Sufficient Statistic

Chia-Min Wei

November 20, 2024

Notation

For an experiment $(\mathcal{X}, \mathcal{F}, \{\mathrm{P}_\theta\}_{\theta\in\Theta})$, $\mathcal{X}$ represents the sample space, $\mathcal{F}$ the σ-algebra of events and Θ the parameter space. For a probability measure λ defined on $\mathcal{F}$, we write $\mathrm{E}_\lambda[Y]$ to represent the expectation of a random variable Y. As for a candidate probability measure P_θ in an experiment, we simply write $\mathrm{E}_\theta[Y]$ instead of $\mathrm{E}_{\mathrm{P}_\theta}[Y]$.

1 Introduction

In statistics, we often summarize what we see from the whole sample. For example, suppose we are estimating the expectation of some distribution, we often report the sample mean as a summary of the whole sample. However, information about the parameter may be lost along the summarizing process. Ways of summarizing the data without losing information about the parameter are called *sufficient statistics*. In other words, to infer the parameter, it is sufficient to see the summary rather than the entire sample.

2 Definition of Sufficient Statistic

Let $\{\mathcal{X}, \mathcal{F}, \{\mathrm{P}_\theta\}_{\theta\in\Theta}\}$ be a statistical experiment generated by the sample X. A statistician's job is to estimate the true $\theta_0 \in \Theta$ after observing X. In this note, we assume that all P_θ are dominated by some measure μ on $\mathcal{F}$ with p.d.f. $f(\cdot \mid \theta)$.

Definition 2.1 (Sufficient Statistic): Let $T : (\mathcal{X}, \mathcal{F}) \to (\mathcal{T}, \mathcal{G})$ be a measurable map defined on the sample space. If $\mathrm{P}_\theta(\cdot \mid T)$ does not depend on θ, then we say T is a sufficient statistic for θ.

T contains all available information concerning θ. At first sight, the condition listed in Definition 2.1 seems quite strange. To gain more intuition, we temporarily shift to the Bayesian paradigm. Suppose θ is a random variable with prior distribution π. Then Definition 2.1 basically says that given T, then X and θ are independent.

$$\theta \perp\!\!\!\perp X \mid T.$$

Suppose a statistician observes the sample $X = x$. Summarizing the sample with a sufficient statistic T, the statistician gives a summary $t = T(x)$. Assume that given any $\theta \in \Theta$, T has a p.d.f. $g(\cdot \,|\, \theta)$. The posterior distribution of θ is

$$\begin{aligned}
\pi(\theta \,|\, x) &= \frac{\pi(\theta) f(x \,|\, \theta)}{\int_{\theta \in \Theta} \pi(\theta) f(x \,|\, \theta)\, d\theta} \\
&= \frac{\pi(\theta) f(x \,|\, t, \theta) g(t \,|\, \theta)}{\int_{\theta \in \Theta} \pi(\theta) f(x \,|\, t, \theta) g(t \,|\, \theta)\, d\theta} \\
&= \frac{\pi(\theta) f(x \,|\, t) g(t \,|\, \theta)}{f(x \,|\, t) \int_{\theta \in \Theta} \pi(\theta) g(t \,|\, \theta)\, d\theta} && \text{(by Definition 2.1)} \\
&= \pi(\theta \,|\, t).
\end{aligned}$$

This means that the statistician will end up with the same posterior distribution of θ if he didn't see x but only saw $t = T(x)$ in the first place.

It is often tedious to distinguish a sufficient statistic by explicitly checking the definition. This can be seen in the following simple example. Let the sample X consists of n i.i.d. observations $X_1, ..., X_n$. The order statistic of X, $T(x_1, ..., x_n) = (t_1, ..., t_n)$ with $t_1 \le t_2 \le ... \le t_n$ is a sufficient statistic. However, checking this fact is indeed quite tedious.

> **Example 2.1:** Suppose $X_1, ..., X_n$ are i.i.d. with a distribution dominated by the Lebesgue measure with p.d.f. $f(x_i \,|\, \theta)$. A sample $x = (x_1, ..., x_n) \in \mathcal{X} = \mathbb{R}^n$ consists of the realizations of $X_1, ..., X_n$. The order statistic, rearranging x_1 to x_n from small to large, $T(x) = (t_1, ..., t_n)$, is a sufficient statistic.

