Notes for Math 281 Mathematical Statistics

Life is either an amazing adventure, or nothing.

Edited By

Haixiao Wang

The University of California
San Diego

$\boxed{\mathcal{HX}}$

2019
Antares

Preface

This lecture note is based on the handout of professor **Ery Arias-Castro**, who was teaching mathematical statistics in Winter and Spring quarter 2019 at UC San Diego, which is aimed at helping PhD students pass the qualifying exam in mathematical statistics. Meanwhile, I wrote down some supplementary materials for beginners, since I am new to Statistics. I majored in Physics and Economics during my undergraduate.

Since there are too much materials covered in the handout, I just focus on the essential parts, such as basic definitions and proof of important theorems. I wrote down some solutions to the practice problems in the book, which are quite helpful to understand the materials better. I have checked this note several times when I was preparing for the qualifying exam. However, there are still some mistakes which I might not find yet.

There are two main reference books, *Theory of Point Estimation, 2nd Edition*(TPE) written by E.L. Lehmann and George Casella, and *Testing Statistical Hypothesis, 3rd Edition*(TSH), written by E.L. Lehmann and Joseph P. Romano. While reading this note, these two books are extremely helpful if you want to learn more about mathematical statistics.

I hope that you will enjoy reading this note, as well as I also enjoyed writing this note.

> We shall not cease from exploration
> And the end of all our exploring
> Will be to arrive where we started
> And know the place for the first time.
> — *T. S. Eliot*

Contents

Chapter 1

Preparations

1.1 The Problem

Statistics is concerned with the collection of data and with their analysis and interpretation. We shall take the data as given and ask what they have to tell us. Depending on background knowledge, we may formalize in different assumptions with which the analysis is entered. There are three principal lines of approach

- **Data analysis.** Here, the data are analyzed on their own terms, essentially without extraneous assumptions. The principal aim is the organization and summarization of the data in ways that bring out their main features and clarify their underlying structure.

- **Classical inference and decision theory.** The observations are now postulated to be the values taken on by random variables which are assumed to follow a joint probability distribution, P, belonging to some known class $\mathcal{P}$. Frequently, the distributions are indexed by a parameter, say θ (not necessarily real-valued), taking values in a set, Ω so that

  $$\mathcal{P} = \{\mathbb{P}_\theta, \theta \in \Omega\} \tag{1.1}$$

  The aim of the analysis is then to specify a plausible value for θ (this is the problem of point estimation), or at least to determine a subset of Ω of which we can plausibly assert that it does, or does not, contain θ (estimation by confidence sets or hypothesis testing). Such a statement about θ can be viewed as a summary of the information provided by the data and may be used as a guide to action.

- **Bayesian analysis.** In this approach, it is assumed in addition that θ is itself a random variable (though unobservable) with a known distribution. This prior distribution (specified according to the problem) is modified in light of the data to determine a posterior distribution (the conditional distribution of θ given the data), which summarizes what can be said about θ on the basis of the assumptions made and the data.

Point estimation is one of the most common forms of statistical inference. One measures a physical quantity in order to estimate its value. Now, we specialize to point estimation. In terms of the model 1.1 , suppose that g is a real-valued function defined over Ω and that we would like to know the value of $g(\theta)$ (which may, of course, be θ itself). Unfortunately, θ, and hence $g(\theta)$, is unknown. However, the data can be used to obtain an estimate of $g(\theta)$, a value that one hopes will be close to $g(\theta)$.

In terms of the model 1.1, suppose that g is a real-valued function defined over Ω, and we would like to know the value of $g(\theta)$. Unfortunately, θ is unknown and hence $g(\theta)$ is unknown. However, the data can be used to obtain δ, an estimate of $g(\theta)$. We hope this value will be close to $g(\theta)$.

The formalization of an estimation problem involves two basic gradients

- A real-valued function g defined over a parameter space Ω, whose value at θ is to be estimated. We shall call this $g(\theta)$ the **estimand**.
- A random variable X (typically vector-valued) taking on values in a sample space $\mathcal{X}$ according to a distribution P_θ, which is known to belong to a family $\mathcal{P}$ as stated in 1.1

The problem is to determine a suitable **estimator**.

Definition 1.1.1. *An estimator is a real-valued function δ defined over the sample space. It is used to estimate an estimand, $g(\theta)$, a real-valued function of the parameter.*

Point estimation is one of the most common forms of statistical inference. One measures a physical quantity in order to estimate its value. Now, we specialize to point estimation. In terms of the model 1.1 , suppose that g is a real-valued function defined over Ω and that we would like to know the value of $g(\theta)$ (which may, of course, be θ itself). Unfortunately, θ, and hence $g(\theta)$, is unknown. However, the data can be used to obtain an estimate of $g(\theta)$, a value that one hopes will be close to $g(\theta)$.

The value of $\delta(x)$ taken on by $\delta(X)$ for the observed value x of X is the estimate of $g(\theta)$, which will be our "educated guess" for the unknown value.

Definition 1.1.2. *(Loss Function)*

A loss function (or cost function) L is a function that maps an event or values of one or more variables onto a real number intuitively representing some "cost" associated with the event. In other words, a loss function measures the consequences of estimating $g(\theta)$.

We shall assume that

$$L(\theta, d) \geq 0, \quad \forall \theta, d$$

and

$$L[\theta, g(\theta)] = 0, \quad \forall \theta$$

which means that the loss is 0 when the correct value is estimated.

Definition 1.1.3. *(Risk Function) The risk function measures the accuracy of an estimator δ*

$$R(\theta, \delta) = \mathbb{E}_\theta\{L[\theta, \delta(X)]\}$$

The long-term average loss resulting from the use of δ.

Ideally, we want the estimator δ that has the lowest risk consistency, i.e.

$$R(\theta, \delta) \leq R(\theta, \delta') \quad \forall \theta \in \Omega \text{ and all estimator } \delta'$$

This is not possible. If we let $\delta_{\theta_0}(x) = g(\theta_0) \quad \forall x \in \mathbb{R}$ at any given point θ_0, then we can reduce the risk at this given point to 0. There is no such uniformly best estimator, *i.e.*, no such estimator which simultaneously minimizes the risk of all $\theta_0 \in \Omega$, unless $g(\theta)$ is a constant, *i.e.*

$$
\begin{aligned}
& \min\{R(\theta_0, \delta_{\theta_0}(x))\} = 0 \implies R(\theta_0, \delta_{\theta_0}(x)) = 0 \quad \forall \theta_0 \in \Omega \\
\implies \quad & L[\theta_0, \delta_{\theta_0}(x)] = 0 \implies \theta_0 = \delta_{\theta_0}(x) \equiv c
\end{aligned}
$$

One way of avoiding this difficulty is to restrict the class of estimators by ruling out estimators that too strongly favor one or more values of θ at the cost of neglecting other possible values. This can be achieved by requiring the estimator to satisfy some condition which enforces a certain degree of impartiality. (TPE Page 5)

1.2 Basic Concepts

Definition 1.2.1. *(Exponential Families)*
A family $\mathbb{P}_\theta$ of distributions is said to form an $s-$dimensional exponential family if the distributions $\mathbb{P}_\theta$ have densities of the form

$$p_\theta(x) = \exp\left[\sum_{i=1}^{s} \eta_i(\theta) T_i(x) - B(\theta)\right] h(x)$$

with respect to some common measure μ. Here, the η_i and B are real-valued functions of the parameters and the T_i are real-valued statistics, and x is a point in the sample space $\mathcal{X}$, the support of the density.

Frequently, it is more convenient to use the η_i as the parameters and write the density in the canonical form

$$p(x|\eta) = \exp\left[\sum_{i=1}^{s} \eta_i T_i(x) - A(\eta)\right] h(x) \tag{1.2}$$

The function p given by equation 1.2 is non-negative and is therefore a probability density with respect to the given μ, provided its integral with respect to μ equals 1. A constant $A(\eta)$ for which this is the case exists if and only if

$$\int e^{\sum_{i=1}^{s} \eta_i T_i(x)} h(x) d\mu(x) < \infty \tag{1.3}$$

*The set Ξ of points $\boldsymbol{\eta} = (\eta_1, \cdots, \eta_s)$ for which equation 1.3 holds is called the **natural parameter space** of the family 1.2. $\boldsymbol{\eta}$ is called the **natural parameter**.*

*A reduction is also possible when the η's satisfy a linear constraint. In the latter case, the natural parameter space will be a convex set which lies in a linear subspace of dimension less than s. If the representation 1.2 is minimal in the sense that neither the T's nor the η's satisfy a linear constraint, the natural parameter space will then be a convex set in E_s containing an open s-dimensional rectangle. If 1.2 is minimal and the parameter space contains an s-dimensional rectangle, the family 1.2 is said to be of **full rank**.*

Example 1: (Normal family) If X has the $N(\xi, \sigma_2)$ distribution, then $\theta = (\xi, \sigma_2)$ and the density with respect to Lebesgue measure is

$$p_\theta(x) = \exp\left[\frac{\xi}{\sigma^2}x - \frac{1}{2\sigma^2}x^2 - \frac{\xi^2}{2\sigma^2}\right]\frac{1}{\sqrt{2\pi}\sigma}$$

a two parameter exponential family with natural parameters $(\eta_1, \eta_2) = (\frac{\xi}{\sigma^2}, -\frac{1}{2\sigma^2})$ and natural parameter space $\mathcal{R} \times (-\infty, 0)$. ♠

Definition 1.2.2. *(Unidentifiable Parameter)*

If X is distributed according to p_θ, then θ is said to be unidentifiable on the basis of X if there exist $\theta_1 \neq \theta_2$ for which $\mathbb{P}_{\theta_1} = \mathbb{P}_{\theta_2}$

Definition 1.2.3. *(Completeness of a Statistic)*

Consider a random variable X whose probability distribution belongs to a parametric model $\mathbb{P}_\theta$ parametrized by θ.

Say T is statistic; that is, the composition of a measurable function with a random sample X1,...,Xn.

The statistic T is said to be complete for the distribution of X if, for every measurable function g,

$$\mathbb{E}\left[g(T)\right] = 0 \implies \mathbb{P}_\theta(g(T) = 0) = 1 \quad \forall\theta$$

The statistic T is said to be boundedly complete for the distribution of X if this implication holds for every measurable function g that is also bounded.

Definition 1.2.4. *(Sufficient statistic)*

A statistic $t = T(X)$ is sufficient for underlying parameter θ precisely if the conditional probability distribution of the data X, given the statistic $t = T(X)$, does not depend on the parameter θ

Theorem 1.2.1. *(Fisher–Neyman factorization theorem) Fisher's factorization theorem or factorization criterion provides a convenient characterization of a sufficient statistic. If the probability density function is $f_\theta(x)$, then T is sufficient for θ if and only if nonnegative functions g and h can be found such that*

$$f_\theta(x) = h(x)g_\theta(T(x))$$

i.e.the density f can be factored into a product such that one factor, h, does not depend on θ and the other factor, which does depend on θ, depends on x only through $T(x)$.

Theorem 1.2.2. ***(Rao-Blackwell Theorem)***

Let X be a random variable with distribution $\mathbb{P}_\theta \in \mathcal{P} = \{\mathbb{P}_{\theta'}, \theta' \in \Omega\}$, and let T be sufficient for $\mathcal{P}$. Let δ be an estimator of an estimand $g(\theta)$, and let the loss function $L(\theta, d)$ be a strictly convex function of d. Then if δ has finite expectation and risk,

$$R(\theta, \delta) = \mathbb{E}\left[L(\theta, \delta(X))\right] < \infty$$

and if

$$\eta(t) = \mathbb{E}\left[\delta(X)|t\right]$$

then the risk of estimator $\eta(T)$ satisfies

$$R(\theta, \eta) < R(\theta, \delta)$$

unless $\delta(X) = \delta(T)$

Theorem 1.2.3. *(Jensen's Inequality)*

If ϕ is a convex function defined over an open interval I , and X is a random variable with $\mathbb{P}(X \in I) = 1$ and finite expectation, then

$$\phi[\mathbb{E}[X]] \leq \mathbb{E}[\phi(X)]$$

If ϕ is strictly convex, the inequality is strict unless X is a constant with probability.

Corollary 1.2.4. *If X is a non-constant positive random variable with finite expectation, then*

$$\frac{1}{\mathbb{E}[X]} < \mathbb{E}\left[\frac{1}{X}\right]$$

and

$$\mathbb{E}[log(X)] < \log(\mathbb{E}[X])$$

Definition 1.2.5. *(Ancillary Statistic)*

An ancillary statistic is a measure of a sample whose distribution does not depend on the parameters of the model. An ancillary statistic is a pivotal quantity that is also a statistic. Ancillary statistics can be used to construct prediction intervals.

Example 2: Suppose $X_1, \cdots, X_n$ are independent and identically distributed, and are normally distributed with unknown expected value μ and known variance 1. Let

$$\bar{X}_n = \frac{X_1 + \cdots + X_n}{n}$$

be the sample mean. Then $\max(X_1, \cdots, X_n) - \min(X_1, \cdots, X_n)$ is ancillary statistic, because its sampling distributions do not change when μ changes. ♠

Theorem 1.2.5. *(Basu's Theorem)*

If T is a complete sufficient statistic for the family $\mathcal{P} = \{\mathbb{P}_\theta, \theta \in \Omega\}$, then any ancillary statistic V is independent of T.

Proof If V is ancillary, then probability $p_A = \mathbb{P}[V \in A]$ is independent of θ for all A. Let $\eta_A(t) = \mathbb{P}[V \in A | T = t]$, then $\mathbb{E}_\theta[\eta_A(T)] = \mathbb{P}[V \in A] = p_A$. Hence, by completeness, $\eta_A(t) = p_A$ a.e. in $\mathcal{P}$. This establishes the independence of V and T. $\square$

Chapter 2

Unbiasedness

As what we have already pointed out, there is no such estimator with uniformly minimum risk.

2.1 UMVU Estimators

Definition 2.1.1. *(Unbiasedness)*
An estimator $\delta(x)$ of $g(\theta)$ is unbiased if

$$\mathbb{E}_\theta\left[\delta(X)\right] = g(\theta) \quad \forall\theta \in \Omega$$

When used repeatedly, an unbiased estimator in the long run will estimate the right value "on the average". This feature is attractive. However, sometimes, the class of unbiased estimators may be empty.
Example 3: (Nonexistence of unbiased estimator)
Let X be distributed according to the binomial distribution $B(n,\theta)$, $\theta \in [0,1]$. The estimand is $g(\theta)$, and the estimator is $\theta : \mathcal{X} \mapsto \mathbb{R}$. The unbiasedness of estimator δ requires that

$$\mathbb{E}_\theta\left[\delta(X)\right] = \sum_{k=0}^{n} \delta(k)\binom{n}{k}\theta^k(1-\theta)^k = g(\theta) \quad \forall\theta \in [0,1] \tag{2.1}$$

However, equation 2.1 will not hold when $g(\theta)$ is not a polynomial of degree at most n. For example, let $g(\theta) = \frac{1}{\theta}$. As $\theta \to 0$,the left hand side of 2.1 will tend to $\delta(0)$ and the right hand side will tend to ∞ ♠

If there exists an unbiased estimator δ of g, the estimand g will be called U-estimable. Even when g is U-estimable, there is no guarantee that any of its unbiased estimators are desirable in other ways.

Lemma 2.1.1. *(Lemma 2.1.4 in TPE)*
If δ_0 is any unbiased estimator of $g(\theta)$, the totality of unbiased estimators is given by $\delta = \delta_0 - U$ where U us any unbiased estimator of zero, i.e., $\mathbb{E}_\theta\left[U\right] = 0 \quad \forall\theta \in \Omega$

To illustrate this approach, suppose the loss function is squared error. The risk of an unbiased estimator δ is then just the variance of δ. Restricting attention to estimator δ_0, δ and U with finite variance, we have, if δ_0 unbiased

$$\mathbb{V}\mathrm{ar}(\delta) = \mathbb{V}\mathrm{ar}(\delta_0 - U) = \mathbb{E}\left[(\delta_0 - U)^2\right] - \{\mathbb{E}\left[\delta_0 - U\right]\}^2 = \mathbb{E}\left[(\delta_0 - U)^2\right] - [g(\theta)]^2$$

So that the variance of δ is minimized by $\mathbb{E}\left[(\delta_0 - U)^2\right]$.

Definition 2.1.2. *(UMVU and LMVU)*
*An unbiased estimator $\delta(x)$ of $g(\theta)$ is the **uniform minimum variance unbiased (UMVU)** estimator of $g(\theta)$ if $\mathbb{V}\mathrm{ar}_\theta(\delta(x)) \le \mathbb{V}\mathrm{ar}_\theta(\delta'(x))$ for all $\theta \in \Omega$, where $\delta'(x)$ is another unbiased estimator of $g(\theta)$. The estimator $\delta(x)$ is **locally minimum variance unbiased (LMVU)** at $\theta = \theta_0$ if $\mathbb{V}\mathrm{ar}_{\theta_0}(\delta(x)) \le \mathbb{V}\mathrm{ar}_{\theta_0}(\delta'(x))$ for any other estimator $\delta'(x)$.*

Example 4: (Locally Best Unbiased Estimation, Example 2.1.5 in TPE)
Let X take on the values $-1, 0, 1, \cdots$ with probabilities $\mathbb{P}\left[X = -1\right] = p$, $\mathbb{P}\left[X = k\right] = q^2p^k, k = 0, 1, 2, \cdots$, where $0 < p < 1$ and $q = 1 - p$. Consider the problem estimating

1. p
2. q^2

Simple unbiased estimators of p and q^2 are

$$\delta_{-1}(X) = \begin{cases} 1, & if\ X = -1 \\ 0, & otherwise \end{cases} \quad \delta_0(X) = \begin{cases} 1, & if\ X = 0 \\ 0, & otherwise \end{cases}$$

Suppose that U is an estimator of 0. U is unbiased if and only if $U(k) = -kU(-1)$, since

$$\begin{aligned} \mathbb{E}\left[U\right] &= \sum_{k=-1}^{\infty} \mathbb{P}\left[X = k\right] U(k) = pU(-1) + \sum_{k=0}^{\infty}(-q^2p^k k)U(-1) \\ &= pU(-1) - U(-1)q^2 \sum_{k=0}^{\infty} kp^k \\ &= U(-1)[p - q^2 \frac{p}{(1-p)^2}] = 0 \end{aligned}$$

or equivalently, $U(k) = ak$ for some a and $k = -1, 0, 1, \cdots$.

Now, we want to obtain an estimator, which minimizes the variance at a specific parameter $\mathbb{P}\left[X = -1\right] = p_0$. We are going to minimizes

$$\sum_{k=-1}^{\infty} \mathbb{P}\left[X = k\right] \left[\delta_i(k) - ak\right]^2, \quad i = -1, 0$$

The corresponding 2 minimizing values of a are

$$a_{-1}^* = \frac{-p_0}{p_0 + q_0^2 \sum_{k=1}^{\infty} k^2 p_0^k}, \quad a_0^* = 0$$

for estimating p and q^2.

Now, the new estimator $\delta_0^*(X) = \delta_0(X) - a_0^*X = \delta_0(X)$ minimizes the variance among all unbiased estimators not only when $p = p_0$, since a_0^* does not depend on p_0. Therefore, $\delta_0(X)$ is UMVU.

On the other hand, $\delta_{-1}^*(X) = \delta_1(X) - a_1^*X$ depends on p_0, it only minimizes the variance at $p = p_0$, so it is LMVU. ♠

Proposition 2.1.2. *If a UMVU estimator exists, then it is unique.*

Proof Suppose that δ_0, δ_1 are unbiased, then $\delta_t = (1-t)\delta_0 + t\delta_1$ is unbiased for all $t \in \mathbb{R}$.

$$\begin{aligned}
\mathbb{V}\mathrm{ar}_\theta(\delta_t) &= (1-t)^2\mathbb{V}\mathrm{ar}_\theta(\delta_0) + 2t(1-t)\mathbb{C}\mathrm{ov}_\theta(\delta_0,\delta_1) + \mathrm{t}^2\mathbb{V}\mathrm{ar}_\theta(\delta_1) \\
& \quad (\text{Since } \mathbb{C}\mathrm{ov}_\theta(\delta_0,\delta_1) \le \sqrt{\mathbb{V}\mathrm{ar}_\theta(\delta_0)\mathbb{V}\mathrm{ar}_\theta(\delta_1)}) = \sigma_0\sigma_1 \\
&\le (1-t)^2\sigma_0^2 + 2t(1-t)\sigma_0\sigma_1 + t^2\sigma_1^2 \\
&= [(1-t)\sigma_0 + t\sigma_1]^2 \\
&\le \max\{\mathbb{V}\mathrm{ar}_\theta(\delta_0), \mathbb{V}\mathrm{ar}_\theta(\delta_1)\} \quad t \in [0,1]
\end{aligned}$$

According to the definition of UMVU, $\mathbb{V}\mathrm{ar}_\theta(\delta_t) = \mathbb{V}\mathrm{ar}_\theta(\delta_0) = \mathbb{V}\mathrm{ar}_\theta(\delta_1)$. Thus

$$\mathbb{C}\mathrm{ov}_\theta(\delta_0,\delta_1) = \sqrt{\mathbb{V}\mathrm{ar}_\theta(\delta_0)\mathbb{V}\mathrm{ar}_\theta(\delta_1)})$$

Consider 2 random variables X, Y with $Y = aX + b$. The condition for $\mathbb{C}\mathrm{ov}_\theta(\mathrm{X},\mathrm{Y}) = \sqrt{\mathbb{V}\mathrm{ar}_\theta(\mathrm{X})\mathbb{V}\mathrm{ar}_\theta(\mathrm{Y})}) = \sigma_\mathrm{X}\sigma_\mathrm{Y}$ is that $b = 0, a = 1$. $\square$

Theorem 2.1.3. *(Theorem 2.1.7 in TPE)*

Let X have distribution $\mathbb{P}_\theta, \theta \in \Omega$. Let δ be an estimator in $\triangle$, and let $\mathcal{U}$ denote the set of all unbiased estimators of 0 which are in $\triangle$. Then, a necessary and sufficient condition, i.e., if and only if, for δ to be a UMVU estimator of its expectation $g(\theta)$ is that

$$\mathbb{E}_\theta[\delta U] = 0 \quad \forall U \in \mathcal{U} \text{ and } \forall \theta \in \Omega$$

Proof of Theorem

Necessity. Suppose δ is UMVU for estimating its expectation $g(\theta)$. Fix $U \in \mathcal{U}, \theta \in \Omega$ and for arbitrary real λ, let $\delta' = \delta + \lambda U$. Then, δ' is also an unbiased estimator of $g(\theta)$, since $\mathbb{E}[U] = 0$. So that

$$\begin{aligned}
\mathbb{V}\mathrm{ar}_\theta(\delta) &\le \mathbb{V}\mathrm{ar}_\theta(\delta + \lambda U) \\
&= \mathbb{V}\mathrm{ar}_\theta(\delta) + \lambda^2\mathbb{V}\mathrm{ar}_\theta(U) + 2\lambda\mathbb{C}\mathrm{ov}_\theta(\delta,\mathrm{U}) \quad \forall\lambda
\end{aligned}$$

Thus

$$\lambda^2\mathbb{V}\mathrm{ar}_\theta(U) + 2\lambda\mathbb{C}\mathrm{ov}_\theta(\delta,\mathrm{U}) \ge 0 \quad \forall\lambda$$

As a result

$$\min_\lambda\{\lambda^2\mathbb{V}\mathrm{ar}_\theta(U) + 2\lambda\mathbb{C}\mathrm{ov}_\theta(\delta,\mathrm{U})\} = -\frac{[\mathbb{C}\mathrm{ov}_\theta(\delta,\mathrm{U})]^2}{\mathbb{V}\mathrm{ar}_\theta(\mathrm{U})} \ge 0$$

Therefore, $\mathbb{C}\mathrm{ov}_\theta(\delta,\mathrm{U}) = \mathbb{E}_\theta[\delta\mathrm{U}] - \mathbb{E}_\theta[\delta]\mathbb{E}_\theta[\mathrm{U}] = \mathbb{E}_\theta[\delta\mathrm{U}] = 0$. Proved.

Sufficiency.(Homework) Suppose that $\mathbb{E}_\theta[\delta U] = 0$ for all $U \in \mathcal{U}$. We want to show that δ is UMVU. Let δ' be any unbiased estimator of $g(\theta)$. There is nothing to show if $\mathbb{V}\mathrm{ar}_\theta(\delta') = \infty$. Therefore, we assume that $\mathbb{V}\mathrm{ar}_\theta(\delta') < \infty$. Then $\delta - \delta' \in \mathcal{U}$, because $\mathbb{E}_\theta[\delta - \delta'] = 0$ and

$$\begin{aligned}
\mathbb{V}\mathrm{ar}_\theta(\delta - \delta') &= \mathbb{V}\mathrm{ar}_\theta(\delta) - 2\mathbb{C}\mathrm{ov}_\theta(\delta,\delta') + \mathbb{V}\mathrm{ar}_\theta(\delta') < \infty \\
& \quad \text{Since } [\mathbb{C}\mathrm{ov}_\theta(\delta,\delta')]^2 \le \mathbb{V}\mathrm{ar}_\theta(\delta)\mathbb{V}\mathrm{ar}_\theta(\delta') < \infty
\end{aligned}$$

Then we know

$$\mathbb{E}_\theta\left[\delta(\delta-\delta')\right]=0 \implies \mathbb{E}_\theta\left[\delta^2\right]=\mathbb{E}_\theta\left[\delta\delta'\right]$$

Since $\mathbb{E}_\theta[\delta]=\mathbb{E}_\theta[\delta']$, then

$$\mathbb{V}\mathrm{ar}_\theta(\delta) = \mathbb{C}\mathrm{ov}_\theta(\delta,\delta') \leq \sqrt{\mathbb{V}\mathrm{ar}_\theta(\delta)\mathbb{V}\mathrm{ar}_\theta(\delta')}$$

The second inequality comes from **Schwarz inequality**

$$\left(\int fg\right)^2 \leq \int f^2 d\mu \int g^2 d\mu, \quad \text{plug in } f=\delta-\mathbb{E}_\theta[\delta] \text{ and } g=\delta'-\mathbb{E}_\theta[\delta']$$

Therefore, proved.

□

There are 3 possible situations:

1. No non-constant U-estimable function has a UMVU estimator.
   Example 5: (Nonexistence of UMVU Estimator)
   Let $X_1,\cdots,X_n$ be a sample from a discrete distribution which assigns probability 1/3 to each of the points $\theta-1,\theta,\theta+1$. Let θ range over the integers. Then no constant function of θ has a $UMVU$ estimator. (Problem 1.9) ♠

2. Some non-constant U-estimable function have UMVU estimators, but not all of them.
   Example 6: Let $X_1,\cdots,X_n \overset{i.i.d.}{\sim} Bernoulli(\theta)$. $T=\sum_{i=1}^n X_i \sim Binomial(n,\theta)$ is sufficient. There are no unbiased estimator of $g(\theta)$, unless g is a polynomial of degree $\leq n$. ♠

3. Every U-estimable function has a UMVU estimator.

Lemma 2.1.4. *(Lemma 2.1.10 in TPE)*
Let X be distributed according to a distribution from $\mathcal{P}=\{\mathbb{P}_\theta,\theta\in\Omega\}$, and let T be a complete sufficient statistic for $\mathcal{P}$. Then every U-estimable function $g(\theta)$ has one and only one unbiased estimator that is a function of T.

Proof of Lemma If δ_1 and δ_2 are two unbiased estimators of $g(\theta)$, their difference $f(T)=\delta_1(T)-\delta_2(T)$ satisfies

$$\mathbb{E}_\theta[f(T)]=0 \quad \forall\theta\in\Omega$$

By the completeness of T, $\delta_1(T)=\delta_2(T)$ a.e. $\mathcal{P}$, as was to be proved. □

Theorem 2.1.5. *(Theorem 2.1.11 in TPE)*
Let X be distributed according to a distribution in $\mathcal{P}=\{\mathbb{P}_\theta,\theta\in\Omega\}$ and suppose that T is a complete sufficient statistic for $\mathcal{P}$.

1. *For every U-estimable function $g(\theta)$, there exists an unbiased estimator that uniformly minimizes the risk for any loss function $L(\theta, d)$ which is convex in its second argument; therefore, this estimator in particular is UMVU.*

2. *The UMVU estimator of (1) is the unique unbiased estimator which is a function of T. It is the unique unbiased estimator with minimum risk, provided its risk is finite and L is strictly convex in d.*

Proof Let $\delta(x)$ be an unbiased estimator for $g(\theta)$. Define $\eta(t) = \mathbb{E}_\theta\left[\delta(X)|T=t\right]$, while θ is any fixed element in Ω. In addition, $\eta(t)$ is unbiased for $g(\theta)$.

$$\mathbb{E}_\theta\left[\eta(T)\right] = \mathbb{E}_\theta\left[\mathbb{E}_\theta\left[\delta(X)|T=t\right]\right] = \mathbb{E}_\theta\left[\delta(X)\right] = g(\theta)$$

By Rao-Blackwell Theorem 1.2.2,

$$\begin{aligned} \mathbb{E}_\theta\left[(\eta(T)-g(\theta))^2\right] &= \mathbb{E}_\theta\left[\left[\mathbb{E}_\theta\left[\delta(X)|T=t\right]-g(\theta)\right]^2\right] \\ \text{Jensen Inequality} &\leq \mathbb{E}_\theta\left[\mathbb{E}_\theta\left[(\delta(X)-g(\theta))^2|T=t\right]\right] \\ &= \mathbb{E}_\theta\left[(\delta(X)-g(\theta))^2\right] \end{aligned}$$

$\eta(T)$ does not depend on the choice of $\delta(X)$. If there exists another unbiased estimator for $g(\theta)$, denoted by δ', then $\eta'(t) = \mathbb{E}_\theta\left[\delta'(X)|T=t\right]$. Thus $\mathbb{E}_\theta\left[\eta(T)-\eta'(T)\right] = 0 \forall\theta \in \Omega$. By completeness of T, $\eta(T) = \eta'(T)$. □

Two main method for finding a UMVU when T is complete sufficient

1. Method 1: By completeness, η is determined by $\mathbb{E}_\theta\left[\eta(T)\right] = g(\theta)$
2. Method 2: Find δ unbiased for $g(\theta)$, and compute $\eta(t) = \mathbb{E}_\theta\left[\delta(X)|T=t\right]$

2.2 Continuous One and Two Sample Problems

Example 7: (1 Normal Sample, Example 2.2.1 in TPE)
$X_1, \cdots, X_n \overset{i.i.d.}{\sim} N(\xi, \sigma^2)$. Assume that ξ and σ are both unknown.

$$\textit{Likelihood} = \frac{e^{-\frac{1}{2\sigma^2}(x_1-\xi)^2}}{\sqrt{2\pi\sigma^2}} \cdots \frac{e^{-\frac{1}{2\sigma^2}(x_n-\xi)^2}}{\sqrt{2\pi\sigma^2}} \propto \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i-\xi)^2\right] \propto \exp\left[-\frac{1}{2\sigma^2}A + \frac{\xi}{\sigma^2}B - \frac{n\xi^2}{2\sigma^2}\right]$$

where $A = \sum_{i=1}^n x_i^2$, $B = \sum_{i=1}^n x_i$. Thus, $X = (X_1, \cdots, X_n) \mapsto T = (A, B)$ is sufficient. Since T is full rank, then $T = (A, B)$ is complete sufficient. As a result, $(\bar{X}, S^2)$ is complete sufficient, where $\bar{X} = \frac{1}{n}\sum_{i=1}^n x_i$, $S^2 = \frac{1}{n-1}\sum_{i=1}^n (x_i - \bar{X})^2$. Then $\mathbb{E}_\theta\left[\bar{X}\right] = \xi$, $\mathbb{E}_\theta\left[S^2\right] = \sigma^2$. Therefore, $\bar{X}$ is a UMVU for ξ, and S^2 is a UMVU for σ^2.

(Homework) Find the UMVU for $\frac{\xi}{\sigma}$

Note that $\frac{(n-1)S^2}{\sigma^2}$ has a chi square distribution with $n-1$ degree of freedom, χ^2_{n-1}. $\bar{X}$ has the $N(\xi, \frac{\sigma^2}{n})$ distribution. $\bar{X}$ and $\frac{(n-1)S^2}{\sigma^2}$ are independent.

Let's evaluate $\mathbb{E}[S]$. To simplify, let $q = \frac{(n-1)S^2}{\sigma^2}$, then $S = \sqrt{\frac{q\sigma^2}{n-1}}$.

$$\begin{aligned}
\mathbb{E}[S] &= \int_0^\infty \sqrt{\frac{\sigma^2}{n-1}}\sqrt{q} f(q, n-1)dq = \sqrt{\frac{\sigma^2}{n-1}} \int_0^\infty \sqrt{q}\frac{q^{\frac{n-1}{2}-1}e^{-\frac{q}{2}}}{2^{\frac{n-1}{2}}\Gamma(\frac{n-1}{2})}dq \\
&= \sqrt{\frac{\sigma^2}{n-1}}\frac{\sqrt{2}}{\Gamma(\frac{n-1}{2})}\int_0^\infty \left(\frac{q}{2}\right)^{\frac{n}{2}-1} e^{-\frac{q}{2}}d(\frac{q}{2}) = \sqrt{\frac{2\sigma^2}{n-1}}\frac{\Gamma(\frac{n}{2})}{\Gamma(\frac{n-1}{2})} \\
&\implies \mathbb{E}\left[\sqrt{\frac{n-1}{2}}\frac{\Gamma(\frac{n-1}{2})}{\Gamma(\frac{n}{2})}S\right] = \sigma
\end{aligned}$$

We use $k_{n,r}$ to denote this constant which is relevant to n, where

$$k_{n,r} = \sqrt{\frac{n^r}{2^r}}\frac{\Gamma(\frac{n}{2})}{\Gamma(\frac{n+r}{2})}$$

Thus, the UMVU estimator for σ is $k_{n-1,1}S$. Actually, for $r > 1-n$,

$$\mathbb{E}[S^r] = \int_0^\infty \sqrt{\frac{\sigma^{2r}}{(n-1)^r}}\sqrt{q^r} f(q, n-1)dq = \sqrt{\frac{\sigma^{2r}}{(n-1)^r}} \int_0^\infty \sqrt{q^r}\frac{q^{\frac{n-1}{2}-1}e^{-\frac{q}{2}}}{2^{\frac{n-1}{2}}\Gamma(\frac{n-1}{2})}dq \tag{2.2}$$

$$= \sqrt{\frac{\sigma^{2r}}{(n-1)^r}}\frac{\sqrt{2^r}}{\Gamma(\frac{n-1}{2})}\int_0^\infty \left(\frac{q}{2}\right)^{\frac{n+r-1}{2}-1} e^{-\frac{q}{2}}d(\frac{q}{2}) = \sqrt{\frac{2^r\sigma^{2r}}{(n-1)^r}}\frac{\Gamma(\frac{n+r-1}{2})}{\Gamma(\frac{n-1}{2})} \tag{2.3}$$

$$\implies \mathbb{E}\left[\sqrt{\frac{(n-1)^r}{2^r}}\frac{\Gamma(\frac{n-1}{2})}{\Gamma(\frac{n+r-1}{2})}S\right] = \sigma^r \tag{2.4}$$

Thus, the UMVU estimator for σ^r is $k_{n-1,r}S^r$. We now compute $\mathbb{E}\left[\frac{\bar{X}}{S}\right]$

$$\mathbb{E}\left[\frac{\bar{X}}{S}\right] = \mathbb{E}\left[\frac{1}{S}\right]\mathbb{E}\left[\bar{X}\right] = \xi\frac{1}{k_{n-1,-1}\sigma}$$

Therefore, the UMVU estimator for $\frac{\xi}{\sigma}$ is $k_{n-1,-1}\frac{\bar{X}}{S}$. ♠

Example 8: (2 Normal Sample)

$X_1, \cdots, X_m \overset{i.i.d.}{\sim} N(\xi, \sigma^2)$, $Y_1, \cdots, Y_n \overset{i.i.d.}{\sim} N(\eta, \tau^2)$. Assume that ξ, σ, η, τ are unknown.

$$\begin{aligned}
\textit{Likelihood} &= \prod_{i=1}^m \frac{e^{-\frac{1}{2\sigma^2}(x_i-\xi)^2}}{\sqrt{2\pi\sigma^2}}\prod_{j=1}^n \frac{e^{-\frac{1}{2\tau^2}(y_j-\eta)^2}}{\sqrt{2\pi\tau^2}} \\
&\propto \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^m x_i^2 + \frac{\xi}{\sigma^2}\sum_{i=1}^m x_i - \frac{1}{2\tau^2}\sum_{j=1}^n y_j^2 + \frac{\eta}{\tau^2}\sum_{j=1}^n y_j\right]
\end{aligned}$$

$(\bar{X}, S^2, \bar{Y}, T^2)$ is complete and sufficient, where

$$S^2 = \frac{1}{m-1}\sum_{i=1}^m (x_i - \bar{X})^2, \quad T^2 = \frac{1}{n-1}\sum_{j=1}^n (y_j - \bar{Y})^2$$

Note that $\frac{(m-1)S^2}{\sigma^2} \sim \chi^2_{m-1}$, $\frac{(n-1)T^2}{\tau^2} \sim \chi^2_{n-1}$, and ξ, σ, η, τ are independent. Suppose that we want to estimate $\frac{\sigma^2}{\tau^2}$, using the result of previous example 2.4,

$$\mathbb{E}\left[\frac{S^2}{T^2}\right] = \mathbb{E}\left[S^2\right]\mathbb{E}\left[\frac{1}{T^2}\right] = (m-1)\sigma^2\frac{1}{k_{n-1,-2}\tau^2}$$

Thus the UMVU estimator for $\frac{\sigma^2}{\tau^2}$ is $\frac{k_{n-1,-2}S^2}{(m-1)T^2}$

(Homework) Find the UMVU estimator for $\frac{\sigma}{\tau}$

$$\mathbb{E}\left[\frac{S}{T}\right] = \mathbb{E}\left[S\right]\mathbb{E}\left[\frac{1}{T}\right] = \frac{\sigma}{k_{m-1,1}}\frac{1}{k_{n-1,-1}\tau}$$

Thus the UMVU estimator for $\frac{\sigma}{\tau}$ is $k_{m-1,1}k_{n-1,-1}\frac{S}{T}$ ♠

2.3 Discrete Distributions

Example 9: (Binomial Experiment, Example 2.3.1 in TPE)

$X_1, \cdots, X_n \overset{i.i.d.}{\sim} Bernoulli(\theta)$, $x_1, \cdots, x_n \in \{0, 1\}$.

$$\mathbb{P}\left[X_1 = x_1, \cdots, X_n = x_n\right] = \prod_{i=1}^{n}\theta^{x_i}(1-\theta)^{1-x_i} = \theta^y(1-\theta)^{n-y}$$

$$\mathbb{P}\left[Y = y\right] = \binom{n}{y}\theta^y(1-\theta)^{n-y} = \binom{n}{y}\exp\left[y\log(\frac{\theta}{1-\theta})\right](1-\theta)^n$$

where $Y = \sum_{i=1}^n X_i \sim Binomial(n, \theta)$ is sufficient by factorization theorem 1.2.1. In this exponential family, it is of full rank. So Y is complete. Suppose we want to estimate θ, $\frac{Y}{n}$ is a UMVU estimator for θ.

