0$, it follows that:

$$n\left[X_{(1)} - k\right] \overset{d}{\to} -k^{-\lambda} g'\left(k^{-\lambda}\right) Y = \frac{k}{\lambda} Y = V \sim \text{Exp}\left(\lambda/k\right),$$

according to Slutsky's theorem in conjunction with the delta method. Alternatively, for $x \in (0, \infty)$, We calculate that:

$$\begin{aligned}\mathbb{P}\left[n\left(X_{(1)} - k\right) \leqslant x\right] &= F_{X_{(1)}}\left(\frac{x}{n} + k\right) = 1 - \left(\frac{k}{x/n + k}\right)^{n\lambda} \\ &= 1 - \left(1 + \frac{x/k}{n}\right)^{-n\lambda} \overset{n\to\infty}{\to} 1 - e^{-\lambda x/k},\end{aligned}$$

which is the CDF of $V \sim \text{Exp}\left(\lambda/k\right)$, so we conclude that $n\left[X_{(1)} - k\right] \overset{d}{\to} V$. □

Example 3.35. Let $X_1, \ldots, X_n \sim \text{Bernoulli}(p)$ be a random sample. We want to show that the statistic $T_n(X) = \min\left\{\overline{X}_n, 1 - \overline{X}_n\right\}$ is a strongly consistent estimator of $g(p) = \min\{p, 1 - p\}$ and calculate its asymptotic distribution. According to the strong law of large numbers, we know that $\overline{X}_n \xrightarrow{\text{a.s.}} \mathbb{E}(X_1) = p$. Since the function $g(p) = \min\{p, 1-p\}$ is continuous on $\Theta = (0,1)$, the statistic $T_n(X)$ is a strongly consistent estimator of $g(p)$, according to the continuous mapping theorem. Furthermore, we know that $\sqrt{n}\left(\overline{X}_n - p\right) \xrightarrow{d} Y \sim \mathcal{N}(0, p(1-p))$, according to the central limit theorem. Since the function $g(p) = \min\{p, 1-p\}$ is continuously differentiable for $p \neq \frac{1}{2}$ with $|g'(p)| = 1$, it follows that:

$$\sqrt{n}\left[T_n(X) - \min\{p, 1-p\}\right] \xrightarrow{d} g'(p)\, Y \sim \mathcal{N}(0, p(1-p)),$$

according to the delta method. For $p = \frac{1}{2}$, we observe that:

$$T_n(X) - \min\{p, 1-p\} = \min\left\{\overline{X}_n - \frac{1}{2}, \frac{1}{2} - \overline{X}_n\right\} = -\left|\overline{X}_n - \frac{1}{2}\right|.$$

Since the function $g(x) = -|x|$ is continuous, it follows that:

$$\sqrt{n}\left[T_n(X) - \min\{p, 1-p\}\right] = -\sqrt{n}\left|\overline{X}_n - \frac{1}{2}\right| \xrightarrow{d} -|Y|,$$

where $Y \sim \mathcal{N}\left(0, \frac{1}{4}\right)$, according to the continuous mapping theorem. □

3.11 Maximum Likelihood Estimation

Definition 3.22. The joint PMF or PDF of the random variables $X_1, \ldots, X_n$ regarded as a function of ϑ is called the *likelihood function* of the sample X for ϑ and is denoted by $\mathcal{L}(\vartheta \mid x) = f(x; \vartheta)$.

Note 3.23. If $X_1, \ldots, X_n$ are independent, then $\mathcal{L}(\vartheta \mid x) = \prod_{i=1}^n f(x_i; \vartheta)$.

Interpretation: The likelihood function expresses how plausible it is to have observed the sample x as a function of the parameter ϑ. Therefore, a "reasonable" estimator of ϑ results from maximizing the likelihood function with respect to ϑ. In this manner, we estimate the unknown parameter by the value of ϑ for which it is most likely to have observed the sample.

Definition 3.23. The statistic $\widehat{\vartheta}(X)$ for which the likelihood function is maximized, i.e. $\widehat{\vartheta}(X) = \arg\max_{\vartheta \in \Theta} \mathcal{L}(\vartheta \mid X)$, is called the *maximum likelihood estimator* (MLE) of ϑ.

Note 3.24. i. For an unknown parameter ϑ there may not exist any MLE, there may exist a unique MLE, or there may exist multiple MLEs, i.e. the likelihood function might have multiple global maxima.

ii. Since the function $g(x) = \log x$ is strictly increasing on $(0, \infty)$, we infer that the maxima of the *log-likelihood function* $\ell(\vartheta \mid x) = \log \mathcal{L}(\vartheta \mid x)$ coincide with the maxima of the likelihood function. For reasons of computational ease and numerical stability (the products turn into sums), maximizing the log-likelihood function is usually preferred.

iii. If the log-likelihood function $\ell(\vartheta \mid x)$ is partially differentiable on an open set $\Theta_0 \subseteq \Theta$, then candidate global maxima of $\ell(\vartheta \mid x)$ are given by solving the system of equations $\frac{\partial \ell(\vartheta|x)}{\partial \vartheta_j} = 0$ for $j = 1, 2, \ldots, s$.

Proposition 3.20. If the statistic $T(X)$ is sufficient for ϑ and $\widehat{\vartheta}(X)$ is the unique MLE of ϑ, then it holds that $\widehat{\vartheta}(X) = \psi(T)$ for some function ψ.

Proof. According to the Fisher - Neyman factorization criterion, we know that:

$$\mathcal{L}(\vartheta \mid x) = f(x; \vartheta) = g(T(x); \vartheta)\, h(x).$$

Since the function $h(x)$ does not depend on the value of ϑ, we deduce that the unique MLE $\widehat{\vartheta}(X)$ of ϑ is the value of ϑ which maximizes the function $g(T(x); \vartheta)$. Therefore, we conclude that the MLE $\widehat{\vartheta}(X)$ of ϑ depends on the sample X through the sufficient statistic $T(X)$ for ϑ. □

Proposition 3.21. Let X be a random sample from a distribution which satisfies the regularity conditions of the Cramér - Rao inequality. If the Fisher information $\mathcal{I}_X(\vartheta)$ is differentiable on Θ and the statistic $T(X)$ is an efficient estimator of ϑ, then $T(X)$ is also the MLE of ϑ.

Proof. According to proposition 3.8 (page 54) for the parametric function $g(\vartheta) = \vartheta$, we know that $\mathcal{S}_X(\vartheta) = \mathcal{I}_X(\vartheta)\,[T(X) - \vartheta]$ $\forall \vartheta \in \Theta$. Since the score function is the derivative of the log-likelihood function with respect to ϑ, setting it equal to 0 and solving the resulting equation should yield the MLE of ϑ. Since $\mathcal{I}_X(\vartheta) \in (0, \infty)$, it follows that $\widehat{\vartheta}(X) = T(X)$. Next, we calculate that:

$$\frac{\partial}{\partial \vartheta}\mathcal{S}_X(\vartheta) = \mathcal{I}_X'(\vartheta)\,[T(X) - \vartheta] - \mathcal{I}_X(\vartheta), \quad \frac{\partial}{\partial \vartheta}\mathcal{S}_X(\widehat{\vartheta}) = -\mathcal{I}_X(\widehat{\vartheta}) < 0.$$

Thus, the efficient estimator $T(X)$ of ϑ is also the MLE of ϑ. □

Note 3.25. The converse is generally not true, i.e. the MLE of ϑ isn't necessarily an efficient estimator of ϑ. In fact, there are cases where the MLE of ϑ is not even an unbiased estimator of ϑ.

Proposition 3.22. (Invariance Property) If the statistic $\widehat{\vartheta}(X)$ is the MLE of ϑ, then $g(\widehat{\vartheta})$ is the MLE of the parametric function $g(\vartheta)$, i.e. it holds that $\widehat{g(\vartheta)} = g(\widehat{\vartheta})$.

Proof. Assume that the function g is injective and let:

$$\mathcal{L}^*(\eta \mid x) = \mathcal{L}\left(g^{-1}(\eta) \mid x\right) = \mathcal{L}(\vartheta \mid x).$$

Then, we observe that:

$$\mathcal{L}^*\left(g(\widehat{\vartheta}) \mid x\right) = \mathcal{L}\left(g^{-1}(g(\widehat{\vartheta})) \mid x\right) = \mathcal{L}\left(\widehat{\vartheta} \mid x\right).$$

Therefore, we conclude that $\widehat{\eta} = \widehat{g(\vartheta)} = g(\widehat{\vartheta})$. The proof for the general case requires the notion of the induced likelihood function and can be found in Casella and Berger, Section 7.2. □

Example 3.36. Let $X_1, \ldots, X_n \sim \text{Exp}(\lambda)$ be a random sample. We calculate that:

$$\ell(\lambda \mid x) = \log \mathcal{L}(\lambda \mid x) = n \log \lambda - \lambda \sum_{i=1}^{n} x_i,$$

$$\frac{\partial \ell(\lambda \mid x)}{\partial \lambda} = \frac{n}{\lambda} - \sum_{i=1}^{n} x_i = 0 \quad \Rightarrow \quad \widehat{\lambda}(x) = \frac{1}{\overline{x}}, \quad \frac{\partial^2 \ell(\lambda \mid x)}{\partial \lambda^2} = -\frac{n}{\lambda^2} < 0, \quad \forall \lambda > 0,$$

i.e. the function $\ell(\lambda \mid x)$ is strictly concave on $\Theta = (0, \infty)$. Therefore, the statistic $\widehat{\lambda}(X) = \frac{1}{\overline{X}}$ is the MLE of λ. □

Example 3.37. Let $X_1, \ldots, X_n \sim \text{Bin}(N, p)$ be a random sample with known N. We calculate that:

$$\ell(p \mid x) = \sum_{i=1}^{n} \log \binom{N}{x_i} + \log p \sum_{i=1}^{n} x_i + \log(1-p)\left(nN - \sum_{i=1}^{n} x_i\right),$$

$$\frac{\partial \ell(p \mid x)}{\partial p} = \frac{1}{p} \sum_{i=1}^{n} x_i - \frac{1}{1-p}\left(nN - \sum_{i=1}^{n} x_i\right) = 0 \quad \Rightarrow$$

$$(1-\widehat{p}) \sum_{i=1}^{n} x_i = \widehat{p}\left(nN - \sum_{i=1}^{n} x_i\right) \quad \Rightarrow \quad \widehat{p}(x) = \frac{1}{nN} \sum_{i=1}^{n} x_i = \frac{1}{N}\overline{x},$$

$$\frac{\partial^2 \ell(p \mid x)}{\partial p^2} = -\frac{1}{p^2} \sum_{i=1}^{n} x_i - \frac{1}{(1-p)^2}\left(nN - \sum_{i=1}^{n} x_i\right) < 0, \quad \forall p \in (0,1),$$

i.e. the function $\ell(p \mid x)$ is strictly concave on $\Theta = (0, 1)$. Therefore, the statistic $\widehat{p}(X) = \frac{1}{N}\overline{X}$ is the MLE of p. If $x_1 = \cdots = x_n = 0$, we infer that $\mathcal{L}(p \mid x) = (1-p)^{nN}$, i.e. the likelihood function is strictly decreasing on $\Theta = (0, 1)$, so the MLE of p doesn't exist. If $x_1 = \cdots = x_n = N$, we observe that $\mathcal{L}(p \mid x) = p^{nN}$, i.e. the likelihood function is strictly increasing on $\Theta = (0, 1)$, so the MLE of p doesn't exist. □

Example 3.38. Let $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \vartheta)$ be a random sample with known μ. We calculate that:

$$\ell(\vartheta \mid x) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\vartheta - \frac{1}{2\vartheta}\sum_{i=1}^{n}(x_i-\mu)^2,$$

$$\frac{\partial \ell(\vartheta \mid x)}{\partial \vartheta} = -\frac{n}{2\vartheta} + \frac{1}{2\vartheta^2}\sum_{i=1}^{n}(x_i-\mu)^2 = 0 \quad\Rightarrow\quad \widehat{\vartheta}(x) = \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2,$$

$$\frac{\partial^2 \ell(\vartheta \mid x)}{\partial \vartheta^2} = \frac{n}{2\vartheta^2} - \frac{1}{\vartheta^3}\sum_{i=1}^{n}(x_i-\mu)^2 = -\frac{n}{2\vartheta^2}\left[\frac{2}{n\vartheta}\sum_{i=1}^{n}(x_i-\mu)^2 - 1\right],$$

$$\frac{\partial^2 \ell(\widehat{\vartheta} \mid x)}{\partial \vartheta^2} = -\frac{n}{2\widehat{\vartheta}^2} < 0,$$

i.e. the function $\ell(\vartheta \mid x)$ has a maximum at $\widehat{\vartheta}$. Next, we calculate that:

$$\lim_{\vartheta\to\infty}\mathcal{L}(\vartheta \mid x) = \lim_{\vartheta\to\infty}(2\pi\vartheta)^{-n/2}\exp\left\{-\frac{1}{2\vartheta}\sum_{i=1}^{n}(x_i-\mu)^2\right\} = 0,$$

$$\lim_{\vartheta\to 0^+}\mathcal{L}(\vartheta \mid x) = \lim_{\vartheta\to 0^+}(2\pi\vartheta)^{-n/2}\exp\left\{-\frac{1}{2\vartheta}\sum_{i=1}^{n}(x_i-\mu)^2\right\} = 0.$$

Therefore, the statistic $\widehat{\vartheta}(X) = \frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2$ is the MLE of ϑ. In the case where $x_1 = \cdots = x_n = \mu$, we observe that $\mathcal{L}(\vartheta \mid x) = (2\pi\vartheta)^{-n/2}$, i.e. the likelihood function is strictly decreasing on $\Theta = (0,\infty)$ and doesn't have any maxima. However, it holds that $\mathbb{P}(X_1 = \cdots = X_n = \mu) = 0$, so the MLE of ϑ exists with probability 1. □

Example 3.39. Let $X_1,\ldots,X_n \sim \mathcal{U}(0,\vartheta)$ be a random sample. We calculate that:

$$\mathcal{L}(\vartheta \mid x) = \frac{1}{\vartheta^n}\prod_{i=1}^{n}\mathbb{1}_{[0,\vartheta]}(x_i) = \frac{1}{\vartheta^n}\mathbb{1}_{[0,\vartheta]}\left(x_{(n)}\right) = \begin{cases}\vartheta^{-n}, & \vartheta \geqslant x_{(n)}\\ 0, & \vartheta < x_{(n)}\end{cases}.$$

For $\vartheta \geqslant x_{(n)}$, the likelihood function is strictly decreasing, so it has a unique global maximum $\widehat{\vartheta}(X) = X_{(n)}$. □

Example 3.40. Let $X_1,\ldots,X_n \sim \mathcal{U}(2\vartheta,3\vartheta)$ be a random sample with $\vartheta > 0$. We calculate that:

$$\begin{aligned}\mathcal{L}(\vartheta \mid x) &= \frac{1}{\vartheta^n}\mathbb{1}_{[2\vartheta,3\vartheta]}\left(x_{(1)}\right)\mathbb{1}_{[2\vartheta,3\vartheta]}\left(x_{(n)}\right) = \frac{1}{\vartheta^n}\mathbb{1}_{[2\vartheta,\infty)}\left(x_{(1)}\right)\mathbb{1}_{(-\infty,3\vartheta]}\left(x_{(n)}\right)\\ &= \begin{cases}\vartheta^{-n}, & 2\vartheta \leqslant x_{(1)} \text{ and } 3\vartheta \geqslant x_{(n)}\\ 0, & \text{otherwise}\end{cases} = \begin{cases}\vartheta^{-n}, & \frac{1}{3}x_{(n)} \leqslant \vartheta \leqslant \frac{1}{2}x_{(1)}\\ 0, & \text{otherwise}\end{cases}.\end{aligned}$$

For $\vartheta \in \left[\frac{1}{3}x_{(n)}, \frac{1}{2}x_{(1)}\right]$, the likelihood function is strictly decreasing, so it has a unique global maximum $\widehat{\vartheta}(X) = \frac{1}{3}X_{(n)}$. □

Example 3.41. Let $X_1, \ldots, X_n \sim \mathcal{U}(\vartheta, \vartheta+1)$ be a random sample with $\vartheta \in \mathbb{R}$. We calculate that:

$$\begin{aligned}\mathcal{L}(\vartheta \mid x) &= \mathbb{1}_{[\vartheta,\vartheta+1]}\left(x_{(1)}\right)\mathbb{1}_{[\vartheta,\vartheta+1]}\left(x_{(n)}\right) = \mathbb{1}_{[\vartheta,\infty)}\left(x_{(1)}\right)\mathbb{1}_{(-\infty,\vartheta+1]}\left(x_{(n)}\right) \\ &= \begin{cases} 1, & \vartheta \leqslant x_{(1)} \text{ and } \vartheta \geqslant x_{(n)} - 1 \\ 0, & \text{otherwise} \end{cases} = \begin{cases} 1, & x_{(n)} - 1 \leqslant \vartheta \leqslant x_{(1)} \\ 0, & \text{otherwise} \end{cases}.\end{aligned}$$

For $\vartheta \in \left[x_{(n)} - 1, x_{(1)}\right]$, the likelihood function is constant, so it has infinitely many global maxima $\widehat{\vartheta}(X) \in \left[X_{(n)} - 1, X_{(1)}\right]$. □

Example 3.42. Let $X_1, \ldots, X_n \sim \mathcal{U}(\vartheta_1, \vartheta_2)$ be a random sample. We calculate that:

$$\begin{aligned}\mathcal{L}(\vartheta_1, \vartheta_2 \mid x) &= \frac{1}{(\vartheta_2 - \vartheta_1)^n}\mathbb{1}_{[\vartheta_1,\infty)}\left(x_{(1)}\right)\mathbb{1}_{(-\infty,\vartheta_2]}\left(x_{(n)}\right) \\ &= \begin{cases} (\vartheta_2 - \vartheta_1)^{-n}, & \vartheta_1 \leqslant x_{(1)} \text{ and } \vartheta_2 \geqslant x_{(n)} \\ 0, & \text{otherwise} \end{cases}.\end{aligned}$$

For $(\vartheta_1, \vartheta_2) \in \left(-\infty, x_{(1)}\right] \times \left[x_{(n)}, \infty\right)$, the likelihood function is strictly increasing with respect to ϑ_1 and strictly decreasing with respect to ϑ_2, so it has a unique global maximum $\widehat{\vartheta}(X) = \left(X_{(1)}, X_{(n)}\right)$. □

Note 3.26. In the case of a parameter vector $\vartheta \in \mathbb{R}^2$, we may endeavor to perform successive maximization of the likelihood function with respect to each unknown parameter separately, that is:

$$\max_{(\vartheta_1,\vartheta_2)\in\Theta} \mathcal{L}(\vartheta_1, \vartheta_2 \mid x) = \max_{\vartheta_2\in\Theta_2}\left\{\max_{\vartheta_1\in\Theta_1} \mathcal{L}(\vartheta_1, \vartheta_2 \mid x)\right\}.$$

This method will only lead to the solution of the joint maximization problem if the maximization with respect to ϑ_1 leads to a global maximum which doesn't depend on the value of ϑ_2.

Example 3.43. Let $X_1, \ldots, X_n$ be a random sample with $f(x; \lambda, k) = \lambda e^{-\lambda(x-k)}$ for $\lambda > 0$, $k \in \mathbb{R}$ and $x \geqslant k$. We calculate that:

$$\mathcal{L}(\lambda, k \mid x) = \lambda^n \exp\left\{-\lambda\sum_{i=1}^{n} x_i + n\lambda k\right\}\mathbb{1}_{[k,\infty)}\left(x_{(1)}\right).$$

First, we fix λ and maximize with respect to k. For $k \leqslant x_{(1)}$, the likelihood function is strictly increasing with respect to k, so it has a unique global maximum $\widehat{k}(X) = X_{(1)}$. Now, we maximize $\ell\left(\lambda, x_{(1)} \mid x\right)$ with respect to λ. We calculate that:

$$\ell\left(\lambda, x_{(1)} \mid x\right) = n\log\lambda - \lambda\sum_{i=1}^{n} x_i + n\lambda x_{(1)},$$

$$\frac{\partial \ell\left(\lambda, x_{(1)} \mid x\right)}{\partial \lambda} = \frac{n}{\lambda} - \sum_{i=1}^{n} x_i + n x_{(1)} = 0 \quad \Rightarrow \quad \widehat{\lambda}(x) = \frac{1}{\overline{x} - x_{(1)}},$$

$$\frac{\partial^2 \ell\left(\lambda, x_{(1)} \mid x\right)}{\partial \lambda^2} = -\frac{n}{\lambda^2} < 0, \quad \forall \lambda > 0,$$

i.e. the function $\ell\left(\lambda, x_{(1)} \mid x\right)$ is strictly concave on $(0,\infty)$. Therefore, the statistic $\widehat{\vartheta}(X) = \left(\frac{1}{\overline{X} - X_{(1)}}, X_{(1)}\right)$ is the MLE of $\vartheta = (\lambda, k)$. If $x_1 = \cdots = x_n$, we observe that $\mathcal{L}\left(\lambda, x_{(1)} \mid x\right) = \lambda^n$, i.e. the likelihood function is strictly increasing on $(0,\infty)$ and it doesn't have any maxima. However, it holds that $\mathbb{P}(X_1 = \cdots = X_n) = 0$, so the MLE of λ exists and is unique with probability 1. □

Example 3.44. Let $X_1, \ldots, X_n \sim \mathcal{N}(\vartheta_1, \vartheta_2)$ be a random sample. We want to calculate the MLE of the parametric function $g(\vartheta) = \frac{\vartheta_1}{\sqrt{\vartheta_2}}$ and compare the MSE of the MLE of ϑ_2 against the MSE of the UMVUE of ϑ_2. We calculate that:

$$\ell(\vartheta_1, \vartheta_2 \mid x) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\vartheta_2 - \frac{1}{2\vartheta_2}\sum_{i=1}^{n}(x_i - \vartheta_1)^2.$$