Proof. It is equivalent to proving that for any L^1 map $\phi : \mathcal{X} \to \mathbb{R}$, $\mathrm{E}_\theta[\phi \,|\, T]$ does not depend on θ. Write $T = (T_1, ..., T_n)$. Define for any $(x_1, ..., x_n) \in \mathcal{X}$ the L^1 random variable H,

$$H(x_1, ..., x_n) = \frac{1}{n!} \sum \phi(x_{j_1}, ..., x_{j_n})$$

where the sum is taken over all permutations $(j_1, ..., j_n)$ of $\{1, 2, 3, ..., n\}$. One can easily see that H can be written as a measurable function of T, and thus $H \in L^1(\mathcal{X}, \sigma(T), \mathrm{P}_\theta)$.

We show that $H = \mathrm{E}_\theta[\phi \mid T]$. It then suffices to prove for any non-negative bounded map $\psi : \mathbb{R}^n \to \mathbb{R}$,

$$\mathrm{E}_\theta[H\psi(T)] = \mathrm{E}_\theta[\phi\psi(T)].$$

Note that for any $(x_1, ..., x_n) \in \mathcal{X}$ and permutation $(j_1, ..., j_n)$,

$$T(x_1, ..., x_n) = T(x_{j_1}, ..., x_{j_n}).$$

Therefore,

$$\begin{aligned}
\mathrm{E}_\theta[H\psi(T)] &= \frac{1}{n!}\sum \int_{\mathbb{R}^n} \phi(x_{j_1}, ..., x_{j_n})\psi(T(x_1, ..., x_n)) \prod_{i=1}^n f(x_i \mid \theta)\, d(x_1, ..., x_n) \\
&= \frac{1}{n!}\sum \int_{\mathbb{R}^n} \phi(x_{j_1}, ..., x_{j_n})\psi(T(x_{j_1}, ..., x_{j_n})) \prod_{i=1}^n f(x_{j_i} \mid \theta)\, d(x_1, ..., x_n) \\
&= \frac{1}{n!} n! \int_{\mathbb{R}^n} \phi(x_1, ..., x_n)\psi(T(x_1, ..., x_n)) \prod_{i=1}^n f(x_i \mid \theta)\, d(x_1, ..., x_n) \\
&= \mathrm{E}_\theta[\phi\psi(T)].
\end{aligned}$$

Since $H = \mathrm{E}_\theta[\phi \mid T]$ does not depend on θ, the proof is done. □

3 Characterization of Sufficient Statistic

Fortunately, we do have a theorem that helps us identify sufficient statistics when $\{\mathrm{P}_\theta\}_{\theta\in\Theta}$ is dominated by a σ-finite measure. It is called the **Factorization Theorem**, and was first proposed by Fisher and Neyman. Before stating and proving the theorem, we need a preliminary result.

Lemma 1: Let $(\mathcal{X}, \mathcal{F}, \{\mathrm{P}_\theta\}_{\theta\in\Theta})$ be an experiment with $\{\mathrm{P}_\theta\}_{\theta\in\Theta}$ being dominated by a σ-finite measure μ. Then there exists a countable subset of Θ, $\{\theta_i\}_{i\in\mathbb{N}}$, and a sequence of positive numbers $\{c_i\}_{i\in\mathbb{N}}$ with $\sum_{i\in\mathbb{N}} c_i = 1$ such that the probability measure $\lambda = \sum_{i\in\mathbb{N}} c_i \mathrm{P}_{\theta_i}$ dominates all P_θ.

Proof. Write $\mathcal{P} = \{\mathrm{P}_\theta\}_{\theta\in\Theta}$. It is without loss of generality to assume that μ is a finite measure, since if $\mathcal{P}$ is dominated by a σ-finite measure, it must be dominated by a finite measure. Let Λ be the collection of probability measure that can be written as $\sum_{i\in\mathbb{N}} c_i \mathrm{P}_i$ for some countable $\{\mathrm{P}_i\} \subset \mathcal{P}$ and positive c_i's such that $\sum_{i\in\mathbb{N}} c_i = 1$. It suffices to prove that there exists $\lambda^* \in \Lambda$ such that λ^* dominates all $\lambda \in \Lambda$. Define the collection of sets

$$\mathcal{A} := \left\{ A \in \mathcal{F} : \exists \lambda \in \Lambda \text{ such that } \lambda(A) > 0 \text{ and } \frac{d\lambda}{d\mu} > 0 \text{ a.e. } \mu \text{ on } A. \right\}$$