Question: What kind of functions of θ admits an unbiased estimator?

Let's consider $g(\theta) = \theta^2$, $\quad (n \geq 2)$.

$$\mathbb{E}_\theta\left[Y^2\right] = \mathbb{V}\mathrm{ar}_\theta(Y) + (\mathbb{E}_\theta\left[Y\right])^2 = n\theta(1-\theta) + (n\theta)^2 = n\theta + n(n-1)\theta^2$$

Thus

$$\mathbb{E}_\theta\left[\frac{Y^2 - Y}{n(n-1)}\right] = \theta^2$$

$\frac{Y^2-Y}{n(n-1)}$ is a UMVU estimator for θ.

We can also use the conditioning method. Note that $\mathbb{E}_\theta\left[X_1X_2\right] = \mathbb{E}_\theta\left[X_1\right]\mathbb{E}_\theta\left[X_2\right] = \theta^2$. Let $\delta(X_1, \cdots, X_n) = X_1X_2$, $\delta \in \{0, 1\}$.

$$\begin{aligned}
\mathbb{P}\left[\delta = 1 | Y = y\right] &= \frac{\mathbb{P}\left[X_1 = 1, X_2 = 1, Y = y\right]}{\mathbb{P}\left[Y = y\right]} \\
\mathbb{P}\left[Y = y\right] &= \binom{n}{y}\theta^y(1-\theta)^{n-y} \\
\mathbb{P}\left[X_1 = 1, X_2 = 1, Y = y\right] &= \mathbb{P}\left[X_1 = 1, X_2 = 1, \sum_{i=3}^{n} X_i = y - 2\right] \\
&= \theta^2\binom{n-2}{y-2}\theta^{y-2}(1-\theta)^{(n-2)-(y-2)}
\end{aligned}$$

$$\mathbb{E}_\theta\left[\delta = 1 | Y = y\right] = \frac{\binom{n-2}{y-2}}{\binom{n}{y}} = \frac{y(y-1)}{n(n-1)}$$

Note $g(\theta) = \frac{1}{\theta}$ does not admit an unbiased estimator, *i.e.*, there is no UMVUE for $\frac{1}{\theta}$. ♠

Example 10: (Inverse Binomial Experiment, Example 2.3.2 in TPE)

$X_1, \cdots, X_n \overset{i.i.d.}{\sim} Bernoulli(\theta)$, $x_1, \cdots, x_n \in \{0, 1\}$. Let $X = Y + m$ denote the number of total trials until m success are observed, where m is given. Therefore, $Y \sim Negative\ Binomial(m, \theta)$, Y is the number of failures.

$$\mathbb{P}_\theta[Y = y] = \binom{m+y-1}{m-1}\theta^m(1-\theta)^y \quad \text{The last trial should be successful.}$$

By factorizationtheorem 1.2.1 Y is sufficient. Y is full rank, thus it is complete. We can use the "negative binomial"

$$\begin{aligned}
\binom{m+y-1}{m-1} &= \frac{(m+y-1)!}{y!(m-1)!} = \frac{(m+y-1)(m+y-2)\cdots(m+1)m}{y!} \\
&= (-1)^y \frac{(-m)(-m-1)\cdots(-m-y+2)(-m-y+1)}{y!} \\
&= (-1)^y \binom{-m}{y}
\end{aligned}$$

Note that by the last expression and the binomial series, for every $0 \le \theta < 1$,

$$(1-\theta)^{-m} = \sum_{y=0}^{\infty} \binom{-m}{y}(-\theta)^y = \sum_{y=0}^{\infty} \binom{m+y-1}{y}\theta^y$$

and $\sum_{y=0}^{\infty} \mathbb{P}_\theta[Y = y] = 1$, we have

$$\begin{aligned}
\mathbb{E}_\theta[Y] &= \sum_{y=0}^{\infty} y\mathbb{P}_\theta[Y = y] = \sum_{y=0}^{\infty} y\binom{m+y-1}{m-1}\theta^m(1-\theta)^y \\
&= \sum_{y=1}^{\infty} \frac{(m+y-1)!}{(y-1)!(m-1)!}\theta^m(1-\theta)^y \\
&= m\frac{1-\theta}{\theta}\sum_{y=1}^{\infty}\binom{m+y-1}{y-1}\theta^{m+1}(1-\theta)^{y-1} \\
(z = y-1) \quad &= m\theta^{m+1}\frac{1-\theta}{\theta}\sum_{z=0}^{\infty}\binom{m+1+z-1}{z}(1-\theta)^z \\
&= m\theta^{m+1}\frac{1-\theta}{\theta}\theta^{-m-1} \\
&= m\frac{1-\theta}{\theta}
\end{aligned}$$

Therefore, $\mathbb{E}_\theta\left[\frac{Y+m}{m}\right] = \frac{1}{\theta}$, $\frac{Y+m}{m}$ is UMVU estimator for $\frac{1}{\theta}$. ♠

Example 11: (Poisson Experiment, Example 2.3.7 in TPE)

Suppose that we observe $X \sim Poisson(\theta)$. Consider estimating $g(\theta) = \theta^m, m \geq 1, m \in \mathbb{N}$. $\mathbb{P}_\theta[X = x] = e^{-\theta}\frac{\theta^x}{x!}$. So X is sufficient and complete. Using method 1,

$$\begin{aligned} \mathbb{E}_\theta\left[\delta(X)\right] &= \sum_{x=0}^{\infty} \delta(x) e^{-\theta}\frac{\theta^x}{x!} = \theta^m \\ \sum_{x=0}^{\infty} \delta(x)\frac{\theta^x}{x!} &= \theta^m e^{-\theta} = \sum_{x=0}^{\infty} \frac{\theta^{x+m}}{x!} \end{aligned}$$

Thus

$$\delta(x) = \begin{cases} 0, & x \leq m-1 \\ x(x-1)\cdots(x-m+1), & x \geq m \end{cases}$$

Consider estimating $g(\theta) = e^{-a\theta}$. Use method 1 again.

$$\begin{aligned} \mathbb{E}_\theta\left[\delta(X)\right] &= \sum_{x=0}^{\infty} \delta(x) e^{-\theta}\frac{\theta^x}{x!} = e^{-a\theta} \\ \sum_{x=0}^{\infty} \delta(x)\frac{\theta^x}{x!} &= e^{(-a+1)\theta} = \sum_{x=0}^{\infty} \frac{(-a+1)^x\theta^x}{x!} \end{aligned}$$

Therefore, $\delta(x) = (1-a)^x$ is UMVU estimator for $e^{-a\theta}$. Since $\mathbb{E}_\theta\left[X\right] = \theta . \mathbb{V}\mathrm{ar}_\theta(X) = \theta$, by Chebyshev's inequality, $\mathbb{P}_\theta(|X-\theta| \geq t\sqrt{\theta}) \leq \frac{1}{t^2}$, $X = \theta + O_p(\sqrt{\theta})$ as $\theta \to \infty$, thus $\frac{X-\theta}{\sqrt{\theta}} \xrightarrow{p} \mathsf{N}(0,1)$.

Suppose $X_1, \cdots, X_n \overset{i.i.d.}{\sim} Poisson(\theta)$, defining $X = \sum_{i=1}^n X_i$. We know that $X \sim Poisson(n\theta)$. Then $\delta(x) = (1-\frac{a}{n})^x$ is a UMVU estimator for $e^{-a\theta}$. ♠

2.4 Nonparametric Families

Example 12: (Problem 1.6.33 in TPE)

Let $\mathcal{F}$ be a class of densities on $\mathbb{R}$. $X_1, \cdots, X_n \overset{i.i.d.}{\sim} f \in \mathcal{F}$. Let $X_{(k)}$ denote the k$^{\text{th}}$ smallest X_i. We know $X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n-1)} \leq X_{(n)}$. Since we can write the likelihood in the form of

$$Likelihood = f(X_1)f(X_2)\cdots f(X_n) = f(X_{(1)})f(X_{(2)})\cdots f(X_{(n)})$$

which is a function of $(X_{(1)}, X_{(2)}, \cdots, X_{(n-1)}, X_{(n)})$, the order statistic is sufficient.

We will use the following theorem and proposition to prove the completeness of order statistics.

Proposition 2.4.1. *If statistic T is complete for $\mathcal{F}_0$, then it is complete for any $\mathcal{F}_1$ such that $\mathcal{F}_0 \subset \mathcal{F}_1$.*

Theorem 2.4.2. *The order statistics are complete for the class of densities with finite moment generative functions.*

Consider the sub-class of densities of the form

$$f_\theta(x) = C(\theta)\exp[\theta_1 x + \theta_2 x^2 + \cdots \theta_n x^n - x^{2n}]$$

where $\theta = (\theta_1, \cdots, \theta_n) \in \mathbb{R}^n$, $X_1, \cdots, X_n \overset{i.i.d.}{\sim} f_\theta(X_1) \in \mathcal{F}$. The term $-x^{2n}$ is to make sure that the densities have finite moment generating functions. Thus the likelihood function can be written as

$$\mathit{Likelihood} = [C(\theta)]^m \exp\left[\theta_1 \sum_{i=1}^n x_i + \theta_2 \sum_{i=1}^n x_i^2 + \cdots + \theta_n \sum_{i=1}^n x_i^n - \sum_{i=1}^n x_i^{2n}\right]$$

Thus, $T = (T_1, \cdots, T_n) = (\sum_{i=1}^n x_i, \cdots, \sum_{i=1}^n x_i^n)$ is complete for $\mathcal{F}_0$ and hence for $\mathcal{F} \supset \mathcal{F}_0$. $(X_{(1)}, \cdots, X_{(n)})$ and $(T_1, \cdots, T_n)$ has one to one correspondence, then the **order statistic is complete and sufficient.** ♠

Example 13:
Let $\mathcal{F}$ be a class of densities with finite second moment. Suppose we want to estimate $\mu = \mu(f) =$ mean of f based on $X_1, \cdots, X_n \overset{i.i.d.}{\sim} f$. We know that $\bar{X}$ is unbiased for μ and $\bar{X}$ is a UMVU estimator for μ. In this case, the finite second moment guarantees that the variance of the estimator is finite, since we want to see its Loss function. ♠

Example 14:
Let $\mathcal{F}$ be a class of densities with finite 4^{th} moment. Suppose we want to estimate $\sigma^2 = \sigma^2(f) =$ variance of f based on $X_1, \cdots, X_n \overset{i.i.d.}{\sim} f$. We know that $S^2 = \frac{1}{n-1}\sum_{i=1}^n (x_i - \bar{X})^2$ is unbiased for σ^2 and S^2 is a UMVU estimator for σ^2. In this case, the finite 4^{th} moment guarantees that the variance of the S^2 is finite, since we want to see its Loss function. ♠

Example 15: (U-statistics)
Suppose that $\gamma(\delta)$ admits an unbiased estimator of the form $\delta(X_1, \cdots, X_r)$, with $r \le n$. Then

$$\mathbb{E}\left[\delta(X_1, \cdots, X_r) | X_{(1)}, \cdots, X_{(n)}\right] = \frac{1}{\binom{n}{r}} \sum_{|\beta|=r, \beta\subset\{1,\ldots,n\}} \delta(X_1, \cdots, X_r)$$

is a UMVU for $\gamma(\delta)$. This is a combination fo method 2 2.1 and U-statistics ??. ♠

2.5 Information Inequality

$X \sim \mathbb{P}_\theta, \theta \in \Omega$ for any $\psi(X, \theta)$ estimating $g(\theta)$. The Cauchy-Schwarz inequality gives

$$\mathbb{V}\mathrm{ar}_\theta(\delta) \ge \frac{[\mathbb{C}\mathrm{ov}_\theta(\delta, \psi)]^2}{\mathbb{V}\mathrm{ar}_\theta(\psi)} \tag{2.5}$$

Theorem 2.5.1. *(TPE theorem 2.5.1, page 113 in TPE)*
For a given ψ, $\mathbb{C}\mathrm{ov}_\theta(\delta, \psi)$ is independent of δ if and only if

$$\mathbb{C}\mathrm{ov}_\theta(\mathrm{U}, \psi) = 0, \quad \mathrm{U} \in \mathcal{U}$$

where $\mathcal{U}$ is the class of statistics defined as

$$\mathcal{U} = \{U : \mathbb{E}_\theta\left[U\right] = 0, \mathbb{E}_\theta\left[U^2\right] < \infty, \quad \forall\theta \in \Omega\}$$

i.e., $\mathcal{U}$ is the class of unbiased estimators of 0*.*

Proof of Theorem To say that $\mathbb{C}\mathrm{ov}_\theta(\delta, \psi)$ depends on δ only through $g(\theta)$ is equivalent to saying that for any two estimators δ_1 and δ_2 with $\mathbb{E}_\theta\left[\delta_1\right] = \mathbb{E}_\theta\left[\delta_2\right]$ for all θ, we have $\mathbb{C}\mathrm{ov}_\theta(\delta_1, \psi) = \mathbb{C}\mathrm{ov}_\theta(\delta_2, \psi)$. The proof of the theorem is then easily established by writing

$$\mathbb{C}\mathrm{ov}_\theta(\delta_1, \psi) - \mathbb{C}\mathrm{ov}_\theta(\delta_2, \psi) = \mathbb{C}\mathrm{ov}_\theta(\delta_1 - \delta_2, \psi) = \mathbb{C}\mathrm{ov}_\theta(\mathrm{U}, \psi)$$

Therefore, $\mathbb{C}\mathrm{ov}_\theta(\delta_1, \psi) = \mathbb{C}\mathrm{ov}_\theta(\delta_2, \psi)$ for all δ_1 and δ_2 if and only if $\mathbb{C}\mathrm{ov}_\theta(\mathrm{U}, \psi)$ for all U. □

Example 16 (Hammersley-Chapman-Robbins inequality, Example 2.5.2 in TPE)**:**
Suppose X is distributed with density $p_\theta = p(x, \theta)$, and for the moment, suppose that $p(x, \theta) > 0$ for all x. If θ and $\theta + \triangle$ are two values for which $g(\theta) \neq g(\theta + \triangle)$, then the function

$$\psi(x, \theta) = \frac{p(x, \theta + \triangle)}{p(x, \theta)} - 1$$

satisfies the conditions of theorem 2.5.1, since $\mathbb{E}_\theta\left[\psi\right] = 0$, and hence

$$\mathbb{C}\mathrm{ov}(\mathrm{U}, \psi) = \mathbb{E}\left[\mathrm{U}\psi\right] = \mathbb{E}_{\theta+\triangle}\left[\mathrm{U}\right] - \mathbb{E}_\theta\left[\mathrm{U}\right] = 0$$

where U is any unbiased estimator for 0. In fact,

$$\mathbb{C}\mathrm{ov}(\delta, \psi) = \mathbb{E}_\theta\left[\delta\psi\right] = \mathrm{g}(\theta + \triangle) - \mathrm{g}(\theta)$$

So that the information inequality 2.5 becomes

$$\begin{aligned}
\mathbb{V}\mathrm{ar}_\theta(\delta) \quad &\geq \quad \frac{\left[g(\theta + \triangle) - g(\theta)\right]^2}{\mathbb{E}_\theta\left[\frac{p(X,\theta+\triangle)}{p(X,\theta)} - 1\right]^2} \\
\triangle \to 0 \quad &= \quad \frac{\left[\frac{g(\theta+\triangle)-g(\theta)}{\triangle}\right]^2}{\mathbb{E}_\theta\left[\frac{1}{p(X,\theta)}\left(\frac{p(X,\theta+\triangle)-p(X,\theta)}{\triangle}\right)\right]^2} \\
\text{Under some assumptions} \quad &= \quad \frac{\left[g'(\theta)\right]^2}{\mathbb{E}_\theta\left[\left(\frac{1}{p(X,\theta)}\frac{\partial p(X,\theta)}{\partial\theta}\right)^2\right]}
\end{aligned}$$

Therefore, we change the definition of $\psi(x, \theta)$, let $p'_\theta = \frac{\partial p(x,\theta)}{\partial\theta}$

$$\psi(x, \theta) = \frac{1}{p(x, \theta)}\frac{\partial p(x, \theta)}{\partial\theta} = \frac{p'_\theta}{p_\theta}$$

Then we still have $\mathbb{E}_\theta[\psi] = 0$. Notice that $g(\theta) = \mathbb{E}_\theta[\delta] = \int \delta p_\theta d\mu$, then the $\mathbb{C}\mathrm{ov}_\theta(\delta, \psi)$ becomes

$$\mathbb{C}\mathrm{ov}_\theta(\delta,\psi) = \mathbb{E}_\theta[\delta\psi] - \mathbb{E}_\theta[\delta]\,\mathbb{E}_\theta[\psi] = \int \delta \frac{\mathrm{p}'_\theta}{\mathrm{p}_\theta}\mathrm{p}_\theta \mathrm{d}\mu = \int \delta \mathrm{p}'_\theta \mathrm{d}\mu = \frac{\partial \mathrm{g}(\theta)}{\partial\theta} = \mathrm{g}'(\theta)$$

The last two equality hold under some relaxed conditions. So that the information inequality 2.5 becomes

$$\mathbb{V}\mathrm{ar}_\theta(\delta) \geq \frac{[g'(\theta)]^2}{\mathbb{V}\mathrm{ar}_\theta\left[\frac{\partial \log p(X,\theta)}{\partial\theta}\right]}$$

♠

Fisher Information
The function $\psi = \frac{p'_\theta}{p_\theta}$ is the relative rate at which the density p_θ changes at x. The average of the square of this rate is denoted by

$$I(\theta) = \mathbb{E}_\theta\left[\left(\frac{\partial \log p(X,\theta)}{\partial\theta}\right)^2\right] = \int \left(\frac{p'_\theta}{p_\theta}\right)^2 p_\theta d\mu$$

It is plausible that the greater this expectation is at a given value θ_0, the easier it is to distinguish θ_0 from neighboring values θ , and, therefore, the more accurately θ can be estimated at $\theta = \theta_0$. The quantity $I(\theta)$ is called the **information (or Fisher information)** that X contains about the parameter θ.

$I(\theta)$ depends on the particular parametrization chosen. In fact, if $\theta = h(\xi)$ and h is differentiable, the information that X contains about ξ is

$$I^*(\theta) = I[h(\xi)]\cdot[h'(\xi)]^2 \tag{2.6}$$

Lemma 2.5.2. *(Properties of Fisher Information, Lemma 2.5.3 in TPE)*
To obtain alternative expressions for $I(\theta)$ that sometimes are more convenient, let us make the following assumptions:

1. *Ω is an open interval(finite, infinite, or semi-infinite)*
2. *The distributions $\mathbb{P}_\theta$ has common support, so that without loss of generality, the set $A = \{x : p_\theta(x) > 0\}$ is independent of θ*
3. *For any $x \in A$ and $\theta \in \Omega$, the derivative $p'_\theta = \frac{\partial p(x,\theta)}{\partial\theta}$ exists and it is finite.*

Then we have the following results

1. *If all the assumptions above hold, and the derivative with respect to θ of the left side of $\int p_\theta(x)d\mu(x) = 1$ can be obtained by differentiating under the integral sign,i.e., $\int \delta(x)p'_\theta(x)d\mu(x) = \frac{\partial}{\partial\theta}[\int \delta(x)p_\theta(x)d\mu(x)]$, then*

$$\mathbb{E}_\theta\left[\frac{\partial}{\partial\theta}\log p_\theta(X)\right] = 0, \quad I(\theta) = \mathbb{V}\mathrm{ar}_\theta\left[\frac{\partial}{\partial\theta}\log p_\theta(X)\right]$$

2. *If, in addition, the second derivative with respect to θ of $\log p_\theta(x)$ exists for all x and all θ, and the second derivative with respect to θ of the left side of $\int p_\theta(x)d\mu(x) = 1$ can be obtained by differentiating under the integral sign, then*

$$I(\theta) = -\mathbb{E}_\theta\left[\frac{\partial^2}{\partial\theta^2}\log p_\theta(X)\right]$$

Proof of Lemma The first part of the proof is obvious. For the second part, we have

$$\frac{\partial^2}{\partial\theta^2}\log p_\theta(x) = \frac{\frac{\partial^2}{\partial\theta^2}p_\theta(x)}{p_\theta(x)} - \left[\frac{\frac{\partial}{\partial\theta}p_\theta(x)}{p_\theta(x)}\right]^2$$

Since we can integrate first then differentiate the integral, we have $\int p'_\theta(x)d\mu(x) = \frac{\partial}{\partial\theta}\int p_\theta(x)d\mu(x) = \frac{\partial(1)}{\partial\theta} = 0$. Therefore, $\mathbb{E}_\theta\left[\frac{p'_\theta}{p_\theta}\right] = \mathbb{E}_\theta\left[\frac{p''_\theta}{p_\theta}\right] = 0$. Then the result follows by taking expectation of both sides. □

Theorem 2.5.3. *(Theorem 2.5.4 in TPE)*
Let X be distributed according to the exponential family 1.2

$$p_\theta(x) = \exp\left[\sum_{i=1}^{s}\eta(\theta)T_i(x) - B(\theta)\right]h(x)$$

with $s = 1$. And let $\tau(\theta) = \mathbb{E}_\theta[T]$, the so-called mean-value parameter. Then, T

$$I[\tau(\theta)] = \frac{1}{\mathbb{V}\mathrm{ar}_\theta(T)}$$

Proof of Theorem The amount of information that X contains about θ, $I(\theta)$, is

$$\begin{aligned} I(\theta) &= \mathbb{V}\mathrm{ar}_\theta\left[\frac{\partial}{\partial\theta}\log p_\theta(X)\right] \\ &= \mathbb{V}\mathrm{ar}_\theta[\eta'(\theta)T(X) - B'(\theta)] \\ &= [\eta'(\theta)]^2\mathbb{V}\mathrm{ar}(T) \end{aligned}$$

According to the composition of information 2.6, $I(\theta) = I(\tau(\theta))[\tau'(\theta)]^2$, thus

$$I(\tau(\theta)) = \frac{I(\theta)}{[\tau'(\theta)]^2} = \left[\frac{\eta'(\theta)}{\tau'(\theta)}\right]^2\mathbb{V}\mathrm{ar}_\theta(T)$$

Since(to be proved)

$$\mathbb{E}_\theta[T] = \int T(x)\exp\left[\eta(\theta)T(x) - B(\theta)\right]h(x)dx = \frac{B'(\theta)}{\eta'(\theta)}$$

$$\mathbb{V}\mathrm{ar}_\theta(T) = \frac{B''(\theta) - \eta''(\theta)\tau(\theta)}{[\eta'(\theta)]^2} = \frac{\eta'(\theta)}{\tau'(\theta)}$$

Thus $I(\tau(\theta)) = \frac{1}{\mathbb{V}\mathrm{ar}_\theta(T)}$. □

Theorem 2.5.4. *(Theorem 2.5.12 in TPE)*
$\delta(X)$ achieves the information bound if and only if the model is of the form $p_\theta(x) = \exp\left[\sum_{i=1}^{s} \eta(\theta)T_i(x) - B(\theta)\right] h(x)$.

Theorem 2.5.5. *(Theorem 2.5.8 in TPE)*
Let X and Y be independently distributed with densities p_θ and q_θ, respectively, with respect to measures μ and ν satisfying the assumptions in lemma 2.5.2 and $\mathbb{E}_\theta\left[\frac{\partial}{\partial\theta}\log p_\theta(X)\right] = 0$. If $I_1(\theta)$, $I_2(\theta)$, and $I(\theta)$ are the information about θ contained in X,Y, and (X, Y), respectively, then

$$I(\theta) = I_1(\theta) + I_2(\theta)$$

Proof of Theorem By definition

$$I(\theta) = \mathbb{E}\left[\frac{\partial}{\partial\theta}\log p_\theta(X) + \frac{\partial}{\partial\theta}\log q_\theta(Y)\right]$$

Using the independence of X and Y, and the fact $\mathbb{E}_\theta\left[\frac{\partial}{\partial\theta}\log p_\theta(X)\right] = 0$, we obtain the result. $\square$

Corollary 2.5.6. *(Corollary 2.5.9 in TPE)*
If $X_1, \cdots, X_n$ are i.i.d,satisfying the assumptions in lemma 2.5.2 and $\mathbb{E}_\theta\left[\frac{\partial}{\partial\theta}\log p_\theta(X)\right] = 0$, and each has information $I(\theta)$, then the information in $X = (X_1, \cdots, X_n)$ is $nI(\theta)$.

Chapter 3

Equivariance

3.1 Location Equivariant Estimator

Example 17: (Estimating Binomial p, Example 3.1.1 in TPE)

Consider n binomial trials with unknown probability $p (0 < p < 1)$ of success which we wish to estimate with loss function $L(p, d)$, for example, $L(p, d) = (d - p)^2$ or $L(p, d) = (d - p)^2/p(1-p)$. If $X_i, i = 1, ..., n$ is 1 or 0 as the i$^{\text{th}}$ trial is a success or failure, the joint distribution of the X's is

$$\mathbb{P}\left[X_1, \cdots, X_n\right] = p^{\sum x_i}(1-p)^{\sum(1-x_i)}$$

Now, suppose that there is another statistician interchanges the definition of success and failure. From his point of view, the probability of success is $p' = 1 - p$, and the indicator of success and failure on the i$^{\text{th}}$ trial is $X'_i = 1 - X_i$. Therefore, the joint distribution of X'_i is

$$\mathbb{P}\left[X'_1, \cdots, X'_n\right] = (p')^{\sum x'_i}(1-p')^{\sum(1-x'_i)} = p^{\sum x_i}(1-p)^{\sum(1-x_i)} = \mathbb{P}\left[X_1, \cdots, X_n\right]$$

In the new terminology, the estimated value d' of p' is $d' = 1 - d$, and the loss function, suggested in the beginning example, satisfies $L(p, d) = L(p', d')$. Under these circumstances, the problem of estimating p with loss function L is said to be **invariant under the transformations**.

Now, we want to use $\delta(\boldsymbol{X})$, where $\boldsymbol{X} = (X_1, \cdots, X_n)$, as an estimator of p under the previous situation.

- It is natural to use $\delta(\boldsymbol{X}') = \delta(1 - X_1, \cdots, 1 - X_n)$ to estimate $p' = 1 - p$.
- It is natural to use $1 - \delta(\boldsymbol{X})$ to estimate $p' = 1 - p$.

It seems desirable that these two estimators should agree and hence that

$$\delta(\boldsymbol{X}') = 1 - \delta(\boldsymbol{X})$$

An estimator satisfying $\delta(\boldsymbol{X}') = 1 - \delta(\boldsymbol{X})$ will be called equivariant under transformation $p' = 1 - p$, $X'_i = 1 - X_i$ and $d' = 1 - d$. ♠

Now we come to a more general case. Let $\boldsymbol{X} = (X_1, \cdots, X_n)$ have joint distribution with probability density

$$f(\boldsymbol{x} - \xi) = f(x_1 - \xi, \cdots, x_n - \xi), \quad -\infty < \xi < \infty$$

where f is known and ξ is an unknown **location parameter**. Suppose that for the problem of estimating ξ with loss function $L(\xi, d)$, we have found a satisfactory estimator $\delta(\boldsymbol{X})$. Consider the following transformations

$$
\begin{aligned}
X_i' &= X_i + a && (3.1)\\
\xi' &= \xi + a && (3.2)\\
f(\boldsymbol{x}' - \xi') &= f(x_1' - \xi', \cdots, x_n' - \xi') = f(x_1 - \xi, \cdots, x_n - \xi) = f(\boldsymbol{x} - \xi) && (3.3)\\
d' &= d + a && (3.4)
\end{aligned}
$$

We require that $L(\xi', d') = L(\xi + a, d + a) = L(\xi, d)$ for all values of a, and this happens if and only if $L(\xi, d)$ depends only on the difference $d - \xi$, that is, L is of the form $L(\xi, d) = \rho(d - \xi)$, since we can put $a = -\xi$ and get $\rho(d - \xi) = L(0, d - \xi)$ We now introduce the formal definition of location invariance.

Definition 3.1.1. *(Location Invariance)*
*A family of densities $f(x|\xi)$, with parameter ξ,and a loss function $L(\xi, d)$ are **location invariant** if, respectively, $f(x'|\xi') = f(x|\xi)$ and $L(\xi, d) = L(\xi', d')$ whenever $\xi = \xi + a$ and $d' = d + a$. If both the densities and the loss function are location invariant, the problem of estimating ξ is said to be location invariant under the transformations 3.1.*

Suppose now that we had decided to use $\delta(\boldsymbol{X})$ as an estimator of ξ, then

- It is natural to use $\delta(\boldsymbol{X}') = \delta(X_1 + a, \cdots, X_n + a)$ to estimate $\xi' = \xi + a$.
- It is natural to use $\delta(\boldsymbol{X}) + a$ to estimate $\xi' = \xi + a$.

As before, it seems desirable that these two estimators should agree and hence that

$$\delta(X_1 + a, \cdots, X_n + a) = \delta(\boldsymbol{X}) + a, \quad \forall a \tag{3.5}$$

Definition 3.1.2. *(Location Equivariant Estimator)*
*An estimator satisfying equation 3.5 will be called equivariant under the transformations 3.1, or **location equivariant**.*

(Homework) Prove that the MLE $\hat{\xi}$ is location equivariant.
Sketch of Proof If $\delta(x) = \text{argmax}_{\xi\in\mathbb{R}} f(x-\xi)$, then $\delta(x+a) = \text{argmax}_{\xi\in\mathbb{R}} f(x+a-\xi)$, $\delta(x)+a = \text{argmax}_{\xi\in\mathbb{R}} f(x-(\xi-a)) = \text{argmax}_{\xi\in\mathbb{R}} f(x-\xi+a)$. Therefore, $\delta(x)+a = \delta(x+a)$. MLE is location equivariant. □

Theorem 3.1.1. *(Theorem 3.1.4 in TPE)*
Let $\boldsymbol{X}$ be distributed with density $f(\boldsymbol{x} - \xi) = f(x_1 - \xi, \cdots, x_n - \xi), -\infty < \xi < \infty$, and let δ be equivariant for estimating ξ with loss function $L(\xi, d) = \rho(d - \xi)$. Then, the bias, risk, and variance of δ are all constant (i.e., do not depend on ξ).

Proof of Theorem Note that if $\boldsymbol{X}$ has density $f(\boldsymbol{x})$ (*i.e.*, $\xi = 0$), then $X + \xi$ has density $f(\boldsymbol{x} - \xi)$. Thus, the bias can be written as

$$b(\xi) = \mathbb{E}_\xi\left[\delta(\boldsymbol{X})\right] - \xi = \mathbb{E}_0\left[\delta(\boldsymbol{X} + \xi)\right] - \xi = \mathbb{E}_0\left[\delta(\boldsymbol{X})\right]$$

which does not depend on ξ. (Homework)The proofs for risk and variance are analogous □

Definition 3.1.3. *(MRE)*

In a location invariant estimation problem, if a location equivariant estimator exists which minimizes the constant risk, it is called the ***minimum risk equivariant (MRE) estimator****.*

If we want to show that $\delta(\boldsymbol{X})$ is a MRE, we need to

- prove it is equivariant
- show it has the minimum risk

Note that there is always a sequence of equivariant estimators $\{\delta_k\}$ such that $R(\delta_k) := R(0, \delta_k) \xrightarrow{k} \infty \inf R(0, \delta) = R(\delta)$, while δ is equivariant.

Lemma 3.1.2. *(Lemma 3.1.6 in TPE)*

If δ_0 is any equivariant estimator, then a necessary and sufficient condition for δ to be equivariant is that

$$\delta(\boldsymbol{X}) = \delta_0(\boldsymbol{X}) + u(\boldsymbol{X})$$

where $u(\boldsymbol{X})$ is any function satisfying $u(\boldsymbol{X}) = u(\boldsymbol{X} + a)$ for all $\boldsymbol{X}$ and a.

Proof of Lemma First we assume the 2 conditions hold, since δ_0 is any equivariant estimator, then

$$\delta(\boldsymbol{X} + a) = \delta_0(\boldsymbol{X} + a) + u(\boldsymbol{X} + a) = \delta_0(\boldsymbol{X}) + a + u(\boldsymbol{X}) = \delta(\boldsymbol{X}) + a$$

Thus $\delta(\boldsymbol{X})$ is equivariant.

Conversely, if $\delta(\boldsymbol{X})$ is equivariant, let $u(\boldsymbol{X}) = \delta(\boldsymbol{X}) - \delta_0(\boldsymbol{X})$, then

$$u(\boldsymbol{X} + a) = \delta(\boldsymbol{X} + a) - \delta_0(\boldsymbol{X} + a) = \delta(\boldsymbol{X}) + a - \delta_0(\boldsymbol{X}) - a = \delta(\boldsymbol{X}) - \delta_0(\boldsymbol{X}) = u(\boldsymbol{X})$$

So that the 2 conditions above hold. □

Theorem 3.1.3. *(Theorem 3.1.8 in TPE)*

If δ_0 is any equivariant estimator, then a necessary and sufficient condition for δ to be equivariant is that there exists a function v of $n-1$ arguments for which

$$\delta(\boldsymbol{X}) = \delta_0(\boldsymbol{X}) - v(\boldsymbol{Y}), \quad \forall \boldsymbol{X}$$

where $\boldsymbol{Y} = (Y_1, \cdots, Y_{n-1})$.