First, we fix ϑ_2 and maximize with respect to ϑ_1:

$$\frac{\partial \ell(\vartheta_1, \vartheta_2 \mid x)}{\partial \vartheta_1} = \frac{1}{\vartheta_2}\sum_{i=1}^{n}(x_i - \vartheta_1) = 0 \quad \Rightarrow \quad \widehat{\vartheta}_1(x) = \overline{x},$$

$$\frac{\partial^2 \ell(\vartheta_1, \vartheta_2 \mid x)}{\partial \vartheta_1^2} = -\frac{n}{\vartheta_2} < 0, \quad \forall \vartheta_1 \in \mathbb{R},$$

i.e. the function $\ell(\vartheta_1, \vartheta_2 \mid x)$ is strictly concave for fixed ϑ_2 and has a unique global maximum $\widehat{\vartheta}_1$. Now, we maximize $\ell(\overline{x}, \vartheta_2 \mid x)$ with respect to ϑ_2. We calculate that:

$$\frac{\partial \ell(\overline{x}, \vartheta_2 \mid x)}{\partial \vartheta_2} = -\frac{n}{2\vartheta_2} + \frac{1}{2\vartheta_2^2}\sum_{i=1}^{n}(x_i - \overline{x})^2 = 0 \quad \Rightarrow \quad \widehat{\vartheta}_2(x) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2,$$

$$\frac{\partial^2 \ell(\overline{x}, \vartheta_2 \mid x)}{\partial \vartheta_2^2} = \frac{n}{2\vartheta_2^2} - \frac{1}{\vartheta_2^3}\sum_{i=1}^{n}(x_i - \overline{x})^2 = -\frac{n}{2\vartheta_2^2}\left[\frac{2}{n\vartheta_2}\sum_{i=1}^{n}(x_i - \overline{x})^2 - 1\right],$$

$$\frac{\partial^2 \ell(\widehat{\vartheta}_1, \widehat{\vartheta}_2 \mid x)}{\partial \vartheta_2^2} = -\frac{n}{2\widehat{\vartheta}_2^2} < 0,$$

i.e. the function $\ell(\overline{x}, \vartheta_2 \mid x)$ has a maximum at $\widehat{\vartheta}_2$. Additionally, we calculate that:

$$\lim_{\vartheta_2 \to \infty} \mathcal{L}\left(\widehat{\vartheta}_1, \vartheta_2 \mid x\right) = \lim_{\vartheta_2 \to \infty} (2\pi\vartheta_2)^{-n/2} \exp\left\{-\frac{1}{2\vartheta_2}\sum_{i=1}^{n}(x_i - \overline{x})^2\right\} = 0,$$

$$\lim_{\vartheta_2 \to 0^+} \mathcal{L}\left(\widehat{\vartheta}_1, \vartheta_2 \mid x\right) = \lim_{\vartheta_2 \to 0^+} (2\pi\vartheta_2)^{-n/2} \exp\left\{-\frac{1}{2\vartheta_2}\sum_{i=1}^{n}(x_i - \overline{x})^2\right\} = 0.$$

Therefore, the statistic $\widehat{\vartheta}(X) = \left(\overline{X}, \frac{n-1}{n}S^2\right)$ is the MLE of $\vartheta = (\vartheta_1, \vartheta_2)$. According to the invariance property, the statistic $g(\widehat{\vartheta}) = \sqrt{\frac{n}{n-1}}\frac{\overline{X}}{S}$ is the MLE of $g(\vartheta) = \frac{\vartheta_1}{\sqrt{\vartheta_2}}$. To understand exactly how important this property of the MLE is, it suffices to consider how arduous the procedure to calculate the UMVUE of $g(\vartheta)$ would be. According to example 3.26 (page 49), the sample variance S^2 is the UMVUE of ϑ_2. According to note 3.12 (page 38), we know that $\mathrm{MSE}_{\vartheta_2}\left(S^2\right) = \mathrm{Var}\left(S^2\right) = \frac{2}{n-1}\vartheta_2^2$. Additionally, we calculate that:

$$\mathbb{E}\left(\widehat{\vartheta}_2\right) = \mathbb{E}\left(\frac{n-1}{n}S^2\right) = \frac{n-1}{n}\vartheta_2, \quad \mathrm{bias}_{\vartheta_2}\left(\widehat{\vartheta}_2\right) = \mathbb{E}\left(\widehat{\vartheta}_2\right) - \vartheta_2 = -\frac{1}{n}\vartheta_2,$$

$$\mathrm{Var}\left(\widehat{\vartheta}_2\right) = \mathrm{Var}\left(\frac{n-1}{n}S^2\right) = \frac{(n-1)^2}{n^2}\frac{2}{n-1}\vartheta_2^2 = \frac{2(n-1)}{n^2}\vartheta_2^2,$$

$$\mathrm{MSE}_{\vartheta_2}\left(\widehat{\vartheta}_2\right) = \mathrm{Var}\left(\widehat{\vartheta}_2\right) + \mathrm{bias}^2_{\vartheta_2}\left(\widehat{\vartheta}_2\right) = \frac{2n-1}{n^2}\vartheta_2^2.$$

We compare the MSEs of the 2 estimators as follows:

$$\mathrm{MSE}_{\vartheta_2}\left(\widehat{\vartheta}_2\right) < \mathrm{MSE}_{\vartheta_2}\left(S^2\right) \quad \Leftrightarrow \quad \frac{2n-1}{n^2} < \frac{2}{n-1} \quad \Leftrightarrow \quad -3n+1 < 0.$$

Therefore, the biased MLE of ϑ_2 has a smaller MSE than the UMVUE of ϑ_2. □

Theorem 3.24.* Let X be a random sample with joint PMF or PDF $f(x;\vartheta)$ for $\vartheta \in \Theta \subseteq \mathbb{R}$ and $x \in S$. Suppose that the following regularity conditions are satisfied:

I. The parameter space Θ is an open subset of $\mathbb{R}$.

II. The support $S = \{x \in \mathbb{R}^n : f(x;\vartheta) > 0\}$ doesn't depend on the value of ϑ.

III. $\frac{\partial}{\partial\vartheta}f(x;\vartheta) < \infty \ \forall x \in S$ and $\forall \vartheta \in \Theta$.

IV. The likelihood function $\mathcal{L}(\vartheta \mid X)$ has a unique global maximum $\widehat{\vartheta}_n(X) \ \forall n \in \mathbb{N}$.

V. The parameter ϑ is *identifiable*, i.e. the likelihood function is injective with respect to ϑ.

Then, the MLE $\widehat{\vartheta}_n(X)$ of ϑ is a consistent estimator of ϑ.

Theorem 3.25.* Suppose that the following additional regularity conditions also hold:

VI. $\frac{\partial^3}{\partial\vartheta^3}f(x;\vartheta) < \infty \ \forall x \in S$ and $\forall \vartheta \in \Theta$.

VII. $\int_S \frac{\partial^3}{\partial\vartheta^3}f(x;\vartheta)dx = \frac{\partial^3}{\partial\vartheta^3}\int_S f(x;\vartheta)dx = 0 \ \forall \vartheta \in \Theta$.

VIII. $\mathcal{I}_X(\vartheta) \in (0,\infty) \ \forall \vartheta \in \Theta$.

IX. For all $\vartheta \in \Theta$, there exist $\delta_\vartheta > 0$ and a function $M(x,\vartheta)$ with $\mathbb{E}_\vartheta\left[M(X,\vartheta)\right] < \infty$

such that:

$$\left| \frac{\partial^3}{\partial \vartheta_*^3} \log f(x; \vartheta_*) \right| \leqslant M(x, \vartheta), \quad \forall x \in S, \quad \forall \vartheta_* \in [\vartheta - \delta_\vartheta, \vartheta + \delta_\vartheta].$$

Then, it holds that $\sqrt{n}\left(\widehat{\vartheta}_n - \vartheta\right) \xrightarrow{d} Y \sim \mathcal{N}\left(0, \mathcal{I}_{X_1}^{-1}(\vartheta)\right)$, i.e. the MLE $\widehat{\vartheta}_n(X)$ of ϑ is an asymptotically efficient estimator of ϑ.

Note 3.27.* Suppose that the distribution of X belongs to the one-parameter multivariate exponential family with $f(x; \vartheta) = h(x)e^{Q(\vartheta)T(x) - A(\vartheta)}$. If the parameter space Θ is an open subset of $\mathbb{R}$ and the function $Q : \Theta \to \mathbb{R}$ is continuously differentiable with $Q'(\vartheta) \neq 0 \ \forall \vartheta \in \Theta$, then the regularity conditions I-III and VI-VIII are satisfied, so it remains to check the validity of the regularity conditions IV, V and IX.

Note 3.28. We observe that the MLE of a parameter has many "good" properties under certain conditions, especially for large samples. Indicatively, we mention that it's asymptotically unbiased, asymptotically efficient, consistent, a function of the sufficient statistic and possesses the invariance property, in contrast with the UMVUE of the unknown parameter. In table 3.4 we summarize the MLEs of the parameters of some widely used distributions.

Bernoulli(p)	$\overline{X}$
Poisson(λ)	
Bin(N, p) with known N	$\overline{X}/N$
Exp(λ)	$1/\overline{X}$
Gamma(k, λ) with known k	$k/\overline{X}$
Beta($\vartheta, 1$)	$-n/\sum_{i=1}^n \log X_i$
Beta($1, \vartheta$)	$-n/\sum_{i=1}^n \log(1 - X_i)$
$\mathcal{N}\left(\mu, \sigma^2\right)$ with known μ	$\sum_{i=1}^n (X_i - \mu)^2/n$
$\mathcal{N}\left(\mu, \sigma^2\right)$	$\left(\overline{X}, (n-1)S^2/n\right)$
$\mathcal{U}(\vartheta_1, \vartheta_2)$	$\left(X_{(1)}, X_{(n)}\right)$

TABLE 3.4: Notable Maximum Likelihood Estimators

3.12 Method of Moments Estimators

Definition 3.24. Let X be a sample from a distribution with unknown parameter ϑ. For $k = 1, 2, \ldots$, we define the following quantities:

i. *Theoretical (raw) moment* of order k: $\mu_k = \mathbb{E}_\vartheta(X_1^k)$.

ii. *Sample (raw) moment* of order k: $M_k = \frac{1}{n}\sum_{i=1}^n X_i^k$.

Additionally, for $k = 2, 3, \ldots$ we define:

iii. *Theoretical central moment* of order k: $\mu_k^* = \mathbb{E}_\vartheta\left[(X_1 - \mu_1)^k\right]$.

iv. *Sample central moment* of order k: $M_k^* = \frac{1}{n}\sum_{i=1}^n (X_i - \mu_1)^k$.

Method of Moments: According to the strong law of large numbers, we know that:

$$M_k = \frac{1}{n}\sum_{i=1}^n X_i^k \overset{\text{a.s.}}{\to} \mathbb{E}_\vartheta\left(X_1^k\right) = \mu_k,$$

$$M_k^* = \frac{1}{n}\sum_{i=1}^n (X_i - \mu_1)^k \overset{\text{a.s.}}{\to} \mathbb{E}_\vartheta\left[(X_1 - \mu_1)^k\right] = \mu_k^*.$$

Considering any of the equations $M_k = \mu_k$ and $M_k^* = \mu_k^*$, in order to obtain a system of a total of s equations which we can solve for the unknown parameter $\vartheta \in \Theta \subseteq \mathbb{R}^s$, we end up with a *method of moments estimator* (MOME) $\widetilde{\vartheta}(X)$ of ϑ. The MOME is obviously not unique, since it depends on the choice of the system of equations.

Note 3.29. We construct a system of equations starting from the lower order moments, which are theoretically easier to compute. We usually work with the central moments, since it holds that $\mu_2^* = \operatorname{Var}\vartheta(X_1)$, which is more readily known than $\mu_2 = \mathbb{E}_\vartheta\left(X_1^2\right)$. If the theoretical first order moment μ_1 in M_k^* isn't known, then it's substituted by the corresponding sample moment $M_1 = \overline{X}$. If the moments μ_k and μ_k^* don't depend on the value of ϑ for some k, then we skip the corresponding equations and move on to the higher order moments.

Example 3.45. Let $X_1, \ldots, X_n \sim \operatorname{Exp}(\lambda)$ be a random sample. We equate the first order moments:

$$\mu_1 = M_1 \quad \Rightarrow \quad \frac{1}{\lambda} = \overline{X} \quad \Rightarrow \quad \widetilde{\lambda}(X) = \frac{1}{\overline{X}}.$$

We observe that the MOME of λ happens to be the same as the MLE of λ. □

Example 3.46. Let $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \vartheta)$ be a random sample with known μ. We observe that the first order theoretical moment $\mu_1 = \mu$ doesn't depend on the value of ϑ, so we skip it. We equate the second order central moments:

$$\mu_2^* = M_2^* \quad \Rightarrow \quad \operatorname{Var}(X_1) = \frac{1}{n}\sum_{i=1}^n (X_i - \mu_1)^2 \quad \Rightarrow \quad \widetilde{\vartheta}(X) = \frac{1}{n}\sum_{i=1}^n (X_i - \mu)^2.$$

We observe that the MOME of ϑ happens to be the same as the MLE of ϑ. □

Example 3.47. Let $X_1, \ldots, X_n \sim \mathcal{N}(\vartheta_1, \vartheta_2)$ be a random sample. We equate the first order moments and the second order central moments:

$$\mu_1 = M_1 \quad \Rightarrow \quad \mathbb{E}(X_1) = \frac{1}{n}\sum_{i=1}^n X_i \quad \Rightarrow \quad \widetilde{\vartheta}_1(X) = \overline{X},$$

$$\mu_2^* = M_2^* \quad \Rightarrow \quad \operatorname{Var}(X_1) = \frac{1}{n}\sum_{i=1}^n (X_i - \mu_1)^2 \quad \Rightarrow \quad \widetilde{\vartheta}_2(X) = \frac{1}{n}\sum_{i=1}^n \left(X_i - \overline{X}\right)^2.$$

We observe that the MOME of ϑ happens to be the same as the MLE of ϑ. □

Example 3.48. Let $X_1, \ldots, X_n \sim \text{Gamma}(k, \lambda)$ be a random sample. We equate the first order moments and the second order central moments:

$$\mu_1 = M_1 \quad \Rightarrow \quad \mathbb{E}(X_1) = \frac{1}{n}\sum_{i=1}^n X_i \quad \Rightarrow \quad \frac{k}{\lambda} = \overline{X},$$

$$\mu_2^* = M_2^* \quad \Rightarrow \quad \operatorname{Var}(X_1) = \frac{1}{n}\sum_{i=1}^n (X_i - \mu_1)^2 \quad \Rightarrow \quad \frac{k}{\lambda^2} = \frac{1}{n}\sum_{i=1}^n \left(X_i - \overline{X}\right)^2 \quad \Rightarrow$$

$$\frac{\overline{X}}{\lambda} = \frac{n-1}{n}S^2 \quad \Rightarrow \quad \widetilde{\lambda}(X) = \frac{n\overline{X}}{(n-1)S^2} \quad \Rightarrow \quad \widetilde{k}(X) = \overline{X}\widetilde{\lambda}(X) = \frac{n\overline{X}^2}{(n-1)S^2}.$$

In contrast with the MLE of $\vartheta = (k, \lambda)$, which doesn't have a closed form solutions and may only be calculated numerically, we observe that the MOME of ϑ can be calculated fairly easily. However, we also observe that it's not a function of the sufficient statistic $T(X) = (\sum_{i=1}^n X_i, \sum_{i=1}^n \log X_i)$ for ϑ. □

Example 3.49. Let $X_1, \ldots, X_n \sim \mathcal{U}(0, \vartheta)$ be a random sample. We equate the first order moments:

$$\mu_1 = M_1 \quad \Rightarrow \quad \frac{\vartheta}{2} = \overline{X} \quad \Rightarrow \quad \widetilde{\vartheta}(X) = 2\overline{X}.$$

We observe that the MOME of ϑ isn't a function of the sufficient statistic $T(X) = X_{(n)}$ for ϑ, since MOMEs are always functions of the sample moments. □

Example 3.50. Let $X_1, \ldots, X_n \sim \mathcal{U}(-\vartheta, \vartheta)$ be a random sample with $\vartheta > 0$. We observe that the first order theoretical moment $\mu_1 = 0$ doesn't depend on the value of ϑ, so we skip it. We calculate that:

$$\mathbb{E}\left(X_1^2\right) = \int_{-\vartheta}^{\vartheta} \frac{x^2}{2\vartheta}dx = \frac{\vartheta^2}{3}.$$

We equate the second order raw moments:

$$\mu_2 = M_2 \quad \Rightarrow \quad \frac{\vartheta^2}{3} = \frac{1}{n}\sum_{i=1}^n X_i^2 \quad \Rightarrow \quad \widetilde{\vartheta}(X) = \sqrt{3M_2(X)}. \quad \square$$

Example 3.51. Let $X_1, \ldots, X_n \sim \mathcal{U}(\vartheta_1, \vartheta_2)$ be a random sample. We equate the first order moments and the second order central moments:

$$\mu_1 = M_1 \quad \Rightarrow \quad \mathbb{E}(X_1) = \frac{1}{n}\sum_{i=1}^n X_i \quad \Rightarrow \quad \frac{\vartheta_1 + \vartheta_2}{2} = \overline{X} \quad \Rightarrow \quad \vartheta_1 + \vartheta_2 = 2\overline{X},$$

$$\mu_2^* = M_2^* \quad \Rightarrow \quad \operatorname{Var}(X_1) = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu_1)^2 \quad \Rightarrow$$

$$\frac{(\vartheta_2 - \vartheta_1)^2}{12} = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2 \quad \Rightarrow \quad \vartheta_2 - \vartheta_1 = 2S\sqrt{\frac{3(n-1)}{n}} \quad \Rightarrow$$

$$\widetilde{\vartheta}_1(X) = \overline{X} - S\sqrt{\frac{3(n-1)}{n}}, \quad \widetilde{\vartheta}_2(X) = \overline{X} + S\sqrt{\frac{3(n-1)}{n}}. \quad \square$$

Example 3.52. Let $X_1, \dots, X_n$ be a random sample with $f(x; \lambda, k) = \lambda e^{-\lambda(x-k)}$ for $\lambda > 0$, $k \in \mathbb{R}$ and $x \geqslant k$. Let $Y_i = X_i - k$ for $i = 1, 2, \dots, n$. For $y > 0$, we calculate that:

$$F_{Y_1}(y) = \mathbb{P}(X_1 - k \leqslant y) = F(y + k; \lambda, k), \quad f_{Y_1}(y) = f(y + k; \lambda, k) = \lambda e^{-\lambda y},$$

i.e. $Y_i \sim \operatorname{Exp}(\lambda)$ for $i = 1, 2, \dots, n$. Therefore, we infer that:

$$\mathbb{E}(X_1) = \mathbb{E}(Y_1) + k = \frac{1}{\lambda} + k, \quad \operatorname{Var}(X_1) = \operatorname{Var}(Y_1) = \frac{1}{\lambda^2}.$$

We equate the first order moments and the second order central moments:

$$\mu_1 = M_1 \quad \Rightarrow \quad \mathbb{E}(X_1) = \frac{1}{n}\sum_{i=1}^{n} X_i \quad \Rightarrow \quad \frac{1}{\lambda} + k = \overline{X} \quad \Rightarrow \quad k = \overline{X} - \frac{1}{\lambda},$$

$$\mu_2^* = M_2^* \quad \Rightarrow \quad \operatorname{Var}(X_1) = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu_1)^2 \quad \Rightarrow \quad \frac{1}{\lambda^2} = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2 \quad \Rightarrow$$

$$\widetilde{\lambda}(X) = \frac{1}{S}\sqrt{\frac{n}{n-1}} \quad \Rightarrow \quad \widetilde{k}(X) = \overline{X} - S\sqrt{\frac{n-1}{n}}. \quad \square$$

Example 3.53. Let $X_1, \dots, X_n \sim \operatorname{Bin}(N, p)$ be a random sample. We equate the first order moments and the second order central moments:

$$\mu_1 = M_1 \quad \Rightarrow \quad \mathbb{E}(X_1) = \frac{1}{n}\sum_{i=1}^{n} X_i \quad \Rightarrow \quad Np = \overline{X} \quad \Rightarrow \quad N = \frac{\overline{X}}{p},$$

$$\mu_2^* = M_2^* \quad \Rightarrow \quad \operatorname{Var}(X_1) = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu_1)^2 \quad \Rightarrow$$

$$Np(1-p) = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2 \quad \Rightarrow \quad \widetilde{p}(X) = 1 - \frac{n-1}{n\overline{X}}S^2 \quad \Rightarrow$$

$$\widetilde{N}(X) = \frac{n\overline{X}^2}{n\overline{X} - (n-1)S^2}.$$

For the statistic $(\widetilde{N}, \widetilde{p})$ to constitute an estimator of (N, p), it needs to take values on the parameter space $\Theta = \mathbb{N} \times (0, 1)$ and to agree with the support of the distribution.

For this reason, we set forth the following restrictions:

- $(n-1)S^2 < n\overline{X}$ which implies that $\widetilde{p}(X) \in (0,1)$;
- $\widetilde{N}(X) \in \mathbb{N}$ and $\widetilde{N}(X) \geqslant X_{(n)}$ which implies that:

$$\widetilde{N}(X) = \max\left\{\left\lfloor \frac{n\overline{X}^2}{n\overline{X}-(n-1)S^2} \right\rfloor, X_{(n)}\right\}. \quad \square$$

Note 3.30. Even though MOMEs are generally easier to calculate than estimators of other kinds, they lack certain "good" properties. For example, they're not necessarily functions of some sufficient statistic and they don't necessarily take values on the parameter space.

Chapter 4

Confidence Intervals

4.1 Introduction

For a given sample x from a distribution with unknown parameter ϑ, we have thus far studied how to calculate a point estimate of ϑ like the MLE $\widehat{\vartheta}(x)$, the UMVUE $\delta(x)$, the efficient estimator $T(x)$ or the MOME $\widetilde{\vartheta}(x)$. These values constitute some "good" estimates of the true value of ϑ, according to the criteria set forth in the previous chapter. However, the mere calculation of a point estimate of ϑ doesn't provide us with any information on the uncertainty we have about our point estimate, i.e. how far away the true value of ϑ could lie from the point estimate we calculated based on our sample. Therefore, we arrive at the idea for the construction of an interval around our point estimate within which the true value of ϑ lies with some prespecified level of "confidence".

Definition 4.1. For given $\alpha \in (0,1)$, we consider a random interval of the form $\mathcal{I}_{g(\vartheta);1-\alpha}(X) = [L(X), U(X)]$ such that:

$$\inf_{\vartheta \in \Theta} \mathbb{P}_{\vartheta}\left[L(X) \leqslant g(\vartheta), U(X) \geqslant g(\vartheta)\right] = \inf_{\vartheta \in \Theta} \mathbb{P}_{\vartheta}\left[L(X) \leqslant g(\vartheta) \leqslant U(X)\right] = 1 - \alpha,$$

which is called a $100(1-\alpha)\%$ *confidence interval* (CI) for the parametric function $g(\vartheta)$. The quantity $1-\alpha$ is called the *confidence level* of the CI.

Interpretation: Assume that we let $\alpha = 0.05$, collect a sample x and construct a 95% CI $\mathcal{I}_{\vartheta;0.95}(x) = [0.9, 1.2]$ for ϑ based on it. According to the previous definition, one could think that the true value of the unknown parameter ϑ lies inside the interval $[0.9, 1.2]$ with 95% probability. However, this interpretation of the CI is **incorrect**. In frequentist statistics, the parameter ϑ is considered to be an unknown constant, so it will either lie or not lie inside the interval $[0.9, 1.2]$. Since ϑ is not a random variable, assigning a probability to the event that $0.9 \leqslant \vartheta \leqslant 1.2$ is meaningless. Some correct interpretations of a 95% CI for ϑ are detailed below.