There exists a sequence $\{A_i\}_{i \in \mathbb{N}} \in \mathcal{A}$ such that

$$\mu(A_i) \to \sup_{A \in \mathcal{A}} \mu(A).$$

Write $A^* = \bigcup_{i \in \mathbb{N}} A_i$ and $\lambda^* = \sum_{i \in 1} 2^{-i} \lambda_i$ where each λ_i corresponds to A_i. One can check that $\lambda^* \in \Lambda$ and that $\lambda^*(A^*) > 0$ with $d\lambda^*/d\mu > 0$ a.e. μ on A^*. This then implies that $A^* \in \mathcal{A}$. Now let $E \in \mathcal{F}$ such that $\lambda^*(E) = 0$ and $\lambda \in \Lambda$. The fact that $\lambda^*(E) = 0$ implies that $\mu(A^* \cap E) = 0$. If $\lambda(E) > 0$, then there exists $E' \subset E$ and $E' \in \mathcal{A}$. But then $A^* \cup E' \in \mathcal{A}$ and $\mu(A^* \cup E') = \mu(A^*) + \mu(E') > \mu(A^*)$, contradicting that fact that $\mu(A^*) = \sup_{A \in \mathcal{A}} \mu(A)$. □

Theorem 1 (Factorization Theorem): Let $(\mathcal{X}, \mathcal{F}, \{\mathrm{P}_\theta\}_{\theta \in \Theta})$ be an experiment and $\{\mathrm{P}_\theta\}_{\theta \in \Theta}$ be dominated by a σ-finite measure μ. Then $T : (\mathcal{X}, \mathcal{F}) \to (\mathcal{T}, \mathcal{G})$ is a sufficient statistic if and only if there exists measurable functions $\{g_\theta\}_{\theta \in \Theta}$ defined on $(\mathcal{T}, \mathcal{G})$ and h defined on $(\mathcal{X}, \mathcal{F})$ such that

$$f(x \mid \theta) = g_\theta(T(x)) h(x).$$

Proof. By Lemma 1, there exists $\lambda = \sum_{i \in \mathbb{N}} c_i \mathrm{P}_{\theta_i}$ that dominates all P_θ in $\{\mathrm{P}_\theta\}_{\theta \in \Theta}$.

Assume that T is sufficient for θ. We show that the Radon-Nikodym derivative of P_θ for $(\mathcal{F}, \lambda)$ can be written as a measurable function of T, $g_\theta(T)$. If this is established, by writing h as the derivative for $(\mathcal{F}, \mu)$, we have for all $\theta \in \Theta$,

$$f(x \mid \theta) = \frac{d\mathrm{P}_\theta}{d\lambda} \frac{d\lambda}{d\mu} = g_\theta(T(x)) h(x).$$

Indeed, for any θ, there exists a measurable function g_θ defined on $(\mathcal{T}, \mathcal{G})$ such that $g_\theta(T)$ is the derivative of P_θ for $(\sigma(T), \lambda)$. We show that $g_\theta(T)$ is also the derivative of P_θ for $(\mathcal{F}, \lambda)$. Let $A \in \mathcal{F}$ and $A_0 \in \sigma(T)$. Since T is sufficient, $P_\theta(A \mid T) = P(A \mid T)$ does not depend on θ. Note that for any θ,

$$\begin{aligned} \int_{A_0} \mathrm{P}(A \mid T) \, d\mathrm{P}_\theta &= \mathrm{E}_\theta[\mathrm{E}_\theta[\mathbf{1}_A \mid T] \, \mathbf{1}_{A_0}] \\ &= \mathrm{E}_\theta[\mathbf{1}_A \, \mathbf{1}_{A_0}] = \mathrm{P}_\theta(A \cap A_0). \end{aligned}$$

Hence,

$$\begin{aligned}\int_{A_0} \mathrm{P}(A \,|\, T)\, d\lambda &= \sum_{i\in\mathbb{N}} c_i \int_{A_0} \mathrm{P}(A \,|\, T)\, d\mathrm{P}_{\theta_i} \\ &= \sum_{i\in\mathbb{N}} c_i \mathrm{P}_{\theta_i}(A \cap A_0) = \lambda(A \cap A_0).\end{aligned}$$