Theorem 3.1.4. *(Theorem 3.1.10 in TPE)*

Let $\boldsymbol{X} = (X_1, \cdots, X_n)$ be distributed with density $f(\boldsymbol{x} - \xi) = f(x_1 - \xi, \cdots, x_n - \xi), -\infty < \xi < \infty$. Let $Y_i = X_1 - X_n$ and $\boldsymbol{Y} = (Y_1, \cdots, Y_{n-1})$. Suppose that the loss function is given by $L(\xi, d) = \rho(d - \xi)$ and that there exists an equivariant estimator δ_0 of ξ with finite risk. Assume that for each $\boldsymbol{y}$ there exists a number $v(\boldsymbol{y}) = v^*(\boldsymbol{y})$ which minimizes

$$\mathbb{E}_0 \left[\rho[\delta_0(\boldsymbol{X}) - v(\boldsymbol{Y})]|\boldsymbol{Y}\right]$$

Then, a ***location equivariant estimator*** *δ* ***of*** *ξ* ***with minimum risk(MRE)*** *exists and is given by*

$$\delta^*(\boldsymbol{X}) = \delta_0(\boldsymbol{X}) - v^*(\boldsymbol{Y})$$

Proof of Theorem The MRE estimator is found by determining v so as to minimize

$$R_\xi(\delta) = \mathbb{E}_\xi\left[\rho[\delta_0(\boldsymbol{X}) - v(\boldsymbol{Y}) - \xi]\right]$$

Since the risk is independent of ξ , it suffices to minimize

$$R_0(\delta) = \mathbb{E}_0\left[\rho[\delta_0(\boldsymbol{X}) - v(\boldsymbol{Y})]\right] = \int \mathbb{E}_0\left[\rho[\delta_0(\boldsymbol{X}) - v(\boldsymbol{Y})]|\boldsymbol{Y}\right] d\mathbb{P}_0(\boldsymbol{Y})$$

The integral is minimized by minimizing the integrand with respect to $\boldsymbol{Y}$. Since δ_0 has finite risk $\mathbb{E}_0\left[\rho[\delta_0(\boldsymbol{X}) - v(\boldsymbol{Y})]|\boldsymbol{Y}\right] < \infty$ a.e. under $\mathbb{P}_0$, thus the minimization of $\mathbb{E}_0\left[\rho[\delta_0(\boldsymbol{X}) - v(\boldsymbol{Y})]|\boldsymbol{Y}\right]$ is meaningful. Therefore, a location equivariant estimator δ of ξ is given by

$$R_0(\delta) = \int \mathbb{E}_0\left[\rho[\delta_0(\boldsymbol{X}) - v(\boldsymbol{Y})]|\boldsymbol{Y}\right] d\mathbb{P}_0(\boldsymbol{Y}) \geq \int \mathbb{E}_0\left[\rho[\delta_0(\boldsymbol{X}) - v^*(\boldsymbol{Y})]|\boldsymbol{Y}\right] d\mathbb{P}_0(\boldsymbol{Y})$$

since

$$R(\delta(\boldsymbol{X})) = \mathbb{E}_0\left[\rho[\delta(\boldsymbol{X})]\right] = \mathbb{E}_0\left[\rho[\delta_0(\boldsymbol{X}) - v(\boldsymbol{Y})]\right] = R(\delta^*(\boldsymbol{X}))$$

□

Corollary 3.1.5. *(Corollary 3.1.11 in TPE)*

Let $\boldsymbol{X} = (X_1, \cdots, X_n)$ be distributed with density $f(\boldsymbol{x} - \xi) = f(x_1 - \xi, \cdots, x_n - \xi), -\infty < \xi < \infty$, let $Y_i = X_i - X_n (i = 1, \cdots, n-1)$ and $\boldsymbol{Y} = (Y_1, \cdots, Y_{n-1})$. Suppose that the loss function is given by $L(\xi, d) = \rho(d - \xi)$ and that there exists an equivariant estimator δ_0 of ξ with finite risk. Assume that for each $\boldsymbol{y}$ there exists a number $v(\boldsymbol{y}) = v^*(\boldsymbol{y})$ which minimizes

$$\mathbb{E}_0\left[\rho[\delta_0(\boldsymbol{X}) - v(\boldsymbol{Y})]|\boldsymbol{Y}\right]$$

suppose that ρ is convex and not monotone. Then, an MRE estimator of ξ exists; it is unique if ρ is strictly convex.

Proof Let ρ be a convex function defined on $(-\infty, \infty)$ and X a random variable such that $\phi(a) = \mathbb{E}\left[\rho(X - a)\right]$ is finite for some a. If ρ is not monotone, $\phi(a)$ takes on its minimum value and the set on which this value is taken is a closed interval. If ϕ is strictly convex, the minimizing value is unique.

Take Z such that $\mathbb{E}\left[\rho(Z)\right] < \infty$, then $Z \sim \delta(X)|Y = y$.

1. $\phi(a) = \mathbb{E}\left[\rho(Z - a)\right]$ is convex. Suppose $a, b \in \mathbb{R}, \lambda \in [0, 1]$, then

$$\begin{aligned} \phi(\lambda a + (1-\lambda)b) &= \mathbb{E}\left[\rho(Z - \lambda a - (1-\lambda)b)\right] = \mathbb{E}\left[\rho(\lambda(Z-a) + (1-\lambda)(Z-b))\right] \\ \rho \text{ is convex} &\leq \mathbb{E}\left[\lambda\rho(Z-a) + (1-\lambda)\rho(Z-b)\right] = \lambda\phi(a) + (1-\lambda)\phi(b) \end{aligned}$$

   Therefore, ϕ is convex.

2. ϕ is not monotone. Note that $\rho(Z) \to \infty$ when $Z \to \pm\infty$. We take $a_n \to \pm\infty$, then $\phi(a_n) = \mathbb{E}\left[\rho(Z - a_n)\right]$. We know $\rho(Z - a_n) \to \infty$ as $a_n \to \pm\infty$, then by Fatou's lemma,

$$\liminf_{n\to\infty} \phi(a_n) = \liminf_{n\to\infty} \mathbb{E}\left[\rho(Z - a_n)\right] \geq \mathbb{E}\left[\liminf_{n\to\infty} \rho(Z - a_n)\right] = \infty$$

3. ϕ is strictly convex when ρ is strictly convex.

The first 2 conditions ensure that there is a minimum value in this function. The last condition ensure that the minimizing value is unique. □

Corollary 3.1.6. *(Corollary 3.1.12 in TPE)*

1. *(Squared Error Loss) Let* $\rho(t) = t^2$ *and suppose that* $\mathbb{E}[X^2] < \infty$*. Let the loss function be* $L(\xi, d) = (\xi - d)^2$*. For* δ_0*, given*

$$\begin{aligned} v^*(y) &= \operatorname*{argmin}_{v\in\mathbb{R}} \mathbb{E}\left[(\delta_0(X) - v)^2 | Y = y\right] \\ (Z \overset{\xi=0}{\sim} \delta_0(X)|Y = y) &= \operatorname*{argmin}_{v\in\mathbb{R}} \mathbb{E}\left[(Z - v)^2\right] \\ &= \mathbb{E}\left[\delta_0(X)|Y = y\right] \end{aligned}$$

Then $\delta^*(X) = \delta_0(X) - \mathbb{E}[\delta_0(X)|Y = y]$

2. *(Absolute Error Loss) Let* $\rho(t) = |t|$ *and suppose that* $\mathbb{E}[|X|] < \infty$*. Let the loss function be* $L(\xi, d) = |\xi - d|$*. For* δ_0*, given* $v^*(y) = \operatorname{argmin}_{v\in\mathbb{R}} \mathbb{E}_0\left[|\delta_0 - v|\big|Y = y\right] = Median_0[\delta_0(X)|Y = y]$*.*

Corollary 3.1.7. *(Corollary 3.1.14 in TPE Page 152)*
Let $\boldsymbol{X} = (X_1, \cdots, X_n)$ *be distributed with density* $f(\boldsymbol{x} - \xi) = f(x_1 - \xi, \cdots, x_n - \xi), -\infty < \xi < \infty$*, let* $Y_i = X_i - X_n (i = 1, \cdots, n-1)$ *and* $\boldsymbol{Y} = (Y_1, \cdots, Y_{n-1})$*. Suppose that the loss function is given by* $L(\xi, d) = \rho(d - \xi)$ *and that there exists an equivariant estimator* δ_0 *of* ξ *with finite risk. Assume that* $0 \le \rho(t) \le M, \forall t \in \mathbb{R}$ *for some* $M \in \mathbb{R}$*, and that* $\rho(t) \to M$ *when* $t \to \pm\infty$*. The density* f *of* X *is continuous a.e. Then, an MRE estimator of* ξ *exists.*

Proof Fix $y \in \mathbb{R}^n$. Let $Z \overset{\xi=0}{\sim} \delta_0(X)|Y = y$. Define $\phi(a) = \mathbb{E}[\rho(Z - a)]$. We are going to show the following things

1. $\phi(a) \to M$ as $a \to \pm\infty$. Suppose that there is a sequence $\{\rho_n(x)\}_{n=1}^{\infty}$ with $|\rho_n(x)| \le M$ a.s. for all n, together with $\mathbb{E}[M] = M < \infty$, $\rho_n(x) \to M$ as $n \to \infty$. Then by Dominated convergence theorem, $\liminf \mathbb{E}[\rho_n(X)] = \mathbb{E}[M] = M$. Thus $\phi(a) \to M$ as $a \to \infty$.

2. $\phi(a)$ is continuous. Suppose that if $f_n, n = 1, 2, \cdots$ and f are probability densities such that $f_n(a) \to f(a)$ a.e. then $\int \rho f_n \to \int \rho f$ for all bounded ρ.

Since $\phi(a)$ is bounded and continuous, it can achieve its minimum when $a \in [-\infty, \infty]$. Therefore, MRE exists. □

Example 18: (MRE under $0-1$ loss, Example 3.1.15 in TPE)
Suppose that

$$\rho(d - \xi) = \begin{cases} 1 & |d - \xi| > k \\ 0 & otherwise \end{cases}$$

Then, v will minimize $\mathbb{E}_0\left[\rho(X-v)\right] = \mathbb{P}\left[|X-v|>k\right]$, provided it maximizes $\mathbb{P}\left[|X-v|\le k\right]$. ♠

Example 19: (Uniform, Example 3.1.19 in TPE Page 154)

Let $X_1,\cdots,X_n \overset{i.i.d.}{\sim} Uniform[\xi-\frac{1}{2},\xi+\frac{1}{2}]$, show that $\frac{1}{2}(X_{(1)}+X_{(n)})$ is MRE when ρ is convex and even, where $X_{(i)}$ is the order statistic, $X_{(1)}\le\cdots\le X_{(n)}$. ♠

Proof Define $\delta_0(X)=\frac{1}{2}(X_{(1)}+X_{(n)})$. We want to use $\delta^*(X)=\delta_0(X)-v^*(y)$ with $Y=(X_1-X_n,\cdots,X_{n-1}-X_n)$ and $v^*(y)=\operatorname{argmin}_{a\in\mathbb{R}}\mathbb{E}_0\left[\rho(\delta_0(X)-a)|Y=y\right]$. We are going to prove that $v^*(y)=0$ when ρ is convex and even.

First, the conditional distribution $\delta_0(X)|\boldsymbol{Y}=\boldsymbol{y}$ depends on $\boldsymbol{y}$ only through the difference $X_{(i)}-X_{(n)}, i=1,\cdots,n-1$. Define $Z_i:=\frac{X_{(i)}-X_{(n)}}{X_{(1)}-X_{(n)}}$. Z_i does not depend on ξ. Therefore, $\boldsymbol{Z}=(Z_2,\cdots,Z_{n-1})$ is ancillary for this model. By Basu's theorem 1.2.5, the pair $(X_{(1)},X_{(n)})$ is independent of $\boldsymbol{Z}$. Define $T=X_{(1)}-X_{(n)}$. We can work with $(\boldsymbol{Z},T)$ instead of $\boldsymbol{Y}$.

Then, we are going to deal the $\delta_0(X)|T=t$ instead of $\delta_0(X)|\boldsymbol{Y}=\boldsymbol{y}$, since $\delta_0(X)|X_{(n)}-X_{(i)}$ is equivalent to $\delta_0(X)|X_{(n)}-X_{(1)}$, and $\boldsymbol{Z}$ is independent of T.

Notice that $\delta_0(X)|T=t$ is symmetric about 0. Obviously, T is invariant under the transformation $x\to -x$ when $\xi=0$, since it is just the length between $X_{(n)}$ and $X_{(1)}$. Thus, $\mathbb{P}_{\xi=0}[T(-X)\in B]=\mathbb{P}_{\xi=0}[T(X)\in B]$. Meanwhile, $\delta_0(X)$ is symmetric about 0 when $\xi=0$. Then $\mathbb{P}_{\xi=0}[\delta_0(-X)\in A]=\mathbb{P}_{\xi=0}[-\delta_0(X)\in A]$. We compute

$$\begin{aligned}
\mathbb{P}\left[\delta_0(X)\in A|T(X)\in B\right] &= \mathbb{P}\left[\delta_0(-X)\in A|T(-X)\in B\right]\\
&= \frac{\mathbb{P}\left[\delta_0(-X)\in A, T(-X)\in B\right]}{\mathbb{P}\left[T(-X)\in B\right]}\\
&= \frac{\mathbb{P}\left[-\delta_0(X)\in A, T(X)\in B\right]}{\mathbb{P}\left[T(X)\in B\right]}\\
&= \mathbb{P}\left[-\delta_0(X)\in A|T(X)\in B\right]
\end{aligned}$$

Thus $\delta_0(X)|T=t$ is symmetric about 0. Then $\operatorname{argmin}_{a\in\mathbb{R}}\mathbb{E}_0\left[\rho(\delta_0(X)-a)|Y=y\right]$ takes its minimum when $a=0$. Therefore, $v^*(y)=0$. $\delta^*(X)=\frac{1}{2}(X_{(1)}+X_{(n)})$ is MRE. □

Example 20: (Normal,Example 3.1.16 in TPE Page 153)

Let $X_1,\cdots,X_n$ be i.i.d according to $\mathsf{N}(\xi,\sigma^2)$, where σ is known. Show that $\delta_0=\bar{X}$ is an MRE when ρ is convex and even. Does this extend to the case when $\rho(a)=|a|$ with $\phi:\mathbb{R}_+\to\mathbb{R}_+$ is non-decreasing? ♠

Proof We know that $\bar{X}$ is equivariant for a location model. The MRE is given by $\delta^*(X)=\delta_0(X)-v^*(y)$ according to theorem 3.1.4. Then we take $\delta_0(X)=\bar{X}$. Let $Y=(X_1-\bar{X},\cdots,X_{n-1}-\bar{X})$. We know the following 2 facts

- When $\xi=0$, $\bar{X}\sim\mathsf{N}(0,\frac{\sigma^2}{n})$ is symmetric about 0.
- $\bar{X}$ is independent of Y. First, $Y=(X_1-\bar{X},\cdots,X_{n-1}-\bar{X})$ is an ancillary statistic, since the μ terms cancelled when the do the minus. $\bar{X}$ is a complete sufficient statistic. Following the Basu's theorem 1.2.5, $\bar{X}$ is independent of $Y=$

$(X_1 - \bar{X}, \cdots, X_{n-1} - \bar{X})$. We can also prove the independence using the Gaussian property, since $\mathbb{Cov}(\bar{X}, X_i - \bar{X}) = 0$ for all $i = 1, \cdots, n-1$, then $\bar{X}$ and Y are independent (This is true only when $\boldsymbol{X}$ is a normal vector).

Let $Z \sim \delta_0(X)|Y = y$. Since $\bar{X}$ and $\boldsymbol{Y}$ are independent, then $Z \sim \delta_0(X)$. Now, we want to minimize $\mathbb{E}_0 [\rho(Z - v)]$ to get $v^*(y)$. Assuming that $\xi = 0$. Define $\phi(a) = \mathbb{E}_0 [\rho(X - a)]$. We know the following 2 facts

- $\phi(a) = \mathbb{E}[\rho(Z - a)]$ is convex. Suppose $a, b \in \mathbb{R}, \lambda \in [0, 1]$, then

$$\begin{aligned}\phi(\lambda a + (1-\lambda)b) &= \mathbb{E}[\rho(Z - \lambda a - (1-\lambda)b)] = \mathbb{E}[\rho(\lambda(Z-a) + (1-\lambda)(Z-b))] \\ \rho \text{ is convex } &\leq \mathbb{E}[\lambda\rho(Z-a) + (1-\lambda)\rho(Z-b)] = \lambda\phi(a) + (1-\lambda)\phi(b)\end{aligned}$$

  Therefore, ϕ is convex.

- $\phi(a) = \mathbb{E}[\rho(Z - a)]$ is even.

$$\begin{aligned}\phi(-a) = \mathbb{E}[\rho(Z + a)] &= \mathbb{E}[\rho(-Z - a)] \quad (\rho \text{ is even}) \\ &= \mathbb{E}[\rho(Z - a)] \quad (Z \text{ is symmetric about } 0) \\ &= \phi(a)\end{aligned}$$

Since $\phi(a)$ is even and convex about 0, then it takes its minimum at 0. That is $\min_{a\in\mathbb{R}} \phi(a) = \phi(0)$. Therefore, $v^*(y) = 0$ and $\delta^*(X) = \bar{X}$.

It is true when $\rho(a) = |a|$ with $\phi : \mathbb{R}_+ \to \mathbb{R}_+$ is non-decreasing, since $\phi(a) = \mathbb{E}[|Z - a|]$ is symmetric about 0 and $\min_{a\in\mathbb{R}} \phi(a) = \phi(0)$. Thus $v^*(y) = 0$ and $\delta^*(X) = \bar{X}$. □

Theorem 3.1.8. *(Theorem 3.1.20 in TPE Page 154)*

Let $\boldsymbol{X} = (X_1, \cdots, X_n)$ be distributed with density $f(\boldsymbol{x} - \xi) = f(x_1 - \xi, \cdots, x_n - \xi), -\infty < \xi < \infty$. Suppose that the loss function is given by $L(\xi, d) = (d - \xi)^2$. Then the Pitman Estimator of ξ, given by

$$\delta^*(x) = \frac{\int_{-\infty}^{\infty} u f(x_1 - u, \cdots, x_n - u) du}{\int_{-\infty}^{\infty} f(x_1 - u, \cdots, x_n - u) du}$$

is MRE.

Proof of Theorem We want to use $\delta^*(\boldsymbol{X}) = \delta_0(\boldsymbol{X}) - v^*(\boldsymbol{y})$ with $\boldsymbol{Y} = (X_1 - X_n, \cdots, X_{n-1} - X_n)$ and $v^*(\boldsymbol{y}) = \text{argmin}_{a\in\mathbb{R}} \mathbb{E}_0 [\rho(\delta_0(\boldsymbol{X}) - a)|\boldsymbol{Y} = \boldsymbol{y}]$. Under the square error loss function assumption, $v^*(\boldsymbol{y}) = \mathbb{E}_0 [\delta_0(\boldsymbol{X})|\boldsymbol{Y} = \boldsymbol{y}]$, a result of projection.

Let $\delta_0(\boldsymbol{X}) = X_n$. To compute $\mathbb{E}_0 [X_n|\boldsymbol{Y} = \boldsymbol{y}]$, let $y_i = x_i - x_n, i = 1, \cdots, n-1$ and $y_n = x_n$. Denote the transformation by T and $\det(T') = 1$. Using the change of variable formula, for any function g, we have

$$\begin{aligned}\mathbb{E}[g(\boldsymbol{Y}, Y_n)] &= \mathbb{E}[g(T(\boldsymbol{X}))] = \int g(T(\boldsymbol{x})) f(x_1 \cdots, x_n) dx_1 \cdots dx_n \\ &= \int g(\boldsymbol{y}, y_n) f(y_1 + y_n, \cdots, y_{n-1} + y_n, y_n) \frac{1}{\det(T')} dy_1 \cdots dy_n \\ &= \int g(\boldsymbol{y}, y_n) f(y_1 + y_n, \cdots, y_{n-1} + y_n, y_n) dy_1 \cdots dy_n\end{aligned}$$

Then the conditional density of $\boldsymbol{y} = (y_1, \cdots, y_n)$ is

$$\frac{f(y_1 + y_n, \cdots, y_{n-1} + y_n, y_n)}{\int_{-\infty}^{\infty} f(y_1 + t, \cdots, y_{n-1} + t, t)dt}$$

Then

$$\begin{aligned}
\mathbb{E}_0\left[X_n|\boldsymbol{y}\right] &= \mathbb{E}_0\left[Y_n|\boldsymbol{y}\right] = \frac{\int_{-\infty}^{\infty} tf(y_1 + t, \cdots, y_{n-1} + t, t)dt}{\int_{-\infty}^{\infty} f(y_1 + t, \cdots, y_{n-1} + t, t)dt} \\
&= \frac{\int_{-\infty}^{\infty} tf(x_1 - x_n + t, \cdots, x_{n-1} - x_n + t, t)dt}{\int_{-\infty}^{\infty} f(x_1 - x_n + t, \cdots, x_{n-1} - x_n + t, t)dt} \\
(u = x_n - t) &= X_n - \frac{\int_{-\infty}^{\infty} uf(x_1 - u, \cdots, x_{n-1} - u, x_n - u)du}{\int_{-\infty}^{\infty} f(x_1 - u, \cdots, x_{n-1} - u, x_n - u)du}
\end{aligned}$$

Therefore,

$$\delta^*(\boldsymbol{X}) = X_n - \mathbb{E}_0\left[X_n|\boldsymbol{y}\right] = \frac{\int_{-\infty}^{\infty} uf(x_1 - u, \cdots, x_{n-1} - u, x_n - u)du}{\int_{-\infty}^{\infty} f(x_1 - u, \cdots, x_{n-1} - u, x_n - u)du}$$

□

Lemma 3.1.9. *(Lemma 3.1.23 in TPE Page 154)*
Let the loss function be squared error.

- *When $\delta(\boldsymbol{X})$ is any equivariant estimator with constant bias b, then $\delta(\boldsymbol{X}) - b$ is equivariant, unbiased, and has smaller risk than $\delta(\boldsymbol{X})$.*
- *The unique MRE estimator is unbiased.*
- *If a UMVU estimator exists and is equivariant, it is MRE.*

Proof of Lemma

1. Obviously, $\delta(\boldsymbol{X} + a) - b = \delta(\boldsymbol{X}) + a - b$, so $\delta(\boldsymbol{X}) - b$ is equivariant. Let the risk be expected squared error. If δ is an unbiased estimator of $g(\theta)$ and $\delta^* = \delta + b$, where the bias b is independent of θ, then δ^* has uniformly larger risk. In fact, $R_{\delta^*}(\theta) = R_\delta(\theta) + b^2$. Thus $\delta(\boldsymbol{X}) - b$ is unbiased. And it has smaller risk than $\delta(\boldsymbol{X})$
2. MRE has the smallest risk, thus $b = 0$, Therefore unbiased.
3. UMVU means that the variance is the smallest. Under squared error loss, it has the smallest risk. Since it is also equivariant, then it is MRE.

□

Definition 3.1.4. *An estimator δ of $g(\theta)$ is said to be risk-unbiased if it satisfies*

$$\mathbb{E}_\theta\left[L(\theta, \delta(\boldsymbol{X}))\right] \le \mathbb{E}_\theta\left[L(\theta', \delta(\boldsymbol{X}))\right] \quad \forall \theta' \neq \theta$$

Example 21: (Mean-unbiasedness, Example 3.1.25 in TPE Page 157)

If the loss function is squared error, then $g(\theta)$ is risk-unbiased if

$$\mathbb{E}_\theta\left[(g(\theta)-\delta(\boldsymbol{X}))^2\right] \leq \mathbb{E}_\theta\left[(g(\theta')-\delta(\boldsymbol{X}))^2\right] \quad \forall \theta' \neq \theta$$

Suppose that $\mathbb{E}_\theta\left[\delta^2\right] < \infty$ and $\mathbb{E}_\theta\left[\delta(\boldsymbol{X})\right] \in \Omega_g = \{g(\theta) : \theta \in \Omega\}$. Then we minimize the right hand side and get the condition $g(\theta') = \mathbb{E}_\theta\left[\delta(\boldsymbol{X})\right]$. Thus the condition of risk unbiasedness reduces to the usual unbiasedness condition $g(\theta) = \mathbb{E}_\theta\left[\delta(\boldsymbol{X})\right]$. ♠

Chapter 4

Average Risk Optimality

In this chapter, we want to minimize the (weighted) average risk for some suitable non-negative weight function and minimizing the maximum risk.

4.1 Introduction

Definition 4.1.1. *(Bayes estimator)*

For $X \sim \mathbb{P}_\theta, \theta \in \Omega$, and the loss function $L(\theta, d)$. The risk function is $R(\theta, \delta(x)) = \mathbb{E}_\theta\left[L(\theta, \delta(X))\right]$. Let Λ be a distribution on Ω, which satisfies $\Theta \sim \Lambda$ and

$$\int d\Lambda(\theta) = 1$$

An estimator δ_Λ satisfies,

$$\delta_\Lambda = \underset{\delta}{\operatorname{argmin}} \int_\Omega R(\theta, \delta) d\Lambda(\theta) = \underset{\delta}{\operatorname{argmin}} \mathbb{E}_\Lambda\left[R(\Theta, \delta)\right] \tag{4.1}$$

is called a Bayes estimator with respect to Λ. We usually denote $r(\Lambda, \delta) = \int_\Omega R(\theta, \delta) d\Lambda(\theta)$

Theorem 4.1.1. *(Theorem 4.1.1 in TPE)*

Let Θ have distribution Λ, and given $\Theta = \theta$, let X have distribution $\mathbb{P}_\theta$. Suppose, in addition, the following assumptions hold for the problem of estimating $g(\theta)$ with non-negative loss function $L(\theta, d)$.

1. *There exists an estimator δ_0 with finite risk.*
2. *For almost all x, there exists a value δ_Λ minimizing $\mathbb{E}\left[L(\Theta, \delta(x)) | X = x\right]$*

*Then, δ_Λ is a **Bayes estimator**.*

Proof of Theorem Let δ be any estimator with finite risk, *i.e.*

$$r(\Lambda, \delta) = \mathbb{E}_\theta\left[\mathbb{E}_\Lambda\left[L(\Theta, \delta(X))\right]\right] < \infty$$

Then we have $\mathbb{E}_\Lambda\left[L(\Theta, \delta(X)) | X = x\right] < \infty$. Thus by definition,

$$\mathbb{E}_\Lambda\left[L(\Theta, \delta(x)) | X = x\right] \geq \mathbb{E}_\Lambda\left[L(\Theta, \delta_\Lambda(x)) | X = x\right] \quad a.s.$$

Taking the expectation of both sides, then the results follows.

$$\mathbb{E}_\theta\left[\mathbb{E}_\Lambda\left[L(\Theta,\delta(x))|X=x\right]\right] \geq \mathbb{E}_\theta\left[\mathbb{E}_\Lambda\left[L(\Theta,\delta_\Lambda(x))|X=x\right]\right] \quad a.s.$$

Meanwhile, X has marginal distribution $\mathbb{P}_\Lambda(A) = \int \mathbb{P}_\theta(A)d\Lambda(\theta)$. □

Corollary 4.1.2. *(Corollary 4.1.2 in TPE)*
Suppose the assumptions of Theorem 4.1.1 holds, and we want to estimate $g(\theta)$

1. *If $L(\theta,d) = (d-g(\theta))^2$, and $\Theta \sim \Lambda$, then*

$$\delta_\Lambda(x) = \mathbb{E}_\Lambda\left[g(\Theta)|X=x\right]$$

More generally, if $L(\theta,d) = w(\theta)(d-g(\theta))^2$, then

$$\delta_\Lambda(x) = \frac{\mathbb{E}_\Lambda\left[w(\Theta)g(\Theta)|X=x\right]}{\mathbb{E}_\Lambda\left[w(\Theta)|X=x\right]} = \frac{\int w(\theta)g(\theta)d\Lambda(\theta|x)}{\int w(\theta)d\Lambda(\theta|x)}$$

Proof *The Bayes estimator is obtain by minimizing $\mathbb{E}_\Lambda[(g(\Theta)-\delta(x))^2|X=x]$ By assumption (1) of Theorem refThm: Bayes estimator Identification, there exists $\delta_0(x)$ for which $r(\Lambda,\delta) = \mathbb{E}_\theta\left[\mathbb{E}_\Lambda\left[L(\Theta,\delta(X))\right]\right] < \infty$ for almost all values of x, and it then follows that $\mathbb{E}_\Lambda[(g(\Theta)-\delta(x))^2|X=x]$ is minimized by $\delta_\Lambda(x) = \mathbb{E}_\Lambda\left[g(\Theta)|X=x\right]$. Similarly for the weighted case.* □

2. *If $L(\theta,d) = |d-g(\theta)|$, then $\delta_\Lambda(x)$ is any median of the conditional distribution of Θ given x.*

3. *If*

$$L(\theta,d) = \begin{cases} 0 \quad , \quad |d-c| \leq c \\ 1 \quad , \quad |d-c| > c \end{cases}$$

then $\delta_\Lambda(x)$ is the midpoint of the interval I of length $2c$ which maximizes $\mathbb{P}\left[\Theta \in I|X=x\right]$.

$$\begin{aligned}
\delta_\Lambda &= \underset{d}{\operatorname{argmin}}\, \mathbb{E}_\Lambda\left[\mathbf{1}_{\{|d-\Theta|>c\}}|X=x\right] = \underset{d}{\operatorname{argmin}}\, \mathbb{P}_\Lambda(|d-\Theta|>c|X=x) \\
&= \underset{d}{\operatorname{argmax}}\, \mathbb{P}_\Lambda(|d-\Theta|\leq c|X=x) \\
&= \textit{ midpoint of } I, \textit{ where } I \textit{ is the interval } \underset{d}{\operatorname{argmax}}\, \mathbb{P}_\Lambda(\Theta\in I|X=x)
\end{aligned}$$

Remark If the map $d \mapsto L(\theta,d)$ is strictly convex, then for every $\theta\in\Omega$, and $\mathbb{P}_\Lambda[A]=0 \implies \mathbb{P}_\theta[A]=0, \forall\theta\in\Omega$, where Λ is supported on Ω, then the Bayes estimator (if exists) is unique.

Example 22: (Poisson, Example 4.1.3 in TPE)
Under squared error loss function $L(\theta,d)=(d-\theta)^2$, if $X_1,\cdots,X_n$ are i.i.d $Poisson(\theta)$, then the likelyhood

$$L(\theta|x) = \prod_{i=1}^n \frac{e^{-\theta}\theta^{x_i}}{x_i!}$$

Meanwhile, θ has the $\Lambda = gamma(\alpha, \beta) = \Gamma(\alpha, \beta)$ prior distribution,

$$p(\theta|\alpha, \beta) = \frac{\beta^{\alpha}\theta^{\alpha-1}e^{-\beta\theta}}{\Gamma(\alpha)} \quad \text{for } \theta > 0 \text{ and } \alpha, \beta > 0$$

$$L(x_1, \cdots, x_n|\theta) = \prod_{i=1}^{n} \frac{e^{-\theta}\theta^{x_i}}{(x_i)!} = \frac{e^{-n\theta}\theta^{\sum_{i=1}^{n} x_i}}{\prod_{i=1}^{n}(x_i)!}$$

1. Thus the posterior

$$\pi(\theta|x_1, \cdots, x_n) \propto \theta^{\alpha-1+\sum_{i=1}^{n} x_i} e^{-(n+\beta)\theta} \quad \text{for } \theta > 0 \text{ and } \alpha, \beta > 0$$

   Thus posterior distribution is $\Theta|X = x \sim \Gamma(\alpha + \sum_{i=1}^{n} x_i, n + \beta)$.

2. By the previous corollary, the Bayes estimator is

$$\delta_\Lambda(x) = \mathbb{E}_\Lambda [g(\Theta)|X = x] = \mathbb{E}_\Lambda [\Theta|X = x]$$

3. How do things change if we observe $X_1, \cdots, X_n \overset{i.i.d}{\sim} Poisson(\theta)$.

♠

Corollary 4.1.3. *(Corollary 4.1.4 in TPE)*
If the loss function $L(\theta, d)$ is squared error, or more generally, if it is strictly convex in d, if provided

1. *the average risk of δ_Λ with respect to Λ is finite, and*
2. *if Q is the marginal distribution of X given by $Q(A) = \int \mathbb{P}_\theta[X \in A]d\Lambda(\theta)$,*

then a.e. Q implies a.e. $\mathcal{P}$, which means that a Bayes solution δ_Λ is unique (a.e. $\mathcal{P}$), where $\mathcal{P}$ is the class of distributions $\mathbb{P}_\theta$, provided

Sketch of Proof For general strictly convex loss functions, the result follows by the same argument from Problem 1.7.26. For squared error loss, it follows the previous corollary. □

Three Kind of Bayes

1. Empirical Bayes. The parameters of the prior distribution are themselves estimated from the data.
2. Hierarchical Bayes. The parameters of the prior distribution are, in turn, modeled by another distribution, sometimes called a hyperprior distribution.
3. Robust Bayes. The performance of an estimator is evaluated for each member of the prior class, with the goal of finding an estimator that performs well (is robust) for the entire class.

Example 23: (Binomial, Example 4.1.5 in TPE)

Suppose that $X \sim Binomial(n, \theta)$, loss function $L(\theta, d) = (d - \theta)^2$. A two-parameter family of prior distributions for θ is beta distribution $Beta(a, b)$. To determine the Bayes estimator of a given estimand $g(\theta)$, let us first obtain the conditional distribution (posterior distribution) of θ given x.

First, the Likelihood

$$L(x|\theta) = \binom{n}{x} \theta^x (1-\theta)^{n-x}$$

and the prior distribution

$$p(\theta|a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \theta^{a-1} (1-\theta)^{b-1}$$

then the joint density of X and θ is

$$\binom{n}{x} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \theta^{a+x-1} (1-\theta)^{n-x+b-1}$$

The conditional density of θ given x is obtained by dividing by the marginal of x, which is a function of x alone. Thus, the conditional density of θ given x has the form

$$\pi(\theta|x) = C(a, b, x) \theta^{a+x-1} (1-\theta)^{n-x+b-1}$$

We can regard this again as beta distribution $Beta(\alpha, \beta)$, with $\alpha = a + x$, $\beta = n - x + b$. Then according to theorem 4.1.1, we have

$$\delta_\Lambda(x) = \mathbb{E}_\Lambda [\Theta | X = x] = \frac{\alpha}{\alpha + \beta} = \frac{a + x}{a + b + n}$$

It is interesting to compare this Bayes estimator with the usual estimator $\frac{X}{n}$.

- Before any observations are taken, the estimator from the Bayesian approach is the expectation of the prior distribution $Beta(a, b)$: $\mathbb{E}[\theta|a, b] = \frac{a}{a+b}$.
- Once X has been observed, the standard non-Bayesian (for example, UMVU) estimator is $\frac{X}{n}$

The estimator $\delta_\Lambda(x) = \frac{a+x}{a+b+n}$ lies between these two. In fact

$$\delta_\Lambda(X) = \frac{a + X}{a + b + n} = \left(\frac{a+b}{a+b+n}\right)\frac{a}{a+b} + \left(\frac{n}{a+b+n}\right)\frac{X}{n}$$

is a weighted average of $\frac{a}{a+b}$(the estimator of p before any observations are taken) and $\frac{X}{n}$(the estimator without consideration of a prior). ♠

(Problem) Derive the Bayes estimator for $L(\theta, d) = \frac{(d-\theta)^2}{\theta(1-\theta)}$

Proof After similar process, we find out the posterior density of θ given x is

$$\pi(\theta|x) = C(a, b, x) \theta^{a+x-1} (1-\theta)^{n-x+b-1}$$

We are going to use the weighted version in corollary 4.1.2. $L(\theta, d) = w(\theta)(d-\theta)^2 = \frac{(d-\theta)^2}{\theta(1-\theta)}$, where $w(\theta) = \frac{1}{\theta(1-\theta)}$, then

$$\begin{aligned}
\delta_\Lambda(x) &= \frac{\mathbb{E}_\Lambda\left[w(\Theta)g(\Theta)|X=x\right]}{\mathbb{E}_\Lambda\left[w(\Theta)|X=x\right]} = \frac{\int w(\theta)g(\theta)d\Lambda(\theta|x)}{\int w(\theta)d\Lambda(\theta|x)} = \frac{\int \frac{\theta}{\theta(1-\theta)}\theta^{a+x-1}(1-\theta)^{n-x+b-1}d\theta}{\int \frac{1}{\theta(1-\theta)}\theta^{a+x-1}(1-\theta)^{n-x+b-1}d\theta} \\
&= \frac{\int \theta^{a+x-1}(1-\theta)^{n-x+b-2}d\theta}{\int \theta^{a+x-2}(1-\theta)^{n-x+b-2}d\theta} \\
&= \frac{B(a+x, n-x+b-1)}{B(a+x-1, n-x+b-1)} = \frac{\Gamma(a+x)\Gamma(n-x+b-1)\Gamma(a+b+n-2)}{\Gamma(a+b+n-1)\Gamma(a+x-1)\Gamma(n-x+b-1)} \\
&= \frac{\Gamma(a+x)}{\Gamma(a+x-1)}\frac{\Gamma(a+b+n-2)}{\Gamma(a+b+n-1)} \quad \text{since } \Gamma(z+1) = z\Gamma(z) \\
&= \frac{(a+x-1)\Gamma(a+x-1)}{\Gamma(a+x-1)}\frac{\Gamma(a+b+n-2)}{(a+b+n-2)\Gamma(a+b+n-2)} \\
&= \frac{a+x-1}{a+b+n-2}
\end{aligned}$$

□

4.2 First Examples

Example 24: (Estimating Normal mean, Example 4.2.2 in TPE)

Let $X_1, \cdots, X_n$ be i.i.d as $\mathsf{N}(\theta, \sigma^2)$, with σ known, and let the estimand be θ. As a prior distribution for Θ, we shall assume the normal distribution $\mathsf{N}(\mu, b^2)$. The joint density of Θ and $\mathbf{X} = (X_1, \cdots, X_n)$ is then proportional to

$$f(\mathbf{x}, \theta) = \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i - \theta)^2\right]\exp\left[-\frac{1}{2b^2}(\theta-\mu)^2\right]$$

To obtain the posterior distribution of $\Theta|\mathbf{x}$, the joint density is divided by the marginal density of $\mathbf{X}$, so that the posterior distribution has the form $C(\mathbf{x})f(\theta|\mathbf{x})$, where $C(\mathbf{x})$ is any function of $\mathbf{x}$ not involving θ. Thus the posterior density of $\Theta|\mathbf{x}$ is

$$\begin{aligned}
\pi(\theta|x) &= C(\mathbf{x})\exp\left[-\frac{1}{2}\theta^2\left(\frac{n}{\sigma^2}+\frac{1}{b^2}\right) - \theta\left(\frac{\sum_{i=1}^n x_i}{\sigma^2} + \frac{\mu}{b^2}\right)\right] \\
&= C(\mathbf{x})\exp\left\{\left[-\frac{1}{2}\left(\frac{n}{\sigma^2}+\frac{1}{b^2}\right)\right]\left[\theta^2 - 2\theta\left(\frac{n\bar{x}}{\sigma^2}+\frac{\mu}{b^2}\right)\left(\frac{n}{\sigma^2}+\frac{1}{b^2}\right)^{-1}\right]\right\}
\end{aligned}$$

Thus the posterior density can be recognized to be the normal density with mean and variance as following

$$\mathbb{E}\left[\Theta|\mathbf{x}\right] = \left(\frac{n\bar{x}}{\sigma^2}+\frac{\mu}{b^2}\right)\left(\frac{n}{\sigma^2}+\frac{1}{b^2}\right)^{-1}, \quad \mathbb{V}\mathrm{ar}(\Theta|\mathbf{x}) = \left(\frac{n}{\sigma^2}+\frac{1}{b^2}\right)^{-1}$$

When the loss is squared error, the Bayes estimator of θ, given by corollary 4.1.2, can be written as

$$\delta_\Lambda(x) = \frac{n}{\sigma^2}\left(\frac{n}{\sigma^2}+\frac{1}{b^2}\right)^{-1}\cdot\bar{x} \quad + \quad \frac{1}{b^2}\left(\frac{n}{\sigma^2}+\frac{1}{b^2}\right)^{-1}\cdot\mu$$

and by Corollary 2.7.19, this result remains true for any loss function $\rho(d-\theta)$ for which ρ is convex and even.(to prove) ♠

Theorem 4.2.1. *(Theorem 4.2.3 in TPE)*

Let $\Theta \sim \Lambda$, and $\mathbb{P}_\theta$ denote the conditional distribution of X given θ. Consider the estimation of $g(\theta)$, under squared error loss function. Then, an unbiased estimator $\delta(X)$ can be a Bayes solution if and only if

$$\mathbb{E}_{\Lambda,\theta}\left[(\delta(X)-g(\Theta))^2\right]=0$$

where the expectation is taken with respect to variation in both X and Θ.