If we repeated the sample collection process K many times and repeated the calculation of the 95% CI for ϑ for each of these samples x_k, then 95% of these intervals would contain the true value of ϑ. In other words, the interval varies across different repetitions, since it depends on the observed sample, and not the unknown parameter, which always remains constant. This interpretation of CIs arises from the strong law of large numbers as follows:

$$\frac{1}{K}\sum_{k=1}^{K}\mathbb{1}_{[L(x_k),U(x_k)]}(\vartheta) \overset{\text{a.s.}}{\to} \mathbb{E}_\vartheta\left[\mathbb{1}_{[L(X),U(X)]}(\vartheta)\right] = \mathbb{P}_\vartheta\left[L(X) \leqslant \vartheta \leqslant U(X)\right],$$

where $\frac{1}{K}\sum_{k=1}^{K}\mathbb{1}_{[L(x_k),U(x_k)]}(\vartheta)$ is precisely the percentage of the computed CIs which contain the true value of ϑ. We observe that 5% of the computed CIs wouldn't contain the true value of ϑ by construction.

There is 95% probability that a CI calculated from a sample collected in the future will contain the true value of ϑ. Note that this probabilistic interpretation still concerns the CI and not the unknown parameter ϑ, which remains constant. Since we haven't yet observed the sample based on which we'll construct the CI, we can assume that it's random. Therefore, the CI which we'll construct based on it is also going to be random, and we can assign the previously stated probabilistic interpretation to it.

4.2 Pivotal Quantity Method

Definition 4.2. A random variable $Q\left(X, g(\vartheta)\right)$ is called a *pivotal quantity* (or pivot) for the parametric function $g(\vartheta)$ if it depends on the value of $g(\vartheta)$ but its distribution doesn't depend on the value of ϑ.

Note 4.1. We observe that the pivot Q doesn't constitute a statistic, since it depends on the value of ϑ. The pivotal quantity method aims at the construction of CIs for which the probability $\mathbb{P}_\vartheta\left[L(X) \leqslant g(\vartheta) \leqslant U(X)\right]$ doesn't depend on the value of ϑ. Therefore, it follows that:

$$\mathbb{P}_\vartheta\left[L(X) \leqslant g(\vartheta) \leqslant U(X)\right] = 1-\alpha, \quad \forall\vartheta \in \Theta.$$

Pivotal Quantity Method $\rightarrow$ We generally heed the following steps:

1. We determine a "good" estimator $T(X)$ or a sufficient statistic $T(X)$ for ϑ.
2. We determine the distribution $T(X)$.
3. We determine a pivotal quantity $Q\left(X, g(\vartheta)\right)$. The method of determining a suitable pivot heavily depends on the distribution of $T(X)$.

4. We determine constants c_1 and c_2 such that $\mathbb{P}(c_1 \leqslant Q \leqslant c_2) = 1 - \alpha$.

5. We solve the inequality $c_1 \leqslant Q \leqslant c_2$ with respect to $g(\vartheta)$ and arrive at an inequality of the form $L(X) \leqslant g(\vartheta) \leqslant U(X)$. The interval $[L(X), U(X)]$ is a $100(1-\alpha)\%$ CI for the parametric function $g(\vartheta)$.

Definition 4.3. If $Z \sim \mathcal{N}(0,1)$ and $X \sim \chi^2_\nu$ are independent random variables, then we define:

$$T = \frac{Z}{\sqrt{X/\nu}} \sim t_\nu.$$

We say that the random variable T follows *Student's t* distribution with ν degrees of freedom.

Definition 4.4. If $X \sim \chi^2_{\nu_1}$ and $Y \sim \chi^2_{\nu_2}$ are independent random variables, then we define:

$$F = \frac{X/\nu_1}{Y/\nu_2} \sim F_{\nu_1,\nu_2}.$$

We say that the random variable F follows *Snedecor's F* distribution with ν_1 and ν_2 degrees of freedom.

Proposition 4.1. i. If $X_n \sim t_n$, then $X_n \xrightarrow{d} Z \sim \mathcal{N}(0,1)$.

ii. If $T \sim t_\nu$, then $T^2 \sim F_{1,\nu}$.

iii. If $F \sim F_{\nu_1,\nu_2}$, then $F^{-1} \sim F_{\nu_2,\nu_1}$.

Proof. i. Let $Y \sim \chi^2_n$ be independent of $Z \sim \mathcal{N}(0,1)$. Then, there exist independent random variables $Y_1, Y_2, \ldots, Y_n \sim \chi^2_1$ such that Y and $\sum_{i=1}^n Y_i$ have the same distribution. Since $\mathbb{E}(Y_1) = 1$, it follows that $\overline{Y} \xrightarrow{p} 1$, according to the weak law of large numbers. Therefore, we conclude that:

$$X_n \stackrel{d}{=} \frac{Z}{\sqrt{Y/n}} \stackrel{d}{=} \frac{Z}{\sqrt{\overline{Y}}} \xrightarrow{d} Z \sim \mathcal{N}(0,1),$$

according to the continuous mapping theorem and Slutsky's theorem.

ii. Consider the independent random variables $Z \sim \mathcal{N}(0,1)$ and $X \sim \chi^2_\nu$. Then, the random variables T and $\frac{Z}{\sqrt{X/\nu}}$ have the same distribution. Since $Z^2 \sim \chi^2_1$, it follows that:

$$T^2 \stackrel{d}{=} \frac{Z^2}{X/\nu} \sim F_{1,\nu}.$$

iii. Consider the independent random variables $X \sim \chi^2_{\nu_1}$ and $Y \sim \chi^2_{\nu_2}$. Then, the random variables F and $\frac{X/\nu_1}{Y/\nu_2}$ have the same distribution. Therefore, we conclude that:

$$F^{-1} \stackrel{d}{=} \frac{Y/\nu_2}{X/\nu_1} \sim F_{\nu_2,\nu_1}.$$

□

Note 4.2. The most involved step in the construction of a CI by use of the pivotal quantity method is the designation of the pivotal quantity itself, since the process of determining it mostly depends on the distribution of $T(X)$. In most cases, we endeavor to transform $T(X)$ into a pivotal quantity $Q\left(X, g(\vartheta)\right)$ which follows one of the following 4 distributions: $\mathcal{N}(0,1)$, χ^2_ν, t_ν, F_{ν_1,ν_2}. In order to determine this transformation, we either use some of the properties of the χ^2 distribution detailed in note 3.11 (page 38) or the definitions of the t_ν and F_{ν_1,ν_2} distributions. Obviously, the choice of a suitable pivotal quantity isn't unique.

Note 4.3. We summarize the most notable cases in which the previous 4 distributions are used in the construction of CIs.

i. $\mathcal{N}(0,1)$ distribution:
   - CIs for the mean of a normal distribution when its variance is known.
   - Asymptotic CIs using the central limit theorem.

ii. χ^2_ν distribution:
   - CIs for the variance of a normal distribution.
   - CIs for a positive parameter of a continuous distribution with support which doesn't depend on the parameter.

iii. t_ν distribution: CIs for the mean of a normal distribution when its variance is unknown.

iv. F_{ν_1,ν_2} distribution:
   - CIs for the ratio of variances of 2 independent normal distributions.
   - CIs for the ratio of 2 positive parameters of 2 independent continuous distributions with supports which don't depend on the values of the parameters.

Note 4.4. i. If we have a random sample $X_1, \ldots, X_n \sim \mathcal{U}(k, \vartheta)$ with known k, we may define the pivotal quantity $Q = \frac{X_{(n)}-k}{\vartheta-k} \sim \text{Beta}(n,1)$.

ii. If we have a random sample $X_1, \ldots, X_n \sim \mathcal{U}(\vartheta, k)$ with known k, we may define the pivotal quantity $Q = \frac{k-X_{(1)}}{k-\vartheta} \sim \text{Beta}(n,1)$.

Definition 4.5. Let X be a random variable with support S and CDF $F(x)$. For given $\alpha \in (0,1)$, the constant $c \in S$ for which it holds that $\mathbb{P}(X > c) = \alpha$ or equivalently $F(c) = 1-\alpha$ is called the *upper α-quantile* of the distribution.

Note 4.5. If the CDF $F(x)$ is continuous, then it's strictly increasing on S, so it's also invertible. Therefore, it holds that $c = F^{-1}(1-a)$. In this case, the upper

α-quantile of the distribution is the point to the right of which the area under the curve of the PDF is equal to α. We denote the upper α-quantiles of the distributions $\mathcal{N}(0,1)$, χ^2_ν, t_ν, F_{ν_1,ν_2} by Z_α, $\chi^2_{\nu;\alpha}$, $t_{\nu;\alpha}$, $F_{\nu_1,\nu_2;\alpha}$ respectively.

Note 4.6. The $\mathcal{N}(0,1)$ and t_ν distributions are symmetric around 0, i.e. it holds that $f(-c) = f(c)$ and $F(-c) = 1 - F(c)$. Hence, we observe that $\mathbb{P}(X > c) = \alpha \Leftrightarrow \mathbb{P}(X > -c) = 1 - \alpha$. That is, c is their upper α-quantile if and only if $-c$ is their upper $(1-\alpha)$-quantile or equivalently $Z_{1-\alpha} = -Z_a$ and $t_{\nu;1-\alpha} = -t_{\nu;\alpha}$. In contrast, the support of the χ^2_ν and F_{ν_1,ν_2} distributions is $(0,\infty)$, and they don't exhibit any symmetry. However, according to the properties of the F_{ν_1,ν_2} distribution, it holds that $F_{\nu_1,\nu_2;1-\alpha} = \frac{1}{F_{\nu_2,\nu_1;\alpha}}$.

Note 4.7. The pivotal quantity method doesn't provide a specific way of calculating the constants c_1, c_2. In theory, this choice could be made in an infinite number of possible ways, but it's usually made in one of the following 2 ways:

i. $\mathbb{P}(Q < c_1) = \mathbb{P}(Q > c_2) = \frac{\alpha}{2}$ which leads to the construction of *equal-tailed* CIs.

ii. Minimization of the statistic $\ell(X) = U(X) - L(X)$ or its expected value $\mathbb{E}[\ell(X)]$ with respect to (c_1, c_2), which leads to the construction of *minimum length* CIs.

Minimum length CIs are better than equal-tailed CIs, but they're also generally more difficult to construct. In some cases, these 2 kinds of CIs may also coincide.

Note 4.8. If the distribution of the pivotal quantity Q is continuous, the constants c_1, c_2 of equal-tailed CIs are chosen so that the area under the curve of the PDF of Q to the left of c_1 is equal to $\frac{\alpha}{2}$ and the area under the curve of the PDF of Q to the right of c_2 is also equal to $\frac{\alpha}{2}$. In this way, the area under the curve of the PDF of Q between c_1 and c_2 is equal to $1-\alpha$, which is the desired confidence level. In other words, c_1 is chosen as the upper $(1-\frac{\alpha}{2})$-quantile of Q and c_2 is chosen as the upper $\frac{\alpha}{2}$-quantile of Q.

Note 4.9. As far as the construction of minimum length CIs is concerned, we distinguish the following 2 important cases:

i. If the length of the CI is a multiple of $c_2 - c_1$, then we specify the constants c_1, c_2 such that the CI will contain the values of Q with the highest density. To achieve this we need to know about the behavior of the graph of the PDF of Q.

ii. Otherwise, we differentiate the constraint $\mathbb{P}(c_1 \leqslant Q \leqslant c_2) = 1-\alpha$ with respect to c_1, paying attention to the fact that c_2 is a function of c_1, and solve with respect to $\frac{\partial c_2}{\partial c_1}$. Next, we differentiate the length of the CI with respect to c_1, substitute the derivative $\frac{\partial c_2}{\partial c_1}$ and infer the monotonicity of the length with respect to c_1.

   - If the length is a strictly decreasing function of c_1, then c_1 must take the minimum possible value on the support of Q and c_2 is specified so that $\mathbb{P}(Q \leqslant c_2) = 1 - \alpha$.

- If the length is a strictly increasing function of c_1, then c_2 must take the maximum possible value on the support of Q and c_1 is specified so that $\mathbb{P}(Q \geqslant c_1) = 1 - \alpha$.

Example 4.1. Let $X_1, \dots, X_n$ be a random sample with $F(x; k) = 1 - e^{-\lambda(x-k)}$ for known $\lambda > 0$, $k \in \mathbb{R}$ and $x \geqslant k$. According to example 3.43 (page 75), the statistic $\widehat{k}(X) = X_{(1)}$ is the MLE of k. According to example 3.52 (page 81), we know that $Y_i = X_i - k \sim \text{Exp}(\lambda)$ for $i = 1, 2, \dots, n$, so $Y_{(1)} = X_{(1)} - k \sim \text{Exp}(n\lambda)$. Since the distribution of the random variable $Y_{(1)}$ doesn't depend on the value of k, it constitutes a suitable pivotal quantity Q. We solve the inequality $c_1 \leqslant Q \leqslant c_2$ with respect to k:

$$c_1 \leqslant X_{(1)} - k \leqslant c_2 \quad \Leftrightarrow \quad X_{(1)} - c_2 \leqslant k \leqslant X_{(1)} - c_1.$$

For the equal-tailed CI, we specify constants c_1, c_2 such that:

$$\mathbb{P}(Q < c_1) = \frac{\alpha}{2} \quad \Rightarrow \quad 1 - e^{-n\lambda c_1} = \frac{\alpha}{2} \quad \Rightarrow \quad c_1 = -\frac{1}{n\lambda} \log\left(1 - \frac{\alpha}{2}\right),$$

$$\mathbb{P}(Q > c_2) = \frac{\alpha}{2} \quad \Rightarrow \quad e^{-n\lambda c_2} = \frac{\alpha}{2} \quad \Rightarrow \quad c_2 = -\frac{1}{n\lambda} \log \frac{\alpha}{2}.$$

Therefore, we arrive at the following equal-tailed CI:

$$\left[X_{(1)} + \frac{1}{n\lambda} \log \frac{\alpha}{2}, X_{(1)} + \frac{1}{n\lambda} \log\left(1 - \frac{\alpha}{2}\right)\right].$$

The length of the CI is equal to $c_2 - c_1$. We observe that the PDF of the pivotal quantity Q is strictly decreasing on $[0, \infty)$. Since we want the CI to attain its minimum length, it's equivalent to require that it contains the values of Q with the highest density. Thus, the CI attains its minimum length for $c_1 = 0$. We specify c_2 such that:

$$\mathbb{P}(Q \leqslant c_2) = 1 - \alpha \quad \Rightarrow \quad 1 - e^{-n\lambda c_2} = 1 - \alpha \quad \Rightarrow \quad c_2 = -\frac{1}{n\lambda} \log \alpha.$$

Therefore, we arrive at the following minimum length CI:

$$\left[X_{(1)} + \frac{1}{n\lambda} \log \alpha, X_{(1)}\right]. \quad \square$$

Example 4.2. Let $X_1, \dots, X_n \sim \mathcal{U}(\vartheta, \vartheta + 1)$ be a random sample with $\vartheta \in \mathbb{R}$. According to example 3.16 (page 40), we know that $T(X) = \big(X_{(1)}, X_{(n)}\big)$ is a sufficient statistic for ϑ. For $x \in [\vartheta, \vartheta + 1]$, we calculate that $F_{X_{(n)}}(x) = (x - \vartheta)^n$. We define a pivotal quantity $Q = X_{(n)} - \vartheta$. For $y \in [0,1]$, we calculate that:

$$F_Q(y) = \mathbb{P}\left[X_{(n)} - \vartheta \leqslant y\right] = F_{X_{(n)}}(y+\vartheta) = y^n,$$

i.e. $Q \sim \text{Beta}(n,1)$. We solve the inequality $c_1 \leqslant Q \leqslant c_2$ with respect to ϑ:

$$c_1 \leqslant X_{(n)} - \vartheta \leqslant c_2 \quad \Leftrightarrow \quad X_{(n)} - c_2 \leqslant \vartheta \leqslant X_{(n)} - c_1.$$

For the equal-tailed CI, we specify constants c_1, c_2 such that:

$$\mathbb{P}(Q < c_1) = \frac{\alpha}{2} \quad \Rightarrow \quad c_1^n = \frac{\alpha}{2} \quad \Rightarrow \quad c_1 = \left(\frac{\alpha}{2}\right)^{1/n},$$

$$\mathbb{P}(Q > c_2) = \frac{\alpha}{2} \quad \Rightarrow \quad 1 - c_2^n = \frac{\alpha}{2} \quad \Rightarrow \quad c_2 = \left(1 - \frac{\alpha}{2}\right)^{1/n}.$$

Therefore, we arrive at the following equal-tailed CI:

$$\left[X_{(n)} - \left(1 - \frac{\alpha}{2}\right)^{1/n}, X_{(n)} - \left(\frac{\alpha}{2}\right)^{1/n}\right].$$

The length of the CI is equal to $c_2 - c_1$. We observe that the PDF of the pivotal quantity Q is strictly increasing on $[0,1]$. Since we want the CI to attain its minimum length, it's equivalent to require that it contains the values of Q with the highest density. Therefore, the CI attains its minimum length for $c_2 = 1$. We specify c_1 such that:

$$\mathbb{P}(Q \geqslant c_1) = 1 - \alpha \quad \Rightarrow \quad 1 - c_1^n = 1 - \alpha \quad \Rightarrow \quad c_1 = \alpha^{1/n}.$$

Therefore, a minimum length CI for ϑ is $\left[X_{(n)} - 1, X_{(n)} - \alpha^{1/n}\right]$. □

Example 4.3. Let $X_1, \ldots, X_n \sim \mathcal{U}(\vartheta, k)$ be a random sample with known k. According to example 3.42 (page 75), the statistic $\widehat{\vartheta}(X) = X_{(1)}$ is the MLE of ϑ. For $x \in [\vartheta, k]$, we calculate that:

$$F_{X_{(1)}}(x) = 1 - \left(\frac{k-x}{k-\vartheta}\right)^n.$$

We define a pivotal quantity $Q = \frac{k - X_{(1)}}{k-\vartheta}$. For $y \in [0,1]$, we calculate that:

$$F_Q(y) = \mathbb{P}\left[\frac{k - X_{(1)}}{k-\vartheta} \leqslant y\right] = 1 - F_{X_{(1)}}(k - (k-\vartheta)y) = \left[\frac{k - k + (k-\vartheta)y}{k-\vartheta}\right]^n = y^n,$$

i.e. $Q \sim \text{Beta}(n,1)$. We solve the inequality $c_1 \leqslant Q \leqslant c_2$ with respect to ϑ:

$$c_1 \leqslant \frac{k - X_{(1)}}{k-\vartheta} \leqslant c_2 \quad \Leftrightarrow \quad k - \frac{k - X_{(1)}}{c_1} \leqslant \vartheta \leqslant k - \frac{k - X_{(1)}}{c_2}.$$

For the equal-tailed CI, we specify constants c_1, c_2 such that:

$$\mathbb{P}(Q < c_1) = \frac{\alpha}{2} \quad \Rightarrow \quad c_1^n = \frac{\alpha}{2} \quad \Rightarrow \quad c_1 = \left(\frac{\alpha}{2}\right)^{1/n},$$

$$\mathbb{P}(Q > c_2) = \frac{\alpha}{2} \quad \Rightarrow \quad 1 - c_2^n = \frac{\alpha}{2} \quad \Rightarrow \quad c_2 = \left(1 - \frac{\alpha}{2}\right)^{1/n}.$$

Therefore, we arrive at the following equal-tailed CI:

$$\left[k - \left(k - X_{(1)}\right)\left(\frac{\alpha}{2}\right)^{-1/n}, k - \left(k - X_{(1)}\right)\left(1 - \frac{\alpha}{2}\right)^{-1/n}\right].$$

The length of the CI is equal to $\left(k - X_{(1)}\right)\left(\frac{1}{c_1} - \frac{1}{c_2}\right)$. We want to minimize the function $\ell(c_1, c_2) = \frac{1}{c_1} - \frac{1}{c_2}$ under the following constraint:

$$\mathbb{P}(c_1 \leqslant Q \leqslant c_2) = 1 - \alpha \quad \Rightarrow \quad F_Q(c_2) - F_Q(c_1) = 1 - \alpha \quad \Rightarrow \quad c_2^n - c_1^n = 1 - \alpha.$$

First, we differentiate the constraint with respect to c_2:

$$nc_2^{n-1} - nc_1^{n-1}\frac{\partial c_1}{\partial c_2} = 0 \quad \Rightarrow \quad \frac{\partial c_1}{\partial c_2} = \left(\frac{c_2}{c_1}\right)^{n-1}.$$

Next, we differentiate ℓ with respect to c_2:

$$\frac{\partial \ell}{\partial c_2} = -\frac{1}{c_1^2}\frac{\partial c_1}{\partial c_2} + \frac{1}{c_2^2} = -\frac{1}{c_1^2}\left(\frac{c_2}{c_1}\right)^{n-1} + \frac{1}{c_2^2} = \frac{c_1^{n+1} - c_2^{n+1}}{c_1^{n+1}c_2^2} < 0.$$

We also know that $c_2 \in [0, 1]$. Since the length of the CI is a strictly decreasing function of c_2, we infer that it attains its minimum length for $c_2 = 1$. We specify c_1 such that:

$$\mathbb{P}(Q \geqslant c_1) = 1 - \alpha \quad \Rightarrow \quad 1 - c_1^n = 1 - \alpha \quad \Rightarrow \quad c_1 = \alpha^{1/n}.$$

Therefore, a minimum length CI for ϑ is $\left[k - \left(k - X_{(1)}\right)\alpha^{-1/n}, X_{(1)}\right]$. □

Example 4.4. Let $X_1, \ldots, X_n \sim \text{Pareto}(k, \lambda)$ be a random sample with $k > 0$, known $\lambda > 0$ and $F(x; k) = 1 - \left(\frac{k}{x}\right)^\lambda$ for $x \geqslant k$. According to example 3.43 (page 75), the statistic $\widehat{k}(X) = X_{(1)}$ is the MLE of k. For $x \geqslant k$, we calculate that:

$$F_{X_{(1)}}(x) = 1 - \left(\frac{k}{x}\right)^{n\lambda},$$

i.e. $X_{(1)} \sim \text{Pareto}(k, n\lambda)$. We define a pivot $Q = \frac{X_{(1)}}{k}$. For $y \geqslant 1$, we calculate that:

$$F_Q(y) = \mathbb{P}\left[\frac{X_{(1)}}{k} \leqslant y\right] = F_{X_{(1)}}(ky) = 1 - \left(\frac{1}{y}\right)^{n\lambda},$$

i.e. $Q \sim \text{Pareto}(1, n\lambda)$. We solve the inequality $c_1 \leqslant Q \leqslant c_2$ with respect to k:

$$c_1 \leqslant \frac{X_{(1)}}{k} \leqslant c_2 \quad \Leftrightarrow \quad \frac{X_{(1)}}{c_2} \leqslant k \leqslant \frac{X_{(1)}}{c_1}.$$