This means that $\mathrm{P}(\cdot \,|\, T)$ also serves as the conditional probability for λ. Now let $A \in \mathcal{F}$ and $\theta \in \Theta$.

$$\begin{aligned}P_\theta(A) &= \mathrm{E}_\theta[\mathbf{1}_A] = \mathrm{E}_\theta[P(A \,|\, T)] \\ &= \int_{\mathcal{X}} P(A \,|\, T)\, d\mathrm{P}_\theta \\ &= \int_{\mathcal{X}} P(A \,|\, T) g_\theta(T)\, d\lambda && \text{(because } P(A \,|\, T) \text{ is } \sigma(T)\text{-measurable)} \\ &= \int_{\mathcal{X}} \mathrm{E}_\lambda[\mathbf{1}_A \,|\, T] g_\theta(T)\, d\lambda && \text{(by the observation above)} \\ &= \int_A g_\theta(T)\, d\lambda.\end{aligned}$$

Assume conversely that there is such g_θ and h that satisfies

$$f(x \,|\, \theta) = g_\theta(T(x))h(x).$$

We have

$$\frac{d\lambda}{d\mu} = \sum_{i\in\mathbb{N}} c_i g_{\theta_i}(T)h = k(T)h.$$

It then follows that

$$\frac{d\mathrm{P}_\theta}{d\lambda}(x) = g_\theta^*(T(x)) = \begin{cases} g_\theta(T(x))/k(T(x)) & \text{when } k(T(x)) > 0, \\ \text{anything} & \text{otherwise.}\end{cases}$$

We now prove that $\mathrm{P}_\lambda(\cdot \,|\, T)$ serves as the conditional probability for all P_θ. For any $A_0 \in \sigma(T)$ and $\theta \in \Theta$,

$$\begin{aligned}\int_{A_0} \mathrm{P}_\lambda(A \,|\, T)\, d\mathrm{P}_\theta &= \int_{A_0} \mathrm{E}_\lambda[\mathbf{1}_A \,|\, T] g_\theta^*(T)\, d\lambda \\ &= \int_{A_0} \mathrm{E}_\lambda[\mathbf{1}_A\, g_\theta^*(T) \,|\, T]\, d\lambda \\ &= \int_{A\cap A_0} g_\theta^*(T)\, d\lambda = \mathrm{P}_\theta(A_0 \cap A).\end{aligned}$$

□

Example 3.1 (Uniform Distribution): Suppose $X_1, X_2, ..., X_n$ are i.i.d. with uniform distribution $U[l, u]$. $\theta = (l, u)$ and set $\Theta = \{(l, u) \in \mathbb{R}^2 : l < u\}$. A sample $x = (x_1, ..., x_n)$ consists of the realizations of $X_1, .., X_n$. $T(x) = (\min x_i, \max x_i)$ is a sufficient statistic.

Proof. Observe

$$\begin{aligned} f(x \mid l, u) &= \begin{cases} \frac{1}{(u-l)^n} & \text{if } l \leq \min x_i \leq \max x_i \leq u \\ 0 & \text{otherwise.} \end{cases} \\ &= (u-l)^{-n} \, \mathbf{1}\{l \leq \min x_i \leq \max x_i \leq u\}. \end{aligned}$$

Using Theorem 1, by setting,

$$\begin{aligned} g_\theta(T(x)) &= (u-l)^{-n} \, \mathbf{1}\{l \leq \min x_i \leq \max x_i \leq u\} \\ h(x) &= 1, \end{aligned}$$

we know that $(\min x_i, \max x_i)$ is a sufficient statistic. □

Example 3.2 (Normal Distribution): Suppose $X_1, X_2, ..., X_n$ are i.i.d. with normal distribution $\mathcal{N}(\mu, \sigma^2)$. $\theta = (\mu, \sigma^2)$ and set $\Theta = \{(\mu, \sigma) \in \mathbb{R}^2 : \sigma > 0\}$. A sample $x = (x_1, ..., x_n)$ consists of the realizations of $X_1, .., X_n$. $T(x) = (\overline{x}, s^2)$ where

$$\begin{aligned} \overline{x} &= \frac{1}{n} \sum_{i=1}^{n} x_i, \\ s^2 &= \frac{1}{n} \sum_{i=1}^{n} (x_i - \overline{x})^2. \end{aligned}$$

is a sufficient statistic.