Proof of Theorem Suppose that $\delta(X)$ is a Bayes estimator and it is unbiased for estimating $g(\theta)$. Since it is under squared error loss, according to corollary 4.1.2, we deduce the Bayes estimator should be

$$\delta(X)=\mathbb{E}_\Lambda\left[g(\Theta)|X\right], \quad a.s.$$

Also, $\delta(X)$ is unbiased, then

$$\mathbb{E}_\theta\left[\delta(X)\right]=g(\theta), \quad \forall\theta\in\Omega$$

- Conditioning on X,

$$\mathbb{E}\left[g(\Theta)\delta(X)\right]=\mathbb{E}\left[\mathbb{E}\left[g(\Theta)\delta(X)|X\right]\right]=\mathbb{E}\left[\delta(X)\mathbb{E}\left[g(\Theta)|X\right]\right]=\mathbb{E}_\theta\left[\delta(X)^2\right]$$

- Conditioning on Θ,

$$\mathbb{E}\left[g(\Theta)\delta(X)\right]=\mathbb{E}\left[\mathbb{E}\left[g(\Theta)\delta(X)|\Theta\right]\right]=\mathbb{E}\left[g(\Theta)\mathbb{E}\left[\delta(X)|\Theta\right]\right]=\mathbb{E}_\Lambda\left[g(\Theta)^2\right]$$

It follows that

$$\mathbb{E}_{\Lambda,\theta}\left[(\delta(X)-g(\Theta))^2\right]=\mathbb{E}_\theta\left[\delta(X)^2\right]+\mathbb{E}_\Lambda\left[g(\Theta)^2\right]-2\mathbb{E}\left[g(\Theta)\delta(X)\right]=0$$

□

Example 25: (Sample means, Example 4.2.4 in TPE)

Problem $X_1,\cdots,X_n \overset{i.i.d}{\sim} \mathbb{P}_\theta$ with $\mathbb{E}\left[X_i\right]=\theta$ and $\mathbb{V}\mathrm{ar}(X_1)=\sigma^2$ where σ^2 is independent of θ, then the risk of $\bar{X}$ when given θ is

$$R(\theta,\bar{X})=\mathbb{E}\left[(\bar{X}-\theta)^2\right]=\frac{\sigma^2}{n}$$

Then for any proper prior distribution on Θ, $R(\Theta,\bar{X})=\frac{\sigma^2}{n}\neq 0$, which doesn't satisfy theorem 4.2.1. Thus $\bar{X}$ is not a Bayes estimator for any proper prior distribution. ♠

The beta and normal prior distributions in the binomial and normal cases are the so-called conjugate families of prior distributions.

Example 26: (Estimating normal variance with known mean, Example 4.2.5 in TPE)

$X_1, \cdots, X_n \overset{i.i.d}{\sim} \mathsf{N}(0, \sigma^2)$, define $\theta = \frac{1}{2\sigma^2}$, $r = \frac{n}{2}$. The joint density of X_i is $C\theta^r e^{-\theta \sum_{i=1}^n x_i^2}$. As conjugate prior for θ , we take the gamma density $\Gamma(b, \frac{1}{\alpha})$, which has the density

$$\frac{\alpha^b}{\Gamma(b)} \theta^{b-1} e^{-\alpha\theta}, \quad x > 0, b > 0, \alpha > 0$$

Also we know some properties of Gamma distribution $\Gamma(b, \frac{1}{\alpha})$

$$\begin{aligned} \mathbb{E}[\theta] &= \frac{b}{\alpha}, \quad \mathbb{E}\left[\theta^2\right] = \frac{b(b+1)}{\alpha^2} \\ \mathbb{E}\left[\frac{1}{\theta}\right] &= \frac{\alpha}{b-1}, \quad \mathbb{E}\left[\frac{1}{\theta^2}\right] = \frac{\alpha^2}{(b-1)(b-2)} \end{aligned}$$

We just want to estimate σ^2, thus $\sum_{i=1}^n x_i^2$ is sufficient. If we write $y = \sum_{i=1}^n x_i^2$, then the posterior density of θ given the x_i is

$$C(y)\theta^{r+b-1} e^{-\theta(y+\alpha)}$$

which can be regarded as the Gamma distribution $\Gamma(r + b, \frac{1}{\alpha+y})$.

1. Under loss function $L(\sigma^2, d) = (d - \sigma^2)^2$, by corollary 4.1.2 and $\sigma^2 = \frac{1}{2\theta}$, the Bayes estimator under squared error loss is

$$\delta_\Lambda(y) = \mathbb{E}\left[\frac{1}{2\theta}\Big| y = \sum_{i=1}^n x_i^2\right] = \int \frac{(\alpha+y)^{r+b}}{\Gamma(r+b)} \theta^{r+b-1} e^{-\theta(y+\alpha)} \frac{1}{2\theta} d\theta = \frac{\alpha+y}{2(r+b-1)} = \frac{\alpha+y}{n+2b-2}$$

   Thus the Bayes estimator is $\delta_\Lambda = \frac{\alpha+Y}{n+2b-2}$.

2. Under loss function $L(\sigma^2, d) = \frac{(d-\sigma^2)^2}{\sigma^4}$, where $w(\sigma) = \frac{1}{\sigma^4}$. since $\sigma^2 = \frac{1}{2\theta}$, then $L(\theta, d) = 4\theta^2(d - \frac{1}{2\theta})^2$. Let $w(\theta) = 4\theta^2$,$g(\theta) = \frac{1}{2\theta}$ then by corollary 4.1.2

$$\begin{aligned} \delta_\Lambda(y) &= \frac{\mathbb{E}_\Lambda\left[w(\Theta)g(\Theta)|Y\right]}{\mathbb{E}_\Lambda\left[w(\Theta)|Y\right]} = \frac{\int 4\theta^2 \frac{1}{2\theta} \frac{(\alpha+y)^{r+b}}{\Gamma(r+b)} \theta^{r+b-1} e^{-\theta(y+\alpha)} d\theta}{\int 4\theta^2 \frac{(\alpha+y)^{r+b}}{\Gamma(r+b)} \theta^{r+b-1} e^{-\theta(y+\alpha)} d\theta} \\ &= \frac{\frac{2(r+b)}{\alpha+y}}{\frac{4(r+b)(r+b+1)}{(\alpha+y)^2}} = \frac{\alpha+y}{2r+2b+2} = \frac{\alpha+y}{n+2b+2} \end{aligned}$$

   Thus the Bayes estimator is $\delta_\Lambda = \frac{\alpha+Y}{n+2b+2}$.

♠

Example 27: (Estimating normal variance with unknown mean, Example 4.2.6 in TPE) $X_1, \cdots, X_n \overset{i.i.d}{\sim} \mathsf{N}(\xi, \sigma^2)$, define $\tau = \frac{1}{2\sigma^2}$, $r = \frac{n}{2}$. The joint density of X_i is

$$C\tau^r e^{-\tau \sum_{i=1}^n (x_i-\xi)^2} = C\tau^r \exp\left[-\tau \sum_{i=1}^n (x_i - \bar{x})^2 - \tau n(\bar{x} - \xi)^2\right]$$

As conjugate prior for τ , we take the gamma density $\Gamma(b, \frac{1}{\alpha})$, which has the density

$$\frac{\alpha^b}{\Gamma(b)}\tau^{b-1}e^{-\alpha\tau}, \quad x>0, b>0, \alpha>0$$

Let $\xi \sim N(0, a^2)$ and ξ be independent of τ. Then the posterior density of (ξ, τ) given $x_1, \cdots, x_n$ is proportional to

$$\tau^{r+b-1}\exp\Big[-\tau\left(\alpha+\sum_{i=1}^{n}(x_i-\bar{x})^2+n(\bar{x}-\xi)^2\right)-\frac{1}{2a^2}\xi^2\Big]$$

Let $z=\sum_{i=1}^{n}(x_i-\bar{x})^2$. We take $a\to\infty$, then (ξ,τ) given $x_1,\cdots,x_n$ is proportional to

$$\tau^{r+b-1}\exp\Big[-\tau\left(\alpha+\sum_{i=1}^{n}(x_i-\bar{x})^2+n(\bar{x}-\xi)^2\right)\Big]$$

1. When estimating $\tau=\frac{1}{2\sigma^2}$.
   By integrating out ξ, using the change of variable while actually integrating $\sqrt{\tau}(\xi-\bar{x})$, it is seen that the posterior distribution of τ is proportional to

   $$\tau^{r+b-\frac{3}{2}}\exp\Big[-\tau\,(\alpha+z)\Big]$$

   then the posterior of τ given z distribution is $\Gamma[r+b-\frac{1}{2},\frac{1}{\alpha+z}]$.

   - Under loss function $L(\sigma^2,d)=(d-\sigma^2)^2$, by corollary 4.1.2, the best Bayes estimator is

   $$\begin{aligned}\delta_\Lambda &= \mathbb{E}\left[\frac{1}{2\tau}\Big|z=\sum_{i=1}^{n}(x_i-\bar{x})^2\right]=\int\frac{(\alpha+z)^{r+b-\frac{1}{2}}}{\Gamma(r+b-\frac{1}{2})}\tau^{r+b-\frac{3}{2}}e^{-\tau(z+\alpha)}\frac{1}{2\tau}d\tau\\ &= \frac{\alpha+z}{2(r+b-\frac{1}{2})}=\frac{\alpha+z}{n+2b-1}\end{aligned}$$

   - Under loss function $L(\sigma^2,d)=\frac{(d-\sigma^2)^2}{\sigma^4}$, where $w(\sigma)=\frac{1}{\sigma^4}$. since $\sigma^2=\frac{1}{2\tau}$, then $L(\tau,d)=4\tau^2(d-\frac{1}{2\tau})^2$. Let $w(\tau)=4\tau^2$,$g(\tau)=\frac{1}{2\tau}$ then by corollary 4.1.2

   $$\begin{aligned}\delta_\Lambda &= \frac{\mathbb{E}_\Lambda\left[w(\Theta)g(\Theta)|Y\right]}{\mathbb{E}_\Lambda\left[w(\Theta)|Y\right]}=\frac{\int 4\tau^2\frac{1}{2\tau}\frac{(\alpha+y)^{r+b-\frac{1}{2}}}{\Gamma(r+b-\frac{1}{2})}\tau^{r+b-\frac{3}{2}}e^{-\tau(y+\alpha)}d\tau}{\int 4\tau^2\frac{(\alpha+y)^{r+b-\frac{1}{2}}}{\Gamma(r+b-\frac{1}{2})}\tau^{r+b-\frac{3}{2}}e^{-\tau(y+\alpha)}d\tau}\\ &= \frac{\frac{2(r+b-\frac{1}{2})}{\alpha+y}}{\frac{4(r+b-\frac{1}{2})(r+b+\frac{1}{2})}{(\alpha+y)^2}}=\frac{\alpha+y}{2r+2b+1}=\frac{\alpha+y}{n+2b+1}\end{aligned}$$

2. When estimating ξ, by integrating out τ, the posterior density of ξ is proportional to

   $$n(\bar{x}-\xi)^{-2}$$

   which is symmetric about $\bar{x}$ under any loss function. Then the Bayes estimator of ξ is $\bar{X}$ regardless of the values of α and b.

♠

Definition 4.2.1. *(Generalized Bayes Estimator, definition 4.2.9 in TPE)*
An estimator $\delta_\pi(x)$ is a generalized Bayes estimator with respect to a measure $\pi(\theta)$ (even if it is not a proper probability distribution) if the posterior expected loss, $\mathbb{E}\left[L(\Theta, \delta(X))|X = x\right]$, *is minimized at $\delta = \delta_\pi$ for all x.*

Definition 4.2.2. *(Limit of Bayes estimators, definition 4.2.10 in TPE)*
A nonrandomized estimator $\delta(x)$ is a limit of Bayes estimators if there exists a sequence of proper priors π_ν and Bayes estimators $\delta(x)^{\pi_\nu}$ such that $\delta(x)^{\pi_\nu} \to \delta(x)$ a.e. as $\nu \to \infty$ [with respect to the density $f(x|\theta)$]

Example 28: (Improper prior Bayes, Example 4.2.8 and 4.2.11 in TPE)
Continuing the case in Example 23, where $X \sim Binomial(n, \theta)$, loss function $L(\theta, d) = (d - \theta)^2$ and $\theta \sim Beta(a, b)$. The best Bayes estimator is

$$\delta_\Lambda(X) = \frac{a + X}{a + b + n} = \left(\frac{a + b}{a + b + n}\right)\frac{a}{a + b} + \left(\frac{n}{a + b + n}\right)\frac{X}{n}$$

For $a = b = 0$, we have the best estimator $\frac{X}{n}$. However, the prior density in this case is $p(\theta) \propto \theta^{-1}(1-\theta)^{-1}$, which is not integrable, $i.e.\int_0^1 \theta^{-1}(1-\theta)^{-1}d\theta = \infty$. Thus this prior density is improper. The posterior distribution, in this case, is proportional to

$$\binom{n}{x}\theta^{x-1}(1-\theta)^{n-x-1}$$

which is a proper posterior distribution if $1 \leq x \leq n - 1$ with X/n the posterior mean. When $X = 0$ or $X = n$, the previous posterior density is no longer proper.

However, for any estimator $\delta(x)$ that satisfies $\delta(0) = 0$ and $\delta(n) = 1$, the posterior expected loss (1.3) is finite and minimized at $\delta(x) = x/n$ (see Problem 2.16 and Example 2.4). Thus, even though the resulting posterior distribution is not proper for all values of x, $\delta(x) = x/n$ can be considered a Bayes estimator.

We shall now show that it is also a limit of Bayes estimators. This follows since

$$\lim_{a\to 0, b\to 0} \frac{a + X}{a + b + n} = \frac{X}{n}$$

and the $Beta(a, b)$ prior is proper if $a > 0, b > 0$. ♠

Let us next examine the connection between Bayes estimation, sufficiency, and the likelihood function.

If $bfX = (X_1, X_2, \cdots, X_n)$ has density $f(x_1, x_2, \cdots, x_n|\theta) = f(\mathbf{x}|\theta)$, the likelihood function is defined by

$$L(\theta|\mathbf{x}) = f(\mathbf{x}|\theta) = f(x_1, x_2, \cdots, x_n|\theta)$$

If we observe $T = \mathbf{t}$, where T is sufficient for θ, then

$$L(\theta|\mathbf{x}) = f(\mathbf{x}|\theta) = f(x_1, x_2, \cdots, x_n|\theta) = g(\mathbf{t}|\theta)h(\mathbf{x})$$

For any prior distribution $\pi(\theta)$, the posterior distribution is

$$\pi(\theta|\mathbf{x}) = \frac{f(\mathbf{x}|\theta)\pi(\theta)}{\int f(\mathbf{x}|\theta')\pi(\theta')d\theta'} = \frac{g(\mathbf{t}|\theta)h(\mathbf{x})\pi(\theta)}{\int g(\mathbf{t}|\theta')h(\mathbf{x})\pi(\theta')d\theta'} = \frac{g(\mathbf{t}|\theta)\pi(\theta)}{\int g(\mathbf{t}|\theta')\pi(\theta')d\theta'}$$

which only depends on $\mathbf{t}$. Thus $\pi(\theta|\mathbf{x}) = \pi(\theta|\mathbf{t})$. That is, $\pi(\theta|\mathbf{x})$ depends on $\mathbf{x}$ only through $\mathbf{x}$, and the posterior distribution of θ is the same whether we compute it on the basis of $\mathbf{x}$ or of $\mathbf{t}$.

Problems

Example 29: $X \sim \mathsf{N}(\theta, 1)$, $\Lambda_b = \mathsf{N}(0, b^2)$, $\theta \in \mathbb{R}$

1. $\delta_b(x) = \frac{b^2}{b^2+1}x \overset{b\to\infty}{\to} x$
2. Λ = Lebesgue measure $\to \delta_\Lambda(x) = x$

♠

Example 30: $X \sim \mathsf{N}(\theta, 1)$, Λ_b has density $e^{b\theta}$

1. $\delta_b(x) = x + b$
2. Show that this estimator is not the limit of proper estimators.

♠

4.3 Single-Prior Bayes

We now denote the prior density by $\pi(\theta|\gamma)$, where the parameter γ can be real- or vector-valued. (Hence, we are implicitly assuming that the prior π is absolutely continuous with respect to a dominating measure $\mu(\theta)$, which, unless specified, is taken to be Lebesgue measure.)

We can then write a Bayes model in a general form as

$$X|\theta \sim f(x|\theta)$$
$$\Theta|\gamma \sim \pi(\theta|\gamma)$$

Thus, conditionally on θ ,X has sampling density $f(x|\theta)$, and conditionally on γ, Θ has prior density $\pi(\theta|\gamma)$. From this model, we calculate the posterior distribution, $\pi(\theta|\gamma,\gamma)$, from which all Bayesian answers would come.

In general, the Bayes estimator under squared error loss is given by

$$\mathbb{E}\left[\Theta|x\right] \frac{\int \theta f(x|\theta)\pi(\theta)d(\theta)}{\int f(x|\theta)\pi(\theta)d(\theta)} \tag{4.2}$$

where $X \sim f(x|\theta)$ is the observed random variable and $\Theta \sim \pi(\theta)$ is the parameter of interest. However, the expression of equation 4.2 is somewhat difficult to deal with. We want to simplify this expression under some conditions for 2 purposes.

1. **Implementation**
   If a Bayes solution is deemed appropriate, and we want to implement it, we must be able to calculate equation 4.2. Thus, we need reasonably straightforward, and general, methods of evaluating these integrals.

2. **Performance**
   By construction, a Bayes estimator minimizes the posterior expected loss and, hence, the Bayes risk. Often, we are interested in its performance, and perhaps optimality under other measures. For example, we might examine its mean squared error (or, more generally, its risk function) in looking for admissible or minimax estimators. We also might examine Bayesian measures using other priors, in an investigation of Bayesian robustness.

Actually, if we are dealing with independent prior distributions, *i.e.*, $X_i \sim f(x|\theta_i), i = 1, \cdots, n$ are independent and the prior satisfies $\pi(\theta_1, \cdots, \theta_n) = \prod_{i=1}^n \pi(\theta_i)$, then the posterior mean of θ_i satisfies

$$\mathbb{E}\left[\theta_i | x_1, \cdots, x_n\right] = \mathbb{E}\left[\theta_i | x_i\right], \quad 1 \leq i \leq n$$

that is, the Bayes estimator of θ_i only depends on the data through xi. Although the simplification provided by 4.2 may prove useful, it can not go any further at this level of generality.

However, for exponential families, evaluation of 4.2 is sometimes possible through alternate representations of Bayes estimators. We can express the Bayes estimator as a function of partial derivatives with respect to $\mathbf{x}$ using the following theorem.

Theorem 4.3.1. *(Theorem 4.3.2 in TPE)*
If η has prior density $\pi(\eta)$ and $\mathbf{X}$ has density

$$p_\eta(\mathbf{x}) = \exp\left[\sum_{i=1}^{s} \eta_i T_i(\mathbf{x}) - A(\eta)\right] h(\mathbf{x})$$

Then for $j = 1, \cdots, n$

$$\mathbb{E}\left[\sum_{i=1}^{s} \eta_i \frac{T_i(\mathbf{x})}{\partial x_j} \Big| \mathbf{x}\right] = \frac{\partial}{\partial x_j} \log m(\mathbf{x}) - \frac{\partial}{\partial x_j} \log h(\mathbf{x})$$

where $m(\mathbf{x}) = \int p_\eta(\mathbf{x}) \pi(\eta) d\eta$ is marginal distribution of $\mathbf{X}$. Alternatively, the posterior expectation can be expressed in the matrix form as

$$\mathbb{E}\left[\mathcal{T}\eta\right] = \nabla \log m(\mathbf{x}) - \nabla \log h(\mathbf{x})$$

where $\mathcal{T} = \{\frac{\partial T_i}{\partial x_j}\}$, i.e.

$$\mathcal{T} = \begin{bmatrix} \frac{T_1(\mathbf{x})}{\partial x_1} & \cdots & \frac{T_s(\mathbf{x})}{\partial x_1} \\ \vdots & \ddots & \vdots \\ \frac{T_1(\mathbf{x})}{\partial x_n} & \cdots & \frac{T_s(\mathbf{x})}{\partial x_n} \end{bmatrix}$$

Corollary 4.3.2. *(Corollary 4.3.3 in TPE) If* $\mathbf{X} = (X_1, \cdots, X_p)$ *has the density*

$$p_\eta(\mathbf{x}) = \exp\Big[\sum_{i=1}^{p} \eta_i x_i - A(\eta)\Big] h(\mathbf{x})$$

and the η *has prior density* $\pi(\eta)$*, the Bayes estimator of* η *under the loss* $L(\eta, \delta) = \sum_{i=1}^{s} (\eta_i - \delta_i)^2$ *is given by*

$$\mathbb{E}\Big[\eta_i \Big| \mathbf{x}\Big] = \frac{\partial}{\partial x_i} \log m(\mathbf{x}) - \frac{\partial}{\partial x_i} \log h(\mathbf{x})$$

Similarly, in the matrix form as

$$\mathbb{E}[\eta] = \nabla \log m(\mathbf{x}) - \nabla \log h(\mathbf{x})$$

Sketch of Proof We just apply $T_i(\mathbf{x}) = x_i$, and applying the previous theorem. Then it is done. $\square$

Theorem 4.3.3. *(Theorem 4.3.5 in TPE) Under the assumptions of previous Corollary, the risk of the Bayes estimator* $\mathbb{E}\Big[\eta_i \Big| \mathbf{x}\Big]$*, under the sum of squared error loss, is*

$$R(\eta, \mathbb{E}\Big[\eta \Big| \mathbf{x}\Big]) = R[\eta, -\nabla \log h(\mathbf{x})] + \sum_{i=1}^{p} \mathbb{E}\Big[2 \frac{\partial^2}{\partial x_i^2} \log m(\mathbf{x}) + \Big(\frac{\partial}{\partial x_i} \log m(\mathbf{x})\Big)^2\Big]$$

Example 31: (Multiple normal model, Example 4.3.4, 4.3.6 in TPE)

$$\begin{aligned} X_i|\theta_i &\sim \mathsf{N}(\theta_i, \sigma^2), \quad i = 1, \cdots, p \text{ independent} \\ \Theta_i &\sim \mathsf{N}(\mu, \tau^2), \quad i = 1, \cdots, p \text{ independent} \end{aligned}$$

where μ, τ, σ are known. The joint density of $\mathbf{X} = (X_1, \cdots, X_n)$ is

$$\begin{aligned} p_\eta(\mathbf{x}) &= (2\pi\sigma^2)^{-n/2} \exp\Big[-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \theta)^2\Big] \\ &= \exp\Big[\sum_{i=1}^{n} \frac{\theta_i}{\sigma^2} x_i - \sum_{i=1}^{n} \frac{\theta_i^2}{2\sigma^2}\Big] (2\pi\sigma^2)^{-n/2} \exp\Big[-\frac{1}{2\sigma^2} \sum_{i=1}^{n} x_i^2\Big] \\ &= \exp\Big[\sum_{i=1}^{n} \frac{\theta_i}{\sigma^2} x_i - \sum_{i=1}^{n} \frac{\theta_i^2}{2\sigma^2}\Big] h(\mathbf{x}) \\ &= \exp\Big[\sum_{i=1}^{n} \eta_i x_i - A(\eta)\Big] h(\mathbf{x}) \end{aligned}$$

where $\eta_i = \theta_i/\sigma^2$, $A(\eta) = \frac{\sigma^2}{2}\sum_{i=1}^n \eta_i^2$, $h(\mathbf{x}) = (2\pi\sigma^2)^{-n/2}\exp\Big[-\frac{1}{2\sigma^2}\sum_{i=1}^n x_i^2\Big]$.

$$\begin{aligned} m(\mathbf{x}) &= \int p_\eta(\mathbf{x})\pi(\eta)d\eta \\ &= h(\mathbf{x})(2\pi\tau^2)^{-n/2}\int \exp\Big[\sum_{i=1}^n \eta_i x_i - A(\eta)\Big]\exp\Big[-\frac{1}{2\tau^2}\sum_{i=1}^n(\eta_i\sigma^2-\mu)^2\Big]d\eta_1\cdots d\eta_n \\ &= (2\pi\tau\sigma)^{-n}\exp\Big[-\frac{1}{2\sigma^2}\sum_{i=1}^n x_i^2\Big]\prod_{i=1}^n\int\exp\Big[\eta_i x_i - \frac{1}{2\sigma^2}\eta_i^2 - \frac{1}{2\tau^2}(\eta_i\sigma^2-\mu)^2\Big]d\eta_i \\ &= [2\pi(\sigma^2+\tau^2)]^{-\frac{n}{2}}\exp\Big[-\frac{1}{2(\sigma^2+\tau^2)}\sum_{i=1}^n(x_i-\mu)^2\Big] \end{aligned}$$

$$\begin{aligned} \frac{\partial}{\partial x_i}\log m(\mathbf{x}) &= -\frac{x_i-\mu}{\sigma^2+\tau^2} \\ \frac{\partial}{\partial x_i}\log h(\mathbf{x}) &= -\frac{x_i}{\sigma^2} \end{aligned}$$

Then

$$\mathbb{E}\left[\Theta_i|\mathbf{X}\right] = \sigma^2\mathbb{E}\left[\eta_i|\mathbf{X}\right] = \sigma^2\Big[\frac{\partial}{\partial x_i}\log m(\mathbf{x}) - \frac{\partial}{\partial x_i}\log h(\mathbf{x})\Big] = \frac{\tau^2}{\sigma^2+\tau^2}x_i + \frac{\sigma^2}{\sigma^2+\tau^2}\mu$$

To evaluate the risk of the Bayes estimator, we also calculate

$$\frac{\partial^2}{\partial x_i^2}\log m(\mathbf{x}) = -\frac{1}{\sigma^2+\tau^2}$$

Hence

$$R(\eta, \mathbb{E}\Big[\eta\Big|\mathbf{x}\Big]) = R[\eta, -\nabla\log h(\mathbf{x})] - \frac{2p}{\sigma^2+\tau^2} + \sum_{i=1}^p \mathbb{E}\left[\Big(\frac{X_i-\mu}{\sigma^2+\tau^2}\Big)^2\right]$$

The best unbiased estimator(UMVU) of $\eta_i = \frac{\theta_i}{\sigma^2}$ is

$$-\frac{\partial}{\partial X_i}\log h(\mathbf{X}) = \frac{X_i}{\sigma^2}$$

with risk $R(\eta, -\nabla\log h(\mathbf{X})) = \frac{p}{\sigma^2}$. If $\eta_i = \mu$ for each i, then the Bayes estimator has smaller risk, whereas the Bayes estimator has infinite risk as $|\eta_i - \mu| \to \infty$ ♠

In exponential families, there is a general expression for the conjugate prior distribution and that use of this conjugate prior results in a simple expression for the posterior mean. For the density

$$p_\eta(\mathbf{x}) = e^{\eta\mathbf{x}-A(\eta)}h(\mathbf{x}), \quad -\infty < x < \infty$$

and the conjugate prior family is

$$\pi(\eta|k,\mu) = c(k,\mu)e^{k\eta\mu - kA(\eta)},$$

where μ can be thought of as a prior mean and k is proportional to a prior variance or a prior sample size. According to the result of Problem 4.3.9 in TPE,

$$\mathbb{E}[X] = A'(\eta), \qquad \mathbb{V}\text{ar}(X) = A''(\eta), \text{ the expectation is with respect to } p_\eta(\mathbf{x})$$
$$\mathbb{E}[A'(\eta)] = \mu, \qquad \mathbb{V}\text{ar}(A(\eta)) = \frac{1}{k}\mathbb{E}[A''(\eta)], \text{ the expectation is with respect to } \pi(\eta|k,\mu)$$

If $X_1, \cdots, X_n$ is a sample from $p_\eta(\mathbf{x})$, the posterior distribution resulting from $\pi(\eta|k,\mu)$ is

$$\pi(\eta|\mathbf{x},k,\mu) \propto \left[e^{n\eta\bar{x}-nA(\eta)}\right]\left[e^{k\eta\mu-kA(\eta)}\right] = e^{\eta(n\bar{x}+k\mu)-(n+k)A(\eta)}$$

which is still the same form of the conjugate prior family with $k' = n+k$, $\mu' = \frac{n\bar{x}+k\mu}{n+k}$

Suppose that now we are interested in estimating the $A'(\eta)$ under squared error loss. Thus, using Problem 4.3.9

$$\mathbb{E}[A'(\eta)|\mathbf{x},k,\mu] = \mu' = \frac{n\bar{x}+k\mu}{n+k}$$

we see that the posterior mean is a convex combination of the sample mean $\bar{x}$ and prior mean μ.

Example 32: (Conjugate gamma, Example 4.3.7 in TPE)

$X_1, \cdots, X_n \overset{i.i.d.}{\sim} Gamma(\alpha, \beta)$, where α is known, the density is

$$\begin{aligned} f(x;\alpha,\beta) &= \frac{x^{\alpha-1}e^{-\frac{x}{\beta}}}{\beta^\alpha\Gamma(\alpha)} \qquad \text{for } x>0 \text{ and } \alpha,\beta>0, \\ &= \exp\left[-\frac{x}{\beta}+\alpha\log(\frac{1}{\beta})\right]\frac{x^{\alpha-1}}{\Gamma(\alpha)} \\ &= \exp\left[\eta x - A(\eta)\right]h(x) \end{aligned}$$

where $\eta = -\frac{1}{\beta}$, $A(\eta) = -\alpha\log(-\eta)$ and $h(x) = \frac{x^{\alpha-1}}{\Gamma(\alpha)}$, so this is in the form of exponential family. If we use a conjugate prior distribution

$$\pi(\eta|k,\mu) = c(k,\mu)e^{k\eta\mu-kA(\eta)}$$

for β, then using the previous result,

$$\mathbb{E}\left[A'(\eta)\Big|\mathbf{x},k,\mu\right] = \mathbb{E}\left[-\frac{\alpha}{\eta}\Big|\mathbf{x},k,\mu\right] = \mu' = \frac{n\bar{x}+k\mu}{n+k}$$

Then the resulting Bayes estimator under squared error loss is

$$\mathbb{E}\left[\beta\Big|\mathbf{x},k,\mu\right] = \frac{\mu'}{\alpha} = \frac{n\bar{x}+k\mu}{\alpha(n+k)}$$

♠

4.4 Equivariant Bayes

If an estimation problem being invariant under a transformation g of the sample space, and the induced transformations $\bar{g}$ and g^* of the parameter and decision spaces, when considering Bayes estimation, it is natural to select a prior distribution which is also invariant.

Definition 4.4.1. *(Invariant Prior Distribution with respect to $\bar{G}$)*
A prior distribution Λ for θ is invariant with respect to $\bar{G}$ if the distribution of $\bar{g}\theta \sim \Lambda$ for all $\bar{g} \in \bar{G}$; that is, for all $\bar{g} \in \bar{G}$ and all measurable B, we have

$$\mathbb{P}_\Lambda(\bar{g}\theta \in B) = \mathbb{P}_\Lambda(\theta \in B), \text{ or equivalently } \quad \Lambda(B) = \Lambda((\bar{g})^{-1}B) \tag{4.3}$$

Suppose now that such a Λ exists and that the Bayes solution δ_Λ with respect to Λ is unique. By definition 4.3 , any δ satisfies

$$r(\Lambda, \delta) = \int_\Omega R(\theta, \delta(X)) d\Lambda(\theta) = \int_\Omega R(\bar{g}\theta, \delta(X)) d\Lambda(\theta)$$

Since

$$\begin{aligned}
R(\bar{g}\theta, \delta) &= \mathbb{E}_{\bar{g}\theta}\left[L(\bar{g}\theta, \delta(X))\right] \\
&= \mathbb{E}_\theta\left[L(\bar{g}\theta, \delta(gX))\right] \quad \text{(property of equivariant estimator)} \\
&= \mathbb{E}_\theta\left[L(\theta, (g^*)^{-1}\delta(gX))\right]
\end{aligned}$$

Then for prior distribution Λ

$$\begin{aligned}
r(\Lambda, \delta_\Lambda) &= \int_\Omega R(\theta, \delta_\Lambda(X)) d\Lambda(\theta) = \int_\Omega R(g\theta, \delta_\Lambda(X)) d\Lambda(\theta) \\
&= \int_\Omega R(g\theta, (g^*)^{-1}\delta_\Lambda(gX)) d\Lambda(\theta)
\end{aligned}$$

if $\delta_\Lambda(X)$ minimizes $r(\Lambda, \delta_\Lambda)$, so does the estimator $(g^*)^{-1}\delta_\Lambda(gX)$. Hence, if the Bayes estimator is unique, the two must coincide, *i.e.*

$$(g^*)^{-1}\delta_\Lambda(gX) = \delta_\Lambda(X)$$

However, the uniqueness can be asserted only up to null sets, *i.e.*sets N with $\mathbb{P}_\theta(N) = 0, \forall\theta \in \Omega$. Moreover, the set N may depend on g. This encourage us to introduce the rigorous definition of Almost Equivariant Estimator.

Definition 4.4.2. *(Almost Equivariant Estimator)*
An estimator satisfying

$$\delta(X) = (g^*)^{-1}\delta(gX) \quad \forall x \notin N_g$$

*is said to be **almost equivariant**, where $\mathbb{P}_\theta(N_g) = 0, \forall\theta \in \Omega$ and N_g is a set which depends on action g.*

Theorem 4.4.1. *(Theorem 4.4.1 in TPE)*
Suppose that an estimation problem is invariant under a group G and that there exists a distribution Λ over Ω such that

$$\mathbb{P}_\Lambda(\bar{g}\theta \in B) = \mathbb{P}_\Lambda(\theta \in B), \text{ or equivalently } \quad \Lambda(B) = \Lambda((\bar{g})^{-1}B)$$

holds for all (measurable) subsets B of Ω and all $g \in G$. Then, if the Bayes estimator δ_Λ is unique, it is almost equivariant, which means that

$$\delta_\Lambda(X) = (g^*)^{-1}\delta_\Lambda(gX) \quad \forall x \notin N_g$$

where $\mathbb{P}_\theta(N_g) = 0, \forall \theta \in \Omega$ and N_g is a set which depends on action g.