For the equal-tailed CI, we specify constants c_1, c_2 such that:

$$\mathbb{P}(Q < c_1) = \frac{\alpha}{2} \quad \Rightarrow \quad 1 - \frac{1}{c_1^{n\lambda}} = \frac{\alpha}{2} \quad \Rightarrow \quad c_1 = \left(1 - \frac{\alpha}{2}\right)^{-1/n\lambda},$$

$$\mathbb{P}(Q > c_2) = \frac{\alpha}{2} \quad \Rightarrow \quad \frac{1}{c_2^{n\lambda}} = \frac{\alpha}{2} \quad \Rightarrow \quad c_2 = \left(\frac{\alpha}{2}\right)^{-1/n\lambda}.$$

Therefore, we arrive at the following equal-tailed CI:

$$\left[X_{(1)} \left(\frac{\alpha}{2}\right)^{1/n\lambda}, X_{(1)} \left(1 - \frac{\alpha}{2}\right)^{1/n\lambda}\right].$$

The length of the CI is equal to $X_{(1)} \left(\frac{1}{c_1} - \frac{1}{c_2}\right)$. We want to minimize the function $\ell(c_1, c_2) = \frac{1}{c_1} - \frac{1}{c_2}$ under the following constraint:

$$\mathbb{P}(c_1 \leqslant Q \leqslant c_2) = 1 - \alpha \quad \Rightarrow \quad F_Q(c_2) - F_Q(c_1) = 1 - \alpha \quad \Rightarrow \quad c_1^{-n\lambda} - c_2^{-n\lambda} = 1 - \alpha.$$

First, we differentiate the constraint with respect to c_1:

$$-\frac{n\lambda}{c_1^{n\lambda+1}} + \frac{n\lambda}{c_2^{n\lambda+1}} \frac{\partial c_2}{\partial c_1} = 0 \quad \Rightarrow \quad \frac{\partial c_2}{\partial c_1} = \left(\frac{c_2}{c_1}\right)^{n\lambda+1}.$$

Next, we differentiate ℓ with respect to c_1:

$$\frac{\partial \ell}{\partial c_1} = -\frac{1}{c_1^2} + \frac{1}{c_2^2} \frac{\partial c_2}{\partial c_1} = -\frac{1}{c_1^2} + \frac{1}{c_2^2} \left(\frac{c_2}{c_1}\right)^{n\lambda+1} = \frac{c_2^{n\lambda-1} - c_1^{n\lambda-1}}{c_1^{n\lambda+1}} > 0.$$

We also know that $c_1 \geqslant 1$. Since the length of the CI is a strictly increasing function of c_1, we infer that it attains its minimum length for $c_1 = 1$. We specify c_2 such that:

$$\mathbb{P}(Q \leqslant c_2) = 1 - \alpha \quad \Rightarrow \quad 1 - \frac{1}{c_2^{n\lambda}} = 1 - \alpha \quad \Rightarrow \quad c_2 = \alpha^{-1/n\lambda}.$$

Therefore, a minimum length CI for k is $\left[\alpha^{1/n\lambda} X_{(1)}, X_{(1)}\right]$. □

Example 4.5. Let $X_1, \ldots, X_n \sim \text{Gamma}(k, \lambda)$ be a random sample with known k. We can easily show that the statistic $T(X) = \sum_{i=1}^n X_i \sim \text{Gamma}(nk, \lambda)$ is sufficient for λ. According to note 3.11 (page 38), we define a pivotal quantity $Q = 2\lambda T \sim \chi^2_{2nk}$. We solve the inequality $c_1 \leqslant Q \leqslant c_2$ with respect to λ:

$$c_1 \leqslant 2\lambda \sum_{i=1}^n X_i \leqslant c_2 \quad \Leftrightarrow \quad \frac{c_1}{2\sum_{i=1}^n X_i} \leqslant \lambda \leqslant \frac{c_2}{2\sum_{i=1}^n X_i}.$$

For the equal-tailed CI, we specify constants c_1, c_2 such that:

$$\mathbb{P}(Q < c_1) = \frac{\alpha}{2} \quad \Rightarrow \quad \mathbb{P}(Q > c_1) = 1 - \frac{\alpha}{2} \quad \Rightarrow \quad c_1 = \chi^2_{2nk;1-\alpha/2},$$

$$\mathbb{P}(Q > c_2) = \frac{\alpha}{2} \quad \Rightarrow \quad c_2 = \chi^2_{2nk;\alpha/2}.$$

Therefore, we arrive at the following equal-tailed CI:

$$\left[\frac{\chi^2_{2nk;1-\alpha/2}}{2\sum_{i=1}^n X_i}, \frac{\chi^2_{2nk;\alpha/2}}{2\sum_{i=1}^n X_i}\right]. \quad \square$$

Example 4.6. Let $X_1, \ldots, X_n \sim \text{Laplace}(\mu, \lambda)$ be a random sample with known $\mu \in \mathbb{R}$, $\lambda > 0$ and $f(x;\lambda) = \frac{\lambda}{2} e^{-\lambda|x-\mu|}$ for $x \in \mathbb{R}$. We can easily show that the statistic $T(X) = \sum_{i=1}^n |X_i - \mu|$ is sufficient for λ. We define $Y_i = |X_i - \mu|$ for $i = 1, 2, \ldots, n$. For $y > 0$, we calculate that:

$$F_{Y_1}(y) = \mathbb{P}\left(|X - \mu| \leqslant y\right) = \mathbb{P}(\mu - y \leqslant X \leqslant \mu + y) = F(\mu + y; \lambda) - F(\mu - y; \lambda),$$

$$f_{Y_1}(y) = f(\mu + y; \lambda) + f(\mu - y; \lambda) = \frac{\lambda}{2} e^{-\lambda|y|} + \frac{\lambda}{2} e^{-\lambda|-y|} = \frac{\lambda}{2} e^{-\lambda y} + \frac{\lambda}{2} e^{-\lambda y} = \lambda e^{-\lambda y},$$

i.e. $Y_i \sim \text{Exp}(\lambda)$ for $i = 1, 2, \ldots, n$, so it follows that $T(X) \sim \text{Gamma}(n, \lambda)$. In exactly the same manner as in the previous example, we define the pivot $Q = 2\lambda T \sim \chi^2_{2n}$ and calculate that $c_1 = \chi^2_{2n;1-\alpha/2}$, $c_2 = \chi^2_{2n;\alpha/2}$. Therefore, we arrive at the following equal-tailed CI:

$$\left[\frac{\chi^2_{2n;1-\alpha/2}}{2\sum_{i=1}^n |X_i - \mu|}, \frac{\chi^2_{2n;\alpha/2}}{2\sum_{i=1}^n |X_i - \mu|}\right]. \quad \square$$

Example 4.7. Let $X_1, \ldots, X_n \sim \text{Beta}(1, \vartheta)$ be a random sample with $\vartheta > 0$ and $f(x;\vartheta) = \vartheta(1-x)^{\vartheta-1}$ for $x \in (0,1)$. We can show that $T(X) = -\sum_{i=1}^n \log(1 - X_i)$ is a sufficient statistic for ϑ. We define $Y_i = -\log(1 - X_i)$ for $i = 1, 2, \ldots, n$. For $y > 0$, we calculate that:

$$F_{Y_1}(y) = \mathbb{P}\left(-\log(1 - X_1) \leqslant y\right) = \mathbb{P}\left(1 - X_1 \geqslant e^{-y}\right) = F\left(1 - e^{-y}; \vartheta\right),$$

$$f_{Y_1}(y) = e^{-y} f\left(1 - e^{-y}; \vartheta\right) = e^{-y}\vartheta\left(1 - 1 + e^{-y}\right)^{\vartheta - 1} = \vartheta e^{-\vartheta y},$$

i.e. $Y_i \sim \text{Exp}(\vartheta)$ for $i = 1, 2, \ldots, n$, so it follows that $T(X) \sim \text{Gamma}(n, \vartheta)$. In exactly the same manner as in the previous example, we define the pivot $Q = 2\vartheta T \sim \chi^2_{2n}$ and calculate that $c_1 = \chi^2_{2n;1-\alpha/2}$, $c_2 = \chi^2_{2n;\alpha/2}$. Therefore, we arrive at the following equal-tailed CI:

$$\left[-\frac{\chi^2_{2n;1-\alpha/2}}{2\sum_{i=1}^n \log(1 - X_i)}, -\frac{\chi^2_{2n;\alpha/2}}{2\sum_{i=1}^n \log(1 - X_i)}\right]. \quad \square$$

Example 4.8. Let $X_1, \ldots, X_n$ be a random sample with $F(x;\lambda) = 1 - e^{-\lambda(x-k)}$ for $\lambda > 0$, known $k \in \mathbb{R}$ and $x \geqslant k$. According to example 4.1 (page 88), we know that $Y_i = X_i - k \sim \text{Exp}(\lambda)$ for $i = 1, 2, \ldots, n$, so $T(X) = \sum_{i=1}^n X_i - nk \sim \text{Gamma}(n, \lambda)$. In exactly the same manner as in the previous example, we define the pivotal quantity $Q = 2\lambda T \sim \chi^2_{2n}$ and calculate that $c_1 = \chi^2_{2n;1-\alpha/2}$, $c_2 = \chi^2_{2n;\alpha/2}$. Therefore, we arrive at the following equal-tailed CI:

$$\left[\frac{\chi^2_{2n;1-\alpha/2}}{2\sum_{i=1}^n X_i - 2nk}, \frac{\chi^2_{2n;\alpha/2}}{2\sum_{i=1}^n X_i - 2nk}\right]. \quad \square$$

Example 4.9. Let $X_1, \ldots, X_n \sim \text{Pareto}(k, \lambda)$ be a random sample with known $k > 0$, $\lambda > 0$, $f(x;\lambda) = \frac{\lambda k^\lambda}{x^{\lambda+1}}$ and $F(x;\lambda) = 1 - \left(\frac{k}{x}\right)^\lambda$ for $x \geqslant k$. We calculate that:

$$\ell(\lambda \mid x) = n \log \lambda + n\lambda \log k - (\lambda + 1)\sum_{i=1}^n \log x_i,$$

$$\frac{\partial \ell(\lambda \mid x)}{\partial \lambda} = \frac{n}{\lambda} + n \log k - \sum_{i=1}^n \log x_i = 0 \quad \Rightarrow$$

$$\widehat{\lambda}(x) = \frac{n}{\sum_{i=1}^n \log x_i - n \log k} = \frac{n}{\sum_{i=1}^n \log \frac{x_i}{k}}, \quad \frac{\partial^2 \ell(\lambda \mid x)}{\partial \lambda^2} = -\frac{n}{\lambda^2} < 0, \quad \forall \lambda > 0.$$

We define $Y_i = \log \frac{X_i}{k}$ for $i = 1, 2, \ldots, n$. For $y > 0$, we calculate that:

$$F_{Y_1}(y) = \mathbb{P}\left(\log \frac{X_1}{k} \leqslant y\right) = F\left(pe^y;\lambda\right) = 1 - \frac{k^\lambda}{k^\lambda e^{\lambda y}} = 1 - e^{-\lambda y},$$

i.e. $Y_i \sim \text{Exp}(\lambda)$ for $i = 1, 2, \ldots, n$ and $T(X) = \sum_{i=1}^n \log \frac{X_i}{k} \sim \text{Gamma}(n, \lambda)$. In exactly the same manner as in the previous example, we define the pivotal quantity $Q = 2\lambda T \sim \chi^2_{2n}$ and calculate that $c_1 = \chi^2_{2n;1-\alpha/2}$, $c_2 = \chi^2_{2n;\alpha/2}$. Therefore, we arrive at the following equal-tailed CI:

$$\left[\frac{\chi^2_{2n;1-\alpha/2}}{2\sum_{i=1}^n \log X_i - 2n \log k}, \frac{\chi^2_{2n;\alpha/2}}{2\sum_{i=1}^n \log X_i - 2n \log k}\right]. \quad \square$$

Example 4.10. Let $X_1, \ldots, X_n \sim \text{Exp}(\lambda_1)$ and $Y_1, \ldots, Y_m \sim \text{Exp}(\lambda_2)$ be 2 independent random samples. We want to construct a CI for the ratio $\frac{\lambda_1}{\lambda_2}$. We know that $T_1(X) = \sum_{i=1}^n X_i \sim \text{Gamma}(n, \lambda_1)$ and $T_2(Y) = \sum_{i=1}^m Y_i \sim \text{Gamma}(m, \lambda_2)$ are sufficient statistics for λ_1 and λ_2 respectively. Let $Q_1 = 2\lambda_1 T_1 \sim \chi^2_{2n}$ and $Q_2 = 2\lambda_2 T_2 \sim \chi^2_{2m}$. Since the 2 samples are independent of each other, we infer that the random variables Q_1 and Q_2 are also independent. Hence, we construct the following pivotal quantity:

$$Q = \frac{Q_1/2n}{Q_2/2m} = \frac{\lambda_1}{\lambda_2}\frac{\overline{X}}{\overline{Y}} \sim F_{2n,2m}.$$

We solve the inequality $c_1 \leqslant Q \leqslant c_2$ with respect to $\frac{\lambda_1}{\lambda_2}$:

$$c_1 \leqslant \frac{\lambda_1}{\lambda_2}\frac{\overline{X}}{\overline{Y}} \leqslant c_2 \quad \Leftrightarrow \quad c_1\frac{\overline{Y}}{\overline{X}} \leqslant \frac{\lambda_1}{\lambda_2} \leqslant c_2\frac{\overline{Y}}{\overline{X}}.$$

For the equal-tailed CI, we specify constants c_1, c_2 such that:

$$\mathbb{P}(Q < c_1) = \frac{\alpha}{2} \quad \Rightarrow \quad \mathbb{P}(Q > c_1) = 1 - \frac{\alpha}{2} \quad \Rightarrow \quad c_1 = F_{2n,2m;1-\alpha/2},$$

$$\mathbb{P}(Q > c_2) = \frac{\alpha}{2} \quad \Rightarrow \quad c_2 = F_{2n,2m;\alpha/2}.$$

Therefore, we arrive at the following equal-tailed CI:

$$\left[F_{2n,2m;1-\alpha/2}\frac{\overline{Y}}{\overline{X}}, F_{2n,2m;\alpha/2}\frac{\overline{Y}}{\overline{X}}\right]. \quad \square$$

4.3 CIs for a Normal Population

Let $X_1, \ldots, X_n \sim \mathcal{N}\left(\mu, \sigma^2\right)$ be a random sample. We want to construct CIs for the parameters μ and σ^2. We distinguish 4 different cases which we present throughout this paragraph.

Example 4.11. The variance σ^2 is known. According to example 3.44 (page 76), the statistic $\overline{X} \sim \mathcal{N}\left(\mu, \frac{1}{n}\sigma^2\right)$ is the MLE of μ. Hence, we define a pivotal quantity $Q = \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)$. We solve the inequality $c_1 \leqslant Q \leqslant c_2$ with respect to μ:

$$c_1 \leqslant \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \leqslant c_2 \quad \Leftrightarrow \quad \overline{X} - c_2\frac{\sigma}{\sqrt{n}} \leqslant \mu \leqslant \overline{X} - c_1\frac{\sigma}{\sqrt{n}}.$$

For the equal-tailed CI, we specify constants c_1, c_2 such that:

$$\mathbb{P}(Q < c_1) = \frac{\alpha}{2} \quad \Rightarrow \quad \mathbb{P}(Q > c_1) = 1 - \frac{\alpha}{2} \quad \Rightarrow \quad c_1 = Z_{1-\alpha/2} = -Z_{\alpha/2},$$

$$\mathbb{P}(Q > c_2) = \frac{\alpha}{2} \quad \Rightarrow \quad c_2 = Z_{\alpha/2}.$$

Therefore, we arrive at the following equal-tailed CI:

$$\left[\overline{X} - Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \overline{X} + Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right].$$

The length of the CI is equal to $\frac{\sigma}{\sqrt{n}}\left(c_2 - c_1\right)$. We observe that the PDF of the pivotal quantity Q is symmetric and unimodal around 0. Since we want the CI to attain its minimum length, it's equivalent to require that it contains the values of Q with the highest density. Therefore, the CI attains its minimum length for $c_2 = -c_1$, which implies that the minimum length CI coincides with the equal-tailed CI. $\square$

Note 4.10. We observe that the length of the previous CI is equal to $\ell = 2Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$, i.e. it doesn't depend on the sample X. We note the following facts:

- The length of the CI is a strictly decreasing function of the sample size n, which means that the CI becomes more and more precise as we collect more observations for our sample.
- The length of the CI is a strictly increasing function of the variance σ^2, which means that the smaller the variation of the observations in the sample is the larger the precision of the constructed CI will be.
- Since it holds that $Z_{\alpha/2} = \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)$ and the inverse of the CDF Φ of the $\mathcal{N}(0,1)$ distribution is a strictly increasing function, we infer that the length of the CI is a strictly decreasing function of α or equivalently a strictly increasing function of $1-\alpha$. In other words, the larger the "confidence" we want to have that the true value of ϑ is going to lie within the CI the wider the CI we need to construct is going to be.

Example 4.12. The variance σ^2 is equal to 4. We want to determine the smallest possible sample size n such that the 99% CI for μ has length at most equal to 0.1. Since $\alpha = 0.01$, we demand the following:

$$\ell = 2Z_{0.005}\frac{\sigma}{\sqrt{n}} \leqslant 0.1 \quad \Rightarrow \quad n \geqslant 4Z_{0.005}^2\frac{\sigma^2}{0.01} \approx 10615.83,$$

which means that the smallest possible sample size we require is $n = 10616$. $\square$

Example 4.13. The variance σ^2 is unknown and we want to construct a CI for the mean μ. The random variable $Z = \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)$ doesn't constitute a pivotal quantity anymore, since it depends on the value of the unknown parameter σ^2. According to example 3.26 (page 49), we know that the statistic $S^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X})^2$ is the UMVUE of σ^2. According to note 3.11, we know that $V = \frac{n-1}{\sigma^2}S^2 \sim \chi^2_{n-1}$. Additionally, the random variables Z and V are independent according to Basu's theorem. Therefore, we construct the following pivotal quantity:

$$Q = \frac{Z}{\sqrt{V/(n-1)}} = \frac{\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}}{S/\sigma} = \frac{\overline{X}-\mu}{S/\sqrt{n}} \sim t_{n-1}.$$

We solve the inequality $c_1 \leqslant Q \leqslant c_2$ with respect to μ:

$$c_1 \leqslant \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \leqslant c_2 \quad \Leftrightarrow \quad \overline{X} - c_2 \frac{S}{\sqrt{n}} \leqslant \mu \leqslant \overline{X} - c_1 \frac{S}{\sqrt{n}}.$$

For the equal-tailed CI, we specify constants c_1, c_2 such that:

$$\mathbb{P}(Q < c_1) = \frac{\alpha}{2} \quad \Rightarrow \quad \mathbb{P}(Q > c_1) = 1 - \frac{\alpha}{2} \quad \Rightarrow \quad c_1 = t_{n-1;1-\alpha/2} = -t_{n-1;\alpha/2},$$

$$\mathbb{P}(Q > c_2) = \frac{\alpha}{2} \quad \Rightarrow \quad c_2 = t_{n-1;\alpha/2}.$$

Therefore, we arrive at the following equal-tailed CI:

$$\left[\overline{X} - t_{n-1;\alpha/2} \frac{S}{\sqrt{n}}, \overline{X} + t_{n-1;\alpha/2} \frac{S}{\sqrt{n}}\right].$$

The length of the CI is equal to $\frac{S}{\sqrt{n}}(c_2 - c_1)$. We observe that the PDF of the pivotal quantity Q is symmetric and unimodal around 0. Since we want the CI to attain its minimum length, it's equivalent to require that it contains the values of Q with the highest density. Therefore, the CI attains its minimum length for $c_2 = -c_1$, which implies that the minimum length CI coincides with the equal-tailed CI. □

Example 4.14. The mean μ is known. According to example 3.38 (page 73), the statistic $\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2$ is the MLE of σ^2. According to note 3.11, we define a pivot $Q = \frac{n}{\sigma^2}\widehat{\sigma}^2 \sim \chi^2_n$. We solve the inequality $c_1 \leqslant Q \leqslant c_2$ with respect to σ^2:

$$c_1 \leqslant \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \mu)^2 \leqslant c_2 \quad \Leftrightarrow \quad \frac{1}{c_2}\sum_{i=1}^{n}(X_i - \mu)^2 \leqslant \sigma^2 \leqslant \frac{1}{c_1}\sum_{i=1}^{n}(X_i - \mu)^2.$$

For the equal-tailed CI, we specify constants c_1, c_2 such that:

$$\mathbb{P}(Q < c_1) = \frac{\alpha}{2} \quad \Rightarrow \quad \mathbb{P}(Q > c_1) = 1 - \frac{\alpha}{2} \quad \Rightarrow \quad c_1 = \chi^2_{n;1-\alpha/2},$$

$$\mathbb{P}(Q > c_2) = \frac{\alpha}{2} \quad \Rightarrow \quad c_2 = \chi^2_{n;\alpha/2}.$$

Therefore, we arrive at the following equal-tailed CI:

$$\left[\frac{1}{\chi^2_{n;\alpha/2}}\sum_{i=1}^{n}(X_i - \mu)^2, \frac{1}{\chi^2_{n;1-\alpha/2}}\sum_{i=1}^{n}(X_i - \mu)^2\right]. \quad \square$$

Example 4.15. The mean μ is unknown and we want to construct an equal-tailed CI for the variance σ^2. We know that the statistic $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$ is the UMVUE of σ^2, so we define a pivot $Q = \frac{n-1}{\sigma^2}S^2 \sim \chi^2_{n-1}$. In exactly the same manner as in the previous example, we calculate that $c_1 = \chi^2_{n-1;1-\alpha/2}$, $c_2 = \chi^2_{n-1;\alpha/2}$.

Therefore, we arrive at the following equal-tailed CI:

$$\left[\frac{1}{\chi^2_{n-1;\alpha/2}} \sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2, \frac{1}{\chi^2_{n-1;1-\alpha/2}} \sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2\right]. \quad \square$$

4.4 CIs for Two Independent Normal Populations

Let $X_1, \ldots, X_n \sim \mathcal{N}\left(\mu_1, \sigma_1^2\right)$ and $Y_1, \ldots, Y_m \sim \mathcal{N}\left(\mu_2, \sigma_2^2\right)$ be 2 independent random samples. We want to construct CIs for the mean difference $\mu_1 - \mu_2$ and the variance ratio $\frac{\sigma_1^2}{\sigma_2^2}$. We distinguish 4 different cases which we present throughout this paragraph.