Proof. Use the relationship

$$\sum_{i=1}^{n} (x_i - \mu)^2 = ns^2 + n(\overline{x} - \mu)^2,$$

we obtain

$$\begin{aligned} f(x \mid \mu, \sigma) &= (\sqrt{2\pi}\sigma)^{-n} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\right) \\ &= (\sqrt{2\pi}\sigma)^{-n} \exp\left(-\frac{1}{2\sigma^2}(ns^2 + n(\overline{x} - \mu)^2)\right). \end{aligned}$$

Using Theorem 1, by setting,

$$\begin{aligned} g_\theta(T(x)) &= (\sqrt{2\pi}\sigma)^{-n} \exp\left(-\frac{1}{2\sigma^2}(ns^2 + n(\overline{x} - \mu)^2)\right) \\ h(x) &= 1, \end{aligned}$$

we know that $T(x) = (\overline{x}, s^2)$ is a sufficient statistic. □

Example 3.3 (Poisson Distribution): Suppose $X_1, , , , ., X_n$ are i.i.d. with Poisson distribution $Poisson(\lambda)$. $\theta = \lambda$ and $\Theta = (0, \infty)$. A sample $x = (x_1, ..., x_n)$ consists of the realizations of $X_1, ..., X_n$. $\mathcal{X} = \{0, 1, 2, ...\}$ and each P_θ is dominated by the uniform measure on $\mathcal{X}$. $T(x) = \overline{x}$ is a sufficient statistic.

Proof. Observe

$$f(x \mid \lambda) = \prod_{i=1}^{n} \frac{\lambda^k e^{-k}}{k!} = \lambda^{n\overline{x}} e^{-\lambda} (\prod_{i=1}^{n} x_i!).$$

Using Theorem 1, by setting

$$\begin{aligned} g_\theta(T(x)) &= \lambda^{n\overline{x}} e^{-\lambda} \\ h(x) &= \prod_{i=1}^{n} x_i!, \end{aligned}$$

we know that $T(x) = \overline{x}$ is a sufficient statistic. □

4 Sufficiency Principle

Since sufficient statistic summarizes the sample without loss of information about the parameter, it is reasonable to require that inferences should depend only on sufficient statistics. This is the so-called **Sufficiency Principle**.

Sufficiency Principle: Let $\{\mathcal{X}, \mathcal{F}, \{\mathrm{P}_\theta\}_{\theta\in\Theta}\}$ be an experiment and T be a sufficient statistic for θ. Then inferences about θ should depend only on T. Namely, if two samples x and y satisfy $T(x) = T(y)$, then they should lead to the same inference on θ.

The Sufficiency Principle can be justified in two ways: **Fisher's thought experiment** and **Rao-Blackwell Theorem**.

Fisher's thought experiment proceeds as follows. Consider an experiment $(\mathcal{X}, \mathcal{F}, \{\mathrm{P}_\theta\}_{\theta\in\Theta})$ and two statisticians, Fisher and Neyman, aiming to estimate θ. After the experiment is conducted, Fisher sees the whole sample $x \in \mathcal{X}$, while Neyman only sees the sufficient statistic $t = T(x)$. Neyman then uses the sufficient statistic $t = T(x)$ and a randomization device to generate a new sample $Y \in \mathcal{X}$ with the distribution $\mathrm{P}(\cdot \,|\, t)$. Note that Neyman doesn't need to know θ to compute $\mathrm{P}_\theta(\cdot \,|\, t)$ by the definition of sufficient statistic.

At first glance, Neyman is running a different experiment than Fisher. However, these two experiments are equivalent in the sense that, given any θ, X and Y have the same *unconditional* probability. This means that Neyman has just as much knowledge about θ as Fisher. Let us write Neyman's experiment as $(\mathcal{X}, \mathcal{F}, \{\mathrm{P}'_\theta\}_{\theta\in\Theta})$.

Proposition 1: For all $\theta \in \Theta$, $\mathrm{P}_\theta = \mathrm{P}'_\theta$.