Example 33: (Equivariant binomial, Example 4.4.2 in TPE)
Suppose we are interested in estimating θ under squared error loss, where $X \sim Binomial(n, \theta)$, where $\theta \in \Omega := [0, 1]$. A common group of transformations which leaves the problem invariant is

$$gX = n - X, \quad \bar{g}\theta = 1 - \theta$$

For a prior Λ satisfying equation 4.3, we must have

$$\mathbb{P}_\Lambda(\bar{g}\theta \leq t) = \mathbb{P}_\Lambda(\theta \leq t), \quad \forall t \in [0, 1]$$

If Λ has density $\gamma(\theta)$, then equation 4.3 implies

$$\int_0^t \gamma(\theta)d\theta = \int_0^t \gamma(1-\theta)d\theta, \quad \forall t \in [0, 1]$$

which, upon differentiating, requires $\gamma(t) = \gamma(1-t)$ for all $t \in [0, 1]$ and, hence, that $\gamma(t)$ must be symmetric about $t = 1/2$. It then follows that, for example, a Bayes rule under a symmetric beta prior is equivariant. ♠

Chapter 5

Minimaxity and Admissibility

In this chapter, we are going to minimize the risk function $R(\theta, \delta)$ in the following ways: minimizing maximum risk.

5.1 Minimax Estimation

Definition 5.1.1. *(Minimax Estimator, definition 5.1.1 in TPE)*
An estimator δ^M of θ, which minimizes the maximum risk, that is, which satisfies

$$\inf_{\delta} \sup_{\theta} R(\theta, \delta) = \sup_{\theta} R(\theta, \delta^M) \tag{5.1}$$

*is called a **minimax estimator**.*

Remark There is a sequence of estimators $\{\delta_k\}_{k=1}^{\infty}$ such that

$$\sup_{\theta \in \Omega} R(\theta, \delta) \overset{k \to \infty}{\longrightarrow} \inf_{\delta} \sup_{\theta} R(\theta, \delta)$$

If we want to apply the Bayesian idea to find the minimax estimator, there is a question: for what prior distribution Λ is the Bayes solution δ_Λ likely to be minimax? One possible answer is that the minimax estimator would be Bayes for the worst possible distribution We denote the average risk (Bayes risk) of the Bayes solution δ_Λ by

$$r_\Lambda = r(\Lambda, \delta_\Lambda) = \int R(\theta, \delta_\Lambda) d\Lambda(\theta)$$

Definition 5.1.2. *(Least Favorable Prior Distribution, definition 5.1.3 in TPE)*
A prior distribution Λ is least favorable if

$$r_\Lambda \geq r_{\Lambda'}$$

for all prior distribution Λ'.

This is the prior distribution which has the greatest average loss. The following theorem provides a simple condition for a Bayes estimator δ_Λ to be minimax.

Theorem 5.1.1. *(Theorem 5.1.4 in TPE)*
Suppose that Λ is a distribution on Θ such that

$$r(\Lambda, \delta_\Lambda) = \int R(\theta, \delta_\Lambda) d\Lambda(\theta) = \sup_\theta R(\theta, \delta_\Lambda)$$

Then

1. *δ_Λ is minimax estimator.*
2. *If δ_Λ is the unique Bayes solution with respect to Λ, it is the unique minimax estimator.*
3. *Λ is least favorable.*

Proof of Theorem

1. Let δ be any other estimator. Then,

$$\sup_\theta R(\theta, \delta) \geq \int R(\theta, \delta) d\Lambda(\theta) \geq \int R(\theta, \delta_\Lambda) d\Lambda(\theta) = \sup_\theta R(\theta, \delta_\Lambda)$$

2. This follows by replacing $\geq$ by $>$ in the second equality of the proof of (1).
3. Let Λ' be some other distribution of θ. Then

$$\begin{aligned} r(\Lambda', \delta_{\Lambda'}) &\leq \int R(\theta, \delta_{\Lambda'}) d\Lambda'(\theta) \leq \int R(\theta, \delta_\Lambda) d\Lambda'(\theta) \\ &\leq \int R(\theta, \delta_\Lambda) d\Lambda(\theta) = \sup_\theta R(\theta, \delta_\Lambda) \\ &= r(\Lambda, \delta_\Lambda) \end{aligned}$$

□

Corollary 5.1.2. *(Corollary 5.1.5 in TPE)*
If a Bayes solution δ_Λ has constant risk, then it is minimax.

Proof If δ_Λ has constant risk, then

$$r(\Lambda, \delta_\Lambda) = \int R(\theta, \delta_\Lambda) d\Lambda(\theta) = \sup_\theta R(\theta, \delta_\Lambda)$$

clearly holds. □

Corollary 5.1.3. *(Corollary 5.1.6 in TPE)*
Let ω_Λ be the set of parameter points at which the risk function of δ_Λ takes on its maximum, that is,

$$\omega_\Lambda = \{\theta : R(\theta, \delta_\Lambda) = \sup_{\theta'} R(\theta', \delta_\Lambda)\}$$

Then, δ_Λ is minimax if and only if $\Lambda(\omega_\Lambda) = 1$.

In other words, a sufficient condition for δ_Λ to be minimax is that there exists a set ω such that $\Lambda(\omega) = 1$ and $R(\theta, \delta_\Lambda)$ attains its maximum at all points of ω.

Sketch of Proof This means that any $\theta, \theta' \in \omega$, $R(\theta, \delta_\Lambda) = R(\theta', \delta_\Lambda)$. Since $\Lambda(\omega) = 1$, then $R(\theta, \delta_\Lambda)$ does not attain its maximum on some null sets. So that we can say δ_Λ has constant risk a.s. and the proof follows the previous corollary. □

Example 34: (Binomial, Example 5.1.7 in TPE)

Suppose that $X \sim \mathit{Binomial}(n, \theta)$, $\theta \in [0, 1]$, loss function $L(d, \theta) = (d - \theta)^2$. Fact: $\frac{X}{n}$ is not minimax. Since $\mathbb{E}_\theta [X] = n\theta$ and its risk function

$$\begin{aligned} R(\frac{X}{n}, \theta) &= \mathbb{E}_\theta \left[L(\frac{X}{n}, \theta) \right] = \mathbb{E}_\theta \left[(\frac{X}{n} - \theta)^2 \right] \\ &= \frac{1}{n^2} \mathbb{E} \left[(X - n\theta)^2 \right] = \frac{1}{n^2} \mathbb{E} \left[(X - \mathbb{E}[X])^2 \right] \\ &= \frac{1}{n^2} \mathbb{V}\mathrm{ar}(X) = \frac{1}{n^2} n\theta(1-\theta) = \frac{\theta(1-\theta)}{n} \end{aligned}$$

has a unique maximum at $\theta = 1/2$. According to previous Corollary, we need to use a prior distribution Λ for θ which assigns probability 1 to $\theta = 1/2$. The corresponding Bayes estimator is $\delta(X) \equiv 1/2$, not $\frac{X}{n}$.

In order to determine a minimax estimator by the method of Theorem 5.1.1, let us utilize the result of Example 23 and try a beta distribution for Λ. If $\Lambda = \mathit{Beta}(a, b)$ the Bayes estimator is given by

$$\delta_\Lambda(X) = \frac{a + X}{a + b + n}$$

The risk function is

$$\begin{aligned} R(\theta, \delta_\Lambda) &= \mathbb{E}_\theta [L(\theta, \delta_\Lambda(X))] = \frac{1}{(a+b+n)^2} \mathbb{E}_\theta \left[(a\theta + b\theta + n\theta - a - X)^2 \right] \\ &= \frac{1}{(a+b+n)^2} \left[n\theta(1-\theta) + (a - (a+b)\theta)^2 \right] \end{aligned}$$

If exist values a and b for which the risk function $R(\theta, \delta_\Lambda)$ is constant, then using the Corollary 5.1.2, then δ_Λ is minimax. We should set the coefficients of θ^2 and θ in $R(\theta, \delta_\Lambda)$ to be 0, that is

$$(a+b)^2 = n, \quad 2a(a+b) = n, \quad a, b > 0$$

Then $a = b = \frac{\sqrt{n}}{2}$. It follows that the estimator

$$\delta_\Lambda(X) = \frac{\sqrt{n}}{1+\sqrt{n}} \frac{X}{n} + \frac{1}{1+\sqrt{n}} \frac{1}{2}$$

is constant risk Bayes and, hence, minimax. Because of the uniqueness of the Bayes estimator, it is seen that $\delta_\Lambda(X)$ is the unique minimax estimator for θ.

However, the estimator $\delta_\Lambda(X)$ is biased, since $\frac{X}{n}$ is the only unbiased estimator for θ that is a function of X. $\delta_\Lambda(X)$ has constant risk $R(\theta, \delta_\Lambda) = \frac{1}{4(1+\sqrt{n})^2}$. $\frac{X}{n}$ has risk $R(\theta, \frac{X}{n}) = \frac{\theta(1-\theta)}{n}$.

Now, we do some comparisons between $\frac{X}{n}$ and $\delta_\Lambda(X)$ when n is large or small. For some interval $I_n = (\frac{1}{2} - c_n, \frac{1}{2} + c_n)$

- $R(\theta, \delta_\Lambda) < R(\theta, \frac{X}{n})$ if $\theta \in I_n = (\frac{1}{2} - c_n, \frac{1}{2} + c_n)$. For small values of n, c_n is close to $\frac{1}{2}$, so that the minimax $\delta_\Lambda(X)$ estimator is better (and, in fact, substantially better) for most of the range of θ.

- $R(\theta,\delta_\Lambda) > R(\theta,\frac{X}{n})$ if $\theta \notin I_n = (\frac{1}{2}-c_n,\frac{1}{2}+c_n)$. As $n \to \infty$, $c_n \to 0$ and I_n shrinks toward the point $\frac{1}{2}$. Furthermore,

$$\sup_\theta \frac{R(\theta,\frac{X}{n})}{R(\theta,\delta_\Lambda)} \longrightarrow 1 \quad \text{as } n \to \infty$$

  so that even at $\theta = \frac{1}{2}$, $\frac{X}{n}$ is better, and the improvement achieved by the minimax estimator is negligible. Thus, for large and even moderate n, $\frac{X}{n}$ dominates the minimax estimator $\delta_\Lambda(X)$.

♠

Remark The least favorable distribution is not unique in the previous example. For any Λ, the Bayes estimator of θ

$$\delta_\Lambda(x) = \frac{\mathbb{E}\left[w(\Theta)g(\Theta)|x\right]}{\mathbb{E}\left[w(\Theta)|x\right]} = \frac{\int_0^1 \theta^{x+1}(1-\theta)^{n-x}d\Lambda(\theta)}{\int_0^1 \theta^{x}(1-\theta)^{n-x}d\Lambda(\theta)}$$

Expansion of $(1-\theta)^{n-x}$ in powers of θ shows that $\delta_\Lambda(x)$ depends on Λ only through the first $n+1$ moments of Λ. Any prior distribution with the same first $n+1$ moments gives the same Bayes solution and, hence, by Theorem 5.1.1 is least favorable.

Example 35: (Problem)
$\frac{X}{n}$ becomes a constant risk estimator under the loss function $L(\theta,d) = \frac{(d-\theta)^2}{\theta(1-\theta)}$ ♠

Proof

$$\begin{aligned}
R(\frac{X}{n},\theta) &= \mathbb{E}_\theta\left[L(\frac{X}{n},\theta)\right] = \mathbb{E}_\theta\left[\frac{(\frac{X}{n}-\theta)^2}{\theta(1-\theta)}\right] \\
&= \frac{1}{n^2\theta(1-\theta)}\mathbb{E}\left[(X-n\theta)^2\right] = \frac{1}{n^2\theta(1-\theta)}\mathbb{E}\left[(X-\mathbb{E}\left[X\right])^2\right] \\
&= \frac{1}{n^2\theta(1-\theta)}\mathbb{V}\text{ar}(X) = \frac{1}{n^2\theta(1-\theta)}n\theta(1-\theta) = \frac{1}{n}
\end{aligned}$$

□

Definition 5.1.3. *(Definition 5.1.11 in TPE)*
A sequence of prior distributions $\{\Lambda_n\}$ is least favorable if for every prior distribution Λ we have

$$r_\Lambda \le r = \lim_{n\to\infty} r_{\Lambda_n}$$

where

$$r_{\Lambda_n} = r(\Lambda_n,\delta_\Lambda) = \int R(\theta,\delta_{\Lambda_n})d\Lambda_n(\theta) = \sup_\theta R(\theta,\delta_{\Lambda_n})$$

Theorem 5.1.4. *(Theorem 5.1.12 in TPE)*
Suppose that $\{\Lambda_n\}$ is a sequence of prior distributions with Bayes risk r_{Λ_n} satisfying

$$r_\Lambda \leq r = \lim_{n\to\infty} r_{\Lambda_n}$$

and δ is an estimator such that

$$\sup_\theta R(\theta, \delta) = r$$

Then

- *δ is minimax;*
- *$\{\Lambda_n\}$ is least favorable.*

Proof

1. Suppose δ' is any other estimator, the for every n

$$\sup_\theta R(\theta, \delta') \geq \int R(\theta, \delta') d\Lambda_n(\theta) \geq r_{\Lambda_n} \geq r_\Lambda = \sup_\theta R(\theta, \delta)$$

   Thus δ is minimax.

2. If Λ is any distribution, and any δ, then

$$r_\Lambda = \int R(\theta, \delta_\Lambda) d\Lambda(\theta) \leq \int R(\theta, \delta) d\Lambda(\theta) \leq \sup_\theta R(\theta, \delta) = r$$

□

Lemma 5.1.5. *(Lemma 5.1.13 in TPE)*
If δ_Λ is the Bayes estimator of $g(\theta)$ with respect to Λ and if

$$r_\Lambda = \mathbb{E}_\Lambda \left[\mathbb{E}_\theta \left[(\delta_\Lambda(\mathbf{X}) - g(\Theta))^2\right]\right]$$

is its Bayes risk, then

$$r_\Lambda = \int \mathbb{E}\left[(\delta_\Lambda(\mathbf{x}) - g(\Theta))^2 | \mathbf{x}\right] d\mathbb{P}(\mathbf{x}) = \int \mathbb{V}\mathrm{ar}[g(\Theta)|\mathbf{x}] d\mathbb{P}(\mathbf{x})$$

In particular, if the posterior variance is independent of $\mathbf{x}$, then

$$r_\Lambda = \mathbb{V}\mathrm{ar}[g(\Theta)|\mathbf{x}]$$

Lemma 5.1.6. *(Lemma 5.1.15 in TPE)*
Let X be a random quantity with distribution F , and let $g(F)$ be a functional defined over a set $\mathcal{F}_1$ of distributions F. Suppose that δ is a minimax estimator of $g(F)$ when F is restricted to some subset $\mathcal{F}_0 \subset \mathcal{F}_1$. Then if

$$\sup_{F\in\mathcal{F}_0} R(F, \delta) = \sup_{F\in\mathcal{F}_1} R(F, \delta)$$

δ is minimax also when F is permitted to vary over $\mathcal{F}_1$

Proof If an estimator δ' existed with smaller sup risk over $\mathcal{F}_1$ than δ, it would also have smaller sup risk over $\mathcal{F}_0$ and thus contradict the minimax property of δ over $\mathcal{F}_0$. $\square$

Example 36: (Estimating Normal Mean, Example 5.1.14 in TPE)

Let $\mathbf{X} = (X_1, \cdots, X_n)$, with $X_i \overset{i.i.d.}{\sim} \mathsf{N}(\theta, \sigma^2)$, under loss squared error $L(\theta, d) = (\theta - d)^2$

1. Suppose that σ^2 is known. By theorem 5.1.4, we shall prove that $\bar{X}$ is minimax by finding a sequence of Bayes estimators δ_b satisfying
$$r_\Lambda \leq r = \lim_{b\to\infty} r_{\Lambda_b}$$
with $r = \frac{\sigma^2}{n}$.
As prior distribution for θ, let us try the conjugate normal distribution $\Lambda_b = \mathsf{N}(\mu, b^2)$. As the result in Example 24, the Bayes estimator is
$$\delta_{\Lambda_b}(x) = \frac{n}{\sigma^2}\left(\frac{n}{\sigma^2} + \frac{1}{b^2}\right)^{-1} \cdot \bar{x} \quad + \quad \frac{1}{b^2}\left(\frac{n}{\sigma^2} + \frac{1}{b^2}\right)^{-1} \cdot \mu$$
The posterior variance, which is independent of $\mathbf{x}$, is given by
$$\mathbb{V}\mathrm{ar}(\Theta|x) = \frac{1}{n/\sigma^2 + 1/b^2}$$
By previous lemma 5.1.5, the risk is
$$r_{\Lambda_b} = \mathbb{E}_{\Lambda_b}\left[\mathbb{E}_\theta\left[(\delta_\Lambda(X) - g(\Theta))^2\right]\right] = \frac{1}{n/\sigma^2 + 1/b^2}$$
As $b \to \infty$, $r_{\Lambda_b} \uparrow \frac{\sigma^2}{n}$, which is a constant. And $\sup_\theta R(\theta, \delta) = \frac{\sigma^2}{n}$, by theorem 5.1.4, $\bar{X}$ is minimax.

2. Suppose that σ^2 is unknown. The maximum risk of every estimator will be infinite unless σ^2 is bounded. We shall assume that $\sigma^2 \leq M$. Under this restriction, the maximum risk of $\bar{X}$ is
$$\sup_{\theta,\sigma^2} \mathbb{E}\left[(\bar{X} - \theta)^2\right] \leq \frac{M}{n}$$
Then $\bar{X}$ is minimax when $\sigma^2 = M$. Since $\theta \in \mathbb{R}$, $\sigma^2 \in [0, M]$, let $\mathcal{F}_0 = \mathbb{R} \times \{M\}$, $\mathcal{F}_1 = \mathbb{R} \times [0, M]$, where $\mathcal{F}_0 \subset \mathcal{F}_1$. Also,
$$\sup_{F\in\mathcal{F}_0} R((\theta, \sigma^2), \delta) = \sup_{\theta,\sigma^2} \mathbb{E}\left[(\bar{X} - \theta)^2\right] = \frac{M}{n} = \sup_{F\in\mathcal{F}_1} R((\theta, \sigma^2), \delta)$$
By previous lemma 5.1.6 , $\bar{X}$ is minimax for $\sigma^2 \in [0, M]$.

♠

Example 37: (Estimating Nonparametric Mean, Example 5.1.16 in TPE)

$X_1, \cdots, X_n \overset{i.i.d}{\sim} F$ with finite expectation θ, while F is the distribution on $\mathbb{R}$ with finite second moment. Consider the problem of estimating $\theta = mean(F)$ under squared error loss $L(\theta, d) = (d - \theta)^2$. If the maximum risk of every estimator of θ is infinite, the minimax problem is meaningless. To rule this out, we shall consider two possible restrictions on F:

1. Bounded variance, $\mathbb{V}\mathrm{ar}_F(X_i) \leq M < \infty$, while M is known.

   We are going to show that $\bar{X}$ is minimax. As what we have done when estimating normal mean, the maximum risk of $\bar{X}$ is

$$\sup_{\theta,\sigma^2} \mathbb{E}\left[(\bar{X}-\theta)^2\right] \leq \frac{M}{n}$$

   Thus $\bar{X}$ is minimax when $\sigma^2 = M$. While $\theta \in \mathbb{R}$, $\sigma^2 \in [0, M]$, let $\mathcal{F}_0 = \mathbb{R} \times \{M\}$, $\mathcal{F}_1 = \mathbb{R} \times [0, M]$, where $\mathcal{F}_0 \subset \mathcal{F}_1$, and we have

$$\sup_{F\in\mathcal{F}_0} R((\theta,\sigma^2),\delta) = \sup_{\theta,\sigma^2} R((\theta,\sigma^2),\delta) = \frac{M}{n} = \sup_{F\in\mathcal{F}_1} R((\theta,\sigma^2),\delta)$$

   By lemma 5.1.6, $\bar{X}$ is a minimax estimator when the variance is bounded.

2. Bounded range, $-\infty < a \leq X_i \leq b < \infty$, *i.e.*, X_i has compact support on $[a, b]$.

   Suppose that without loss of generality, let $a = 1$, $b = 0$. Let $\mathcal{F}_1 = \{F : F(1) - F(0) = 1\}$, where F is the distribution function, *i.e.*, CDF. For a special case, we consider $\mathcal{F}_0 = \{F : F = Bernoulli(n,\theta), 0 < \theta < 1\}$, where $\mathcal{F}_0 \subset \mathcal{F}_1$. Let $Y = \sum_{i=1}^n X_i$, then $Y \sim Binomial(n,\theta)$, following the example 34, this suggests the minimax estimator

$$\delta_\Lambda(Y) = \frac{\sqrt{n}}{\sqrt{n}+1}\frac{Y}{n} + \frac{1}{2}\frac{1}{\sqrt{n}+1}$$

   with risk $R(\theta, \delta_\Lambda(Y)) = \frac{1}{4(\sqrt{n}+1)^2}$. Then we shall now prove that $\delta_\Lambda(Y)$ is, indeed, a minimax estimator of θ over $\mathcal{F}_1$, which by lemma 5.1.6, enough to prove that the risk function of the estimator $\delta_\Lambda(Y)$ takes on its maximum over $\mathcal{F}_1$.

   For any distribution $F_1 \in \mathcal{F}_1$, we have $\mathbb{E}_{\theta_1}[Y] = n\theta_1$, and since $0 \leq X_1 \leq 1$, $(X_1)^2 \leq X_1$, then

$$\mathbb{V}\mathrm{ar}_{\theta_1}(Y) \overset{i.i.d.}{=} n\mathbb{V}\mathrm{ar}_{\theta_1}(X_1) = n\mathbb{E}_{\theta_1}\left[(X_1)^2\right] - n(\mathbb{E}_{\theta_1}[X_1])^2 \leq n\mathbb{E}_{\theta_1}[X_1] - n\theta_1^2 = n\theta_1 - n\theta_1^2$$

   Then we write the risk function as

$$\begin{aligned}
R(F_1, \delta_\Lambda(Y)) &= \mathbb{E}\left[\frac{\sqrt{n}}{\sqrt{n}+1}\frac{Y}{n} + \frac{1}{2}\frac{1}{\sqrt{n}+1} - \theta_1\right]^2 \\
&= \frac{1}{n(\sqrt{n}+1)^2}\left[\mathbb{V}\mathrm{ar}_{\theta_1}(Y) + n(\frac{1}{2}-\theta_1)^2\right] \\
&\leq \frac{1}{(\sqrt{n}+1)^2}\left[\theta_1 - \theta_1^2 + \theta_1^2 - \theta_1 + \frac{1}{4}\right] = \frac{1}{4(\sqrt{n}+1)^2}
\end{aligned}$$

   Thus we have

$$\sup_{F\in\mathcal{F}_0} R((\theta,\sigma^2),\delta) = \sup_{\theta,\sigma^2} R((\theta,\sigma^2),\delta) = \frac{M}{n} = \sup_{F\in\mathcal{F}_1} R((\theta,\sigma^2),\delta)$$

   By lemma 5.1.6, $\bar{X}$ is a minimax estimator over $\mathcal{F}_1 = \{F : F(1) - F(0) = 1\}$.

♠

5.2 Admissibility and Minimaxity in Exponential Families

Definition 5.2.1. *(Dominate)*
An estimator δ_1 dominates another estimator δ_2 if

$$\begin{aligned} \forall\theta \in \Omega, \qquad & R(\theta, \delta_1) \le R(\theta, \delta_2) \\ \exists\theta \in \Omega, s.t. \qquad & R(\theta, \delta_1) < R(\theta, \delta_2) \end{aligned}$$

If δ_1 dominates δ_2, then δ_2 would be eliminated from consideration.

If an estimator δ is not dominated by any estimator, then we may want to at least consider it

Definition 5.2.2. *(Admissibility)*
An estimator δ is admissible if it is not dominated.

Lemma 5.2.1. *(Lemma 5.2.1 in TPE)*
Let the range of the estimand $g(\theta)$ be an interval with endpoints a and b, and suppose that the loss function $L(\theta, d)$ is positive when $d \neq g(\theta)$ and zero when $d = g(\theta)$, and that for any fixed θ, $L(\theta, d)$ is increasing as d moves away from $g(\theta)$ in either direction. Then, any estimator δ taking on values outside the closed interval $[a, b]$ with positive probability is inadmissible.

Proof δ is dominated by the estimator δ', which is $\delta' = a$ or $\delta' = b$ when $\delta < a$ or $\delta > a$, and which otherwise is equal to δ. □

Theorem 5.2.2. *(Theorem 5.2.4 in TPE)*
Any unique Bayes estimator is admissible.

Proof If δ_Λ is unique Bayes with respect to the prior distribution Λ, then we have

$$\delta_\Lambda = \operatorname*{argmin}_{\delta} \int R(\theta, \delta(x)) d\Lambda(\theta)$$

If δ_Λ is dominated by δ', then

$$\int R(\theta, \delta') d\Lambda(\theta) \le \int R(\theta, \delta_\Lambda) d\Lambda(\theta)$$

From definition of Bayes estimator, the previous inequality must take equality. However, if $\delta' \neq \delta_\Lambda$ *a.s.*, it will contradict to the uniqueness of δ_Λ. Therefore, $\delta' = \delta_\Lambda$ *a.s.* and δ_Λ is admissible. □

Theorem 5.2.3. *(Theorem 5.2.6 in TPE)*
Let X be a random variable with mean θ and variance σ^2. Then, $aX + b$ is an inadmissible estimator of θ under squared error loss whenever

1. *$a > 1$, or*

2. $a < 0$, *or*

3. $a = 1$ *and* $b \neq 0$

Proof We write the risk of $aX + b$ as a function of a and b,

$$\rho(a, b) = \mathbb{E}\left[(aX + b - \theta)^2\right] = a^2\sigma^2 + [(a-1)\theta + b]^2$$

1. If $a > 1$, then

$$\rho(a, b) = a^2\sigma^2 + [(a-1)\theta + b]^2 \geq a^2\sigma^2 > \sigma^2 = \rho(1, 0)$$

so that $aX + b$ is dominated by X

2. If $a < 0$, then $(a-1)^2 > 1$ and hence

$$\rho(a, b) = a^2\sigma^2 + [(a-1)\theta + b]^2 \geq [(a-1)\theta + b]^2 > (a-1)^2[\theta + \frac{b}{a-1}]^2 > \rho(0, -\frac{b}{a-1})$$

so that $aX + b$ is dominated by the constant estimator $\delta = -\frac{b}{a-1}$

3. If $a = 1$ and $b \neq 0$, then $aX + b = X + b$

$$R(\theta, X + b) = \mathbb{E}\left[(X + b - \theta)^2\right] = \sigma^2 + b^2 > \sigma^2 = \mathbb{E}\left[(X - \theta)^2\right] = R(\theta, X)$$

so that $X + b$ is dominated by X.

□

Example 38: (Admissibility of linear estimators, Example 5.2.7 in TPE)
Let $X_1, \cdots, X_n$ be independent, each distributed according to a $\mathsf{N}(\theta, \sigma^2)$, with σ^2 known. Combining the results of 5.2.3, we see that the estimator $a\bar{X} + b$ is admissible in the strip $0 < a < 1$ in the (a, b) plane, that it is inadmissible to the left $(a < 0)$ and to the right $(a > 1)$.
The left boundary $a = 0$ corresponds to the constant estimators $\delta = b$ which are admissible since $\delta = b$ is the only estimator with zero risk at $\theta = b$. Finally, the right boundary $a = 1$ is inadmissible by (iii) of 5.2.3, with the possible exception of the point $a = 1, b = 0$. ♠

We have thus settled the admissibility of $a\bar{X} + b$ for all cases except $\bar{X}$ itself, which was the estimator of primary interest. In the next example, we shall prove that $\bar{X}$ is indeed admissible.

Example 39: (Admissibility of $\bar{X}$, Example 5.2.8 in TPE)
Let $X_1, \cdots, X_n$ be independent, each distributed according to a $\mathsf{N}(\theta, \sigma^2)$, with σ^2 known. In the location model, $\bar{X}$ is admissible. ♠

(First proof by Blyth's method)

Proof Assume $\sigma^2 = 1$, then $R(\theta, \bar{X}) = \frac{\sigma^2}{n} = \frac{1}{n}$. Suppose that $\bar{X}$ is not admissible, thus there is another estimator δ dominates $\bar{X}$, *i.e.*

$$\begin{aligned} \forall \theta \in \Omega, \quad R(\theta, \delta) \leq \frac{1}{n} = R(\theta, \bar{X}) \\ \exists \theta \in \Omega, \quad R(\theta, \delta) < \frac{1}{n} = R(\theta, \bar{X}) \end{aligned}$$

Now, $R(\theta, \delta)$ is a continuous function of θ for every δ so that there exists $\epsilon > 0$ and $\theta_0 < \theta_1$ such that

$$R(\theta, \delta) < \frac{1}{n} - \epsilon, \quad \theta \in (\theta_0, \theta_1)$$

Let r_τ^* be the average risk of δ with respect to the prior distribution $\Lambda_\tau = \mathsf{N}(0, \tau^2)$, and let r_τ be the Bayes risk, that is, the average risk of the Bayes solution with respect to Λ_τ. Then, by Example 36 with $\sigma = 1$ and τ in place of b,

$$\begin{aligned} r_\tau^* &= \frac{1}{\sqrt{2\pi\tau^2}} \int_{-\infty}^{\infty} R(\theta, \delta) \exp\left(-\frac{\theta^2}{2\tau^2}\right) d\theta = \frac{1}{n} - \frac{\epsilon}{\sqrt{2\pi\tau^2}} \int_{\theta_0}^{\theta_1} \exp\left(-\frac{\theta^2}{2\tau^2}\right) d\theta \\ r_\tau &= \frac{\tau^2}{1 + n\tau^2} = \frac{1}{n} - \frac{1}{n(1 + n\tau^2)} \end{aligned}$$

and

$$\begin{aligned} \frac{\frac{1}{n} - r_\tau^*}{\frac{1}{n} - r_\tau} &= \frac{\frac{1}{\sqrt{2\pi\tau^2}} \int_{-\infty}^{\infty} \left[\frac{1}{n} - R(\theta, \delta)\right] \exp\left(-\frac{\theta^2}{2\tau^2}\right) d\theta}{\frac{1}{n} - \frac{\tau^2}{1+n\tau^2}} \\ &\geq \frac{n(1 + n\tau^2)\epsilon}{\sqrt{2\pi\tau^2}} \int_{\theta_0}^{\theta_1} \exp\left(-\frac{\theta^2}{2\tau^2}\right) d\theta > 0 \\ &\overset{\tau \to \infty}{\rightarrow} \infty \end{aligned}$$

Thus, there exists some τ_0 which is large enough, so that

$$\frac{\frac{1}{n} - r_{\tau_0}^*}{\frac{1}{n} - r_{\tau_0}} > 1 \Longleftrightarrow r_{\tau_0} > r_{\tau_0}^*$$

which contracts to the fact that r_{τ_0} is Bayes risk of Λ_{τ_0}, *i.e.*, $r_{\tau_0} = \min \int_\Omega R(\theta, \delta_{\tau_0}) d\Lambda_{\tau_0}(\theta)$. Therefore, $\bar{X}$ is admissible. □

(Second proof by the Information Inequality Method)

Proof Another useful tool for establishing admissibility is based on the information inequality and solutions to a differential inequality, a method due to Hodges and Lehmann (1951). It follows from the information inequality

$$R(\theta, \delta) \geq \frac{[1 + b'(\theta)]^2}{nI(\theta)} + b^2(\theta)$$

and the fact that

$$R(\theta, \delta) = \mathbb{E}\left[(\delta - \theta)^2\right] = \mathbb{V}\mathrm{ar}_\theta(\delta) + b^2(\theta)$$

where $b(\theta)$ is the bias of δ, and the first term on the right is the information inequality variance bound for estimators with expected value $\theta + b(\theta)$. Note that, in the present case with $\sigma^2 = 1$, $I(\theta) = 1$. Suppose, now, that δ is any estimator satisfying $R(\theta, \delta) \leq \frac{1}{n}$ for all θ, and hence

$$\frac{[1 + b'(\theta)]^2}{nI(\theta)} + b^2(\theta) \leq R(\theta, \delta) \leq \frac{1}{n}$$

We shall then show that $b(\theta) \equiv 0$.

1. Since $b^2(\theta) \leq \frac{1}{n}$ and $|b(\theta)| \leq \frac{1}{\sqrt{n}}$, the function b is bounded.

2. From the fact that $\frac{[1+b'(\theta)]^2}{nI(\theta)} \leq \frac{1}{n}$, it follows that $b'(\theta) < 0$, so that b is nonincreasing.

3. We shall show, next, that there exists a sequence of values θ_i tending to ∞ and such that $b'(\theta_i) \to 0$. Suppose that $b'(\theta)$ were bounded away from 0 as $\theta \to \infty$, say $b'(\theta) \leq -\epsilon$ for all $\theta > \theta_0$. Then $b(\theta)$ cannot be bounded as $\theta \to \infty$, which contradicts (1).

4. Analogously, it is seen that there exists a sequence of values $\theta_i \to -\infty$ and such that $b'(\theta_i) \to 0$.

From (2), $b(\theta) = b'(\theta) = 0$, thus $R(\theta, \delta) \geq \frac{1}{n}$, and hence that $R(\theta, \delta) = \frac{1}{n} = R(\theta, \bar{X})$. Therefore, $\bar{X}$ is admissible and minimax. □

Example 40: (Truncated normal mean, Example 5.2.9 in TPE)

In Example 39, suppose it is known that $\theta > \theta_0$. By lemma 5.2.1, $\bar{X}$ is no longer admissible, since $\bar{X}$ may take value beside θ_0, it will be dominated by θ_0 when $\bar{X} < \theta_0$.

However, assuming that $\sigma^2 = 1$ and using the method of the second proof of Example 39, it is easy to show that $\bar{X}$ continues to be minimax. If it were not, there would exist an estimator δ and an $\epsilon > 0$ such that

$$R(\theta, \delta) \leq \frac{1}{n} - \epsilon, \quad \forall \theta > \theta_0$$

and hence

$$\frac{[1 + b'(\theta)]^2}{n} + b^2(\theta) \leq \frac{1}{n} - \epsilon \quad \forall \theta > \theta_0$$

where $b(\theta)$ is the bias of δ, with

$$R(\theta, \delta) = \mathbb{E}\left[(\delta - \theta)^2\right] = \mathbb{V}\mathrm{ar}_\theta(\delta) + b^2(\theta)$$

and

$$R(\theta, \delta) \geq \frac{[1 + b'(\theta)]^2}{nI(\theta)} + b^2(\theta)$$

As a result, $|b(\theta)| < \frac{1}{\sqrt{n}}$ and $1 + 2b'(\theta) + [b'(\theta)]^2 < 1 - \epsilon n$, which indicates that $b'(\theta) < -\frac{\epsilon n}{2}$. However, this contradicts to the fact $|b(\theta)| < \frac{1}{\sqrt{n}}$. Hence δ is minimax.