Example 4.16. The variances σ_1^2 and σ_2^2 are known. We know that the statistics $\overline{X} \sim \mathcal{N}\left(\mu_1, \frac{1}{n}\sigma_1^2\right)$ and $\overline{Y} \sim \mathcal{N}\left(\mu_2, \frac{1}{m}\sigma_2^2\right)$ are the MLEs of μ_1 and μ_2 respectively. Since the 2 samples are independent, we infer that the statistics $\overline{X}$ and $\overline{Y}$ are also independent, so it follows that $\overline{X} - \overline{Y} \sim \mathcal{N}\left(\mu_1 - \mu_2, \frac{1}{n}\sigma_1^2 + \frac{1}{m}\sigma_2^2\right)$. We construct the following pivotal quantity:

$$Q = \frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{1}{n}\sigma_1^2 + \frac{1}{m}\sigma_2^2}} \sim \mathcal{N}(0,1).$$

In exactly the same manner as in example 4.11 (page 94), we infer that $c_1 = -Z_{\alpha/2}$, $c_2 = Z_{\alpha/2}$. Therefore, we arrive at the following equal-tailed CI:

$$\left[\overline{X} - \overline{Y} - Z_{\alpha/2}\sqrt{\frac{1}{n}\sigma_1^2 + \frac{1}{m}\sigma_2^2}, \overline{X} - \overline{Y} + Z_{\alpha/2}\sqrt{\frac{1}{n}\sigma_1^2 + \frac{1}{m}\sigma_2^2}\right].$$

We note that the minimum length CI for the mean difference $\mu_1 - \mu_2$ coincides with the above equal-tailed CI. $\square$

Example 4.17. The variances σ_1^2 and σ_2^2 are unknown but equal to some common variance σ^2. In exactly the same manner as in the previous example, we define the following random variable:

$$Z = \frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{\sigma\sqrt{\frac{1}{n} + \frac{1}{m}}} \sim \mathcal{N}(0,1),$$

which doesn't constitute a pivot, since it depends on the value of the unknown parameter σ^2. We know that $S_1^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$ and $S_2^2 = \frac{1}{m-1}\sum_{i=1}^{m}\left(Y_i - \overline{Y}\right)^2$ are 2 different UMVUEs of σ^2 based on the samples X and Y respectively, so it follows that the *pooled sample variance* $S_p^2 = \frac{(n-1)S_1^2 + (m-1)S_2^2}{n+m-2}$ is the UMVUE of σ^2 based on the 2 samples put together. We also know that $V_1 = \frac{n-1}{\sigma^2}S_1^2 \sim \chi^2_{n-1}$ and $V_2 = \frac{m-1}{\sigma^2}S_2^2 \sim \chi^2_{m-1}$. Since the 2 samples are also independent, we infer that the random variables V_1 and V_2 are independent. According to note 3.11, we infer that:

$$W = \frac{n+m-2}{\sigma^2} S_p^2 = \frac{(n-1)S_1^2 + (m-1)S_2^2}{\sigma^2} = V_1 + V_2 \sim \chi^2_{n+m-2}.$$

According to Basu's theorem, the random variables Z and W are independent, so we construct the following pivotal quantity:

$$Q = \frac{Z}{\sqrt{W/(n+m-2)}} = \frac{\frac{\overline{X}-\overline{Y}-(\mu_1-\mu_2)}{\sigma\sqrt{\frac{1}{n}+\frac{1}{m}}}}{S_p/\sigma} = \frac{\overline{X}-\overline{Y}-(\mu_1-\mu_2)}{S_p\sqrt{\frac{1}{n}+\frac{1}{m}}} \sim t_{n+m-2}.$$

In exactly the same manner as in example 4.13 (page 95), $c_1 = -t_{n+m-2;\alpha/2}$ and $c_2 = t_{n+m-2;\alpha/2}$. Therefore, we arrive at the following equal-tailed CI:

$$\left[\overline{X}-\overline{Y}-t_{n+m-2;\alpha/2}S_p\sqrt{\frac{1}{n}+\frac{1}{m}}, \overline{X}-\overline{Y}+t_{n+m-2;\alpha/2}S_p\sqrt{\frac{1}{n}+\frac{1}{m}}\right].$$

We note that the minimum length CI for the mean difference $\mu_1 - \mu_2$ coincides with the above equal-tailed CI. □

Example 4.18. The means μ_1 and μ_2 are known. We know that the statistics $\widehat{\sigma}_1^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i-\mu_1)^2$ and $\widehat{\sigma}_2^2 = \frac{1}{m}\sum_{i=1}^{m}(Y_i-\mu_2)^2$ are the MLEs of σ_1^2 and σ_2^2 respectively. We also know that $V_1 = \frac{n}{\sigma_1^2}\widehat{\sigma}_1^2 \sim \chi^2_n$ and $V_2 = \frac{m}{\sigma_2^2}\widehat{\sigma}_2^2 \sim \chi^2_m$. Since the 2 samples are independent, we infer that the random variables V_1 and V_2 are also independent. Therefore, we construct the following pivotal quantity:

$$Q = \frac{V_1/n}{V_2/m} = \frac{\widehat{\sigma}_1^2}{\widehat{\sigma}_2^2}\frac{\sigma_2^2}{\sigma_1^2} = \frac{\sum_{i=1}^{n}(X_i-\mu_1)^2}{\sum_{i=1}^{m}(Y_i-\mu_2)^2}\frac{\sigma_2^2}{\sigma_1^2} \sim F_{n,m}.$$

We solve the inequality $c_1 \leqslant Q \leqslant c_2$ with respect to $\frac{\sigma_1^2}{\sigma_2^2}$:

$$c_1 \leqslant \frac{\widehat{\sigma}_1^2}{\widehat{\sigma}_2^2}\frac{\sigma_2^2}{\sigma_1^2} \leqslant c_2 \quad \Leftrightarrow \quad \frac{1}{c_2}\frac{\widehat{\sigma}_1^2}{\widehat{\sigma}_2^2} \leqslant \frac{\sigma_2^2}{\sigma_1^2} \leqslant \frac{1}{c_1}\frac{\widehat{\sigma}_1^2}{\widehat{\sigma}_2^2}.$$

For the equal-tailed CI, we specify constants c_1, c_2 such that:

$$\mathbb{P}(Q < c_1) = \frac{\alpha}{2} \quad \Rightarrow \quad \mathbb{P}(Q > c_1) = 1 - \frac{\alpha}{2} \quad \Rightarrow \quad c_1 = F_{n,m;1-\alpha/2},$$

$$\mathbb{P}(Q > c_2) = \frac{\alpha}{2} \quad \Rightarrow \quad c_2 = F_{n,m;\alpha/2}.$$

Therefore, we arrive at the following equal-tailed CI:

$$\left[\frac{1}{F_{n,m;\alpha/2}}\frac{\widehat{\sigma}_1^2}{\widehat{\sigma}_2^2}, \frac{1}{F_{n,m;1-\alpha/2}}\frac{\widehat{\sigma}_1^2}{\widehat{\sigma}_2^2}\right] = \left[F_{m,n;1-\alpha/2}\frac{\widehat{\sigma}_1^2}{\widehat{\sigma}_2^2}, F_{m,n;\alpha/2}\frac{\widehat{\sigma}_1^2}{\widehat{\sigma}_2^2}\right]. \quad \square$$

Example 4.19. The means μ_1 and μ_2 are unknown and we want to construct an equal-tailed CI for the variance ratio $\frac{\sigma_1^2}{\sigma_2^2}$. We know that $S_1^2 = \frac{1}{n-1}\sum_{i=1}^n \left(X_i - \overline{X}\right)^2$ and $S_2^2 = \frac{1}{m-1}\sum_{i=1}^m \left(Y_i - \overline{Y}\right)^2$ are the UMVUEs of σ_1^2 and σ_2^2 respectively. We also know that $V_1 = \frac{n-1}{\sigma_1^2}S_1^2 \sim \chi^2_{n-1}$ and $V_2 = \frac{m-1}{\sigma_2^2}S_2^2 \sim \chi^2_{m-1}$. Since the 2 samples are independent, we infer that the random variables V_1 and V_2 are also independent. Therefore, we construct the following pivotal quantity:

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

In exactly the same manner as in the previous example, $c_1 = F_{n-1,m-1;1-\alpha/2}$ and $c_2 = F_{n-1,m-1;\alpha/2}$. Therefore, we arrive at the following equal-tailed CI:

$$\left[\frac{1}{F_{n-1,m-1;\alpha/2}}\frac{S_1^2}{S_2^2}, \frac{1}{F_{n-1,m-1;1-\alpha/2}}\frac{S_1^2}{S_2^2}\right] = \left[F_{m-1,n-1;1-\alpha/2}\frac{S_1^2}{S_2^2}, F_{m-1,n-1;\alpha/2}\frac{S_1^2}{S_2^2}\right]. \quad \square$$

4.5 Asymptotic Confidence Intervals

Definition 4.6. For given $\alpha \in (0,1)$, we consider a random interval of the form $\mathcal{I}_{g(\vartheta);1-\alpha}(X) = [L_n(X), U_n(X)]$ such that:

$$\lim_{n\to\infty}\inf_{\vartheta\in\Theta}\mathbb{P}_\vartheta\left[L_n(X) \leqslant g(\vartheta) \leqslant U_n(X)\right] = 1-\alpha,$$

which is called a $100(1-\alpha)\%$ *asymptotic confidence interval* for $g(\vartheta)$.

Note 4.11. For the construction of an asymptotic CI it suffices to determine a sequence of random variables $Q_n(X, g(\vartheta))$ which depends on the value of the parametric function $g(\vartheta)$ and converges in distribution to some random variable whose distribution doesn't depend on the value of ϑ. For this reason, we make use of the asymptotic results presented in paragraph 3.10.

Example 4.20. Let $X_1, \ldots, X_n \sim \text{Pareto}(k,\lambda)$ be a random sample with $k > 0$, known $\lambda > 2$ and $F(x;k) = 1 - \left(\frac{k}{x}\right)^\lambda$ for $x \geqslant k$. According to example 3.34 (page 70), we know that $n\left[X_{(1)} - k\right] \xrightarrow{d} Y \sim \text{Exp}(\lambda/k)$. According to Slutsky's theorem, it follows that:

$$Q_n = n\left[\frac{X_{(1)}}{k} - 1\right] \xrightarrow{d} \frac{1}{k}Y = V \sim \text{Exp}(\lambda).$$

We solve the inequality $c_1 \leqslant Q_n \leqslant c_2$ with respect to k:

$$c_1 \leqslant n\left[\frac{X_{(1)}}{k} - 1\right] \leqslant c_2 \quad \Leftrightarrow \quad \frac{X_{(1)}}{1+c_2/n} \leqslant k \leqslant \frac{X_{(1)}}{1+c_1/n}.$$

For the asymptotic equal-tailed CI, we specify constants c_1, c_2 such that:

$$\lim_{n\to\infty} \mathbb{P}(Q_n < c_1) = \frac{\alpha}{2} \quad \Rightarrow \quad \mathbb{P}(V < c_1) = \frac{\alpha}{2} \quad \Rightarrow \quad c_1 = -\frac{1}{\lambda}\log\left(1 - \frac{\alpha}{2}\right),$$

$$\lim_{n\to\infty} \mathbb{P}(Q_n > c_2) = \frac{\alpha}{2} \quad \Rightarrow \quad \mathbb{P}(V > c_2) = \frac{\alpha}{2} \quad \Rightarrow \quad c_2 = -\frac{1}{\lambda}\log\frac{\alpha}{2}.$$

Therefore, we arrive at the following asymptotic equal-tailed CI:

$$\left[\frac{X_{(1)}}{1 - \log\left(\alpha/2\right)/n\lambda}, \frac{X_{(1)}}{1 - \log\left(1 - \alpha/2\right)/n\lambda}\right]. \quad \square$$

Example 4.21. Let $X_1, \ldots, X_n \sim \text{Exp}(\lambda)$ be a random sample. According to example 3.33 (page 69), we know that $\sqrt{n}\left(\frac{1}{\overline{X}_n} - \lambda\right) \xrightarrow{d} Y \sim \mathcal{N}\left(0, \lambda^2\right)$. According to Slutsky's theorem, it follows that:

$$Q_n = \sqrt{n}\left(\frac{1}{\lambda\overline{X}_n} - 1\right) \xrightarrow{d} \frac{1}{\lambda}Y = Z \sim \mathcal{N}\left(0, 1\right).$$

We solve the inequality $c_1 \leqslant Q_n \leqslant c_2$ with respect to λ:

$$c_1 \leqslant \sqrt{n}\left(\frac{1}{\lambda\overline{X}_n} - 1\right) \leqslant c_2 \quad \Leftrightarrow \quad \frac{1}{\overline{X}_n\left(1 + c_2/\sqrt{n}\right)} \leqslant \lambda \leqslant \frac{1}{\overline{X}_n\left(1 + c_1/\sqrt{n}\right)}.$$

For the asymptotic equal-tailed CI, we specify constants c_1, c_2 such that:

$$\lim_{n\to\infty} \mathbb{P}(Q_n < c_1) = \frac{\alpha}{2} \quad \Rightarrow \quad \mathbb{P}(Z > c_1) = 1 - \frac{\alpha}{2} \quad \Rightarrow \quad c_1 = Z_{1-\alpha/2} = -Z_{\alpha/2},$$

$$\lim_{n\to\infty} \mathbb{P}(Q_n > c_2) = \frac{\alpha}{2} \quad \Rightarrow \quad \mathbb{P}(Z > c_2) = \frac{\alpha}{2} \quad \Rightarrow \quad c_2 = Z_{\alpha/2}.$$

Therefore, we arrive at the following asymptotic equal-tailed CI:

$$\left[\frac{1}{\overline{X}_n\left(1 + Z_{\alpha/2}/\sqrt{n}\right)}, \frac{1}{\overline{X}_n\left(1 - Z_{\alpha/2}/\sqrt{n}\right)}\right]. \quad \square$$

Example 4.22. Let $X_1, \ldots, X_n \sim \text{Bernoulli}(p)$ be a random sample. According to the central limit theorem, we know that $\sqrt{n}\left(\overline{X}_n - p\right) \xrightarrow{d} Y \sim \mathcal{N}\left(0, p(1-p)\right)$. According to the weak law of large numbers, we also know that $\overline{X}_n \xrightarrow{p} p$. According to Slutsky's theorem, it follows that:

$$Q_n = \frac{\sqrt{n}\left(\overline{X}_n - p\right)}{\sqrt{\overline{X}_n\left(1 - \overline{X}_n\right)}} \xrightarrow{d} \frac{1}{\sqrt{p(1-p)}}Y = Z \sim \mathcal{N}(0, 1).$$

We solve the inequality $c_1 \leqslant Q_n \leqslant c_2$ with respect to p:

$$c_1 \leqslant \frac{\sqrt{n}\left(\overline{X}_n - p\right)}{\sqrt{\overline{X}_n\left(1-\overline{X}_n\right)}} \leqslant c_2 \quad \Leftrightarrow$$

$$\overline{X}_n - c_2\sqrt{\frac{1}{n}\overline{X}_n\left(1-\overline{X}_n\right)} \leqslant p \leqslant \overline{X}_n - c_1\sqrt{\frac{1}{n}\overline{X}_n\left(1-\overline{X}_n\right)}.$$

In exactly the same manner as in the previous example, it follows that $c_1 = -Z_{\alpha/2}$, $c_2 = Z_{\alpha/2}$. Therefore, we arrive at the following asymptotic equal-tailed CI:

$$\left[\overline{X}_n - Z_{\alpha/2}\sqrt{\frac{1}{n}\overline{X}_n\left(1-\overline{X}_n\right)}, \overline{X}_n + Z_{\alpha/2}\sqrt{\frac{1}{n}\overline{X}_n\left(1-\overline{X}_n\right)}\right].$$

We note that the above asymptotic CI for $p \in (0,1)$ tends to cover wider and wider intervals outside of the parameter space as p tends towards 0 or 1. □

Example 4.23. Let $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ be a random sample. We want to construct an asymptotic CI for the mean μ. According to example 3.32 (page 68), we know that:

$$Q_n = \frac{\overline{X}_n - \mu}{S_n/\sqrt{n}} \xrightarrow{d} Z \sim \mathcal{N}(0,1).$$

In exactly the same manner as in the previous example, it follows that $c_1 = -Z_{\alpha/2}$, $c_2 = Z_{\alpha/2}$. Therefore, we arrive at the following asymptotic equal-tailed CI:

$$\left[\overline{X}_n - Z_{\alpha/2}\frac{S_n}{\sqrt{n}}, \overline{X}_n + Z_{\alpha/2}\frac{S_n}{\sqrt{n}}\right]. \quad \square$$

Chapter 5

Statistical Hypothesis Testing

5.1 Introduction

In statistical data analysis we are often called to make a decision about whether a formulated statistical hypothesis is mistaken or not. This claim whose validity is called into question is called the *null hypothesis* and is denoted by H_0. The designation of the null hypothesis leads to the formulation of an *alternative hypothesis*, which is denoted by H_1. The decision we are called to make is whether to reject the null hypothesis H_0 or not in favor of the alternative hypothesis H_1. The statistic according to which we make a proper decision is called a *statistical hypothesis test.*

More precisely, the statistical hypotheses H_0 and H_1 concern the CDF F of a random variable X, which belongs to a class of CDFs $\mathcal{F}$. The hypotheses H_0 and H_1 take the form $H_0 : F \in \mathcal{F}_0$ vs. $H_1 : F \in \mathcal{F}_1$, where $\mathcal{F}_0, \mathcal{F}_1 \subset \mathcal{F}$ with $\mathcal{F}_0 \cap \mathcal{F}_1 = \varnothing$. The decision we make is based on a sample x from the CDF F.

In the framework of parametric statistics, the class of CDFs $\mathcal{F}$ is parameterized by an unknown parameter ϑ, so it takes the form $\mathcal{F}_\vartheta = \{F(x;\vartheta) : \vartheta \in \Theta\}$. Hence, the hypotheses H_0 and H_1 specifically concern the value of the unknown parameter ϑ. In other words, the hypotheses H_0 and H_1 take the form $H_0 : \vartheta \in \Theta_0$ vs. $H_1 : \vartheta \in \Theta_1$, where $\Theta_0, \Theta_1 \subset \Theta$ with $\Theta_0 \cap \Theta_1 = \varnothing$.

A statistical hypothesis is called *simple* if it fully determines the CDF $F(x;\vartheta)$. For example, the null hypothesis $H_0 : \vartheta \in \Theta_0$ is simple if the set Θ_0 coincides with a singleton $\{\vartheta_0\}$. Otherwise, it's called a *composite* hypothesis.

Example 5.1. i. $H_0 : X \sim \mathcal{N}(0, 1)$ vs. $H_1 : X \sim \text{Laplace}(0, 1)$ is a test of a simple hypothesis vs. a simple hypothesis.

ii. If $X \sim \mathcal{N}(\mu, \sigma^2)$ with known σ^2, then $H_0 : \mu = \mu_0$ vs. $H_1 : \mu = \mu_1$ is a test of a simple hypothesis vs. a simple hypothesis, since $\Theta_0 = \{\mu_0\}$ and $\Theta_1 = \{\mu_1\}$.

iii. If $X \sim \mathcal{N}(\mu, \sigma^2)$ with known σ^2, then $H_0 : \mu = \mu_0$ vs. $H_1 : \mu > \mu_0$ is a test of a simple hypothesis vs. a *one-sided* composite hypothesis, since $\Theta_0 = \{\mu_0\}$ and $\Theta_1 = (\mu_0, \infty)$.

iv. If $X \sim \mathcal{N}(\mu, \sigma^2)$ with known σ^2, then $H_0 : \mu = \mu_0$ vs. $H_1 : \mu \neq \mu_0$ is a test of a simple hypothesis vs. a *tow-sided* composite hypothesis, since $\Theta_0 = \{\mu_0\}$ and $\Theta_1 = \mathbb{R} \setminus \{\mu_0\}$.

v. If $X \sim \mathcal{N}(\mu, \sigma^2)$ with σ^2 unknown, then $H_0 : \mu = \mu_0$ vs. $H_1 : \mu = \mu_1$ is a test of a composite hypothesis vs. a composite hypothesis, since $\Theta_0 = \{\mu_0\} \times (0, \infty)$ and $\Theta_1 = \{\mu_1\} \times (0, \infty)$.

Definition 5.1. A statistic $\varphi(X) : S \to [0, 1]$ which determines the decision about whether to reject a null hypothesis H_0 or not in favor of an alternative hypothesis H_1 is called a *statistical test.* If the function φ takes the following form:

$$\varphi(x) = \begin{cases} 1, & \text{reject } H_0 \\ 0, & \text{don't reject } H_0 \end{cases},$$

then the test is called *non-randomized.* Otherwise, if it takes the following form:

$$\varphi(x) = \begin{cases} 1, & \text{reject } H_0 \\ \gamma, & \text{reject } H_0 \text{ with probability } \gamma \in (0, 1) \\ 0, & \text{don't reject } H_0 \end{cases},$$

then the test is called *randomized.*

Note 5.1. A non-randomized test partitions the support of the distribution of the sample x into 2 disjoint subsets R and A, i.e. $S = R \cup A$ with $R \cap A = \varnothing$. It holds that:

- If $x \in R$, then we reject the null hypothesis H_0. The subset R is called the *critical region* (or rejection region) of the test.
- If $x \in A$, then we don't reject the null hypothesis H_0. The subset $A = S \setminus R$ is called the *acceptance region* of the test.

Note 5.2. When we conduct a hypothesis test, then we might make the correct decision or we might commit one of the following 2 errors:

- Type I Error $\to$ Reject H_0 when it's in fact true. It holds that:

$$\mathbb{P}_\vartheta(\text{Type I Error}) = \mathbb{P}_\vartheta(X \in R), \quad \vartheta \in \Theta_0.$$

- Type II Error → Fail to reject H_0 when it's in fact untrue. It holds that:

$$\mathbb{P}_\vartheta(\text{Type II Error}) = \mathbb{P}_\vartheta(X \in A), \quad \vartheta \in \Theta_1.$$

	Do not reject H_0	Reject H_0
H_0 True	True Negative	Type I Error
H_0 Not True	Type II Error	True Positive

Table 5.1: Summary of a Hypothesis Test's Possible Outcomes

Definition 5.2. i. The following function:

$$\begin{aligned}\beta_\varphi(\vartheta) &= \mathbb{P}_\vartheta(\text{Correct Rejection of } H_0) = \mathbb{P}_\vartheta(X \in R)\\ &= 1 - \mathbb{P}_\vartheta(\text{Type II Error}), \quad \vartheta \in \Theta_1,\end{aligned}$$

is called the *power* of a test φ.

ii. The following function:

$$\begin{aligned}\pi_\varphi(\vartheta) &= \mathbb{E}_\vartheta\left[\varphi(X)\right] = \mathbb{P}_\vartheta(\text{Reject } H_0) = \mathbb{P}_\vartheta(X \in R)\\ &= \begin{cases}\mathbb{P}_\vartheta(\text{Type I Error}), & \vartheta \in \Theta_0\\ \beta_\varphi(\vartheta), & \vartheta \in \Theta_1\end{cases},\end{aligned}$$

is called the *power function* of a test φ.

iii. The following quantity:

$$\sup_{\vartheta \in \Theta_0} \pi_\varphi(\vartheta) = \sup_{\vartheta \in \Theta_0} \mathbb{P}_\vartheta(X \in R) = \sup_{\vartheta \in \Theta_0} \mathbb{P}_\vartheta(\text{Type I Error}),$$

is called the *size* of a test φ.