Proof. By the process that Y is generated, for any $A \in \mathcal{F}$,

$$\mathrm{P}'_\theta(A \,|\, T) = \mathrm{P}(A \,|\, T) = \mathrm{P}_\theta(A \,|\, T).$$

Therefore,

$$\begin{aligned}
\mathrm{P}'_\theta(A) &= \int_A \mathrm{P}'_\theta(A \,|\, T(x))\, d\mathrm{P}_\theta(x) \\
&= \int_A \mathrm{P}_\theta(A \,|\, T(x))\, d\mathrm{P}_\theta(x) \\
&= \mathrm{P}_\theta(A).
\end{aligned}$$

□

If Fisher uses method $\mathcal{I}$ to infer on θ from $X = x$, Neyman can also use the same method, since he is running an experiment which brings just as much information about θ as the original one. Moreover, it is expected that $\mathcal{I}(x) = \mathcal{I}(y)$ because X and Y have the same probability distribution given any θ. Namely, method $\mathcal{I}$ should satisfy Sufficiency Principle.

The other way of justifying the Sufficiency Principle is from a decision-theoretic point of view. Let $\delta : \mathcal{X} \to \Theta$ denote an estimator for θ and let $L(\delta; \theta)$ denote the loss incurred when we estimate θ with δ. $L : \Theta \times \Theta \to \mathbb{R}_+$ is called the loss function of estimating θ.

Assume $\Theta \subset \mathbb{R}^k$. We say that the loss function L is convex if $L(\delta; \theta)$ is of the form $l(\delta - \theta)$ with $l(\cdot)$ being convex on $\mathbb{R}^k$. **Rao-Blackwell Theorem** says that under a convex loss, any estimator δ is dominated by an estimator δ^* which is a function of T. Also, if δ is unbiased, then δ^* can be chosen to be unbiased.

Theorem 2 (Rao-Blackwell): Let $(\mathcal{X}, \mathcal{F}, \{\mathrm{P}_\theta\}_{\theta \in \Theta})$ be an experiment and T a sufficient statistic. Suppose $\Theta \subset \mathbb{R}^k$ and the loss function L is convex. Let δ be an estimator for θ. Then for any $\theta \in \Theta$,

$$\mathrm{E}_\theta[L(\delta^*; \theta)] \leq \mathrm{E}_\theta[L(\delta; \theta)],$$

where $\delta^* = \mathrm{E}_\theta[\delta \,|\, T] = \mathrm{E}[\delta \,|\, T]$.

Proof. Since T is sufficient, $\mathrm{E}_\theta[\delta \,|\, T]$ is the same across all θ and is written as $\mathrm{E}[\delta \,|\, T]$. Fix $\theta \in \Theta$,

$$\begin{aligned}
\mathrm{E}_\theta[L(\delta; \theta)] &= \mathrm{E}_\theta[l(\delta - \theta)] \\
&= \mathrm{E}_\theta[\mathrm{E}_\theta[l(\delta - \theta) \,|\, T]] \\
&\geq \mathrm{E}_\theta[l(\mathrm{E}_\theta[\delta - \theta \,|\, T])] \\
&= \mathrm{E}_\theta[l(\mathrm{E}_\theta[\delta \,|\, T] - \theta)] \\
&= \mathrm{E}_\theta[l(\delta^* - \theta)] = \mathrm{E}_\theta[L(\delta^*; \theta)].
\end{aligned}$$

The inequality in the middle holds by Jenson's Inequality for conditional expectations. It is easy to see that if δ is unbiased, then δ^* is also unbiased. □

5 Short History of Sufficient Statistic

The concept of sufficient statistic is first proposed by R.A. Fisher in his paper, *On the mathematical foundations of theoretical statistics*, in 1920. Two years later, he established the factorization condition as a sufficient condition for sufficient statistics. Years later, Neyman demonstrated, under certain conditions, the factorization condition also serves as a necessary condition for sufficient statistics in 1935. The general factorization theorem (*Theorem* 1) posed in this note is proposed and proved by Halmos and Savage in their paper, *Application* *of the Radon-Nikodym theorem to the theory of sufficient statistics*, in 1949.

References

Ibragimov, I. A., & Has'minskii, R. Z. (1981). *Statistical estimation: Asymptotic theory*. Springer.

Casella, G., & Berger, R. L. (2002). *Statistical inference* (2nd ed.). Duxbury.

Shao, J. (2003). *Mathematical statistics* (2nd ed.). Springer.

Lehmann, E. L., & Romano, J. P. (2022). *Testing statistical hypothesis* (4th ed.). Springer.