If θ is further restricted to $\theta \in [a, b]$, $\bar{X}$ is not only inadmissible but also no longer minimax. If $\bar{X}$ were minimax, the same would be true of its provement, MLE

$$\delta^*(X) = \begin{cases} a & , \bar{X} < a \\ \bar{X} & , a \leq \bar{X} \leq b \\ b & , \bar{X} > b \end{cases}$$

so that

$$\sup_{a \leq \theta \leq b} R(\theta, \delta^*(X)) = R(\theta, \bar{X}) = \frac{1}{n}$$

However, $R(\theta, \delta^*(X)) < R(\theta, \bar{X})$ for all $a \leq \theta \leq b$ and $R(\theta, \delta^*(X))$ is a continuous function of θ and hence takes on its maximum at some point $a \leq \theta_0 \leq b$, with $\sup_{a \leq \theta \leq b} R(\theta, \delta^*(X)) = R(\theta_0, \delta^*(X)) < R(\theta, \bar{X})$, which contradicts to previous equality. ♠

Example 41: (Linear minimax risk, Example 5.2.9 in TPE, continuing previous example)

Suppose that we want to focus on linear estimators, $\delta^{(a,b)} = a\bar{X} + b$, with $\sigma^2 = 1$. Under squared error loss,

$$R(\theta, a\bar{X} + b) = \mathbb{E}_\theta\left[(a\bar{X} + b - \theta)^2\right] = a^2 \mathbb{V}\mathrm{ar}_\theta(\bar{X}) + [(a-1)\theta + b]^2 = \frac{a^2}{n} + [(a-1)\theta + b]^2$$

By theorem 5.2.3, we only need to consider the case $0 \leq a \leq 1$. It is straightforward to establish

$$\max_{\theta \in [-m,m]} R(\theta, a\bar{X} + b) = \max\{R(-m, a\bar{X} + b), R(m, a\bar{X} + b)\}$$

and that $\delta^* = a^*\bar{X}$ with $a^* = \frac{m^2}{\frac{1}{n} + m^2}$ ♠

Theorem 5.2.4. *(Karlin's Theorem, Theorem 5.2.14 in TPE) Let X have probability density*

$$p_\theta(x) = \beta(\theta)e^{\theta T(x)} \quad (\theta, T(x) real - valued)$$

with respect to μ and let Ω be the natural parameter space. Then, Ω is an interval, with endpoints, say, θ_1 and θ_2 $(-\infty \leq \theta_1 \leq \theta \leq \theta_2 \leq \infty)$. For estimating $\mathbb{E}_\theta[T]$, the estimator $aT + b$ is inadmissible if $a < 0$ or $a > 1$ and is a constant for $a = 0$. It is convenient to write the estimator as

$$\delta_{\lambda,\gamma}(x) = \frac{1}{1+\lambda}T + \frac{\gamma\lambda}{1+\lambda}$$

with $0 \leq \lambda < \infty$ corresponding to $0 < a \leq 1$

Under the above assumptions, a sufficient condition for the admissibility of the estimator for estimating $g(\theta) = \mathbb{E}_\theta[T]$ with squared error loss is that the integral of $e^{-\gamma\lambda\theta}[\beta(\theta)]^{-\lambda}$ diverges at θ_1 and θ_2; that is, that for some (and hence for all) $\theta_1 < \theta_0 < \theta_2$, the two integrals

$$\int_{\theta_0}^{\theta^*} \frac{e^{-\gamma\lambda\theta}}{[\beta(\theta)^\lambda]} d\theta \qquad \int_{\theta^*}^{\theta_0} \frac{e^{-\gamma\lambda\theta}}{[\beta(\theta)^\lambda]} d\theta$$

tend to infinity as θ^* tends to θ_2 and θ_1, respectively.

Example 42: (Binomial, Example 5.2.17 in TPE)

♠

Proof

□

Corollary 5.2.5. *(Corollary 5.2.18 in TPE)*
If the natural parameter space of $p_\theta(x) = \beta(\theta)e^{\theta T(x)}$ *($\theta, T(x)$ real-valued) is the whole real line so that* $\theta_1 = -\infty$, $\theta_2 = \infty$, *then* $T(x)$ *is admissible for estimating* $\mathbb{E}_\theta[T]$ *with squared error loss.*

Proof With $\lambda = 0$ and $\gamma = 1$, the two integrals

$$\int_{\theta_0}^{\theta^*} \frac{e^{-\gamma\lambda\theta}}{[\beta(\theta)^\lambda]} d\theta \qquad \int_{\theta^*}^{\theta_0} \frac{e^{-\gamma\lambda\theta}}{[\beta(\theta)^\lambda]} d\theta$$

clearly tend toward infinity as $\theta \to \pm\infty$ □

Lemma 5.2.6. *(Lemma 5.2.19 in TPE)*
If an estimator has constant risk and is admissible, it is minimax.

Proof If it were not minimax, another estimator would have smaller maximum risk and, hence, uniformly smaller risk □

Corollary 5.2.7. *(Corollary 5.2.20 in TPE)*
Under the assumptions of Corollary 5.2.5, T *is the unique minimax estimator of* $g(\theta) = \mathbb{E}_\theta[T]$ *for the loss function* $[d - g(\theta)]^2/\mathbb{V}\mathrm{ar}_\theta(T)$.

Proof For this loss function, T is a constant risk estimator which is admissible by corollary 5.2.5. It is unique since the loss function is strictly convex in d. □

Lemma 5.2.8. *(Lemma 5.2.21 in TPE) If an estimator is unique minimax, it is admissible.*

Proof If it were not admissible, another estimator would dominate it in risk and, hence, would be minimax. □

Example 43: (Two binomials, Example 5.2.23 in TPE)

♠

5.3 Admissibility and Minimaxity in Group Families

Let us now consider the corresponding problems for group families. In these families there typically exists an MRE estimator δ_0 for any invariant loss function, and it is a constant risk estimator. If δ_0 is also a Bayes estimator, it is minimax by Corollary 5.1.2 and admissible if it is unique Bayes.

Recall Theorem 4.4.1, where it was shown that a Bayes estimator under an invariant prior is (almost) equivariant. It follows that under the assumptions of that theorem, there exists an almost equivariant estimator which is admissible. Theorem 4.4.1 does not require $\bar{G}$ to be transitive over Ω. If we add the assumption of transitivity, we get a stronger result.

Theorem 5.3.1. *(Theorem 5.3.1 in TPE)*
Under the conditions of Theorem 4.4.1, if $\bar{G}$ is transitive over Ω, then the MRE estimator is admissible and minimax.

Theorem 5.3.2. *(Theorem 5.3.5 in TPE)*
Suppose $X = (X_1, \cdots, X_n)$ is distributed according to the density

$$f(\boldsymbol{x} - \theta) = f(x_1 - \theta, \cdots, x_n - \theta)$$

and that the Pitman estimator δ^*, where δ^* is the Bayes estimator of $g(\theta)$ with respect to Λ, and the variance given by lemma 5.1.5

$$r_\Lambda = \mathbb{V}\mathrm{ar}[g(\Theta)|\boldsymbol{x}]$$

is finite. Then, δ^* is minimax under squared error loss.

Chapter 6

Uniformly Most Powerful Tests

6.1 The Problem

6.1.1 Hypothesis test

The statistical model is

$$X \sim \mathbb{P}_\theta, \quad \theta \in \Omega, \quad \text{where } X \text{ takes values } X = x \in \mathcal{X}$$

The distribution $\mathbb{P}_\theta$ of X is belong to the class $\mathcal{P} = \{\mathbb{P}_{\theta'} : \theta' \in \Omega\}$. Divide the class $\mathcal{P}$ and the parameter space Ω into two pieces

- H: class of hypothesis; K: class of alternatives
- $H \subset \mathcal{P}$, corresponding with $\Omega_H \subset \Omega$, where the hypothesis is true.
- $K \subset \mathcal{P}$, corresponding with $\Omega_H \subset \Omega$, where the hypothesis is false.
- $H \cup K = \mathcal{P}$, $H \cap K = \emptyset$, $\Omega_H \cup \Omega_K = \Omega$, $\Omega_H \cap \Omega_K = \emptyset$.

Mathematically, the hypothesis H, is equivalent to the statement that $\mathbb{P}_\theta \in H$, $\theta \in \Omega_H$.

Our goal is to deduce where $\theta \in \Omega_H$ vs $\theta \in \Omega_K$ based on the value of X. A decision procedure, which aims at deciding whether to accept or reject the hypothesis, is called a test of the hypothesis.

Let d_0 denote the decision accepting hypothesis and d_1 denote the decision rejecting hypothesis. A nonrandomized test procedure is to assign the decision, d_0 or d_1, to each possible value x of X, and thereby divides the sample space into two complementary regions $\mathcal{X}_0$ and $\mathcal{X}_1$,

- $X \in \mathcal{X}_0$: $\delta(X) = d_0$, accept hypothesis H. $\mathcal{X}_0$ is called the region of acceptance.
- $X \in \mathcal{X}_1$: $\delta(X) = d_1$, reject hypothesis H. $\mathcal{X}_1$ is called the region of rejection, or critical region.

Definition 6.1.1. *(Test)*

If we observe $X = x$, a random experiment is performed with two possible outcomes R and $\bar{R}$. A measurable function $\phi : \mathcal{X} \to [0, 1]$ with the following property is called the critical function of the test

- $\phi(x) = \mathbb{P}(R) =$ *probability of rejection*

- $1 - \phi(x) = \mathbb{P}(\bar{R})$ = *probability of acceptance*
- $0 \le \phi(x) \le 1$ *for all* x

If $\phi(x)$ takes on only the values 1 and 0, one is back in the case of a nonrandomized test. The set of points x for which $\phi(x) = 1$ is then just the region of rejection, so that in a nonrandomized test $\phi(x)$ is simply the indicator function of the critical region.
Example 44: Let $U \sim Uniform[0,1]$, which is independent of X. We define

$$\bar{\phi} : \mathcal{X} \times [0,1] \to [0,1], \quad \text{where } \bar{\phi}(x,u) = \mathbf{1}_{\{u \le \phi(x)\}}$$

We deduce Ω_K if we observe $X = x$, $U = u$ with $u \le \phi(x)$. $\bar{\phi}(x,u)$ is the indicator function of the critical region. ♠

6.1.2 Two kinds of error

When performing the test, there are two types of errors

- 1st kind: rejecting the hypothesis H when it is true
- 2nd kind: accepting the hypothesis H when it is false

It desirable to keep both probabilities of two errors to a minimum. However, it is not possible to control these two simultaneously. Therefore, we assign a bound to the probability of incorrectly rejecting H when it is true H(1st kind error), and to attempt to minimize the probability of incorrectly accepting the hypothesis H (2nd kind error). Therefore, we introduce the concept of level of significance

Definition 6.1.2. *(Level of Significance, Size of Test)*
Level of significance $\alpha \in [0,1]$ *is largest probability that a test incorrectly rejecting* H*, which satisfying*

$$\mathbb{P}_\theta[\delta(X) = d_1] = \mathbb{P}_\theta[X \in \mathcal{X}_1] \le \alpha, \quad \forall \theta \in \Omega_H \tag{6.1}$$

The LHS of equation (6.1)

$$\sup_{\theta \in \Omega_H} \mathbb{P}_\theta[X \in \mathcal{X}_1]$$

is called the size of the test. Under the nonrandomized context, with $\phi(x) \in \{0,1\}$*, the size of test can be written as*

$$\sup_{\theta \in \Omega_H} \mathbb{P}_\theta[X \in \mathcal{X}_1] = \sup_{\theta \in \Omega_H} \mathbb{E}_\theta[\phi(X)] = \sup_{\theta \in \Omega_H} \mathbb{P}_\theta(\phi(X) = 1)$$

A low significance level results in the hypothesis being rejected only for a set of values of the observations whose total probability under hypothesis is small, so that such values would be most unlikely to occur if H were true.

While controlling the level of significance, we attempt to minimize the probability of incorrectly accepting the hypothesis H (2nd kind error), which is $\mathbb{P}_\theta(\delta(X) = d_0)$, $\forall \theta \in \Omega_K$, or equivalently, to maximize

$$\mathbb{P}_\theta[\delta(X) = d_1] = \mathbb{P}_\theta[X \in \mathcal{X}_1], \quad \forall \theta \in \Omega_K$$

Definition 6.1.3. *(Power Function of Test)*

Restricting consideration to test whose size does not exceed the given level of significance α, the probability of rejection $\mathbb{P}_\theta[X \in \mathcal{X}_1]$, $\forall\theta \in \Omega_K$ evaluated for a given θ in Ω_K is called the power of the test against the alternative θ.

Considered as a function of θ,

$$\beta(\theta) = \mathbb{P}_\theta[\delta(X) = d_1] = \mathbb{P}_\theta[X \in \mathcal{X}_1] \overset{\textit{non-randomized}}{=} \mathbb{E}_\theta\left[\phi(X)\right] = \mathbb{P}_\theta(\phi(X) = 1), \quad \forall\theta \in \Omega_K$$

is called the power function of the test

Under the nonrandomized context, if the distribution of X is $\mathbb{P}_\theta$, with critical function $\phi(x) \in \{0, 1\}$, the probability of rejection is

$$\mathbb{E}_\theta\left[\phi(X)\right] = \int \phi(x) d\mathbb{P}_\theta(x)$$

which is the conditional probability $\phi(x)$ of rejection given x, integrated with respect to the probability distribution $\mathbb{P}_\theta$ of X.

Mathematically speaking, the problem is to select $\phi(x)$ to maximize the power function

$$\beta_\phi(\theta) = \mathbb{E}_\theta\left[\phi(X)\right], \quad \forall\theta \in \Omega_K$$

subject to the condition

$$\sup_{\theta\in\Omega_H} \mathbb{E}_\theta\left[\phi(X)\right] = \alpha$$

Most Powerful(MP) Test at level α: A test $\phi(x)$, which has size $\sup_{\theta\in\Omega_H} \mathbb{E}_\theta\left[\phi(X)\right] = \alpha$, maximizes $\beta_\phi(\theta) = \mathbb{E}_\theta\left[\phi(X)\right]$ for $\theta \in \Omega_K$ among test $\phi'(x)$ with size α

6.2 The Neyman–Pearson Fundamental Lemma

A class of distributions is called simple if it contains a single distribution, and otherwise it is said to be composite. We now consider the case $\Omega_H = \{\theta_0\}$ and $\Omega_K = \{\theta_1\}$ are singletons.

Under a simple hypothesis H and alternative K, let the distributions be $\mathbb{P}_0$ and $\mathbb{P}_1$, corresponding with density f_0 and f_1 with respect to a measure μ, which can be taken as $\mu = \mathbb{P}_0 + \mathbb{P}_1$.

The following question is which kind points $x \in \mathcal{X}$ should be included in the critical region $\mathcal{X}_1$, to maximize $\sum_{x\in\mathcal{X}_1} \mathbb{P}_1(X = x)$ under the restriction $\sum_{x\in\mathcal{X}_1} \mathbb{P}_0(X = x) \leq \alpha$? An economic way is to include the points with highest value of

$$r(x) = \frac{\mathbb{P}_1(x)}{\mathbb{P}_0(x)}$$

The points are therefore ordered according to the value of $r(x)$ and selected for $\mathcal{X}_1$ in this order, as many as one can afford under restriction $\sum_{x\in\mathcal{X}_1} \mathbb{P}_0(X = x) \leq \alpha$. Formally this means that $\mathcal{X}_1$ is the set of all points x for which $r(x) > c$, where c is a constant determined by the condition

$$\mathbb{P}_0(X \in \mathcal{X}_1) = \sum_x \mathbf{1}_{\{r(x)>c\}} \mathbb{P}_0(X = x)$$

However, there is a problem arising. Sometimes α can not be achieved exactly. It is possible that $\mathbb{P}_0(X \in \mathcal{X}_1) < \alpha$ before including x_0 into $\mathcal{X}_1$, but $\mathbb{P}_0(X \in \mathcal{X}_1) > \alpha$ after including x_0 into $\mathcal{X}_1$. The common solution, instead, is to adopt a value of α that can be attained exactly and therefore does not present this problem, which does lead to a simple explicit solution neither the breaking of the $r(x)$-order nor randomization. These considerations are formalized in the following theorem, the fundamental lemma of Neyman and Pearson.

Theorem 6.2.1. *(Fundamental lemma of Neyman and Pearson, theorem 3.2.1 in TSH) Let $\mathbb{P}_0$ and $\mathbb{P}_1$ be the distributions, corresponding with density f_0 and f_1 with respect to a measure μ under hypothesis H and alternative K.*

1. *Existence. For testing $H : f_0$ against the alternative $K : f_1$, at level α, there exists $\phi(x)$ and a constant c such that*

$$\mathbb{E}_0\left[\phi(X)\right] = \alpha \tag{6.2}$$

and

$$\phi(x) = \begin{cases} 1, & \dfrac{f_1(x)}{f_0(x)} > c \\ 0, & \dfrac{f_1(x)}{f_0(x)} < c \end{cases} \tag{6.3}$$

2. *Sufficient condition for a most powerful test. If a test $\phi(x)$ satisfies equation (6.2) and (6.3) for some constant c, then this $\phi(x)$ is most powerful for testing $f_0(x)$ again $f_1(x)$ at level α.*

3. *Necessary condition for a most powerful test. If $\phi(x)$ is most powerful for testing $f_0(x)$ again $f_1(x)$ at level α, then $\phi(x)$ satisfies (6.3) $\mu-a.s.$ for some constant c. $\phi(x)$ also satisfies (6.2) unless there exists a test $\phi'(x)$ of size $< \alpha$ with power $\beta(\theta) = 1$.*

Proof

□

Corollary 6.2.2. *(Corollary 3.2.1 in TSH)*
Let $\beta_M(\theta)$ denote the power of the most powerful level-α test $(0 < \alpha < 1)$ $\phi_M(x)$ for testing $\mathbb{P}_0$ against $\mathbb{P}_1$. Then $\alpha < \beta$ unless $\mathbb{P}_0 = \mathbb{P}_1$.

Proof Consider the test $\phi(x) \equiv \alpha$, then $\beta_\phi(\theta) = \mathbb{E}_\theta\left[\phi(X)\right] = \alpha$. We know $\phi_M(x)$ is most powerful, then $\beta_M(\theta) \geq \beta_\phi(\theta) = \alpha$. If $\alpha = \beta_M(\theta)$, then by theorem 6.2.1, $\phi_M(x)$ has to be in the form of

$$\phi(x) = \begin{cases} 1, & L > c \\ \gamma, & L = c \\ 0, & L < c \end{cases}$$

where $\mathbb{E}_0\left[\phi(X)\right] = \mathbb{P}_0(L > c) + \gamma\mathbb{P}_0(L = c) = \alpha$. However, to satisfy $\beta_M(\theta) = \alpha$,$\mathbb{P}_0(L > c) = \mathbb{P}_0(L < c) = 0$, which means $f_0 = f_1$ a.s., which gives us $\mathbb{P}_0 = \mathbb{P}_1$ a.s. □

Example 45: (Example 3.2.1 in TSH)

Suppose $X \sim \mathsf{N}(\xi, \sigma^2)$, with σ^2 known. The null hypothesis $H : \xi = 0$ and the alternative specifies $K : \xi = \xi_1$ at level α.

- Suppose that $\xi_1 > \xi_0 = 0$, then the likelihood ratio is given by

$$L(x) = \frac{f_1(x)}{f_0(x)} = \frac{\exp[-\frac{1}{2\sigma^2}(x-\xi_1)^2]}{\exp[-\frac{1}{2\sigma^2}x^2]} = \exp\left[\frac{\xi_1 x}{\sigma^2} - \frac{\xi_1^2}{2\sigma^2}\right]$$

According to theorem 6.2.1, a most powerful test $\phi(x)$ at level α is of the form

$$\phi(x) = \begin{cases} 1, & L(x) > c \\ 0, & L(x) < c \end{cases}$$

where c is defined by

$$\mathbb{P}_0(\phi(X) = 1) = \mathbb{P}_0(L(X) > c) \leq \alpha \leq \mathbb{P}_0(L(X) \geq c)$$

Since $L(x)$ is continuous and strictly increasing under $\mathbb{P}_0$ with $\xi_1 > 0$, then c is uniquely defined by $\mathbb{P}_0(L(X) > c) = \alpha$. Thus the set of $\{x : L(x) > c\} \iff \{x : x > L^{-1}(c)\}$. We rewrite the test $\phi(x)$ as

$$\phi(x) = \begin{cases} 1, & x \geq t = \xi_0 + \sigma z_{1-\alpha}, \text{ equivalently } \mathbb{P}_0(X \geq t) = \alpha \\ 0, & x < t = \xi_0 + \sigma z_{1-\alpha} \end{cases}$$

where $z_{1-\alpha}$ is the $(1-\alpha)$ quantile of the standard normal distribution. This test rejects large values of $X \geq t$.

- Suppose that $\xi_1 < \xi_0 = 0$, then $L(x)$ is continuous and strictly decreasing under $\mathbb{P}_0$ with $\xi_1 < 0$, thus c is uniquely defined by $\mathbb{P}_0(L(X) > c) = \alpha$ and the set of $\{x : L(x) > c\} \iff \{x : x < L^{-1}(c)\}$. We rewrite the test $\phi(x)$

$$\phi(x) = \begin{cases} 1, & x \leq t = \xi_0 + \sigma z_{\alpha}, \text{ equivalently } \mathbb{P}_0(X \leq t) = \alpha \\ 0, & x < t = \xi_0 + \sigma z_{\alpha} \end{cases}$$

where z_α is the α quantile of the standard normal distribution. This test rejects small values of $X \leq t$.

♠

6.3 Distribution with Monotone Likelihood Ratio

Definition 6.3.1. *(Uniformly Most Powerful Test)*

A test is uniformly powerful(UMP) at level α if it is most powerful(MP) against any particular alternative $\theta_k \in \Omega_K$.

In the simplest situation, the distributions depend on a single real-valued parameter θ, and the hypothesis is one-sided, say $H : \theta \leq \theta_0$. In general, the most powerful test of H against an alternative $\theta_1 > \theta_0$ depends on θ_1 and is then not UMP(Uniformly Most Powerful Test). However, a UMP test does exist if an additional assumption is satisfied, which is Monotone Likelihood Ratio property.

Definition 6.3.2. *(Monotone Likelihood Ratio)*
The real parameter family of densities $\{f_\theta, \theta \in \Omega\}$ is said to be Monotone Likelihood Ratio(MLR), if there exists a real-valued function $T(x)$, such that for any $\theta < \theta'$, while the distributions $\mathbb{P}_\theta$ and $\mathbb{P}_{\theta'}$ is distinct, the ratio $L = \frac{f_{\theta'}}{f_\theta}$ is an nondecreasing function of $T(x)$, i.e., $L(T(x))$ is nondecreasing.

Theorem 6.3.1. *(Theorem 3.4.1 in TSH)*
Let θ be a real parameter, and let the random variable $X \sim f_\theta(x)$ with monotone likelihood ratio in $T(x)$, i.e., $L(T(x)) = \frac{f_{\theta'}(T(x))}{f_\theta(T(x))}$ is an nondecreasing function of $T(x)$

1. *For testing $H : \theta \leq \theta_0$ against $K : \theta > \theta_0$, there exists a UMP test $\phi(x)$, which is given by*
$$\phi(x) = \begin{cases} 1, & T(x) > C \\ \gamma, & T(x) = C \\ 0, & T(x) < C \end{cases} \tag{6.4}$$
where C and γ are determined by $\mathbb{E}_{\theta_0}[\phi(X)] = \alpha$.

2. *The power function*
$$\beta(\theta) = \mathbb{E}_\theta[\phi(X)]$$
of this test strictly increasing for all points θ for which $0 < \beta < 1$.

3. *For all θ', the test determined by equation (6.5) and $\mathbb{E}_{\theta'}[\phi(X)] = \alpha$ is UMP for testing $H' : \theta \leq \theta'$ against $K' : \theta > \theta'$ at level $\alpha' = \beta(\theta')$.*

4. *For any $\theta < \theta_0$, the test $\phi(x)$ minimizes $\beta(\theta)$(the probability of a first kind error) among all tests satisfying $\mathbb{E}_{\theta_0}[\phi(X)] = \alpha$.*

Example 46: (Example 3.4.1 (Hypergeometric) in TSH)
From a lot containing N items of a manufactured product, a sample of size n is selected at random, and each item in the sample is inspected, while N and n are known. If the total number of defective items in the lot is D, the number X of defectives found in the sample has the hypergeometric distribution

$$\mathbb{P}_D(X = x) = \frac{\binom{D}{x}\binom{N-D}{n-x}}{\binom{N}{n}}, \quad \max(0, n + D - N) \leq x \leq \min(n, D)$$

We want to test $H : D \leq D_0$ against $K : D > D_0$. Note that

$$\frac{\mathbb{P}_{D+1}(X = x)}{\mathbb{P}_D(X = x)} = \begin{cases} \infty, & x = D + 1 \\ \dfrac{D+1}{N-D} \dfrac{N-D-n+x}{D+1-x}, & n + D + 1 - N \leq x \leq D \\ 0, & x = n + D - N \end{cases}$$

Thus the distributions satisfy the assumption of monotone likelihood ratios with $T(x) = x$. Therefore, by theorem 6.3.1, there exists a UMP test for testing the hypothesis, which rejects H when X is too large. ♠

Example 47: (Example 3.4.2 (Binomial) in TSH)

The binomial distributions $Bin(n, \theta)$ with density $f_\theta(x) = \binom{n}{x}\theta^x(1-\theta)^{n-x}$, we want to test $H : \theta \le \theta_0$ vs $K : \theta > \theta_0$ at level α. The likelihood is given by

$$L(x) = \frac{f_{\theta_1}(x)}{f_{\theta_2}(x)} = \frac{\theta_1^x(1-\theta_1)^{n-x}}{\theta_2^x(1-\theta_2)^{n-x}} = \frac{(1-\theta_1)^n}{(1-\theta_2)^n}\left(\frac{\theta_1/(1-\theta_1)}{\theta_2/(1-\theta_2)}\right)^x$$

If $\theta_1 > \theta_2$, then $L(x) = b\alpha^x$ with $\alpha > 1$, thus the family is MLR with respect to $T(x) = x$. Then by theorem 6.3.1, there exists a UMP test $\phi(x)$, which is given by

$$\phi(x) = \begin{cases} 1, & T(x) > C \\ \gamma, & T(x) = C \\ 0, & T(x) < C \end{cases} \tag{6.5}$$

where C and γ are determined by $\mathbb{E}_{\theta_0}[\phi(X)] = \alpha$. The test $\phi(x)$ rejects H when X is too large. ♠

Example 48: (Problem 3.2 in TSH, UMP test for $U(0, \theta)$)

Let $X = (X_1, \cdots, X_n)$ be a sample from the uniform distribution on $(0, \theta)$.

1. For testing $H : \theta \le \theta_0$ against $K : \theta > \theta_0$ at level α. Show that any test $\phi(x)$ is UMP at level α for which $\mathbb{E}_{\theta_0}(\phi(X)) = \alpha$, $\sup_{\theta \le \theta_0} E_\theta(\phi(X)) \le \alpha$, and $\phi(x) = 1$ when $T(X) = \max(X_1, \cdots, X_n) > \theta_0$.

   Remark Note that the family is MLR with respect to $T(x) = \max(x_1, \cdots, x_n)$. For uniform distribution,

$$f(x) = \begin{cases} \frac{1}{\theta}, & x \in [0, \theta] \\ 0, & x \notin [0, \theta] \end{cases}$$

$$f_\theta(T(x)) = \frac{1}{\theta^n}\mathbf{1}_{\{T(x)<\theta\}} = \prod_{i=1}^n \frac{1}{\theta}\mathbf{1}_{\{x_i<\theta\}}$$

   Suppose that $\theta < \theta'$, then $L(T(x))$ is non-decreasing, since

$$L(T(x)) = \frac{f_{\theta'}(T(x))}{f_\theta(T(x))} = \begin{cases} \infty, & T(x) \ge \theta \\ \left(\frac{\theta}{\theta'}\right)^n, & T(x) < \theta \end{cases}$$

   Proof First fix θ_1, while $\theta_1 > \theta_0$. Consider testing $H : \theta = \theta_0$ vs $K : \theta = \theta_1$, by NP lemma (theorem 6.2.1), any test of the form

$$\phi(x) = \begin{cases} 1, & L(x) = \frac{f_{\theta_1}(x)}{f_{\theta_0}(x)} > c \\ 0, & L(x) = \frac{f_{\theta_1}(x)}{f_{\theta_0}(x)} < c \end{cases}$$

is most powerful at level $\alpha \in (0,1)$ for $H : \theta = \theta_0$ vs $K : \theta = \theta_1$ with $\alpha = \mathbb{E}_{\theta_0}[\phi(X)]$, which is used to determine c. However, when determining c, note that

$$L(x) = \begin{cases} \infty, & T(x) \geq \theta_0 \\ \left(\dfrac{\theta_0}{\theta_1}\right)^n, & T(x) < \theta_0 \end{cases}$$

$$\mathbb{P}_{\theta_0}(T(X) > \theta_0) = 1 - \mathbb{P}_{\theta_0}(T(X) \leq \theta_0) = 1 - [\mathbb{P}_{\theta_0}(X_1 \leq \theta_0)]^n = 1 - \left[\frac{\theta_0}{\theta_0}\right]^n = 0$$

- if $c > \left(\frac{\theta_0}{\theta_1}\right)^n$, then

$$\mathbb{E}_{\theta_0}[\phi(X)] = \mathbb{P}_{\theta_0}(\phi(X) = 1) = \mathbb{P}_{\theta_0}(L(X) > c) = \mathbb{P}_{\theta_0}(L(X) = \infty) = \mathbb{P}_{\theta_0}(T(X) \geq \theta_0) = 0$$

- if $c < \left(\frac{\theta_0}{\theta_1}\right)^n$, then

$$\mathbb{E}_{\theta_0}[\phi(X)] = \mathbb{P}_{\theta_0}(\phi(X) = 1) = \mathbb{P}_{\theta_0}(L(X) > c) = 1$$

which contradicts to the fact $\alpha \in (0,1)$. Then we must have $c = \left(\frac{\theta_0}{\theta_1}\right)^n$.

Now, consider testing $H : \theta \leq \theta_0$ vs $K : \theta = \theta_1$. Since $\sup_{\theta \leq \theta_0} \mathbb{E}_\theta[\phi(X)] \leq \alpha$, then ϕ is MP at level α. Note that we can write the test as

$$\phi(x) = \begin{cases} 1, & T(x) \geq \theta_0 \\ 0, & T(x) < \theta_0 \end{cases}$$

which does not depend on the choice of θ_1. Since we pick θ_1 randomly, we can go through all $\theta_1 > \theta_0$. Therefore, $\phi(x)$ is UMP for $H : \theta \leq \theta_0$ vs $K : \theta > \theta_0$ among the test ϕ' with size $\leq \alpha$. □

2. For testing $H : \theta = \theta_0$ against $K : \theta \neq \theta_0$ at level α. Show that a unique UMP test $\psi(x)$ exists, which is given by $T(X) = \max(X_1, \cdots, X_n)$, and

$$\psi(x) = \begin{cases} 1, & T(X) > \theta_0 \text{ or } T(X) \leq \theta_0 \alpha^{1/n} \\ 0, & \text{otherwise} \end{cases}$$

Proof The likelihood of uniform distribution is

$$f_\theta(x) = \begin{cases} \dfrac{1}{\theta}, & x \in [0, \theta] \\ 0, & x \notin [0, \theta] \end{cases}$$

(a) For $H : \theta = \theta_0$ vs $K_1 : \theta > \theta_0$, define $T(X) = \max(X_1, \cdots, X_n)$, for $\theta_1 > \theta_0$, the likelihood ratio is

$$L(x) = \frac{f_{\theta_1(x)}}{f_{\theta_0(x)}} = \begin{cases} \left(\dfrac{\theta_0}{\theta_1}\right)^n, & T(X) \leq \theta_0 \\ \infty, & T(X) > \theta_0 \end{cases}$$

By NP lemma 6.2.1, the MP test is of the form

$$\phi(x) = \begin{cases} 1, & L > c \\ 0, & L < c \end{cases}$$

where c is determined by $\mathbb{E}_{\theta_0}[\phi(X)] = \alpha$.

- If $c < \left(\frac{\theta_0}{\theta_1}\right)^n$, $\phi(x) \equiv 1 \implies \mathbb{E}_{\theta_0}[\phi(X)] = 1$, contradicting to $\alpha \in (0,1)$.
- If $c > \left(\frac{\theta_0}{\theta_1}\right)^n$, $\mathbb{E}_{\theta_0}[\phi(X)] = \mathbb{P}_{\theta_0}(T(X) > \theta_0) = 1 - \mathbb{P}_{\theta_0}(T(X) \leq \theta_0) = 1 - (\frac{\theta_0}{\theta_0})^n = 0$, contradicting to $\alpha \in (0,1)$.

Thus $c = \left(\frac{\theta_0}{\theta_1}\right)^n$, the MP test is given by $\mathbb{E}_{\theta_0}[\phi(X)] = \alpha$,

$$\phi(x) = \begin{cases} 1, & T(X) > \theta_0 \\ 0, & \text{otherwise} \end{cases}$$

Since $\phi(x)$ does not depend on θ_1, we can go through all $\theta_1 > \theta_0$. Then $\phi(x)$ is UMP test for $H : \theta = \theta_0$ vs $K_1 : \theta > \theta_0$.

(b) For $H : \theta = \theta_0$ vs $K_2 : \theta < \theta_0$, define $T(X) = \max(X_1, \cdots, X_n)$, for $\theta_1 < \theta_0$, the likelihood ratio is

$$L(x) = \frac{f_{\theta_1(x)}}{f_{\theta_0(x)}} \begin{cases} \left(\frac{\theta_0}{\theta_1}\right)^n, & T(X) \leq \theta_1 \\ 0, & T(X) > \theta_1 \end{cases}$$

By NP lemma 6.2.1, the MP test is of the form

$$\phi(x) = \begin{cases} 1, & L(X) > c \\ 0, & L(X) < c \end{cases}$$

where c is determined by $\mathbb{E}_{\theta_0}[\phi(X)] = \alpha$.

- If $c > \left(\frac{\theta_0}{\theta_1}\right)^n$, $\phi(x) \equiv 0 \implies \mathbb{E}_{\theta_0}[\phi(X)] = 0$, contradicting to $\alpha \in (0,1)$.
- If $c = \left(\frac{\theta_0}{\theta_1}\right)^n$, this gives us the first test, (1) $\phi = 0$ if $T(X) > \theta_1$.
- If $0 < c < \left(\frac{\theta_0}{\theta_1}\right)^n$, this gives us a test, by $\mathbb{E}_{\theta_0}[\phi(X)] = \alpha \implies c = \theta_0 \alpha^{1/n}$

$$\phi(x) = \begin{cases} 1, & L > \theta_0 \alpha^{1/n} \\ 0, & L < \theta_0 \alpha^{1/n} \end{cases}$$

- If $c = 0$, this gives us a second test,(2) $\phi = 1$ if $T(X) \leq \theta_1$.

Depending on the position between θ_1 and $\theta_0 \alpha^{1/n}$, we choose

- $\theta_1 \geq \theta_0 \alpha^{1/n}$, choose first test (1) $\phi = 0$ if $T(X) > \theta_1 \geq \theta_0 \alpha^{1/n}$.
- $\theta_1 < \theta_0 \alpha^{1/n}$, choose second test (2) $\phi = 1$ if $T(X) \leq \theta_1 < \theta_0 \alpha^{1/n}$.

The MP test is given by $\mathbb{E}_{\theta_0}[\phi(X)] = \alpha$,

$$\phi(x) = \begin{cases} 1, & T(X) < \theta_0 \alpha^{1/n} \\ 0, & \text{otherwise} \end{cases}$$

Since $\phi(x)$ does not depend on θ_1, we can go through all $\theta_1 < \theta_0$. Then $\phi(x)$ is UMP test for $H : \theta = \theta_0$ vs $K_2 : \theta < \theta_0$.