Note 5.3. For finite sample sizes it's not possible to minimize $\mathbb{P}_\vartheta(\text{Type I Error})$ and $\mathbb{P}_\vartheta(\text{Type II Error})$ simultaneously. In fact, as one decreases the other usually increases. Because the null hypothesis H_0 is the hypothesis we lean on when designing the test, its erroneous rejection usually entails the largest risk. For this reason, we prespecify an upper limit α for the probability of committing a type I error, and we try to minimize the probability of committing a type II error, or equivalently we try to maximize the power of the test under this constraint. In other words, we want to maximize the function β_φ under the constraint $\sup_{\vartheta \in \Theta_0} \pi_\varphi(\vartheta) \leqslant \alpha$.

Definition 5.3. The upper limit α on the size of a test is called the *statistical significance level* of the test.

Definition 5.4. A test φ of size α, i.e. for which it holds that $\sup_{\vartheta \in \Theta_0} \pi_\varphi(\vartheta) = \alpha$, is called a *uniformly most powerful* (UMP) test if for every other test φ^* at significance level α it holds that $\beta_\varphi(\vartheta) \geqslant \beta_{\varphi^*}(\vartheta)\ \forall \vartheta \in \Theta_1$.

Note 5.4. i. If the distribution of the sample is continuous and the null hypothesis is simple, i.e. $\Theta_0 = \{\vartheta_0\}$, it's easy to determine a test of size α, since it follows that $\sup_{\vartheta \in \Theta_0} \pi_\varphi(\vartheta) = \mathbb{P}_{\vartheta_0}(X \in R)$.

ii. If the distribution of the sample is discrete, it's not always feasible to construct a non-randomized test of a specific size. In this case, randomized tests are usually utilized.

Example 5.2. Let $X_1, \ldots, X_n \sim \mathcal{U}(0, \vartheta)$ be a random sample. If the critical region of the test for the hypotheses $H_0 : \vartheta = 0.5$ vs. $H_1 : \vartheta = 0.25$ at statistical significance level $\alpha = 5\%$ is of the form $R = \left\{x \in (0, \vartheta)^n : x_{(n)} < c\right\}$ and it holds that $\mathbb{P}_{0.25}(\text{Type II Error}) = 0.2$, then we want to specify the constant c and the sample size n. For $x \in (0, \vartheta)$, we know that $F_{X_{(n)}}(x) = \left(\frac{x}{\vartheta}\right)^n$. First, we calculate that:

$$\mathbb{E}_{0.5}\left[\varphi(X)\right] = \mathbb{P}_{0.5}(X \in R) = \mathbb{P}_{0.5}\left[X_{(n)} < c\right] = (2c)^n = \alpha \quad \Rightarrow \quad c = \frac{1}{2}0.05^{1/n}.$$

Furthermore, we know that:

$$\mathbb{P}_{0.25}(\text{Type II Error}) = \mathbb{P}_{0.25}(X \notin R) = \mathbb{P}_{0.25}(X_{(n)} \geqslant c) = 1 - (4c)^n = 0.2 \quad \Rightarrow$$

$$c = \frac{1}{4}0.8^{1/n} \quad \Rightarrow \quad \frac{1}{2}0.05^{1/n} = \frac{1}{4}0.8^{1/n} \quad \Rightarrow \quad 16^{1/n} = 2 \quad \Rightarrow$$

$$n = 4 \quad \Rightarrow \quad c \approx 0.24. \quad \square$$

5.2 Fundamental Neyman - Pearson Lemma

Theorem 5.1. (Fundamental Neyman - Pearson Lemma) We want to specify a test of the simple hypotheses $H_0 : \vartheta = \vartheta_0$ vs. $H_1 : \vartheta = \vartheta_1$.

- **Existence of UMP Test**: For given $\alpha \in (0, 1)$, the following statistic:

$$\varphi(x) = \begin{cases} 1, & \mathcal{L}(\vartheta_0 \mid x)/\mathcal{L}(\vartheta_1 \mid x) < c \\ \gamma, & \mathcal{L}(\vartheta_0 \mid x)/\mathcal{L}(\vartheta_1 \mid x) = c \\ 0, & \mathcal{L}(\vartheta_0 \mid x)/\mathcal{L}(\vartheta_1 \mid x) > c \end{cases},$$

  where $c > 0$ and $\gamma \in [0, 1]$ are constants such that $\pi_\varphi(\vartheta_0) = \alpha$, is a UMP test of size α.

- **Uniqueness of UMP Test**: If φ^* is another UMP test at significance level α, then it follows that $\varphi^*(x) = \varphi(x)$ for all $x \in S$ such that $\mathcal{L}(\vartheta_0 \mid x) \neq c\mathcal{L}(\vartheta_1 \mid x)$.

Proof. Without loss of generality, assume that the distribution of the sample is continuous. Let φ^* denote another test of the simple hypotheses $H_0 : \vartheta = \vartheta_0$ vs. $H_1 : \vartheta = \vartheta_1$ at significance level α. Since $\varphi^*(x) \in [0, 1]$, we notice that:

$$\varphi(x) = \begin{cases} 1, & \mathcal{L}(\vartheta_0 \mid x) - c\mathcal{L}(\vartheta_1 \mid x) < 0 \\ 0, & \mathcal{L}(\vartheta_0 \mid x) - c\mathcal{L}(\vartheta_1 \mid x) > 0 \end{cases} \Rightarrow$$

$$[\varphi^*(x) - \varphi(x)] [\mathcal{L}(\vartheta_0 \mid x) - c\mathcal{L}(\vartheta_1 \mid x)] \geqslant 0.$$

Then, we calculate that:

$$\begin{aligned} 0 &\leqslant \int_S [\varphi^*(x) - \varphi(x)] [\mathcal{L}(\vartheta_0 \mid x) - c\mathcal{L}(\vartheta_1 \mid x)] \, dx \\ &= \int_{R^*} \mathcal{L}(\vartheta_0 \mid x) - c\mathcal{L}(\vartheta_1 \mid x) dx - \int_R \mathcal{L}(\vartheta_0 \mid x) - c\mathcal{L}(\vartheta_1 \mid x) dx \\ &= \pi_{\varphi^*}(\vartheta_0) - c\beta_{\varphi^*}(\vartheta_1) - [\pi_\varphi(\vartheta_0) - c\beta_\varphi(\vartheta_1)] \\ &= \pi_{\varphi^*}(\vartheta_0) - \alpha + c [\beta_\varphi(\vartheta_1) - \beta_{\varphi^*}(\vartheta_1)] \leqslant c [\beta_\varphi(\vartheta_1) - \beta_{\varphi^*}(\vartheta_1)] , \end{aligned} \tag{5.1}$$

since it holds that $\pi_{\varphi^*}(\vartheta_0) \leqslant \alpha$. Hence, we deduce that $\beta_\varphi(\vartheta_1) \geqslant \beta_{\varphi^*}(\vartheta_1)$ because of the fact that $c > 0$. Since the test φ has greater power than any other arbitrary test at significance level α, we conclude that φ is a UMP test of size α.

Now, let φ^* denote another UMP test of the simple hypotheses $H_0 : \vartheta = \vartheta_0$ vs. $H_1 : \vartheta = \vartheta_1$ at significance level α. Then, it must hold that $\beta_{\varphi^*}(\vartheta_1) = \beta_\varphi(\vartheta_1)$. According to equation 5.1, we know that:

$$0 \leqslant \pi_{\varphi^*}(\vartheta_0) - \alpha + c [\beta_\varphi(\vartheta_1) - \beta_{\varphi^*}(\vartheta_1)] = \pi_{\varphi^*}(\vartheta_0) - \alpha,$$

which implies that $\pi_{\varphi^*}(\vartheta_0) \geqslant \alpha$. Hence, we infer that the UMP test φ^* is also of size α and the inequality given by equation 5.1 actually holds as an equality. Since the function $[\varphi^*(x) - \varphi(x)] [\mathcal{L}(\vartheta_0 \mid x) - c\mathcal{L}(\vartheta_1 \mid x)]$ is non-negative and its integral over S is 0, this implies that the function is actually 0 over S. Therefore, we conclude that $\varphi^*(x) = \varphi(x)$ for all $x \in S$ such that $\mathcal{L}(\vartheta_0 \mid x) \neq c\mathcal{L}(\vartheta_1 \mid x)$. □

Note 5.5. We usually work on the log scale, so we define the following UMP test:

$$\varphi(x) = \begin{cases} 1, & \ell(\vartheta_0 \mid x) - \ell(\vartheta_1 \mid x) < c \\ \gamma, & \ell(\vartheta_0 \mid x) - \ell(\vartheta_1 \mid x) = c \\ 0, & \ell(\vartheta_0 \mid x) - \ell(\vartheta_1 \mid x) > c \end{cases} ,$$

where $c > 0$ and $\gamma \in [0, 1]$ are constants such that $\pi_\varphi(\vartheta_0) = \alpha$.

Note 5.6. In order to specify the constant c, we follow a similar procedure to the pivotal quantity method for the construction of CIs. More precisely, we solve the inequality $\ell(\vartheta_0 \mid X) - \ell(\vartheta_1 \mid X) < c$ with respect to some statistic $T(X)$ whose distribution doesn't depend on the value ϑ_0 under the null hypothesis $H_0 : \vartheta = \vartheta_0$. The statistic $T(X)$ is called a *test statistic*.

Example 5.3. Let $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ be a random sample with known σ^2. We want to find a UMP test for the hypotheses $H_0 : \mu = \mu_0$ vs. $H_1 : \mu = \mu_1$ with $\mu_1 > \mu_0$ and calculate its power. We know that:

$$\ell(\mu \mid x) = -\frac{n}{2} \log\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2.$$

The critical region of the test is given as follows:

$$\ell(\mu_0 \mid x) - \ell(\mu_1 \mid x) < c \quad \Leftrightarrow \quad \frac{1}{2\sigma^2} \left[\sum_{i=1}^{n} (x_i - \mu_1)^2 - \sum_{i=1}^{n} (x_i - \mu_0)^2 \right] < c \quad \Leftrightarrow$$

$$n\left(\mu_1^2 - \mu_0^2\right) - 2(\mu_1 - \mu_0) \sum_{i=1}^{n} x_i < c^* = 2\sigma^2 c \quad \Leftrightarrow$$

$$2(\mu_1 - \mu_0) \sum_{i=1}^{n} x_i > c^{**} = n\left(\mu_1^2 - \mu_0^2\right) - c^* \quad \overset{\mu_1 > \mu_0}{\Leftrightarrow} \quad \overline{x} > c^{***} = \frac{c^{**}}{2n(\mu_1 - \mu_0)} \quad \Leftrightarrow$$

$$T(x) = \frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}} > c_\alpha = \frac{c^{***} - \mu}{\sigma/\sqrt{n}}.$$

It remains to specify the constant c_α, so that the test is of size α, i.e. so that $\pi_\varphi(\mu_0) = \alpha$. Under the null hypothesis H_0, i.e. given that $X_1, \ldots, X_n \sim \mathcal{N}(\mu_0, \sigma^2)$, we know that $T(X) = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1)$. Therefore, we calculate that:

$$\mathbb{E}_{\mu_0}[\varphi(X)] = \alpha \quad \Rightarrow \quad \mathbb{P}_{\mu_0}[T(X) > c_\alpha] = \alpha \quad \Rightarrow \quad c_\alpha = Z_\alpha.$$

According to the fundamental Neyman - Pearson lemma, we arrive at the following UMP test:

$$\varphi(x) = \begin{cases} 1, & \frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}} > Z_\alpha \\ 0, & \frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}} \leqslant Z_\alpha \end{cases}.$$

The power of the above test is calculated as follows:

$$\begin{aligned} \beta_\varphi(\mu_1) &= \mathbb{P}_{\mu_1}(X \in R) = \mathbb{P}_{\mu_1}\left(\frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} > Z_\alpha \right) \\ &= \mathbb{P}_{\mu_1}\left(\frac{\overline{X} - \mu_1}{\sigma/\sqrt{n}} > Z_\alpha + \frac{\mu_0 - \mu_1}{\sigma/\sqrt{n}} \right) = 1 - \Phi\left(Z_\alpha - \frac{\mu_1 - \mu_0}{\sigma/\sqrt{n}} \right), \end{aligned}$$

since $\frac{\overline{X} - \mu_1}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1)$ under H_1, i.e. given that $X_1, \ldots, X_n \sim \mathcal{N}(\mu_1, \sigma^2)$. □

Note 5.7. We observe that the critical region of the previous test doesn't depend on the value μ_1, but only on the direction of the inequality $\mu_1 > \mu_0$, which we used to specify it. In other words, the test is UMP for every simple alternative hypothesis $H_1 : \mu = \mu_1^*$ with $\mu_1^* > \mu_0$. Hence, we infer that it's also UMP for the one-sided alternative hypothesis $H_1^* : \mu > \mu_0$. More generally, the following statements hold:

i. If the critical region of a UMP test φ for the simple hypotheses $H_0 : \vartheta = \vartheta_0$ vs. $H_1 : \vartheta = \vartheta_1$ with $\vartheta_1 > \vartheta_0$ doesn't depend on the value ϑ_1, then the test φ is also UMP for the hypotheses $H_0 : \vartheta = \vartheta_0$ vs. $H_1^* : \vartheta > \vartheta_0$.

ii. If the critical region of a UMP test φ for the simple hypotheses $H_0 : \vartheta = \vartheta_0$ vs. $H_1 : \vartheta = \vartheta_1$ with $\vartheta_1 < \vartheta_0$ doesn't depend on the value ϑ_1, then the test φ is also UMP for the hypotheses $H_0 : \vartheta = \vartheta_0$ vs. $H_1^* : \vartheta < \vartheta_0$.

Note 5.8. We observe that the power of the previous test is a strictly increasing function of the statistical significance level α, a strictly increasing function of the mean difference $\mu_1 - \mu_0$, a strictly decreasing function of the variance σ^2 of the observations in the sample and a strictly increasing function of the sample size n.

Example 5.4. In the setting of the previous example, we want to specify the smallest sample size n, so that the type II error is at most equal to 0.01, if it's known that $\sigma^2 = 4$, $\mu_1 = \mu_0 + 2$ and $\alpha = 1\%$. We demand the following:

$$\mathbb{P}(\text{Type II Error}) = 1 - \beta_\varphi(\mu_1) = \Phi\left(Z_\alpha - \frac{\mu_1 - \mu_0}{\sigma/\sqrt{n}}\right) \leqslant 0.01 \quad \Rightarrow$$

$$Z_{0.01} - \sqrt{n} \leqslant \Phi^{-1}(0.01) = Z_{0.99} = -Z_{0.01} \quad \Rightarrow \quad n \geqslant 4Z_{0.01}^2 \approx 21.65.$$

Therefore, the smallest sample size we require is $n = 22$. □

Example 5.5. Let $X_1, \ldots, X_n \sim \text{Exp}(\lambda)$ be a random sample. We want to find a UMP test for the hypotheses $H_0 : \lambda = \lambda_0$ vs. $H_1 : \lambda < \lambda_0$ and calculate its type II error. We consider the simple alternative hypothesis $H_1^* : \lambda = \lambda_1$ with $\lambda_1 < \lambda_0$, so that we can apply the fundamental Neyman - Pearson lemma. We know that:

$$\ell(\lambda \mid x) = n \log \lambda - \lambda \sum_{i=1}^{n} x_i.$$

The critical region of the test is given as follows:

$$\ell(\lambda_0 \mid x) - \ell(\lambda_1 \mid x) < c \quad \Leftrightarrow \quad n\left(\log \lambda_0 - \log \lambda_1\right) - (\lambda_0 - \lambda_1) \sum_{i=1}^{n} x_i < c \quad \Leftrightarrow$$

$$(\lambda_0 - \lambda_1) \sum_{i=1}^{n} x_i > c^* = n\left(\log \lambda_0 - \log \lambda_1\right) - c \quad \overset{\lambda_1 < \lambda_0}{\Leftrightarrow}$$

$$\sum_{i=1}^{n} x_i > c^{**} = \frac{c^*}{\lambda_0 - \lambda_1} \quad \Leftrightarrow \quad T(x) = 2\lambda_0 \sum_{i=1}^{n} x_i > c_\alpha = 2\lambda_0 c^{**}.$$

Under the null hypothesis H_0, i.e. given that $X_1, \ldots, X_n \sim \text{Exp}(\lambda_0)$, we know that $T(X) = 2\lambda_0 \sum_{i=1}^{n} X_i \sim \chi^2_{2n}$. Therefore, we calculate that:

$$\mathbb{E}_{\lambda_0}[\varphi(X)] = \alpha \quad \Rightarrow \quad \mathbb{P}_{\lambda_0}[T(X) > c_\alpha] = \alpha \quad \Rightarrow \quad c_\alpha = \chi^2_{2n;\alpha}.$$

According to the fundamental Neyman - Pearson lemma, we arrive at the following UMP test:

$$\varphi(x) = \begin{cases} 1, & 2\lambda_0 \sum_{i=1}^{n} x_i > \chi^2_{2n;\alpha} \\ 0, & 2\lambda_0 \sum_{i=1}^{n} x_i \leqslant \chi^2_{2n;\alpha} \end{cases}.$$

Since the critical region of the test doesn't depend on the specific value λ_1, but only on the direction of the inequality $\lambda_1 < \lambda_0$, which we used to specify it, we infer that it's also UMP for the one-sided alternative hypothesis $H_1 : \lambda < \lambda_0$. For $\lambda < \lambda_0$, we calculate that:

$$\begin{aligned} \mathbb{P}_\lambda(\text{Type II Error}) = \mathbb{P}_\lambda(X \notin R) &= \mathbb{P}_\lambda\left(2\lambda_0 \sum_{i=1}^{n} X_i \leqslant \chi^2_{2n;\alpha}\right) \\ &= \mathbb{P}_\lambda\left(2\lambda \sum_{i=1}^{n} X_i \leqslant \frac{\lambda}{\lambda_0}\chi^2_{2n;\alpha}\right) = F_{\chi^2_{2n}}\left(\frac{\lambda}{\lambda_0}\chi^2_{2n;\alpha}\right). \end{aligned}$$

We observe that the type II error is a strictly increasing function of λ, i.e. it increases as λ tends towards the value λ_0. □

Example 5.6. Let X be a sample of size 1 with $f(x;\vartheta) = 1 + \vartheta(x - 0.5)$ for $\vartheta \in (-2, 2)$ and $x \in (0, 1)$. We want to find a UMP test for the hypotheses $H_0 : \vartheta = 0$ vs. $H_1 : \vartheta = 1$ at statistical significance level $\alpha = 10\%$ and calculate its power. The critical region of the test is given as follows:

$$\frac{\mathcal{L}(0 \mid x)}{\mathcal{L}(1 \mid x)} < c \quad \Leftrightarrow \quad \frac{1}{1 + x - 0.5} < c \quad \Leftrightarrow \quad x > c_\alpha = \frac{1}{c} - \frac{1}{2}.$$

Under the null hypothesis H_0, i.e. given that $\vartheta = 0$, we observe that $X \sim \mathcal{U}(0, 1)$. Therefore, we calculate that:

$$\mathbb{E}_0[\varphi(X)] = \alpha \quad \Rightarrow \quad \mathbb{P}_0(X > c_\alpha) = \alpha \quad \Rightarrow \quad c_\alpha = 1 - \alpha = 0.9.$$

According to the fundamental Neyman - Pearson lemma, we arrive at the following UMP test:

$$\varphi(x) = \begin{cases} 1, & x > 0.9 \\ 0, & x \leqslant 0.9 \end{cases}.$$

Finally, we calculate that:

$$\beta_\varphi(1) = \mathbb{P}_1(X \in R) = \mathbb{P}_1(X > 0.9) = \int_{0.9}^{1} (1 + x - 0.5)\, dx = 0.145. \quad \square$$

Example 5.7. Let X be a sample of size 1. We want to find a UMP test for the hypotheses $H_0 : X \sim \mathcal{N}(0,1)$ vs. $H_1 : X \sim \text{Laplace}\left(0, \frac{1}{2}\right)$ at statistical significance level $\alpha = 2\%$ and calculate its power. We know that:

$$L_0(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}, \quad L_1(x) = \frac{1}{4} e^{-|x|/2}.$$

The critical region of the test is given as follows:

$$\ell_0(x) - \ell_1(x) < c \quad \Leftrightarrow \quad -\frac{1}{2}\log(2\pi) - \frac{x^2}{2} + 2\log 2 + \frac{|x|}{2} < c \quad \Leftrightarrow$$

$$x^2 - |x| > c^* = 2\left[2\log 2 - \frac{1}{2}\log(2\pi) - c\right] \quad \Leftrightarrow$$

$$|x| > \frac{1 + \sqrt{1 + 4c^*}}{2} = c_\alpha \quad \text{or} \quad |x| < \frac{1 - \sqrt{1 + 4c^*}}{2} = 1 - \frac{1 + \sqrt{1 + 4c^*}}{2} = 1 - c_\alpha.$$

Under the null hypothesis H_0, i.e. given that $X \sim \mathcal{N}(0,1)$, we observe that:

$$\begin{aligned}\mathbb{P}_0\left(|X| > c_\alpha\right) &= \mathbb{P}_0\left(X > c_\alpha\right) + \mathbb{P}_0\left(X < -c_\alpha\right) = 1 - \Phi(c_\alpha) + \Phi(-c_\alpha) \\ &= 1 - \Phi(c_\alpha) + 1 - \Phi(c_\alpha) = 2\left[1 - \Phi(c_\alpha)\right] \leqslant \alpha \quad \Rightarrow\end{aligned}$$

$$c_\alpha \geqslant \Phi^{-1}\left(1 - \frac{\alpha}{2}\right) = Z_{\alpha/2} = Z_{0.01} \approx 2.33 \quad \Rightarrow \quad 1 - c_\alpha < 0 \quad \Rightarrow$$

$$\mathbb{P}_0\left(|X| < 1 - c_\alpha\right) = 0 \quad \Rightarrow \quad \mathbb{P}_0\left(|X| > c_\alpha\right) = \alpha \quad \Rightarrow \quad c_\alpha \approx 2.33.$$

According to the fundamental Neyman - Pearson lemma, we arrive at the following UMP test:

$$\varphi(x) = \begin{cases} 1, & |x| > 2.33 \\ 0, & |x| \leqslant 2.33 \end{cases}.$$

Finally, we calculate that:

$$\begin{aligned}\beta_\varphi &= \mathbb{P}_1(|x| > c_\alpha) = 1 - \mathbb{P}_1(|x| \leqslant c_\alpha) = 1 - \mathbb{P}_1(-c_\alpha \leqslant x \leqslant c_\alpha) \\ &= 1 - \int_{-c_\alpha}^{c_\alpha} \frac{1}{4} e^{-|x|/2} dx = 1 - \int_{0}^{c_\alpha} \frac{1}{2} e^{-x/2} dx = e^{-c_\alpha/2} \approx 0.31. \quad \square\end{aligned}$$

Example 5.8. Let $X_1, \ldots, X_6 \sim \text{Bernoulli}(p)$ be a random sample. We want to find a UMP test for the hypotheses $H_0 : p = 0.2$ vs. $H_1 : p = 0.5$ at statistical significance level $\alpha = 5\%$. We know that:

$$\ell(p \mid x) = \log p \sum_{i=1}^{6} x_i + \log(1-p)\left(6 - \sum_{i=1}^{6} x_i\right) = \log \frac{p}{1-p} \sum_{i=1}^{6} x_i + 6\log(1-p).$$