(c) For $H : \theta = \theta_0$ vs $K_3 : \theta \neq \theta_0$, Combining the result above together, then the UMP test is given by $\mathbb{E}_{\theta_0}[\phi(X)] = \alpha$,

$$\phi(x) = \begin{cases} 1, & T(X) < \theta_0 \alpha^{1/n} \text{ or } T(X) > \theta_0 \\ 0, & \text{otherwise} \end{cases}$$

Check $\mathbb{E}_{\theta_0}[\phi(X)] = \alpha$. By independence of $\{X_i\}_{i=1}^n$, we know

$$\mathbb{P}_{\theta_0}(T(X) \leq \theta_0 \alpha^{1/n}) = [\mathbb{P}_{\theta_0}(X_1 \leq \theta_0 \alpha^{1/n})]^n = \left[\frac{\theta_0 \alpha^{1/n}}{\theta_0}\right]^n = \alpha$$

$$\mathbb{P}_{\theta_0}(T(X) > \theta_0) = 1 - \mathbb{P}_{\theta_0}(T(X) \leq \theta_0) = 1 - [\mathbb{P}_{\theta_0}(X_1 \leq \theta_0)]^n = 1 - \left[\frac{\theta_0}{\theta_0}\right]^n = 0$$

Then

$$\mathbb{E}_{\theta_0}[\phi(X)] = 1 \cdot \mathbb{P}_{\theta_0}(T(X) > \theta_0) + 1 \cdot \mathbb{P}_{\theta_0}(T(X) \leq \theta_0 \alpha^{1/n}) = \alpha$$

where is the uniqueness ? □

♠

Example 49: (Problem 3.3 in TSH)

$X_1, \cdots, X_n \overset{i.i.d.}{\sim} f$ on $\mathbb{R}$, while f is known. We only observe

$$Y = \max(X_1, \cdots, X_n)$$

We want to test the number of random variables, $H : n \leq n_0$ vs $K : n > n_0$. Is there a UMP test of size α ? ♠

Proof Let

$$F(x) = \int_{-\infty}^{x} f(t)dt = \mathbb{P}[X \leq x]$$

Then

$$\begin{aligned} \mathbb{P}_n[Y \leq y] &= \mathbb{P}[\max(X_1, \cdots, X_n) \leq y] = \mathbb{P}[X_1 \leq y, \cdots, X_n \leq y] \\ &\overset{i.i.d.}{=} \prod_{i=1}^{n} \mathbb{P}[X_i \leq y] = [F(y)]^n \end{aligned}$$

Then Y has the density $g_n(y) = n[f(y)F(y)]^{n-1}$, then

$$\frac{g_{n+1}(y)}{g_n(y)} = \frac{n+1}{n} F(y)$$

is non-decreasing with respect to y, $g_n(y)$ has MLR property, $T(y) = y$. Then by theorem 6.3.1, there exists UMP test, which is given by

$$\phi(y) = \begin{cases} 1, & y > C \\ \gamma, & y = C \\ 0, & y < C \end{cases}$$

where $\mathbb{P}_{n_0}(Y = C) = 0$, since the likelihood ratio is continuous.

$$\begin{aligned} \alpha &= \mathbb{E}_{n_0}[\phi(Y)] = 1 \cdot \mathbb{P}_{n_0}(Y > C) + \gamma \cdot \mathbb{P}_{n_0}(Y = C) \\ &= \mathbb{P}_{n_0}(Y > C) = 1 - \mathbb{P}_{n_0}[Y \le C] = 1 - [F(C)]^n \end{aligned}$$

thus $C = F^{-1}[(1-\alpha)^{\frac{1}{n_0}}]$. $\square$

Example 50: (Problem 3.12 in TSH, Exercise)

Let $X_1, \cdots, X_n$ be independently distributed, each uniformly over the integers $1, 2, \cdots, \theta$. Determine whether there exists a UMP test for testing $H : \theta = \theta_0$, at level $\alpha = \frac{1}{\theta_0^n}$ against the alternatives

1. $\theta > \theta_0$
2. $\theta < \theta_0$
3. $\theta \neq \theta_0$

♠

Proof The likelihood of uniform distribution is

$$f_\theta(x) = \begin{cases} \dfrac{1}{\theta}, & x \in [0, \theta] \\ 0, & x \notin [0, \theta] \end{cases}$$

1. For $H : \theta = \theta_0$ vs $K_1 : \theta > \theta_0$, define $Y = \max(X_1, \cdots, X_n)$, for $\theta_1 > \theta_0$, the likelihood ratio is

   $$L(x) = \frac{f_{\theta_1(x)}}{f_{\theta_0(x)}} \begin{cases} \left(\dfrac{\theta_0}{\theta_1}\right)^n, & Y \le \theta_0 \\ \infty, & Y > \theta_0 \end{cases}$$

   By NP lemma 6.2.1, the MP test is of the form

   $$\phi(x) = \begin{cases} 1, & L > c \\ 0, & L < c \end{cases}$$

   where c is determined by $\mathbb{E}_{\theta_0}[\phi(X)] = \alpha$.

   - If $c < \left(\frac{\theta_0}{\theta_1}\right)^n$, $\phi(x) \equiv 1 \implies \mathbb{E}_{\theta_0}[\phi(X)] = 1$, contradicting to $\alpha \in (0, 1)$.

- If $c > \left(\frac{\theta_0}{\theta_1}\right)^n$, $\mathbb{E}_{\theta_0}[\phi(X)] = \mathbb{P}_{\theta_0}(Y > \theta_0) = 1 - \mathbb{P}_{\theta_0}(Y \leq \theta_0) = 1 - (\frac{\theta_0}{\theta_0})^n = 0$, contradicting to $\alpha \in (0,1)$.

Thus $c = \left(\frac{\theta_0}{\theta_1}\right)^n$, the MP test is given by $\mathbb{E}_{\theta_0}[\phi(X)] = \alpha$,

$$\phi(x) = \begin{cases} 1, & Y > \theta_0 \\ 0, & \text{otherwise} \end{cases}$$

Since $\phi(x)$ does not depend on θ_1, we can go through all $\theta_1 > \theta_0$. Then $\phi(x)$ is UMP test for $H : \theta = \theta_0$ vs $K_1 : \theta > \theta_0$.

2. For $H : \theta = \theta_0$ vs $K_2 : \theta < \theta_0$, define $Y = \max(X_1, \cdots, X_n)$, for $\theta_1 < \theta_0$, the likelihood ratio is

$$L(x) = \frac{f_{\theta_1(x)}}{f_{\theta_0(x)}} \begin{cases} \left(\frac{\theta_0}{\theta_1}\right)^n, & Y \leq \theta_1 \\ 0, & Y > \theta_1 \end{cases}$$

By NP lemma 6.2.1, the MP test is of the form

$$\phi(x) = \begin{cases} 1, & L > c \\ 0, & L < c \end{cases}$$

where c is determined by $\mathbb{E}_{\theta_0}[\phi(X)] = \alpha$.

- If $c > \left(\frac{\theta_0}{\theta_1}\right)^n$, $\phi(x) \equiv 0 \implies \mathbb{E}_{\theta_0}[\phi(X)] = 0$, contradicting to $\alpha \in (0,1)$.
- If $c = \left(\frac{\theta_0}{\theta_1}\right)^n$, this gives us the first test, (1) $\phi = 0$ if $Y > \theta_1$.
- If $0 < c < \left(\frac{\theta_0}{\theta_1}\right)^n$, this gives us a test, by $\mathbb{E}_{\theta_0}[\phi(X)] = \alpha \implies c = \theta_0 \alpha^{1/n}$

$$\phi(x) = \begin{cases} 1, & L > \theta_0 \alpha^{1/n} \\ 0, & L < \theta_0 \alpha^{1/n} \end{cases}$$

- If $c = 0$, this gives us a second test,(2) $\phi = 1$ if $Y \leq \theta_1$.

Depending on the position between θ_1 and $\theta_0 \alpha^{1/n}$, note that $\alpha = \frac{1}{\theta_0^n}$, we choose

- $\theta_1 \geq \theta_0 \alpha^{1/n}$, choose first test (1) $\phi = 0$ if $Y > \theta_1 \geq \theta_0 \alpha^{1/n} = 1$.
- $\theta_1 < \theta_0 \alpha^{1/n}$, choose second test (2) $\phi = 1$ if $Y \leq \theta_1 < \theta_0 \alpha^{1/n} = 1$.

The MP test is given by $\mathbb{E}_{\theta_0}[\phi(X)] = \alpha$,

$$\phi(x) = \begin{cases} 1, & Y < 1 \\ 0, & \text{otherwise} \end{cases}$$

Since $\phi(x)$ does not depend on θ_1, we can go through all $\theta_1 < \theta_0$. Then $\phi(x)$ is UMP test for $H : \theta = \theta_0$ vs $K_2 : \theta < \theta_0$.

3. For $H : \theta = \theta_0$ vs $K_3 : \theta \neq \theta_0$, combining the result above together, the UMP test is given by $\mathbb{E}_{\theta_0}[\phi(X)] = \alpha$,

$$\phi(x) = \begin{cases} 1, & Y < 1 \text{ or } Y > \theta_0 \\ 0, & \text{otherwise} \end{cases}$$

Check $\mathbb{E}_{\theta_0}[\phi(X)] = \alpha$. By independence of $\{X_i\}_{i=1}^n$, we know

$$\mathbb{P}_{\theta_0}(Y < 1) = [\mathbb{P}_{\theta_0}(X_1 \leq 1)]^n = \left[\frac{1}{\theta_0}\right]^n = \alpha$$

$$\mathbb{P}_{\theta_0}(Y > \theta_0) = 1 - \mathbb{P}_{\theta_0}(Y \leq \theta_0) = 1 - [\mathbb{P}_{\theta_0}(X_1 \leq \theta_0)]^n = 1 - \left[\frac{\theta_0}{\theta_0}\right]^n = 0$$

Then

$$\mathbb{E}_{\theta_0}[\phi(X)] = 1 \cdot \mathbb{P}_{\theta_0}(Y > \theta_0) + 1 \cdot \mathbb{P}_{\theta_0}(Y \leq \theta_0 \alpha^{1/n}) = \alpha$$

Therefore, $\phi(x)$ is a UMP test for $H : \theta = \theta_0$ vs $K : \theta \neq \theta_0$ at level α.

□

Example 51: (Problem 3.31 in TSH)

Let $X = (X_1, \cdots, X_n)$ be a sample from the uniform distribution $U(\theta, \theta + 1)$.

1. For testing $H : \theta \leq \theta_0$ against $K : \theta > \theta_0$ at level α, there exists a UMP test which rejects H when $\min(X_1, \cdots, X_n) > \theta_0 + C(\alpha)$ or $\max(X_1, \cdots, X_n) > \theta_0 + 1$ for suitable $C(\alpha)$.

   Proof Order statistic $X_{(1)} \leq \cdots \leq X_{(n)}$, any $\theta' \in (\theta, \theta + 1)$, $\mathbb{P}_{\theta'}(X_{(n)} > \theta_0 + 1) = 1 - \mathbb{P}_{\theta'}(X_{(n)} \leq \theta_0 + 1) = 1 - (\frac{\theta_0}{\theta_0})^n = 0$, then

$$\mathbb{E}_{\theta_0}[\phi(X)] = \mathbb{P}_{\theta_0}(X_{(1)} > \theta_0 + C(\alpha)) = [\mathbb{P}_{\theta_0}(X_1 > \theta_0 + C(\alpha))]^n = (1 - C(\alpha))^n = \alpha$$

   which implies $C(\alpha) = 1 - \alpha^{1/n}$. For any $\theta \leq \theta_0$,

$$\begin{aligned} \mathbb{E}_\theta[\phi(X)] &= \mathbb{P}_\theta(X_{(1)} > \theta_0 + C(\alpha)) + \mathbb{P}_\theta(X_{(n)} > \theta_0 + 1) \\ &= \begin{cases} (\theta + 1 - \theta_0 - C(\alpha))^n, & \theta_0 + C(\alpha) \leq \theta + 1 \\ 0, & \text{otherwise} \end{cases} \\ &= (\theta + 1 - \theta_0 - C(\alpha))_+^n \leq (1 - C(\alpha))^n = \alpha \end{aligned}$$

   Therefore $\phi(x)$ has size $\leq \alpha$. The likelihood is

$$f(x) = \begin{cases} 1, & x \in [\theta, \theta + 1] \\ 0, & x \notin [\theta, \theta + 1] \end{cases}$$

   Take $\theta_1 > \theta_0$,

   - If $\theta_1 \geq \theta_0 + 1$, trivial, rejecting all the time(why?)

- If $\theta_1 < \theta_0 + 1$, then the likelihood ratio is

$$L(x) = \begin{cases} 0, & X_{(1)} \in [\theta_0, \theta_1) \\ \infty, & X_{(n)} \in (\theta_0 + 1, \theta_1 + 1] \\ 1, & \text{otherwise} \end{cases}$$

by theorem 6.2.1, the test of the form

$$\varphi(x) = \begin{cases} 1, & L(x) > a \\ 0, & L(x) < a \end{cases}$$

such that $\mathbb{E}_{\theta_0}[\varphi(X)] = \alpha$ is MP for $\theta = \theta_0$ vs $\theta = \theta_1$ at level α

  - If $\alpha < (\theta_0 + 1 - \theta_1)^n \iff \theta_1 < \theta_0 + 1 - \alpha^{1/n} = \theta_0 + C(\alpha)$, choose $a = 1$, which gives

$$\varphi(x) = \begin{cases} 1, & X_{(n)} \geq \theta_0 + 1 \\ 0, & \text{otherwise} \end{cases}$$

  - If $\alpha \geq (\theta_0 + 1 - \theta_1)^n \iff \theta_1 \geq \theta_0 + 1 - \alpha^{1/n} = \theta_0 + C(\alpha)$, choose $a = 0$, which gives

$$\varphi(x) = \begin{cases} 1, & X_{(1)} \geq \theta_0 + C(\alpha) \\ 0, & \text{otherwise} \end{cases}$$

□

Remark A UMP test rarely exists for two sided problems such as $\theta = \theta_0$ vs $\theta \neq \theta_0$, or $\theta \in [\theta_1, \theta_2]$ vs $\theta \notin [\theta_1, \theta_2]$

2. The family $U(\theta, \theta + 1)$ does not have monotone likelihood ratio. How? **Proof**

□

♠

6.4 Two-Sided Hypotheses

UMP tests exist not only for one-sided but also for certain two-sided hypotheses of the form

$$H : \theta \leq \theta_1 or \theta \geq \theta_2 \quad (\theta_1 < \theta_2)$$

Theorem 6.4.1. *(Theorem 3.7.1 in TSH)*

1. *For testing the hypothesis $H : \theta \leq \theta_1$ or $\theta \geq \theta_2 (\theta_1 < \theta_2)$ against the alternatives $K : \theta_1 < \theta < \theta_2$ in the one-parameter exponential family, there exists a UMP test given by*

$$\phi(x) = \begin{cases} 1, & C_1 < T(x) < C_2 (C_1 < C_2) \\ \gamma_i, & T(x) = C_i, i = 1, 2 \\ 0, & T(x) < C_1 or T(x) > C_2 (C_1 < C_2) \end{cases}$$

where the C_1, C_2 and γ_1, γ_2 are determined by

$$\mathbb{E}_{\theta_1}[\phi(X)] = \mathbb{E}_{\theta_2}[\phi(X)] = \alpha$$

2. *This test minimizes $\mathbb{E}_\theta[\phi(X)]$ subject to $\mathbb{E}_{\theta_1}[\phi(X)] = \mathbb{E}_{\theta_2}[\phi(X)] = \alpha$ for all $\theta < \theta_1$ or $\theta > \theta_2 (\theta_1 < \theta_2)$*

3. *For $0 < \alpha < 1$ the power function of this test has a maximum at a point θ_0 between θ_1 and θ_2 and decreases strictly as θ tends away from θ_0 in either direction, unless there exist two values t_1 , t_2 such that $\mathbb{P}_\theta[T(X) \in \{t_1, t_2\}] = 1$ for all θ, see figure 6.1.*

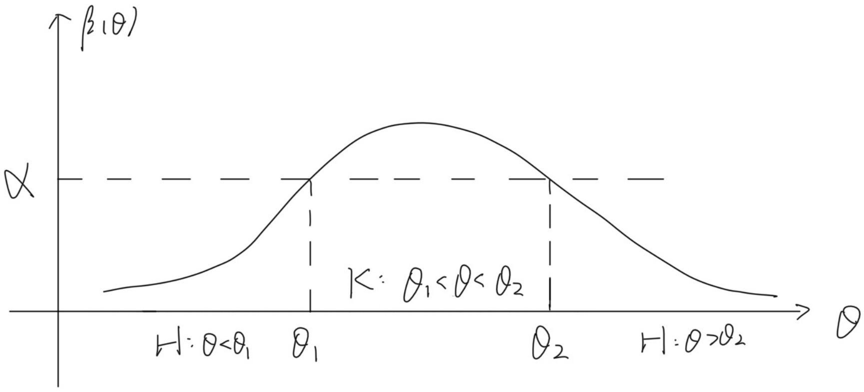

Figure 6.1: Power Function of the UMP Test for Two-Sided Hypotheses

6.5 Least Favorable Distributions

Now we consider the case of a Euclidean sample space; probability densities $\{f_\theta, \theta \in \omega\}$ and g with respect to a measure μ, and the problem of testing $H : \{f_\theta, \theta \in \omega\}$ against the simple alternative $K : \{g\}$. The existence of a most powerful level α test then follows from the weak compactness theorem for critical(test) functions.

Theorem 6.5.1. *(The Weak Compactness Theorem)*
Suppose μ is a σ finite measure on a Euclidean space, then the set of measurable test functions ϕ with $0 \le \phi \le 1$ is compact with respect to the weak convergence, meaning that given $\{\phi_n\}_{n\in\mathbb{N}}$, there exist $\{\phi_n\}_{n\in M}$ and a test function ϕ, such that

$$\int \phi_n f d\mu \overset{n\in M, n\to\infty}{\to} \int \phi f d\mu, \quad \textit{where } M \subset \mathbb{N}$$

for all f which is μ integrable, i.e.$\int |f| d\mu < \infty$

Corollary 6.5.2. *(Existence of a most powerful level α test)*
Suppose that we are testing

$$H = \{f_\theta, \theta \in \omega\} \quad vs \quad K = \{g\}$$

f_θ,g are densities with respect to μ. Let C be class of test functions

$$C = \{\psi : \text{all test functions such that } \sup_{\theta\in\omega} \mathbb{E}_\theta \left[\psi(X)\right] \leq \alpha\}$$

where C is the class of test functions at level α and g is a μ-integrable function. Then, there is ϕ in C, maximizing

$$\int \phi g d\mu = \sup_{\psi\in C} \int \psi g d\mu$$

Proof We want to maximize $\mathbb{E}_g \left[\psi(X)\right]$ over $\psi \in C$, where

$$\mathbb{E}_\theta \left[\psi(X)\right] = \int \psi f_\theta d\mu \quad \in [0,1]$$
$$\mathbb{E}_g \left[\psi(X)\right] = \int \psi g d\mu \quad \in [0,1]$$

By completeness of the real numbers, there exists a sequence of test functions $\{\psi_n, \psi_n \in C\}$ such that

$$\mathbb{E}_g \left[\psi_n(X)\right] \overset{n\to\infty}{\to} \sup_{\psi\in C} \mathbb{E}_g \left[\psi\right]$$

By the weak compactness theorem, there exists a subsequence of $\{\psi_m, \psi_m \in C\}_{m\in M}, M \subset \mathbb{N}$ and a test function ϕ such that

$$\int \psi_n f d\mu \overset{n\in M, n\to\infty}{\to} \int \phi f d\mu, \qquad \text{where } M \subset \mathbb{N}$$

which implies

1. check level α: any $\theta \in \omega$,

$$\mathbb{E}_\theta \left[\psi(X)\right] = \int \psi(x) f_\theta(x) d\mu = \lim_{n\in M, n\to\infty} \int \psi_n(x) f_\theta(x) d\mu \leq \alpha$$

2. check maximizing power function

$$\mathbb{E}_g \left[\phi(X)\right] = \int \phi(x) g(x) d\mu(x) = \lim_{m\in M, m\to\infty} \int \psi_m(x) g d\mu(x) = \sup_{\psi\in C} \mathbb{E}_\theta \left[\psi(X)\right] \leq \alpha$$

□

From another point of view the method of attack, as throughout the theory of hypothesis testing, is to reduce the composite hypothesis to a simple one. This achieved by considering weighted averages of the distributions of $H : \{f_\theta, \theta \in \omega\}$. The composite hypothesis H is replaced by the simple hypothesis $H_\Lambda = \{h_\Lambda(x)\}$ that the probability density of X is given by

$$h_\Lambda(x) = \int_\omega f_\theta(x) d\Lambda(\theta)$$

where Λ is a probability distribution over ω. The problem of finding a suitable Λ is frequently made easier by the following consideration. Since H provides no information concerning θ and since H_Λ is to be equivalent to H for the purpose of testing against g, knowledge of the distribution Λ should provide as little help for this task as possible. To make this precise suppose that θ is known to have a distribution Λ. Then the maximum power β_Λ that can be attained against g is that of the most powerful test $\phi(\Lambda)$ for testing $H_\Lambda = \{h_\Lambda(x)\}$ against g.

Definition 6.5.1. *(Least Favorable Distribution)*

The distribution Λ is said to be least favorable (at level α) if for all Λ' the inequality $\beta_\Lambda \leq \beta_{\Lambda'}$ holds.

Theorem 6.5.3. *(Theorem 3.8.1 in TSH)*

Let a σ-field be defined over ω such that the densities $f_\theta(x)$ are jointly measurable in θ and x. Suppose that over this σ-field there exist a probability distribution Λ such that the most powerful level-α test ϕ_Λ for testing $H_\Lambda = \{h_\Lambda(x)\}$ against $K : g$ is of size $\leq \alpha$ also with respect to the original hypothesis H.

1. *The test ϕ_Λ is most powerful for testing H against g.*
2. *If ϕ_Λ is the unique most powerful level-α for testing $H_\Lambda = \{h_\Lambda(x)\}$ against g, it is also the unique most powerful test of H against g.*
3. *The distribution Λ is least favorable, i.e., for any prior Λ', $\int \phi_\Lambda g d\mu \leq \int \phi_{\Lambda'} g d\mu$*

Proof

1. Take any test ϕ, which is at level α for H vs g, *i.e.*, $\mathbb{E}_\theta[\phi(X)] \leq \alpha, \forall \theta \in H$, then it is also a test at level α for H_Λ, this is because

$$\begin{aligned} &\int \phi(x) h_\Lambda(x) d\mu(x) = \int \phi(x) \left[\int_\omega f_\theta(x) d\Lambda(\theta)\right] d\mu(x) \\ \overset{Fubini}{=} &\int \left[\int \phi(x) f_\theta(x) d\mu(x)\right] d\Lambda(\theta) = \int \mathbb{E}_\theta[\phi(X)] d\Lambda(\theta) \\ &\leq \alpha \int d\Lambda(\theta) = \alpha \end{aligned}$$

Therefore, the power of $\phi(x)$ can not exceed the power of $\phi_\Lambda(x)$, since $\phi_\Lambda(x)$ is MP for testing $H_\Lambda = \{h_\Lambda(x)\}$ vs g. As a result, $\phi_\Lambda(x)$ is most powerful for testing H against g. If $\phi_\Lambda(x)$ is unique for testing $H_\Lambda = \{h_\Lambda(x)\}$ against g, then it is unique for testing H against g

2. (Exercise) Take another distribution Λ', $h_{\Lambda'}(x)$ is defined by $h_{\Lambda'}(x) = \int_w f_\theta(x) d\Lambda'(\theta)$. Since $\phi_\Lambda(x)$ is a level α for H, then it also a level α for $H_{\Lambda'}$. Let $\phi_{\Lambda'}(x)$ be MP test for $H_{\Lambda'}$ against g, we have

$$\int \phi_\Lambda h_{\Lambda'}(x) d\mu(x) \leq \alpha$$

$$\begin{aligned} \beta_\Lambda(\theta) &= \int \phi_\Lambda(x) g(x) d\mu(x) \\ &\leq \max_{\psi \in \{\text{level } \alpha \text{ test for } H_{\Lambda'}\}} \int \psi(x) g(x) d\mu(x) = \int \phi_{\Lambda'}(x) g(x) d\mu(x) = \beta_{\Lambda'}(\theta) \end{aligned}$$

Therefore, Λ is least favorable.

□

Corollary 6.5.4. *(Corollary 3.8.1 in TSH)*

Suppose that Λ is a probability distribution over ω and that ω' is a subset of ω with $\Lambda(\omega') = 1$. Let ϕ_Λ be a test such that

$$\phi_\Lambda = \begin{cases} 1, & \dfrac{g(x)}{h_\Lambda(x)} > c \\ 0, & \dfrac{g(x)}{h_\Lambda(x)} < c \end{cases}$$

Then ϕ_Λ is a most powerful level-α for testing $H = \{f_\theta, \theta \in \omega\}$ against $K = \{g\}$ provided

$$\mathbb{E}_{\theta'}\left[\phi_\Lambda(X)\right] = \sup_{\theta\in\omega} \mathbb{E}_\theta\left[\phi_\Lambda(X)\right] = \alpha, \quad \theta' \in \omega'$$

Example 52: (Sign test, Example 3.8.1 in TSH)

$X_1, \cdots, X_n \overset{i.i.d.}{\sim} F$ on $\mathbb{R}$, where no knowledge is assumed about F. Given $u \in \mathbb{R}$, define $p = \mathbb{P}\left[X \le u\right] = F(u)$. We want to test $H : p \le p_0$ vs $K : p > p_0$, where $p_0 \in [0, 1)$ is given. ♠

Proof Define

- F_-: any distribution of X given $X \le u$, with density f_- with respect to $\mu = F_- + F_+$
- F_+: any distribution of X given $X > u$, with density f_+ with respect to $\mu = F_- + F_+$

Note the fact that distribution F on the real line can be characterized by the probability p together with the conditional probability distributions F_- and F_+ of X given $X \le u$ and $X > u$ respectively, which isolate p from F, since F_- and F_+ are chosen arbitrarily. That is, F can be equivalently given by the triple (p, F_-, F_+).

Now, we fix an alternative, $K_1 = \{p_1\}$, meaning $(p_1, F_-^{(1)}, F_+^{(1)})$ where $p_1 > p_0$, and $F_-^{(1)}$ and $F_+^{(1)}$ be arbitrary distribution on $(-\infty, u]$ and (u, ∞).

Let Λ be a point mass prior distribution at $(p_0, F_-^{(1)}, F_+^{(1)})$, then we average the null $H = \{p \le p_0\}$ as $H_\Lambda = \{h_\Lambda\} = \{(p_0, F_-^{(1)}, F_+^{(1)})\}$. Then we are going to derive a MP test for $H_\Lambda = \{(p_0, F_-^{(1)}, F_+^{(1)})\}$ vs $K = \{(p_1, F_-^{(1)}, F_+^{(1)})\}$.

If sample point $x = (x_1, \cdots, x_n)$ satisfying

$$x_{i_1}, \cdots, x_{i_m} \le u < x_{j_1}, \cdots, x_{j_{n-m}}$$

the likelihood at $(p_1, F_-^{(1)}, F_+^{(1)})$ with respect to $\mu = F_- + F_+$ is

$$p_1^m (1-p_1)^{n-m} f_-^{(1)}(x_{i_1}) \cdots f_-^{(1)}(x_{i_m}) f_+^{(1)}(x_{j_1}) \cdots f_+^{(1)}(x_{j_{n-m}})$$

The form of likelihood for $(p_0, F_-^{(1)}, F_+^{(1)})$ is the same. Then we write down the likelihood ratio

$$L(m) = \frac{p_1^m(1-p_1)^{n-m}}{p_0^m(1-p_0)^{n-m}} = \frac{(1-p_1)^n}{(1-p_0)^n}\left(\frac{p_1/(1-p_1)}{p_0/(1-p_0)}\right)^m$$

The likelihood ratio is increasing with respect to m, and it is discrete, thus we need to randomize the test to ensure the uniqueness. The likelihood ratio test, which rejects large values of $T(x) = \#\{i : x_i \le u\} = m$, is of the form

$$\phi_\Lambda(x) = \begin{cases} 1, & T(x) > c \\ \gamma, & T(x) = c \\ 0, & T(x) < c \end{cases}$$

where c, γ is determined by $\alpha = \mathbb{E}_{p_0}[\phi(X)] = 1\cdot\mathbb{P}[Binomial(n,p_0) > c] + \gamma\cdot\mathbb{P}[Binomial(n,p_0) = c]$, noticing that $T(x) \sim Binomial(n,p)$, which just depends on p.

Now we are going to check the condition that this test is of level α for the original null $H = \{p \le p_0\}$. This is true by theorem 6.3.1, together with the property that likelihood ratio has MLR.

Then by theorem 6.5.3, ϕ_Λ is most powerful at level α for $H = \{p \le p_0\}$ vs $K_1 = \{p_1\}$.

The definition of ϕ_Λ does not depend on p_1. If we pick another alternative $p_2 > p_0$, we could get the same ϕ_Λ. Then we can go through all alternatives for $K = \{p > p_0\}$. Therefore, ϕ_Λ is UMP test for $H = \{p \le p_0\}$ vs $K = \{p > p_0\}$ at level α. □

Example 53: (One-sided tests of variance, Example 3.9.1 in TSH)

Let $X_1, \cdots, X_n$ be i.i.d sample from $\mathsf{N}(\xi, \sigma^2)$. Consider test $H_1 : \sigma \ge \sigma_0$ vs $K_1 : \sigma < \sigma_0$ first.

1. Fix an alternative (ξ_1, σ_1^2), where $\xi_1 \in \mathbb{R}$, $\sigma_1 < \sigma_0$
2. Choose a point mass prior Λ at (ξ_1, σ_0^2)
3. Testing $H_{\Lambda_1} = \{\mathsf{N}(\xi_1, \sigma_0^2)\}$ vs $K_{11} = \{\mathsf{N}(\xi_1, \sigma_1^2)\}$, the likelihood ratio is

$$L(x_1, \cdots, x_n) = \frac{\prod_{i=1}^n \left(\frac{1}{\sqrt{2\pi\sigma_1^2}}\right) \exp\left(-\frac{(x_i-\xi_1)^2}{2\sigma_1^2}\right)}{\prod_{i=1}^n \left(\frac{1}{\sqrt{2\pi\sigma_0^2}}\right) \exp\left(-\frac{(x_i-\xi_1)^2}{2\sigma_0^2}\right)} \propto \exp\left[\left(\frac{1}{2\sigma_0^2} - \frac{1}{2\sigma_1^2}\right)\sum_{i=1}^n (x_i - \xi_1)^2\right]$$

   since $\sigma_0 > \sigma_1$, then $\left(\frac{1}{2\sigma_0^2} - \frac{1}{2\sigma_1^2}\right) < 0$. Let $T = \sum_{i=1}^n (x_i - \xi_1)^2$, $L(T)$ is decreasing with respect to T. The likelihood ratio test is of the form

$$\phi_{\Lambda_1}(x) = \begin{cases} 1, & T \le c \\ 0, & \text{otherwise} \end{cases}$$

   where c is determined by $\mathbb{E}_{\sigma_0}[\phi_{\Lambda_1}(X)] = \mathbb{P}_{\sigma_0}(T \le c) = \alpha$, $c = \sigma_0^2 \chi^2_{n-1,\alpha}$, since $\frac{T}{\sigma_0^2} \sim \chi^2_{n-1}$ By NP lemma, $\phi_{\Lambda_1}(x)$ is unique MP test for H_{Λ_1} vs K_{11} at level α.
4. Then we are going to prove that this test is at level α for $H : \sigma \ge \sigma_0$. Actually,

$$\mathbb{E}_{\sigma\ge\sigma_0}[\phi_{\Lambda_1}(X)] = \mathbb{P}_{\sigma\ge\sigma_0}(T \le c) = \mathbb{P}_{\sigma\ge\sigma_0}\left(\frac{T}{\sigma^2} \le \frac{\sigma_0^2}{\sigma^2}\chi^2_{n-1,\alpha}\right) \le \mathbb{P}_{\sigma\ge\sigma_0}\left(\frac{T}{\sigma^2} \le \chi^2_{n-1,\alpha}\right) = \alpha$$

   Note the fact that $\mathbb{P}_{\xi\in\mathbb{R},\sigma\ge\sigma_0}\left(T \le \sigma_0^2\chi^2_{n-1,\alpha}\right)$ is maximized at $\xi = \xi_1$, $\sigma = \sigma_0$.

5. Since the test $\phi_{\Lambda_1}(x)$ is unique determined by ξ_1, then there is no uniform powerful test for $H_1 : \sigma \geq \sigma_0$ vs $K_1 : \sigma < \sigma_0$.

Now, we consider testing $H_2 : \sigma \leq \sigma_0$ vs $K_2 : \sigma > \sigma_0$.

1. Restrict our attention to the jointly sufficient statistics for (ξ, σ^2)

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

the likelihood is of the form

$$Cu^{(n-3)/2}\exp\left(-\frac{n}{2\sigma^2}(y-\xi)^2 - \frac{u}{2\sigma^2}\right)$$

Note that

$$Y \sim \mathsf{N}(\xi, \frac{\sigma^2}{n}) = \xi + \frac{\sigma}{\sqrt{n}}\mathsf{N}(0,1), \quad U \sim \sigma^2\chi^2_{n-1}, \quad Y \perp U$$

2. Fix an alternative $K_{21} = \{(\xi_1, \sigma_1^2)\}$, where $\xi_1 \in \mathbb{R}$, $\sigma_1 > \sigma_0$

3. Take a prior $\Lambda_2 = \lambda \otimes \delta_{\sigma_0}$ on (ξ, σ^2), where ξ is generated from λ. We choose $\lambda = \mathsf{N}(\xi_1, \frac{\sigma_1^2 - \sigma_0^2}{n}) \perp \mathsf{N}(0,1)$, $\sigma_1 > \sigma_0$. Under H_{Λ_2}, we write $Y = T + \frac{\sigma_0}{\sqrt{n}}Z$, where $T \sim \lambda$, $Z \sim \mathsf{N}(0,1)$, $T \perp Z$ and $U \sim \sigma_0^2\chi^2_{n-1}$, $Y \perp U$. Under $K_{21} = \{(\xi_1, \sigma_1^2)\}$, $Y \sim \mathsf{N}(\xi_1, \frac{\sigma_1^2}{n})$, $U \sim \sigma_1^2\chi^2_{n-1}$ and $Y \perp U$.

4. The likelihood ratio for H_{Λ_2} vs K_{21}

$$L(u, y) = \exp\left(\frac{1}{2}\left(\frac{1}{\sigma_0^2} - \frac{1}{\sigma_1^2}\right)u\right)$$

the test is of the form

$$\phi_{\Lambda_2}(u, y) = \begin{cases} 1, & u \geq c \\ 0, & \text{otherwise} \end{cases}$$

where c is determined by $\mathbb{P}_{H_{\Lambda_2}}(U \geq c) = \mathbb{P}_{H_{\Lambda_2}}(\sigma_0^2\chi^2_{n-1} \geq c) = \alpha$, $c = \sigma_0^2\chi^2_{n-1,1-\alpha}$.

5. Then check $\phi_{\Lambda_2}(x)$ is at level α for H_2.

$$\mathbb{E}_{\sigma\leq\sigma_0}\left[\phi_{\Lambda_2}(X)\right] = \mathbb{P}_{\sigma\leq\sigma_0}(U \geq c) = \mathbb{P}_{\sigma\leq\sigma_0}\left(\frac{U}{\sigma^2} \geq \frac{\sigma_0^2}{\sigma^2}\chi^2_{n-1,1-\alpha}\right) \leq \mathbb{P}_{\sigma\leq\sigma_0}\left(\frac{U}{\sigma^2} \geq \chi^2_{n-1,1-\alpha}\right) = \alpha$$

6. By theorem 6.5.3, $\phi_{\Lambda_2}(x)$ is the (unique) MP test for $H_2 : \sigma \leq \sigma_0$ vs $K_{21} = \sigma = \sigma_1 > \sigma_0$, then $\phi_{\Lambda_2}(x)$ is UMP for $H_2 : \sigma \leq \sigma_0$ vs $K_2 : \sigma > \sigma_0$.

♠

Example 54: (One-sided tests of a combination of means, Example 3.9.2 in TSH)

Assume $\boldsymbol{X} = (X_1, \cdots, X_k)^T$ is multivariate normal with unknown mean $\boldsymbol{\xi} = (\xi_1, \cdots, \xi_k)^T$ and known covariance matrix $\boldsymbol{\Sigma}$. Given data $\boldsymbol{X}^1, \cdots, \boldsymbol{X}^n \overset{i.i.d.}{\sim} \mathsf{N}_k(\boldsymbol{\xi}, \boldsymbol{\Sigma})$. Assume $\boldsymbol{a} = (a_1, \cdots, a_k)^T$ is a fixed vector with $\boldsymbol{a}^T\boldsymbol{\Sigma}\boldsymbol{a} > 0$. The problem is to test

$$H : \boldsymbol{a}^T\boldsymbol{\xi} = \sum_{i=1}^{k} a_i\xi_i \leq \delta \quad vs \quad K : \boldsymbol{a}^T\boldsymbol{\xi} = \sum_{i=1}^{k} a_i\xi_i > \delta$$

1. Reduction by sufficiency, considering
$$\bar{\boldsymbol{X}} = \frac{1}{n}\boldsymbol{X}^i \sim \mathsf{N}_k(\boldsymbol{\xi}, \frac{1}{n}\boldsymbol{\Sigma})$$
without loss of generality, assume that $n = 1$, and our data is $\boldsymbol{X} = (X_1, \cdots, X_k)^T \sim \mathsf{N}_k(\boldsymbol{\xi}, \boldsymbol{\Sigma})$

2. Since $\boldsymbol{\Sigma}$ is known, $\boldsymbol{\Sigma}$ is symmetric and positive semi-definite, *i.e.*, $\boldsymbol{u}^T\boldsymbol{\Sigma}\boldsymbol{u} \geq 0$, $\forall \boldsymbol{u} \in \mathbb{R}^k$, we can decompose it as $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}^{1/2}\boldsymbol{\Sigma}^{1/2}$. Thus $\boldsymbol{\Sigma}^{-1/2}\boldsymbol{X} \sim \mathsf{N}_k(\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\xi}, \boldsymbol{\Sigma}^{-1/2}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{-1/2}) = \mathsf{N}_k(\boldsymbol{\gamma}, \mathbb{I}_k)$, where $\boldsymbol{\gamma} = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{\xi}$, then $\boldsymbol{a}^T\boldsymbol{\xi} = \boldsymbol{b}^T\boldsymbol{\gamma}$, where $\boldsymbol{b} = \boldsymbol{\Sigma}^{1/2}\boldsymbol{a}$

3. Without loss of generality, assume that $\boldsymbol{\Sigma} = \mathbb{I}_k$, $\delta = 0$, $\|\boldsymbol{a}\|_2 = 1$. Notice that we can take an orthogonal matrix $\boldsymbol{R}$ to transform $\boldsymbol{a}$ such that $\boldsymbol{R}\boldsymbol{a} = \boldsymbol{e}_1 = (1, \cdots, 0) \in \mathbb{R}^k$, $k \geq 2$, which lead to $\boldsymbol{a}^T\boldsymbol{\xi} = \boldsymbol{a}^T\boldsymbol{R}^T\boldsymbol{R}\boldsymbol{\xi} = \boldsymbol{c}^T\boldsymbol{\lambda}$, to simplify discussion. Thus we further assume that $\boldsymbol{a} = \boldsymbol{e}_1 = (1, \cdots, 0)$.