The critical region of the test is given as follows:

$$\ell(0.2 \mid x) - \ell(0.5 \mid x) < c \quad \Leftrightarrow \quad \log \frac{0.2/(1-0.2)}{0.5/(1-0.5)} \sum_{i=1}^{6} x_i + 6\log \frac{1-0.2}{1-0.5} < c \quad \Leftrightarrow$$

$$\log 4 \sum_{i=1}^{6} x_i > c^* = 6\log \frac{8}{5} - c \quad \Leftrightarrow \quad T(x) = \sum_{i=1}^{6} x_i > c_\alpha = \frac{c^*}{\log 4}.$$

Under the null hypothesis H_0, i.e. given that $X_1, \ldots, X_6 \sim \text{Bernoulli}(0.2)$, we know that $T(X) = \sum_{i=1}^{6} X_i \sim \text{Bin}(6, 0.2)$. Therefore, we calculate that:

$$\mathbb{P}_{0.2}\left(\sum_{i=1}^{6} X_i > c_\alpha\right) = 1 - F_T(c_\alpha),$$

$$F_T(2) = \sum_{k=0}^{2} \binom{6}{k} 0.2^k 0.8^{6-k} \approx 0.9 \quad \Rightarrow \quad \mathbb{P}_{0.2}\left(\sum_{i=1}^{6} X_i > 2\right) > \alpha,$$

$$F_T(3) = \sum_{k=0}^{3} \binom{6}{k} 0.2^k 0.8^{6-k} \approx 0.98 \quad \Rightarrow \quad \mathbb{P}_{0.2}\left(\sum_{i=1}^{6} X_i > 3\right) < \alpha.$$

Therefore, we set $c_\alpha = 3$ and specify the constant $\gamma \in (0,1)$ so that:

$$\mathbb{E}_{0.2}\left[\varphi(X)\right] = \mathbb{P}_{0.2}\left(\sum_{i=1}^{6} X_i > 3\right) + \gamma \mathbb{P}_{0.2}\left(\sum_{i=1}^{6} X_i = 3\right) = \alpha \quad \Rightarrow$$

$$\gamma = \frac{F_T(3) - (1-\alpha)}{F_T(3) - F_T(2)} \approx 0.4.$$

According to the fundamental Neyman - Pearson lemma, we arrive at the following UMP test:

$$\varphi(x) = \begin{cases} 1, & \sum_{i=1}^{6} x_i > 3 \\ 0.4, & \sum_{i=1}^{6} x_i = 3 \\ 0, & \sum_{i=1}^{6} x_i < 3 \end{cases}.$$

If $\sum_{i=1}^{6} x_i = 3$, then we reject the null hypothesis H_0 with probability 0.4. □

5.3 Monotone Likelihood Ratio Property

If we want to specify a test for the one-sided composite hypotheses $H_0 : \vartheta \leqslant \vartheta_0$ vs. $H_1 : \vartheta > \vartheta_0$, then we may first apply the fundamental Neyman - Pearson lemma to specify a UMP test φ for the simple hypotheses $H_0 : \vartheta = \vartheta_0$ vs. $H_1 : \vartheta = \vartheta_1$ with $\vartheta_1 > \vartheta_0$. If we show that the critical region of the test φ does not depend on the value ϑ_1 and $\sup_{\vartheta \leqslant \vartheta_0} \pi_\varphi(\vartheta) = \pi_\varphi(\vartheta_0)$, i.e. the power function $\pi_\varphi(\vartheta)$ is increasing with respect to ϑ on $(-\infty, \vartheta_0]$, then φ is a UMP test for the composite hypotheses $H_0 : \vartheta \leqslant \vartheta_0$ vs. $H_1 : \vartheta > \vartheta_0$. The same also applies to the one-sided hypotheses $H_0 : \vartheta \geqslant \vartheta_0$ vs. $H_1 : \vartheta < \vartheta_0$, i.e. it suffices to apply the fundamental Neyman - Pearson lemma to specify a UMP test φ for the simple hypotheses $H_0 : \vartheta = \vartheta_0$ vs. $H_1 : \vartheta = \vartheta_1$ with $\vartheta_1 < \vartheta_0$. Then, it suffices to check that the critical region of the test φ does not depend on the value ϑ_1 and the power function $\pi_\varphi(\vartheta)$ is decreasing with respect to ϑ on $[\vartheta_0, \infty)$.

Definition 5.5. We say that the distribution of the sample X has the *monotone likelihood ratio* (MLR) property with respect to some statistic $T(X)$ if the likelihood ratio $\lambda(x) = \frac{\mathcal{L}(\vartheta_2|x)}{\mathcal{L}(\vartheta_1|x)}$ is an increasing function with respect to $T(x)$ on the set $\{x \in S : \mathcal{L}(\vartheta_1 \mid x) > 0 \text{ or } \mathcal{L}(\vartheta_2 \mid x) > 0\}$ for every pair $\vartheta_1, \vartheta_2 \in \Theta$ with $\vartheta_1 < \vartheta_2$.

Note 5.9. If the likelihood ratio $\lambda(x)$ is a decreasing function with respect to some statistic $T(x)$, then it's obviously an increasing function with respect to $-T(x)$, so the distribution of the sample has the MLR property with respect to $T^*(X) = -T(X)$.

Proposition 5.1. If the joint distribution of the sample X belongs to the one-parameter multivariate exponential family with $f(x; \vartheta) = h(x)e^{Q(\vartheta)T(x)-A(\vartheta)}$ and the function $Q : \Theta \to \mathbb{R}$ is strictly increasing, then the distribution of the sample has the MLR property with respect to the statistic $T(X)$.

Proof. Let $\vartheta_1, \vartheta_2 \in \Theta$ with $\vartheta_1 < \vartheta_2$. Then, we calculate that:

$$\lambda(x) = \frac{\mathcal{L}(\vartheta_2 \mid x)}{\mathcal{L}(\vartheta_1 \mid x)} = \frac{h(x)e^{Q(\vartheta_2)T(x)-A(\vartheta_2)}}{h(x)e^{Q(\vartheta_1)T(x)-A(\vartheta_1)}} = \underbrace{e^{A(\vartheta_1)-A(\vartheta_2)-[Q(\vartheta_1)-Q(\vartheta_2)]T(x)}}_{g(T(x))}.$$

Since Q is strictly increasing, we know that $Q(\vartheta_1) - Q(\vartheta_2) < 0$. Let $t_1 \leqslant t_2$. Then, we infer that:

$$[Q(\vartheta_1) - Q(\vartheta_2)]\, t_1 \geqslant [Q(\vartheta_1) - Q(\vartheta_2)]\, t_2 \quad \Rightarrow \quad g(t_1) \leqslant g(t_2).$$

Therefore, we conclude that the likelihood ratio $\lambda(X)$ is an increasing function with respect to the statistic $T(x)$. □

Note 5.10. If the function $Q : \Theta \to \mathbb{R}$ is strictly decreasing, then $Q^*(\vartheta) = -Q(\vartheta)$ is obviously a strictly increasing function, so the distribution of the sample has the MLR property with respect to $T^*(X) = -T(X)$.

Lemma 5.1. If h_1, h_2 are 2 increasing functions and X is a random variable, then it holds that $\mathrm{Cov}\left[h_1(X), h_2(X)\right] \geqslant 0$.

Proof. Let X_1, X_2 be 2 independent random variables with the same distribution as X. If $X_1 \leqslant X_2$, then it holds that $h_1(X_1) - h_1(X_2) \leqslant 0$ and $h_2(X_1) - h_2(X_2) \leqslant 0$. Similarly, if $X_1 \geqslant X_2$, then it holds that $h_1(X_1) - h_1(X_2) \geqslant 0$ and $h_2(X_1) - h_2(X_2) \geqslant 0$. In both cases, it follows that $\left[h_1(X_1) - h_1(X_2)\right]\left[h_2(X_1) - h_2(X_2)\right] \geqslant 0$. Since the random variables X_1, X_2 have the same distribution as X, we calculate that:

$$\begin{aligned} 0 &\leqslant \mathbb{E}\left[\left(h_1(X_1) - h_1(X_2)\right)\left(h_2(X_1) - h_2(X_2)\right)\right] \\ &= \mathbb{E}\left[h_1(X_1)h_2(X_1)\right] + \mathbb{E}\left[h_1(X_2)h_2(X_2)\right] \\ &\qquad - \mathbb{E}\left[h_1(X_1)\right]\mathbb{E}\left[h_2(X_2)\right] - \mathbb{E}\left[h_1(X_2)\right]\mathbb{E}\left[h_2(X_1)\right] \\ &= 2\mathbb{E}\left[h_1(X)h_2(X)\right] - 2\mathbb{E}\left[h_1(X)\right]\mathbb{E}\left[h_2(X)\right] = 2\mathrm{Cov}\left[h_1(X), h_2(X)\right]. \end{aligned}$$

□

Theorem 5.2. (Karlin - Rubin) Suppose that the distribution of the sample X has the MLR property with respect to some statistic $T(X)$.

i. We want to specify a test for the hypotheses $H_0 : \vartheta \leqslant \vartheta_0$ vs. $H_1 : \vartheta > \vartheta_0$. For given $\alpha \in (0, 1)$, a UMP test of size α is given by:

$$\varphi(x) = \begin{cases} 1, & T(x) > c \\ \gamma, & T(x) = c \\ 0, & T(x) < c \end{cases}.$$

ii. We want to specify a test for the hypotheses $H_0 : \vartheta \geqslant \vartheta_0$ vs. $H_1 : \vartheta < \vartheta_0$. For given $\alpha \in (0, 1)$, a UMP test of size α is given by:

$$\varphi(x) = \begin{cases} 1, & T(x) < c \\ \gamma, & T(x) = c \\ 0, & T(x) > c \end{cases}.$$

The constants $c \in \mathbb{R}$ and $\gamma \in [0, 1]$ are specified so that $\pi_\varphi(\vartheta_0) = \alpha$.

Proof. Without loss of generality, assume that the distribution of the sample is continuous and we are interested in the first case. Let $\vartheta < \vartheta_0$. Then, we calculate that:

$$\pi_\varphi(\vartheta_0) - \pi_\varphi(\vartheta) = \mathbb{E}_{\vartheta_0}[\varphi(X)] - \mathbb{E}_\vartheta[\varphi(X)] = \int_S \varphi(x)\left[f(x;\vartheta_0) - f(x;\vartheta)\right]dx$$
$$= \int_S \varphi(x)\left[\frac{f(x;\vartheta_0)}{f(x;\vartheta)} - 1\right] f(x;\vartheta)dx = \mathbb{E}_\vartheta\left[\varphi(X)\left(\frac{f(X;\vartheta_0)}{f(X;\vartheta)} - 1\right)\right].$$

We observe that the test φ is an increasing function with respect to $T(X)$. Since $\vartheta < \vartheta_0$, we also know that the likelihood ratio $\frac{f(X;\vartheta_0)}{f(X;\vartheta)}$ is an increasing function with respect to $T(X)$. Furthermore, we observe that:

$$\mathbb{E}_\vartheta\left[\frac{f(X;\vartheta_0)}{f(X;\vartheta)}\right] = \int_S \frac{f(x;\vartheta_0)}{f(x;\vartheta)} f(x;\vartheta)dx = \int_S f(x;\vartheta_0)dx = 1 \quad \Rightarrow$$

$$\mathbb{E}_\vartheta\left[\frac{f(X;\vartheta_0)}{f(X;\vartheta)} - 1\right] = 0.$$

According to the previous lemma, we infer that:

$$0 \leqslant \mathrm{Cov}_\vartheta\left[\varphi(X), \frac{f(X;\vartheta_0)}{f(X;\vartheta)} - 1\right] = \mathbb{E}_\vartheta\left[\varphi(X)\left(\frac{f(X;\vartheta_0)}{f(X;\vartheta)} - 1\right)\right] = \pi_\varphi(\vartheta_0) - \pi_\varphi(\vartheta),$$

which implies that $\pi_\varphi(\vartheta) \leqslant \pi_\varphi(\vartheta_0) = \alpha$ for $\vartheta < \vartheta_0$. Therefore, we conclude that φ is a UMP test of size α. □

Note 5.11. If we applied the fundamental Neyman - Pearson lemma for the simple hypotheses $H_0 : \vartheta = \vartheta_0$ vs. $H_1 : \vartheta = \vartheta_1$ with $\vartheta_1 < \vartheta_0$, we would need to solve the inequality $\mathcal{L}(\vartheta_0 \mid x) < c\mathcal{L}(\vartheta_1 \mid x)$ to specify the critical region of the test. Since $\vartheta_1 < \vartheta_0$, the likelihood ratio $\frac{\mathcal{L}(\vartheta_0|x)}{\mathcal{L}(\vartheta_1|x)}$ is an increasing function with respect to $T(X)$ according to the MLR property. Therefore, it holds that $\mathcal{L}(\vartheta_0 \mid x) < c\mathcal{L}(\vartheta_1 \mid x)$ if and only if $T(X) > c_\alpha$ for some other constant c_α. A similar observation can be made in the first case.

Example 5.9. Let $X_1, \ldots, X_n \sim \mathcal{N}\left(\mu, \sigma^2\right)$ be a random sample with known μ. We want to specify a UMP test for the hypotheses $H_0 : \sigma^2 \geqslant \sigma_0^2$ vs. $H_1 : \sigma^2 < \sigma_0^2$. We observe that:

$$f\left(x;\sigma^2\right) = (2\pi)^{-n/2}\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i - \mu)^2 - \frac{n}{2}\log\sigma^2\right\},$$

where $Q(\sigma^2) = -\frac{1}{2\sigma^2}$ is a strictly increasing function and $T(x) = \sum_{i=1}^n (x_i - \mu)^2$, so the distribution of the sample has the MLR property with respect to the statistic $T(X)$. According to the Karlin - Rubin theorem, the critical region of the test is given by $T(x) < c$. It remains to specify the constant c so that $\pi_\varphi\left(\sigma_0^2\right) = \alpha$. Given that $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma_0^2)$, we know that:

$$Q(X) = \frac{1}{\sigma_0^2} T(X) = \frac{1}{\sigma_0^2} \sum_{i=1}^{n} (X_i - \mu)^2 \sim \chi_n^2.$$

Therefore, we calculate that:

$$T(X) < c \quad \Leftrightarrow \quad Q(X) = \frac{1}{\sigma_0^2} T(X) < c_\alpha = \frac{c}{\sigma_0^2},$$

$$\mathbb{E}_{\sigma_0^2}\left[\varphi(X)\right] = \alpha \quad \Rightarrow \quad \mathbb{P}_{\sigma_0^2}\left(Q < c_\alpha\right) = \alpha \quad \Rightarrow \quad c_\alpha = \chi^2_{n;1-\alpha}.$$

Finally, we arrive at the following UMP test:

$$\varphi(x) = \begin{cases} 1, & \sum_{i=1}^n (x_i - \mu)^2 < \sigma_0^2 \chi^2_{n;1-\alpha} \\ 0, & \sum_{i=1}^n (x_i - \mu)^2 \geqslant \sigma_0^2 \chi^2_{n;1-\alpha} \end{cases}. \quad \square$$

Example 5.10. Let $X_1, \ldots, X_n \sim \text{Beta}(1, \vartheta)$ be a random sample with $\vartheta > 0$ and $f(x; \vartheta) = \vartheta(1-x)^{\vartheta - 1}$ for $x \in (0, 1)$. We want to specify a UMP test for the hypotheses $H_0 : \vartheta \leqslant \vartheta_0$ vs. $H_1 : \vartheta > \vartheta_0$. We observe that:

$$f(x; \vartheta) = \exp\left\{(1 - \vartheta) \sum_{i=1}^n \log \frac{1}{1 - x_i} + n \log \vartheta\right\},$$

where $Q(\vartheta) = 1 - \vartheta$ is a strictly decreasing function and $T(x) = -\sum_{i=1}^n \log(1 - x_i)$, so the distribution of the sample has the MLR property with respect to the statistic $T^*(X) = -T(X)$. According to the Karlin - Rubin theorem, the critical region of the test is given by $T^*(x) > c$. Given that $X_1, \ldots, X_n \sim \text{Beta}(1, \vartheta_0)$, we know that:

$$T(X) = -\sum_{i=1}^n \log(1 - X_i) \sim \text{Gamma}(n, \vartheta_0), \quad Q(X) = 2\vartheta_0 T(X) \sim \chi^2_{2n},$$

according to example 4.7 (page 92). Therefore, we calculate that:

$$T^*(X) > c \quad \Leftrightarrow \quad Q(X) = 2\vartheta_0 T(X) = -2\vartheta_0 T^*(X) < c_\alpha = -2\vartheta_0 c,$$

$$\mathbb{E}_{\vartheta_0}\left[\varphi(X)\right] = \alpha \quad \Rightarrow \quad \mathbb{P}_{\vartheta_0}\left(Q < c_\alpha\right) = \alpha \quad \Rightarrow \quad c_\alpha = \chi^2_{2n;1-\alpha}.$$

Finally, we arrive at the following UMP test:

$$\varphi(x) = \begin{cases} 1, & -2\vartheta_0 \sum_{i=1}^n \log(1 - x_i) < \chi^2_{2n;1-\alpha} \\ 0, & -2\vartheta_0 \sum_{i=1}^n \log(1 - x_i) \geqslant \chi^2_{2n;1-\alpha} \end{cases}. \quad \square$$

Example 5.11. Let $X_1, \ldots, X_n \sim \text{Pareto}(k, \lambda)$ be a random sample with $k > 0$, known $\lambda > 0$, $f(x;k) = \frac{\lambda k^\lambda}{x^{\lambda+1}}$ and $F(x;k) = 1 - \left(\frac{k}{x}\right)^\lambda$ for $x > k$. We want to specify a UMP test for the hypotheses $H_0 : k \geqslant k_0$ vs. $H_1 : k < k_0$. For $k_1 < k_2$, we calculate the following likelihood ratio:

$$\frac{\mathcal{L}(k_2 \mid x)}{\mathcal{L}(k_1 \mid x)} = \left(\frac{k_2}{k_1}\right)^{n\lambda} \frac{\mathbb{1}_{(k_2,\infty)}\left(x_{(1)}\right)}{\mathbb{1}_{(k_1,\infty)}\left(x_{(1)}\right)} = \lambda(T), \quad T(x) = x_{(1)}.$$

Let $t_1, t_2 \in (k_1, \infty)$ with $t_1 \leqslant t_2$. We distinguish the following cases:

- For $k_1 < t_1 \leqslant t_2 < k_2$, it holds that $\lambda(t_1) = 0 = \lambda(t_2)$.
- For $k_1 < t_1 < k_2 < t_2$, it holds that $\lambda(t_1) = 0 < \left(\frac{k_2}{k_1}\right)^{n\lambda} = \lambda(t_2)$.
- For $k_1 < k_2 < t_1 \leqslant t_2$, it holds that $\lambda(t_1) = \left(\frac{k_2}{k_1}\right)^{n\lambda} = \lambda(t_2)$.

Therefore, the function $\lambda(t)$ is increasing on (k_1, ∞), i.e. the distribution of the sample has the MLR property with respect to the statistic $T(X) = X_{(1)}$. According to the Karlin - Rubin theorem, the critical region of the test is given by $T(x) < c$. Given that $X_1, \ldots, X_n \sim \text{Pareto}(k_0, \lambda)$, we know that:

$$Q(X) = \frac{1}{k_0} T(X) = \frac{1}{k_0} X_{(1)} \sim \text{Pareto}(1, n\lambda),$$

according to example 4.4 (page 90). Therefore, we calculate that:

$$T(X) < c \quad \Leftrightarrow \quad Q(X) = \frac{1}{k_0} T(X) < c_\alpha = \frac{c}{k_0},$$

$$\mathbb{E}_{k_0}\left[\varphi(X)\right] = \alpha \quad \Rightarrow \quad \mathbb{P}_{k_0}\left(Q < c_\alpha\right) = \alpha \quad \Rightarrow$$
$$1 - \frac{1}{c_\alpha^{n\lambda}} = \alpha \quad \Rightarrow \quad c_\alpha = (1-\alpha)^{-1/n\lambda}.$$

Finally, we arrive at the following UMP test:

$$\varphi(x) = \begin{cases} 1, & x_{(1)} < k_0 (1-\alpha)^{-1/n\lambda} \\ 0, & x_{(1)} \geqslant k_0 (1-\alpha)^{-1/n\lambda} \end{cases}. \quad \square$$

Example 5.12. Let $X_1, \ldots, X_n \sim \mathcal{U}(0, \vartheta)$ be a random sample. We want to specify a UMP test for the hypotheses $H_0 : \vartheta \leqslant \vartheta_0$ vs. $H_1 : \vartheta > \vartheta_0$. For $\vartheta_1 < \vartheta_2$, we calculate the following likelihood ratio:

$$\frac{\mathcal{L}(\vartheta_2 \mid x)}{\mathcal{L}(\vartheta_1 \mid x)} = \left(\frac{\vartheta_1}{\vartheta_2}\right)^{n} \frac{\mathbb{1}_{(0,\vartheta_2)}\left(x_{(n)}\right)}{\mathbb{1}_{(0,\vartheta_1)}\left(x_{(n)}\right)} = \lambda(T), \quad T(x) = x_{(n)}.$$

Let $t_1, t_2 \in (0, \vartheta_2)$ with $t_1 \leqslant t_2$. We distinguish the following cases:

- For $0 < t_1 \leqslant t_2 < \vartheta_1 < \vartheta_2$, it holds that $\lambda(t_1) = \left(\frac{\vartheta_1}{\vartheta_2}\right)^n = \lambda(t_2)$.

- For $0 < t_1 < \vartheta_1 < t_2 < \vartheta_2$, it holds that $\lambda(t_1) = \left(\frac{\vartheta_1}{\vartheta_2}\right)^n < \infty = \lambda(t_2)$.
- For $0 < \vartheta_1 < t_1 \leqslant t_2 < \vartheta_2$, it holds that $\lambda(t_1) = \infty = \lambda(t_2)$.

Therefore, the function $\lambda(t)$ is increasing on $(0, \vartheta_2)$, i.e. the distribution of the sample has the MLR property with respect to the statistic $T(X) = X_{(n)}$. According to the Karlin - Rubin theorem, the critical region of the test is given by $T(x) > c$. Given that $X_1, \ldots, X_n \sim \mathcal{U}(0, \vartheta_0)$, we know that:

$$Q(X) = \frac{1}{\vartheta_0} T(X) = \frac{1}{\vartheta_0} X_{(n)} \sim \text{Beta}(n, 1),$$

according to note 4.4 (page 86). Therefore, we calculate that:

$$T(X) > c \quad \Leftrightarrow \quad Q(X) = \frac{1}{\vartheta_0} T(X) > c_\alpha = \frac{c}{\vartheta_0},$$

$$\mathbb{E}_{\vartheta_0}\left[\varphi(X)\right] = \alpha \quad \Rightarrow \quad \mathbb{P}_{\vartheta_0}\left(Q > c_\alpha\right) = \alpha \quad \Rightarrow \quad 1 - c_a^n = \alpha \quad \Rightarrow \quad c_\alpha = (1-\alpha)^{1/n}.$$

Finally, we arrive at the following UMP test:

$$\varphi(x) = \begin{cases} 1, & x_{(n)} > \vartheta_0 (1-\alpha)^{1/n} \\ 0, & x_{(n)} \leqslant \vartheta_0 (1-\alpha)^{1/n} \end{cases}. \quad \square$$

Theorem 5.3.* Suppose that the joint distribution of the sample X belongs to the one-parameter multivariate exponential family with $f(x; \vartheta) = h(x) e^{Q(\vartheta) T(x) - A(\vartheta)}$ and the function $Q : \Theta \to \mathbb{R}$ is strictly monotone. We want to specify a test for the two-sided composite hypotheses $H_0 : \vartheta \leqslant \vartheta_1$ or $\vartheta \geqslant \vartheta_2$ vs. $H_1 : \vartheta_1 < \vartheta < \vartheta_2$. For given $\alpha \in (0,1)$, a UMP test of size α is given by:

$$\varphi(x) = \begin{cases} 1, & c_1 < T(x) < c_2 \\ \gamma_1, & T(x) = c_1 \\ \gamma_2, & T(x) = c_2 \\ 0, & T(x) < c_1 \text{ or } T(x) > c_2 \end{cases}.$$

The constants $c_1, c_2 \in \mathbb{R}$, $\gamma_1, \gamma_2 \in [0,1]$ are specified so that $\pi_\varphi(\vartheta_1) = \pi_\varphi(\vartheta_2) = \alpha$.

5.4 Generalized Likelihood Ratio Tests

Definition 5.6. Consider the general hypotheses $H_0 : \vartheta \in \Theta_0$ vs. $H_1 : \vartheta \in \Theta_1$, where $\Theta_0 \cup \Theta_1 = \Theta$ and $\Theta_0 \cap \Theta_1 = \varnothing$. The following statistic:

$$\lambda^*(x) = \frac{\sup_{\vartheta \in \Theta_0} \mathcal{L}(\vartheta \mid x)}{\sup_{\vartheta \in \Theta} \mathcal{L}(\vartheta \mid x)},$$

is called the *generalized likelihood ratio.*

Note 5.12. It obviously holds that $0 \leqslant \lambda^*(x) \leqslant 1 \; \forall x \in S$. If the MLEs $\widehat{\vartheta}$ of ϑ and $\widehat{\vartheta}_0 = \arg\max_{\vartheta \in \Theta_0} \mathcal{L}(\vartheta \mid x)$ of ϑ under the null hypothesis $H_0 : \vartheta \in \Theta_0$ exist, then it follows that:

$$\lambda^*(x) = \frac{\mathcal{L}(\widehat{\vartheta}_0 \mid x)}{\mathcal{L}(\widehat{\vartheta} \mid x)}.$$

Generalized Likelihood Ratio Criterion: A test of size α for the general hypotheses $H_0 : \vartheta \in \Theta_0$ vs. $H_1 : \vartheta \in \Theta_1$, where $\Theta_0 \cup \Theta_1 = \Theta$ and $\Theta_0 \cap \Theta_1 = \varnothing$, is given by:

$$\varphi(x) = \begin{cases} 1, & \lambda^*(x) < c \\ \gamma, & \lambda^*(x) = c \\ 0, & \lambda^*(x) > c \end{cases}.$$

The constants $c, \gamma \in [0,1]$ are specified so that $\sup_{\vartheta \in \Theta_0} \pi_\varphi(\vartheta) = \alpha$.

Note 5.13. Intuitively, the numerator of the ratio $\lambda^*(x)$ expresses the maximum likelihood under the null hypothesis, while the denominator expresses the maximum likelihood as a whole. If the numerator is much smaller than the denominator, i.e. the ratio $\lambda^*(x)$ is close to 0, then it's not very probable that the sample X follows a distribution with parameter value ϑ which belongs to the set Θ_0, so we reject H_0. If the numerator is close enough to the denominator, i.e. the ratio $\lambda^*(x)$ is close to 1, then we cannot distinguish how probable it is that the sample X follows a distribution with parameter value ϑ which belongs to the set Θ_0 compared to a parameter value which belongs to the entire parameter space Θ, so we don't reject H_0.

Example 5.13. Let $X_1, \ldots, X_n \sim \mathcal{U}(0, \vartheta)$ be a random sample. We want to specify a test for the hypotheses $H_0 : \vartheta = \vartheta_0$ vs. $H_1 : \vartheta \neq \vartheta_0$ and calculate its power. According to example 3.42 (page 75), we know that the statistic $\widehat{\vartheta}(X) = X_{(n)}$ is the MLE of ϑ. Since the null hypothesis H_0 is simple, we infer that $\widehat{\vartheta}_0 = \vartheta_0$. Therefore, we calculate that:

$$\lambda^*(x) = \frac{\mathcal{L}(\vartheta_0 \mid x)}{\mathcal{L}(\widehat{\vartheta} \mid x)} = \frac{\vartheta_0^{-n} \mathbb{1}_{[0,\vartheta_0]}\left(x_{(n)}\right)}{\left[x_{(n)}\right]^{-n} \mathbb{1}_{[0,x_{(n)}]}\left(x_{(n)}\right)} = \begin{cases} \left[x_{(n)}/\vartheta_0\right]^n, & x_{(n)} \leqslant \vartheta_0 \\ 0, & x_{(n)} > \vartheta_0 \end{cases}.$$

According to the generalized likelihood ratio criterion, the critical region of the test is given by:

$$\lambda^*(x) < c \quad \Leftrightarrow \quad \left[\frac{x_{(n)}}{\vartheta_0}\right]^n < c \text{ or } x_{(n)} > \vartheta_0 \quad \Leftrightarrow \quad x_{(n)} < \vartheta_0 c^{1/n} \text{ or } x_{(n)} > \vartheta_0.$$

Under the null hypothesis H_0, i.e. given that $X_1, \dots, X_n \sim \mathcal{U}(0, \vartheta_0)$, we know that:

$$Q(X) = \frac{1}{\vartheta_0}\widehat{\vartheta}(X) = \frac{1}{\vartheta_0}X_{(n)} \sim \text{Beta}(n, 1),$$

according to note 4.4 (page 86). Hence, we calculate that:

$$X_{(n)} < \vartheta_0 c^{1/n} \text{ or } X_{(n)} > \vartheta_0 \quad \Leftrightarrow \quad Q < c_\alpha = c^{1/n} \text{ or } Q > 1,$$

$$\mathbb{E}_{\vartheta_0}[\varphi(X)] = \alpha \quad \Rightarrow \quad \mathbb{P}_{\vartheta_0}(Q < c_\alpha) + \mathbb{P}_{\vartheta_0}(Q > 1) = \alpha \quad \Rightarrow \quad c_\alpha = \alpha^{1/n}.$$

Therefore, we arrive at the following test:

$$\varphi(x) = \begin{cases} 1, & x_{(n)} < \vartheta_0\alpha^{1/n} \text{ or } x_{(n)} > \vartheta_0 \\ 0, & \vartheta_0\alpha^{1/n} \leqslant x_{(n)} \leqslant \vartheta_0 \end{cases}.$$

For $\vartheta > \vartheta_0$, we calculate that:

$$\begin{aligned} \beta_\varphi(\vartheta) &= \mathbb{P}_\vartheta\left(X_{(n)} < \vartheta_0\alpha^{1/n}\right) + \mathbb{P}_\vartheta\left(X_{(n)} > \vartheta_0\right) \\ &= \mathbb{P}_\vartheta\left(\frac{1}{\vartheta}X_{(n)} < \alpha^{1/n}\frac{\vartheta_0}{\vartheta}\right) + \mathbb{P}_\vartheta\left(\frac{1}{\vartheta}X_{(n)} > \frac{\vartheta_0}{\vartheta}\right) \\ &= \alpha\left(\frac{\vartheta_0}{\vartheta}\right)^n + 1 - \left(\frac{\vartheta_0}{\vartheta}\right)^n = 1 - (1-\alpha)\left(\frac{\vartheta_0}{\vartheta}\right)^n. \end{aligned}$$

For $\vartheta < \vartheta_0$, we calculate that:

$$\mathbb{P}_\vartheta\left(X_{(n)} > \vartheta_0\right) = \mathbb{P}_\vartheta\left(\frac{1}{\vartheta}X_{(n)} > \frac{\vartheta_0}{\vartheta}\right) = 0,$$

$$\mathbb{P}_\vartheta\left(X_{(n)} < \vartheta_0\alpha^{1/n}\right) = \mathbb{P}_\vartheta\left(\frac{1}{\vartheta}X_{(n)} < \alpha^{1/n}\frac{\vartheta_0}{\vartheta}\right) = \begin{cases} \alpha\left(\vartheta_0/\vartheta\right)^n, & \vartheta > \vartheta_0\alpha^{1/n} \\ 1, & \vartheta \leqslant \vartheta_0\alpha^{1/n} \end{cases}.$$

Finally, we conclude that:

$$\beta_\varphi(\vartheta) = \begin{cases} 1, & \vartheta \leqslant \vartheta_0\alpha^{1/n} \\ \alpha\left(\vartheta_0/\vartheta\right)^n, & \vartheta_0\alpha^{1/n} < \vartheta \leqslant \vartheta_0 \\ 1 - (1-\alpha)\left(\vartheta_0/\vartheta\right)^n, & \vartheta > \vartheta_0 \end{cases}. \quad \square$$

Proposition 5.2.* Suppose we want to specify a test for the hypotheses $H_0 : \vartheta = \vartheta_0$ vs. $H_1 : \vartheta \neq \vartheta_0$ or the hypotheses $H_0 : \vartheta_1 \leqslant \vartheta \leqslant \vartheta_2$ vs. $H_1 : \vartheta < \vartheta_1$ or $\vartheta > \vartheta_2$. Then, the generalized likelihood ratio criterion leads to a test of size α of the following form:

$$\varphi(x) = \begin{cases} 1, & T(x) < c_1 \text{ or } T(x) > c_2 \\ \gamma_1, & T(x) = c_1 \\ \gamma_2, & T(x) = c_2 \\ 0, & c_1 < T(x) < c_2 \end{cases} .$$

The constants $c_1, c_2 \in \mathbb{R}$ and $\gamma_1, \gamma_2 \in [0,1]$ are specified so that $\pi_\varphi(\vartheta_0) = \alpha$ or $\pi_\varphi(\vartheta_1) = \pi_\varphi(\vartheta_2) = \alpha$ respectively.

Theorem 5.4.* (Wilks) We want to specify a test for the hypotheses $H_0 : \vartheta \in \Theta_0$ vs. $H_1 : \vartheta \in \Theta_1$, where $\Theta_0 \cup \Theta_1 = \Theta$ and $\Theta_0 \cap \Theta_1 = \varnothing$. Suppose that the regularity conditions for the asymptotic efficiency of the MLE of ϑ are satisfied. If d is the number of restrictions that the null hypothesis H_0 sets on the parameter space Θ, then it follows that:

$$D_n(X) = -2\log\lambda_n^*(X) = -2\left[\ell(\widehat{\vartheta}_0 \mid X) - \ell(\widehat{\vartheta} \mid X)\right] \xrightarrow{d} Y \sim \chi_d^2.$$

Therefore, we arrive at the following asymptotic test of size α:

$$\varphi(x) = \begin{cases} 1, & -2\log\lambda_n^*(x) > \chi^2_{d;\alpha} \\ 0, & -2\log\lambda_n^*(x) \leqslant \chi^2_{d;\alpha} \end{cases} .$$

5.5 Statistical Hypothesis Tests for a Normal Population

Let $X_1, \ldots, X_n \sim \mathcal{N}\left(\mu, \sigma^2\right)$ be a random sample. We want to specify tests for the hypotheses $H_0 : \mu = \mu_0$ vs. $H_1 : \mu \neq \mu_0$. We distinguish 3 cases which we present throughout this paragraph.

Example 5.14. The variance σ^2 is known. We know that $\overline{X} \sim \mathcal{N}\left(\mu, \frac{1}{n}\sigma^2\right)$ is the MLE of μ. We calculate that:

$$\begin{aligned} \log\lambda^*(x) = \ell(\mu_0 \mid x) - \ell(\widehat{\mu} \mid x) &= -\frac{1}{2\sigma^2}\left[\sum_{i=1}^n (x_i - \mu_0)^2 - \sum_{i=1}^n (x_i - \overline{x})^2\right] \\ &= -\frac{1}{2\sigma^2}\left(n\mu_0^2 - 2\mu_0\sum_{i=1}^n x_i + 2\overline{x}\sum_{i=1}^n x_i - n\overline{x}^2\right) \\ &= -\frac{n}{2\sigma^2}\left(\mu_0^2 - 2\mu_0\overline{x} + 2\overline{x}^2 - \overline{x}^2\right) = -\frac{n}{2\sigma^2}\left(\overline{x} - \mu_0\right)^2. \end{aligned}$$

According to the generalized likelihood ratio criterion, the critical region of the test is given by:

$$\lambda^*(x) < c \quad \Leftrightarrow \quad \frac{(\overline{x}-\mu_0)^2}{\sigma^2/n} > c^* = -2\log c \quad \Leftrightarrow \quad \left|\frac{\overline{x}-\mu_0}{\sigma/\sqrt{n}}\right| > c_\alpha = \sqrt{c^*}.$$

Under the null hypothesis H_0, i.e. given that $X_1,\ldots,X_n \sim \mathcal{N}(\mu_0,\sigma^2)$, we know that $Z = \frac{\overline{X}-\mu_0}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)$. Therefore, we calculate that:

$$\begin{aligned}\mathbb{E}_{\mu_0}[\varphi(X)] &= \mathbb{P}_{\mu_0}\left(\left|\frac{\overline{X}-\mu_0}{\sigma/\sqrt{n}}\right| > c_\alpha\right) = \mathbb{P}_{\mu_0}(Z > c_\alpha) + \mathbb{P}_{\mu_0}(Z < -c_\alpha)\\ &= 1-\Phi(c_\alpha)+\Phi(-c_\alpha) = 1-\Phi(c_\alpha)+1-\Phi(c_\alpha) = 2\left[1-\Phi(c_\alpha)\right] = \alpha,\end{aligned}$$

$$\Phi(c_\alpha) = 1-\frac{\alpha}{2} \quad \Rightarrow \quad c_\alpha = \Phi^{-1}\left(1-\frac{\alpha}{2}\right) = Z_{\alpha/2}.$$

Finally, we arrive at the following test of size α:

$$\varphi(x) = \begin{cases} 1, & \left|\frac{\overline{x}-\mu_0}{\sigma/\sqrt{n}}\right| > Z_{\alpha/2} \\ 0, & \left|\frac{\overline{x}-\mu_0}{\sigma/\sqrt{n}}\right| \leqslant Z_{\alpha/2} \end{cases}. \quad \square$$

Note 5.14. In the previous test, we observe that:

$$\begin{aligned} A &= \left\{x \in \mathbb{R}^n : \left|\frac{\overline{x}-\mu_0}{\sigma/\sqrt{n}}\right| \leqslant Z_{\alpha/2}\right\} \\ &= \left\{x \in \mathbb{R}^n : \overline{x} - Z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leqslant \mu_0 \leqslant \overline{x} + Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right\} \\ &= \left\{x \in \mathbb{R}^n : \mu_0 \in \mathcal{I}_{\mu;1-\alpha}(x) = \left[\overline{x} - Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \overline{x} + Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right]\right\}, \end{aligned}$$

where $\mathcal{I}_{\mu;1-\alpha}(x)$ is the $100(1-\alpha)\%$ equal-tailed CI for the mean μ. In other words, we don't reject $H_0 : \mu = \mu_0$ against $H_1 : \mu \neq \mu_0$ at statistical significance level α if and only if the value μ_0 lies inside the $100(1-\alpha)\%$ equal-tailed CI for μ. This connection between CIs and tests with two-sided alternative hypotheses provides us with an alternative method of specifying the critical region of tests with two-sided alternative hypotheses.

Example 5.15. The variance σ^2 is unknown. According to example 3.44 (page 76), we know that $\widehat{\mu} = \overline{x}$ and $\widehat{\sigma}^2 = \frac{n-1}{n}S^2$. Under the null hypothesis H_0, i.e. given that $\mu = \mu_0$, we know that $\widehat{\sigma}_0^2 = \frac{1}{n}\sum_{i=1}^n (x_i-\mu_0)^2$, according to example 3.38 (page 73).

Therefore, we calculate that:

$$\begin{aligned}
\widehat{\sigma}_0^2 &= \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x} + \overline{x} - \mu_0)^2 \\
&= \frac{1}{n}\left[\sum_{i=1}^{n}(x_i - \overline{x})^2 + n(\overline{x} - \mu_0)^2 + 2(\overline{x} - \mu_0)\sum_{i=1}^{n}(x_i - \overline{x})\right] \\
&= \widehat{\sigma}^2 + (\overline{x} - \mu_0)^2 + \frac{2}{n}(\overline{x} - \mu_0)\left(\sum_{i=1}^{n} x_i - n\overline{x}\right) = \widehat{\sigma}^2 + (\overline{x} - \mu_0)^2,
\end{aligned}$$

$$\begin{aligned}
\log \lambda^*(x) &= \ell\left(\mu_0, \widehat{\sigma}_0^2 \mid x\right) - \ell\left(\widehat{\mu}, \widehat{\sigma}^2 \mid x\right) \\
&= -\frac{n}{2}\log\frac{\widehat{\sigma}_0^2}{\widehat{\sigma}^2} - \frac{1}{2\widehat{\sigma}_0^2}\sum_{i=1}^{n}(x_i - \mu_0)^2 + \frac{1}{2\widehat{\sigma}^2}\sum_{i=1}^{n}(x_i - \overline{x})^2 \\
&= -\frac{n}{2}\log\left[1 + \frac{(\overline{x} - \mu_0)^2}{\widehat{\sigma}^2}\right] - \frac{n}{2} + \frac{n}{2} = -\frac{n}{2}\log\left[1 + \frac{n(\overline{x} - \mu_0)^2}{(n-1)s^2}\right].
\end{aligned}$$

According to the generalized likelihood ratio criterion, the critical region of the test is given by:

$$\lambda^*(x) < c \quad \Leftrightarrow \quad -\frac{n}{2}\log\left[1 + \frac{n(\overline{x} - \mu_0)^2}{(n-1)s^2}\right] < c^* = \log c \quad \Leftrightarrow$$

$$1 + \frac{n(\overline{x} - \mu_0)^2}{(n-1)s^2} > c^{**} = e^{-2c^*/n} \quad \Leftrightarrow \quad \frac{(\overline{x} - \mu_0)^2}{s^2/n} > c^{***} = (n-1)\left(c^{**} - 1\right) \quad \Leftrightarrow$$

$$\left|\frac{\overline{x} - \mu_0}{s/\sqrt{n}}\right| > c_\alpha = \sqrt{c^{***}}.$$

Under the null hypothesis H_0, we know that $T = \frac{\overline{X} - \mu_0}{S/\sqrt{n}} \sim t_{n-1}$, according to example 4.13 (page 95). Therefore, we calculate that:

$$\begin{aligned}
\mathbb{E}_{\mu_0}\left[\varphi(X)\right] = \mathbb{P}_{\mu_0}\left(\left|\frac{\overline{X} - \mu_0}{S/\sqrt{n}}\right| > c_\alpha\right) &= \mathbb{P}_{\mu_0}(T > c_\alpha) + \mathbb{P}_{\mu_0}(T < -c_\alpha) \\
&= 1 - F_T(c_\alpha) + F_T(-c_\alpha) = 2\left[1 - F_T(c_\alpha)\right] = \alpha,
\end{aligned}$$

$$F_T(c_\alpha) = 1 - \frac{\alpha}{2} \quad \Rightarrow \quad c_\alpha = F_T^{-1}\left(1 - \frac{\alpha}{2}\right) = t_{n-1;\alpha/2}.$$

Finally, we arrive at the following test of size α:

$$\varphi(x) = \begin{cases} 1, & \left|\frac{\overline{x} - \mu_0}{s/\sqrt{n}}\right| > t_{n-1;\alpha/2} \\ 0, & \left|\frac{\overline{x} - \mu_0}{s/\sqrt{n}}\right| \leqslant t_{n-1;\alpha/2} \end{cases}. \quad \square$$

Example 5.16. The variance σ^2 is unknown and we want to specify an asymptotic test. Under the null hypothesis H_0, we know that:

$$T_n(X) = \frac{\overline{X}_n - \mu_0}{S_n/\sqrt{n}} \xrightarrow{d} Z \sim \mathcal{N}(0,1),$$

according to example 4.23 (page 101). Therefore, we arrive at the following asymptotic test of size α:

$$\varphi(x) = \begin{cases} 1, & \left|\frac{\overline{x}-\mu_0}{s/\sqrt{n}}\right| > Z_{\alpha/2} \\ 0, & \left|\frac{\overline{x}-\mu_0}{s/\sqrt{n}}\right| \leqslant Z_{\alpha/2} \end{cases}.$$

Alternatively, we know that:

$$D_n(X) = -2\log\lambda_n^*(X) = n\log\left[1 + \frac{(\overline{X}-\mu_0)^2}{\widehat{\sigma}^2}\right] \xrightarrow{d} Y \sim \chi_1^2,$$

according to Wilks' theorem. Hence, we arrive at the following asymptotic test of size α:

$$\varphi(x) = \begin{cases} 1, & n\log\left[1 + \frac{(\overline{x}-\mu_0)^2}{\widehat{\sigma}^2}\right] > \chi_{1;\alpha}^2 \\ 0, & n\log\left[1 + \frac{(\overline{x}-\mu_0)^2}{\widehat{\sigma}^2}\right] \leqslant \chi_{1;\alpha}^2 \end{cases}. \quad \square$$

Bibliography

Casella, George and Roger L. Berger. *Statistical Inference. Second Edition.* Cengage, 2001.

Ferguson, Thomas S. *A Course in Large Sample Theory. Texts in Statistical Science.* CRC Press, 1996.

Hogg, Robert, Joseph McKean, and Allen Craig. *Introduction to Mathematical Statistics. Eighth Edition.* Pearson Education, 2018.

Keener, Robert W. *Theoretical Statistics. Topics for a Core Course.* Springer Science & Business Media, 2010.

Lehmann, Erich L. *Elements of Large-Sample Theory. Corrected Edition.* Springer Science & Business Media, 1998.

Lehmann, Erich L. and George Casella. *Theory of Point Estimation. Second Edition.* Springer Science & Business Media, 1998.

Lehmann, Erich L. and Joseph P. Romano. *Testing Statistical Hypotheses. Fourth Edition.* Springer Science & Business Media, 2022.

Van der Vaart, Aad. *Asymptotic Statistics. Cambridge Series in Statistical and Probabilistic Mathematics.* Cambridge University Press, 2000.