4. Now, after simplification, given data $\boldsymbol{X} = (X_1, \ldots, X_n) \sim \mathsf{N}_k(\boldsymbol{\xi}, \mathbb{I}_k)$, where $\boldsymbol{a} = \boldsymbol{e}_1 = (1, \cdots, 0)$, $\boldsymbol{\xi} = (\xi_1, \cdots, \xi_k)$. We are actually testing
$$H : \xi_1 \leq 0 \quad vs \quad K : \xi_1 > 0$$
Note that here $X_2, \cdots, X_k$ are nuisance parameters. To use theorem 6.3.1, the likelihood ratio need to only depend on X_1. However, there are other nuisance parameters. Thus we can not apply theorem 6.3.1 directly.

5. Fix an alternative $K^* : \boldsymbol{\xi}^* = (\xi_1^*, \cdots, \xi_k^*)$ such that $\xi_1^* > 0$, $\xi_2^* \in \mathbb{R}$, $\cdots$, $\xi_k^* \in \mathbb{R}$. Choose point mass prior Λ at $(0, \xi_2^*, \cdots, \xi_k^*)$. The likelihood ratio for $H_\Lambda = (0, \xi_2^*, \cdots, \xi_k^*)$ vs $K = (\xi_1^*, \cdots, \xi_k^*)$ is
$$L(x_1, \cdots, x_k) = \frac{\exp\left(-\frac{1}{2}(x_1 - \xi_1^*)^2\right)}{\exp\left(-\frac{1}{2}x_1^2\right)} \propto \exp\left(x_1\xi_1^*\right)$$
Note that $\xi_1^* > 0$, and the likelihood ratio is continuous, the MP test is of the form
$$\phi_\Lambda(X_1, \cdots, X_k) = \begin{cases} 1, & X_1 > c \\ 0, & \text{otherwise} \end{cases}$$
where c is determined by $\mathbb{E}_{\xi_1=0}\left[\phi_\Lambda(X)\right] = \mathbb{P}_{\xi_1=0}(X_1 > c) = \alpha$. Under H_Λ, $X_1 \sim \mathsf{N}(0, 1)$, thus $c = Z_{1-\alpha}$, where $Z_{1-\alpha}$ is quantile of standard normal distribution.

6. Now, we are going to check that $\phi_\Lambda(x)$ is at level α for H. For any $\xi_1 < 0$, $\mathbb{E}_{\xi_1<0}\left[\phi_\Lambda(X)\right] = \mathbb{P}_{\xi_1<0}(X_1 > c) = \mathbb{P}_0(N(0,1) > c - \xi_1) \leq \mathbb{P}_0(N(0,1) > c) = \alpha$.

7. By theorem 6.5.3, $\phi_\Lambda(x)$ is most powerful for H vs K^*.

8. $\phi_\Lambda(x)$ does not depend on K^*, therefore $\phi_\Lambda(x)$ is UMP for H vs K.

♠

Example 55:
We are testing $H : \mathbb{P}_0$ vs $K : \{\mathbb{P}_1, \mathbb{P}_2\}$ at level α. Find UMP test or show there is no UMP test depending on different values of α. ♠

Proof First, we write down the likelihood ratio

	A	B	C	D
$\mathbb{P}_0$	0.01	0.03	0.02	0.94
$\mathbb{P}_1$	0.01	0.02	0.03	0.94
$\mathbb{P}_2$	0.02	0.01	0.03	0.94

	A	B	C	D
$L_1 = \frac{\mathbb{P}_1}{\mathbb{P}_0}$	1	$\frac{2}{3}$	$\frac{3}{2}$	1
$L_2 = \frac{\mathbb{P}_2}{\mathbb{P}_0}$	2	$\frac{1}{3}$	$\frac{3}{2}$	1

1. $0 < \alpha < 0.01$. Considering $\mathbb{P}_0$ vs $\mathbb{P}_1$. The most powerful test is of the form

$$\phi_1 = \begin{cases} 1, & L_1 > t \\ 0, & L_1 < t \end{cases}$$

However, any event A, B, C, D can not be assigned $\phi = 1$, since $\mathbb{P}_0(A) = 0.01$, $\mathbb{P}_0(B), \mathbb{P}_0(C), \mathbb{P}_0(D) > 0.01$, then $\mathbb{E}_0[\phi_1] \geq 0.01 > \alpha$. Meanwhile, we should consider the event with highest likelihood ratio value first, thus we have to let $t = \frac{3}{2}$, and randomize the test in the following way, not exceeding α

$$\phi_1 = \begin{cases} \gamma_1, & \text{if } C \\ 0, & \text{A,B,D} \end{cases}$$

where γ_1 is determined by $\mathbb{E}_0[\phi_1] = \alpha = 0.02\gamma_1 \implies \gamma_1 = \frac{\alpha}{0.02}$.

Considering $\mathbb{P}_0$ vs $\mathbb{P}_2$. Following the same argument above, the most powerful test is of the form

$$\phi_2 = \begin{cases} \gamma_2, & \text{if } A \\ 0, & \text{B,C,D} \end{cases}$$

where γ_2 is determined by $\mathbb{E}_0[\phi_2] = \alpha = 0.01\gamma_2 \implies \gamma_2 = \frac{\alpha}{0.01}$.

We can see that ϕ_1 and ϕ_2 are totally different test, therefore there is no UMP test at level $\alpha \in (0, 0.01)$ for $H = \{\mathbb{P}_0\}$ vs $K = \{\mathbb{P}_1, \mathbb{P}_2\}$.

2. $0.01 \leq \alpha < 0.03$. Considering $\mathbb{P}_0$ vs $\mathbb{P}_1$, in order to satisfy $\mathbb{E}_0[\phi_1] = \alpha \in [0.01, 0.03)$, we have to always reject C, and randomize at A and D. If we also reject A, $\mathbb{E}_0[\phi_1] \geq 0.03$. Thus the MP test is of the form

$$\phi_1 = \begin{cases} 1, & \text{if } C \\ \gamma_A, & \text{if } A \\ 0, & \text{B,D} \end{cases}$$

where γ_A is determined by $\mathbb{E}_0[\phi_1] = \alpha$ and $\gamma_A \in (0, 1)$.

Considering $\mathbb{P}_0$ vs $\mathbb{P}_2$, in order to satisfy $\mathbb{E}_0[\phi_1] = \alpha \in [0.01, 0.03)$, we have to always reject A, and randomize at C. If we also reject C, $\mathbb{E}_0[\phi_1] \geq 0.03$. Thus the MP test is of the form

$$\phi_2 = \begin{cases} 1, & \text{if } A \\ \gamma_C, & \text{if } C \\ 0, & \text{B,D} \end{cases}$$

where γ_C is determined by $\mathbb{E}_0[\phi_1] = \alpha$ and $\gamma_C \in (0, 1)$.

We can see that ϕ_1 and ϕ_2 are totally different test, therefore there is no UMP test at level $\alpha \in [0.01, 0.03)$ for $H = \{\mathbb{P}_0\}$ vs $K = \{\mathbb{P}_1, \mathbb{P}_2\}$.

3. $0.03 \leq \alpha < 0.97$. Considering $\mathbb{P}_0$ vs $\mathbb{P}_1$, to satisfy $\mathbb{E}_0[\phi_1] = \alpha \geq 0.03$, we should definitely reject C and randomize at A and D since $L_1(A) = L_1(D) = 1$, the MP test is of the form

$$\phi_1 = \begin{cases} 1, & C \\ \gamma_A, & A \\ \gamma_D, & D \\ 0, & \text{B} \end{cases}$$

where γ_A and γ_D is determined by

$$\mathbb{E}_{\mathbb{P}_0}[\phi_1] = 0.02 + 0.01\gamma_A + 0.94\gamma_D = \alpha, \quad \gamma_A \in [0, 1], \gamma_D \in [0, 1]$$

As we can see, there are infinitely many choices for γ_A and γ_D.

Considering $\mathbb{P}_0$ vs $\mathbb{P}_2$, to satisfy $\mathbb{E}_0[\phi_1] = \alpha \geq 0.03$, we should definitely reject A and C, since we need to include the points with higher likelihood ratio first, then randomize D, the MP test is of the form

$$\phi_2 = \begin{cases} 1, & A \text{ or } C \\ \gamma_D, & D \\ 0, & \text{B} \end{cases}$$

where γ_D is determined by

$$\mathbb{E}_{\mathbb{P}_0}[\phi_1] = 0.01 + 0.02 + 0.94\gamma_D = \alpha, \quad \gamma_D \in [0, 1]$$

As we can see, there is a unique choice for $\gamma_D = \frac{\alpha - 0.03}{0.94}$.

ϕ_1 and ϕ_2 could be the same test if they take same γ_D, and then γ_A is determined. Therefore, there is a unique UMP test at level $\alpha \in [0.03, 0.97)$ for $H = \{\mathbb{P}_0\}$ vs $K = \{\mathbb{P}_1, \mathbb{P}_2\}$.

4. $0.97 \leq \alpha < 1$. Considering $\mathbb{P}_0$ vs $\mathbb{P}_1$, we reject C, randomize at A and D. For $\mathbb{P}_0$ vs $\mathbb{P}_2$, we reject A and C, randomize at D. As it has been shown in part 3, there is a unique UMP test at level $\alpha \in [0.97, 0.1)$ for $H = \{\mathbb{P}_0\}$ vs $K = \{\mathbb{P}_1, \mathbb{P}_2\}$.

□

Example 56:

	A	B	C	D
$\mathbb{P}_0$	0.05	0.01	0.04	0.90
$\mathbb{P}_1$	0.01	0.02	0.04	0.93
$\mathbb{Q}$	0.30	0.10	0.40	0.20

We are testing $H : \{\mathbb{P}_0, \mathbb{P}_1\}$ vs $K : \mathbb{Q}$ at level α. Find UMP test or show there is no UMP test at level $\alpha = 0.05$. ♠

Proof We first average the null hypothesis by

$$\mathbb{P}_\lambda = (1-\lambda)\mathbb{P}_0 + \lambda\mathbb{P}_1$$

Plot the $L_A(\lambda)$, $L_B(\lambda)$, $L_C(\lambda)$, $L_D(\lambda)$ with $\lambda \in (0,1)$. The intersection of $L_A(\lambda)$ and

	A	B	C	D
$\mathbb{P}_\lambda$	$0.05 - 0.04\lambda$	$0.01 + 0.01\lambda$	0.04	$0.90 + 0.03\lambda$
$\mathbb{Q}$	0.30	0.10	0.40	0.20
$L(\lambda) = \frac{\mathbb{Q}}{\mathbb{P}_\lambda}$	$\frac{30}{5-4\lambda}$	$\frac{10}{1+\lambda}$	10	$\frac{20}{90+3\lambda}$

$L_B(\lambda)$ is $\lambda_1 = \frac{2}{7}$, the intersection of $L_A(\lambda)$ and $L_C(\lambda)$ is $\lambda_2 = \frac{1}{2}$. Then we discuss with respect to values of λ, to determine which points should be included in rejection region.

1. $\lambda \in (0, \lambda_1)$, likelihood ratio $C > B > A > D$, to satisfy $\alpha = 0.05$, the test must be in the form of

$$\phi_\lambda = \begin{cases} 1, & C \\ \gamma_B, & B \\ 0, & A, D \end{cases}$$

   where $\gamma \in [0,1]$ is given by

$$\begin{aligned} \mathbb{E}_0\left[\phi_\lambda\right] &= 0.04 + \gamma_B 0.01 \leq \alpha = 0.05 \\ \mathbb{E}_1\left[\phi_\lambda\right] &= 0.04 + \gamma_B 0.02 \leq \alpha = 0.05 \\ \mathbb{E}_\lambda\left[\phi_\lambda\right] &= (1-\lambda)\mathbb{E}_0\left[\phi_\lambda\right] + \lambda\mathbb{E}_1\left[\phi_\lambda\right] = \alpha \end{aligned}$$

   while there is no solution to this problem. Therefore, no UMP.

2. $\lambda = \lambda_1$, likelihood ratio $C > B = A > D$, to satisfy $\alpha = 0.05$, the test must be in the form of

$$\phi_\lambda = \begin{cases} 1, & C \\ \gamma_A, & A \\ \gamma_B, & B \\ 0, & D \end{cases}$$

   where $\gamma \in [0,1]$ is given by

$$\begin{aligned} \mathbb{E}_0\left[\phi_\lambda\right] &= 0.04 + \gamma_A 0.05 + \gamma_B 0.01 \leq \alpha = 0.05 \\ \mathbb{E}_1\left[\phi_\lambda\right] &= 0.04 + \gamma_A 0.01 + \gamma_B 0.02 \leq \alpha = 0.05 \\ \mathbb{E}_\lambda\left[\phi_\lambda\right] &= (1-\lambda)\mathbb{E}_0\left[\phi_\lambda\right] + \lambda\mathbb{E}_1\left[\phi_\lambda\right] = \alpha \end{aligned}$$

   while there is one possible solution to this problem, $\gamma_A = \frac{1}{9}, \gamma_B = \frac{4}{9}$. Therefore, ϕ_λ is UMP.

3. $\lambda \in (\lambda_1, \lambda_2)$, likelihood ratio $C > A > B > D$, to satisfy $\alpha = 0.05$, the test must be in the form of

$$\phi_\lambda = \begin{cases} 1, & C \\ \gamma, & A \\ 0, & B, D \end{cases}$$

where $\gamma \in [0, 1]$ is given by

$$\begin{aligned} \mathbb{E}_0 [\phi_\lambda] &= 0.04 + \gamma 0.05 \le \alpha \\ \mathbb{E}_1 [\phi_\lambda] &= 0.04 + \gamma 0.01 \le \alpha \\ \mathbb{E}_\lambda [\phi_\lambda] &= (1 - \lambda)\mathbb{E}_0 [\phi_\lambda] + \lambda \mathbb{E}_1 [\phi_\lambda] = \alpha \end{aligned}$$

while there is no solution to this problem. Therefore, no UMP.

4. $\lambda = \lambda_2$, likelihood ratio $C = A > B > D$, to satisfy $\alpha = 0.05$, the test must be in the form of

$$\phi_\lambda = \begin{cases} \gamma_C, & C \\ \gamma_A, & A \\ 0, & B, D \end{cases}$$

where $\gamma_A, \gamma_C \in [0, 1]$ is given by

$$\begin{aligned} \mathbb{E}_0 [\phi_\lambda] &= 0.04\gamma_C + 0.05\gamma_A \le \alpha \\ \mathbb{E}_1 [\phi_\lambda] &= 0.04\gamma_C + 0.01\gamma_A \le \alpha \\ \mathbb{E}_\lambda [\phi_\lambda] &= (1 - \lambda)\mathbb{E}_0 [\phi_\lambda] + \lambda \mathbb{E}_1 [\phi_\lambda] = \alpha \end{aligned}$$

while there is no solution to this problem. Therefore, no UMP.

5. $\lambda \in (\lambda_2, 1)$, likelihood ratio $A > C > B > D$, to satisfy $\alpha = 0.05$, the test must be in the form of

$$\phi_\lambda = \begin{cases} 1, & A \\ \gamma, & C \\ 0, & B, D \end{cases}$$

where $\gamma \in [0, 1]$ is given by

$$\begin{aligned} \mathbb{E}_0 [\phi_\lambda] &= 0.05 + \gamma 0.04 \le \alpha \\ \mathbb{E}_1 [\phi_\lambda] &= 0.01 + \gamma 0.04 \le \alpha \\ \mathbb{E}_\lambda [\phi_\lambda] &= (1 - \lambda)\mathbb{E}_0 [\phi_\lambda] + \lambda \mathbb{E}_1 [\phi_\lambda] = \alpha \end{aligned}$$

while there is no solution to this problem. Therefore, no UMP.

$\square$

Chapter 7

Unbiasedness: Theory and First Applications

7.1 Unbiasedness For Hypothesis Testing

Definition 7.1.1. *(Unbiased Test)*
A test $\phi(x)$, for which the power function $\beta_\phi(\theta) = \mathbb{E}_\theta\left[\phi(X)\right]$ satisfies

$$\begin{aligned} \beta_\phi(\theta) &\leq \alpha, \qquad \theta \in \Omega_H \\ \beta_\phi(\theta) &\geq \alpha, \qquad \theta \in \Omega_K \end{aligned}$$

is said to be unbiased at level α.

Remark Whenever a UMP test exists, it is unbiased, since its power can't fall below the of the test $\phi(x) \equiv \alpha$, by Corollary 6.2.2, that is $\beta_\phi(\theta) = \mathbb{E}_\theta\left[\phi(X)\right] \geq \mathbb{E}_\theta\left[\psi(X)\right] = \alpha$, where $\psi(X) \equiv \alpha$, for all $\theta \in \Omega_K$.

Definition 7.1.2. *(UMP Unbiased Test)*
UMPU test is the uniformly most powerful test among unbiased test at level α.

Remark For large class problems where a UMP test does not exist, there does exist a UMP unbiased test.

When $\beta_\phi(\theta)$ is continuous with respect to θ, the unbiasedness implies

$$\beta_\phi(\theta) = \alpha, \quad \forall \theta \in \omega.$$

where ω is the common boundary of Ω_H and Ω_K, that is, the sets of points θ that are points or limit points in $\Omega_H \cap \Omega_K$. Tests satisfying the condition are said to be similar on the boundary(of H and K).

Lemma 7.1.1. *(Lemma 4.1.1 in TSH)*
If the distributions $\mathbb{P}_\theta$ are such that power function of every test is continuous, and if ϕ_0 is a level α test of H, and ϕ_0 is UMP among all the tests satisfying

$$\beta_\phi(\theta) = \alpha, \quad \forall \theta \in \omega.$$

then ϕ_0 is UMP unbiased.

Proof The class of tests satisfying conditions above contains the class of unbiased tests, and hence ϕ_0 is uniformly at least as powerful as any unbiased test. On the other hand, ϕ_0 is unbiased, since it is uniformly at least as powerful as $\phi(x) \equiv \alpha$. □

7.2 One-Parameter Exponential Families

Theorem 7.2.1. *(TSH Page 111)*

Let $\boldsymbol{X} = (X_1, \cdots, X_n)$ be a random vector with probability density(with respect to some probability measure μ), where $\theta \in \Omega \subset \mathbb{R}$

$$p_\theta(x) = C(\theta)e^{\theta T(x)}h(x)$$

1. *Given $\theta_1 < \theta_2$, consider testing $H : \theta \in [\theta_1, \theta_2]$ vs $K : \theta \notin [\theta_1, \theta_2]$, there is a UMPU test at level α in the form of*

$$\phi(x) = \begin{cases} 1, & T(x) < c_1 \text{ or } T(x) > c_2 \\ \gamma_1, & T(x) = c_1 \\ \gamma_2, & T(x) = c_2 \\ 0, & \text{otherwise} \end{cases}$$

where c_1, c_2, γ_1, γ_2 are determined by $\mathbb{E}_{\theta_1}[\phi(X)] = \mathbb{E}_{\theta_2}[\phi(X)] = \alpha$

2. *Given $\theta_0 \in \Omega$, consider testing $H : \theta = \theta_0$ vs $K : \theta \neq \theta_0$, then there is a UMPU test at level α in the form of*

$$\phi(x) = \begin{cases} 1, & T(x) < c_1 \text{ or } T(x) > c_2 \\ \gamma_1, & T(x) = c_1 \\ \gamma_2, & T(x) = c_2 \\ 0, & \text{otherwise} \end{cases}$$

where c_1, c_2, γ_1, γ_2 are determined by $\mathbb{E}_{\theta_0}[\phi(X)] = \alpha$ and $\mathbb{E}_{\theta_0}[T(X)]\,\alpha = \mathbb{E}_{\theta_0}[T(X)\phi(X)]$

Example 57: (Example 4.2.2 (Normal variance) in TSH)

Let $X_1, \cdots, X_n \overset{i.i.d.}{\sim} \mathsf{N}(0, \sigma^2)$. We want to test $\sigma = \sigma_0$ vs $\sigma \neq \sigma_0$. The likelihood is

$$\left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^n x_i^2\right)$$

Then $T(X) = \sum_{i=1}^n X_i^2$ is sufficient for σ^2, and has probability density $\frac{1}{\sigma^2} f_n(\frac{y}{\sigma^2})$, where

$$f_n(y) = \frac{1}{2^{n/2}\Gamma(n/2)} y^{(n/2)-1} e^{(y/2)}, \quad y > 0$$

is the density of a χ^2_n with n degrees of freedom. By theorem 7.2.1, there is a UMPU test in the form of

$$\phi(x) = \begin{cases} 1, & T(x) \leq c_1 \text{ or } T(x) \geq c_2 \\ 0, & \text{otherwise} \end{cases}$$

where c_1, c_2 are determined by

$$\begin{aligned}\mathbb{E}_{\sigma_0}\left[\phi(X)\right] &= \alpha \\ \mathbb{E}_{\sigma_0}\left[T(X)\phi(X)\right] &= \mathbb{E}_{\sigma_0}\left[T(X)\right]\alpha\end{aligned}$$

Under $H_0 : \sigma = \sigma_0$, $T(X) \sim \sigma_0^2\chi_n^2$. Set $c_1 = \sigma_0^2 B_1$, $c_2 = \sigma_0^2 B_2$, then

$$\begin{aligned}\int_{B_1}^{B_2} f_n(t)dt &= 1-\alpha \\ \int_{B_1}^{B_2} tf_n(t)dt &= \frac{(1-\alpha)\mathbb{E}_{\sigma_0}\left[\sum_{i=1}^n X_i^2\right]}{\sigma_0^2} = n(1-\alpha)\end{aligned}$$

Actually, unless n is very small or σ_0 very close to 0 or ∞, the equaltails test given by

$$\mathbb{P}_{\sigma_0}(T \leq c_1) = \int_0^{B_1} f_n(t)dt = \mathbb{P}_{\sigma_0}(T \geq c_2) = \int_{B_2}^{\infty} f_n(t)dt = \frac{\alpha}{2}$$

which gives us $c_1 = \sigma_0^2\chi^2_{n,\alpha/2}$, $c_2 = \sigma_0^2\chi^2_{n,1-\alpha/2}$. ♠

Example 58: (Problem 4.7 in TSH)

Let X and Y be independently distributed according to one-parameter exponential families, which means

$$\begin{aligned}X &\sim f_\theta(x) \propto e^{\theta T(x)}, \qquad \theta \in \Omega \subset \mathbb{R} \\ Y &\sim g_\gamma(y) \propto e^{\gamma U(y)}, \qquad \gamma \in \Gamma \subset \mathbb{R}\end{aligned}$$

so that their joint distribution is given by

$$d\mathbb{P}_{\theta,\gamma}(x,y) = C(\theta)e^{\theta T(x)}d\mu(x)K(\gamma)e^{\gamma U(y)}d\nu(y)$$

Suppose that with probability 1 the statistics $T(X)$ and $U(Y)$ each take on at least three values and that (θ_0, γ_0) is an interior point of the natural parameter space. Then there is no UMP unbiased test for $H : \theta = \theta_0$ and $\gamma = \gamma_0$ vs $K : \theta \neq \theta_0$ or $\gamma \neq \gamma_0$ at level α if T and U have continuous distributions.

why? [The most powerful unbiased tests against the alternatives $\theta \neq \theta_0, \gamma \neq \gamma_0$ have acceptance regions $C_1 < T(x) < C_2$ and $K_1 < U(y) < K_2$ respectively. These tests are also unbiased against the wider class of alternatives $K : \theta \neq \theta_0$ or $\gamma \neq \gamma_0$ or both.] ♠

Example 59: (Problem 4.8 in TSH) Let (X, Y) be distributed according to the exponential family

$$d\mathbb{P}_{\theta,\gamma}(x,y) = C(\theta,\gamma)e^{\theta x+\gamma y}d\mu(x,y)$$

The only unbiased test for testing $H : \theta \leq \theta_0$ and $\gamma \leq \gamma_0$ against $K : \theta > \theta_0$ or $\gamma > \gamma_0$ or both is $\phi(x,y) \equiv \alpha$.

why? [Take $\theta_0 = \gamma_0 = 0$, and let $\beta(\theta,\gamma)$ be the power function of any level-α test. Unbiasedness implies $\beta(0,\gamma) = \alpha$ for $\gamma < 0$ and hence for all γ, since $\beta(0,\gamma)$ is an analytic function of γ. For fixed $\gamma > 0$, $\beta(\theta, \gamma)$ considered as a function of θ therefore has a minimum at $\theta = 0$, so that $\partial\beta(\theta, \gamma)/\partial\theta$ vanishes at $\theta = 0$ for all positive γ, and hence for all γ. By considering alternatively positive and negative values of γ and using the fact that the partial derivatives of all orders of $\beta(\theta, \gamma)$ with respect to γ are analytic, one finds that for each fixed γ these derivatives all vanish at $\theta = 0$ and hence that the function $\beta(\theta, \gamma)$ must be a constant. Because of the completeness of (X, Y), $\beta(\theta, \gamma) \equiv \alpha$ implies $\phi(x, y) \equiv \alpha$.] ♠

7.3 Similarity and Completeness

Theorem 7.3.1. *(Theorem 4.3.1 in TSH)*
Let X be a random vector with probability distribution

$$d\mathbb{P}_\theta(x) = C(\theta) \exp\left[\sum_{j=1}^{k} \theta_j T_j(x)\right] d\mu(x)$$

and let $\mathcal{P}^T$ be the family of distributions of $T = (T_1(X), \cdots, T_k(X))$ as θ ranges over the set ω. Then $\mathcal{P}^T$ is complete provided ω contains a $k-$ dimensional rectangle.

Theorem 7.3.2. *(Theorem 4.3.2 in TSH)*
Let X be a random vector with probability distribution $\mathbb{P} \in \mathcal{P}$, and let T be a sufficient statistic for $\mathcal{P}$. Then a necessary and sufficient condition for all similar tests to have Neyman structure with respect to T is that the family $\mathcal{P}^T$ of distributions of T is boundedly complete.

7.4 UMP Unbiased Tests for Multiparameter Exponential Families

Theorem 7.4.1. *(Theorem 4.4.1 in TSH)*
Let X be distributed according to

$$d\mathbb{P}^X_{\theta,\xi}(x) = C(\theta, \xi) \exp\left[\theta U(x) + \sum_{i=1}^{k} \xi_i T_i(x)\right] d\mu(x) = C(\theta, \xi) \exp\left[\theta U(x) + \langle \xi, T(x)\rangle\right] d\mu(x)$$

with $\theta \in \mathbb{R}$, $\xi = (\xi_1, \cdots, \xi_k) \in \mathbb{R}^k$ and $T = (T_1, \cdots, T_k) \in \mathbb{R}^k$. Assume that Ω is convex and Ω has non-empty interior, i.e., it is not contained in a linear space of dimension $< k + 1$. Consider the following testing problems at level α

$$\begin{array}{lcl}
H_1 : & \theta \leq \theta_0 & K_1 : \theta > \theta_0 \\
H_2 : & \theta \leq \theta_1 \text{ or } \theta \geq \theta_2 & K_2 : \theta_1 < \theta < \theta_2 \\
H_3 : & \theta_1 \leq \theta \leq \theta_2 & K_3 : \theta < \theta_1 \text{ or } \theta > \theta_2 \\
H_4 : & \theta = \theta_0 & K_4 : \theta \neq \theta_0
\end{array}$$

The idea is to conditioning on T to estimate ξ. Note that

- *(U,T) is sufficient statistic for (θ,ξ), since the joint distribution*

$$d\mathbb{P}_{\theta,\xi}^{U,T}(u,t) = C(\theta,\xi)\exp\left[\theta U(x) + \sum_{i=1}^{k}\xi_i t_i\right] d\nu(u,t)$$

Once $T=t$ is given, U is the only remaining variable and the conditional distribution of U given t constitutes an exponential family

$$d\mathbb{P}_{\theta}^{U|t}(u) = C_t(\theta)e^{\theta u}d\nu_t(u)$$

1. *For $H_1:\theta\le\theta_0$ vs $K_1:\theta>\theta_0$, the UMPU is in the form of*

$$\phi_1(u,t) = \begin{cases} 1, & u > C_0(t) \\ \gamma_0(t), & u = C_0(t) \\ 0, & u < C_0(t) \end{cases}$$

   where $C_0(t)$ and $\gamma_0(t)$ are determined by

$$\mathbb{E}_{\theta_0}\left[\phi_1(U,T)|t\right] = \alpha \quad \forall t$$

2. *For $H_2:\theta\le\theta_1$ or $\theta\ge\theta_2$ vs $K_2:\theta_1<\theta<\theta_2$, the UMPU is in the form of*

$$\phi_2(u,t) = \begin{cases} 1, & C_1(t) < u < C_2(t) \\ \gamma_1(t), & u = C_1(t) \\ \gamma_2(t), & u = C_2(t) \\ 0, & u < C_1(t) \text{ or } u > C_2(t) \end{cases}$$

   where $C_1(t)$, $C_2(t)$, $\gamma_1(t)$and $\gamma_2(t)$ are determined by

$$\mathbb{E}_{\theta_1}\left[\phi_2(U,T)|t\right] = \mathbb{E}_{\theta_2}\left[\phi_2(U,T)|t\right] = \alpha \quad \forall t$$

3. *For $H_3:\theta_1\le\theta\le\theta_2$ vs $K_3:\theta<\theta_1$ or $\theta>\theta_2$, the UMPU is in the form of*

$$\phi_3(u,t) = \begin{cases} 1, & u < C_1(t) \text{ or } u > C_2(t) \\ \gamma_1(t), & u = C_1(t) \\ \gamma_2(t), & u = C_2(t) \\ 0, & C_1(t) < u < C_2(t) \end{cases}$$

   where $C_1(t)$, $C_2(t)$, $\gamma_1(t)$and $\gamma_2(t)$ are determined by

$$\mathbb{E}_{\theta_1}\left[\phi_3(U,T)|t\right] = \mathbb{E}_{\theta_2}\left[\phi_3(U,T)|t\right] = \alpha \quad \forall t$$

4. *For $H_4:\theta=\theta_0$ vs $K_4:\theta\ne\theta_0$, the UMPU is in the form of*

$$\phi_4(u,t) = \begin{cases} 1, & u < C_1(t) \text{ or } u > C_2(t) \\ \gamma_1(t), & u = C_1(t) \\ \gamma_2(t), & u = C_2(t) \\ 0, & C_1(t) < u < C_2(t) \end{cases}$$

   where $C_1(t)$, $C_2(t)$, $\gamma_1(t)$and $\gamma_2(t)$ are determined by

$$\mathbb{E}_{\theta_0}\left[\phi_4(U,T)|t\right] = \alpha \quad \forall t$$
$$\mathbb{E}_{\theta_0}\left[U\phi_4(U,T)|t\right] = \alpha\mathbb{E}_{\theta_0}\left[U|t\right] \quad \forall t$$

Remark Through a transformation of parameters, theorem 7.4.1 can be extended to cover hypotheses concerning parameters of the form

$$\theta^* = a\theta + \sum_{i=1}^{k} b_i \xi_i$$

which means that we can test linear combination of parameters such as $\theta^* = a\theta + b^T\xi$ with $a \neq 0$ and $b \in \mathbb{R}^k$ because $(U/a, T_1 - \frac{b_1 U}{a}, \cdots, T_k - \frac{b_k U}{a})$ satisfies the same conditions as (U, T), which gives us the following lemma.

Lemma 7.4.2. *(Lemma 4.4.1 in TSH)*
The exponential family of distributions can also be written as

$$d\mathbb{P}^X_{\theta,\xi}(x) = K(\theta^*, \xi)\exp\left[\theta^* U^*(x) + \sum_{i=1}^{k} \xi_i T_i^*(x)\right] d\mu(x) = K(\theta^*, \xi)\exp\left[\theta^* U^*(x) + \langle \xi, T^*(x)\rangle\right] d\mu(x)$$

where

$$U^* = \frac{U}{a}, \quad T_i^* = T_i - \frac{b_i}{a}U$$

Application of theorem 7.4.1 to the form of the distributions given in the lemma leads to UMP unbiased tests of the hypothesis $H_1^* : \theta^* \leq \theta_0$ and the analogously defined hypotheses H_2^*, H_3^* and H_4^*.

7.5 Comparing Two Poisson or Binomial Populations

Example 60: (Example in TSH Page 125)

Let $X_1, \cdots, X_m \overset{i.i.d.}{\sim} Poisson(\lambda_0)$ and $Y_1, \cdots, Y_n \overset{i.i.d.}{\sim} Poisson(\mu_0)$, $\{X_i\}_{i=1}^m$ and $\{Y_i\}_{i=1}^n$ are independent. We want to test $H : \mu_0 \leq \lambda_0$ vs $K : \mu_0 > \lambda_0$. Let $\lambda = m\lambda_0$ and $\mu = n\mu_0$

By sufficiency, we may work with (X, Y), which is defined by

$$X := \sum_{i=1}^{m} X_i \sim Poisson(m\lambda_0) = Poisson(\lambda)$$
$$Y := \sum_{i=1}^{n} Y_i \sim Poisson(n\mu_0) = Poisson(\mu)$$

Then $\mu_0 \leq \lambda_0 \iff \frac{\mu}{\lambda} \leq \frac{n}{m} \iff \theta = \ln(\mu) - \ln(\lambda) \leq \ln(\frac{n}{m}) = \theta_0$. Then we transform the hypothesis to $H : \theta \leq \theta_0$ vs $K : \theta > \theta_0$. Consider the probability

$$\mathbb{P}_{\lambda,\mu}(X = x, Y = y) = e^{-\lambda}\frac{\lambda^x}{x!}e^{-\mu}\frac{\mu^y}{y!} = \exp\left[x\ln\lambda + y\ln\mu\right]\frac{e^{-\lambda}e^{-\mu}}{x!y!}$$

Let $\xi = \ln\lambda$, $u = y$, $t = x + y$, then

$$x\ln\lambda + y\ln\mu = y(\ln\mu - \ln\lambda) + (x + y)\ln\lambda = u\theta + t\xi$$

Then by first item in theorem 7.4.1, for $H : \theta \le \theta_0$ vs $K : \theta > \theta_0$, the UMPU is in the form of

$$\phi_1(u,t) = \begin{cases} 1, & u > C_0(t) \\ \gamma_0(t), & u = C_0(t) \\ 0, & u < C_0(t) \end{cases}$$

where $C_0(t)$ and $\gamma_0(t)$ are determined by

$$\mathbb{E}_{\theta_0}\left[\phi_1(U,T)|t\right] = \alpha \quad \forall t$$

Since

$$\begin{aligned} \mathbb{P}_{\lambda,\mu}(U = u|T = t) &= \frac{\mathbb{P}_{\lambda,\mu}(U = u, T = t)}{\mathbb{P}_{\lambda,\mu}(T = t)} = \frac{\mathbb{P}_{\lambda,\mu}(Y = u, X + Y = t)}{\mathbb{P}_{\lambda,\mu}(X + Y = t)} \\ &= \frac{\mathbb{P}_{\lambda,\mu}(Y = u, X = t - u)}{\mathbb{P}_{\lambda,\mu}(X + Y = t)} = \frac{e^{-\lambda}\lambda^{t-u}}{(t-u)!}\frac{e^{-\mu}\mu^u}{u!}\left(\frac{e^{-(\lambda+\mu)}(\lambda+\mu)^t}{t!}\right)^{-1} \\ &= \binom{t}{u}\left(\frac{\mu}{\lambda+\mu}\right)^u\left(1 - \frac{\mu}{\lambda+\mu}\right)^{t-u} \end{aligned}$$

Under $\mu_0 = \lambda_0$, we know $\frac{\mu}{\lambda} = \frac{n}{m}$, and

$$U|T = t \sim Binomial(t, \frac{\mu}{\lambda+\mu}) = Binomial(t, \frac{n}{m+n})$$

Thus it is equivalent to test $Binomial(t, p)$ for $H : p \le \frac{n}{m+n}$ and $K : p > \frac{n}{m+n}$. ♠

Example 61: (Example in TSH Page 126)

Let X and Y be independent binomial variables, where $X \sim Binomial(m, p)$ and $Y \sim Binomial(n, q)$, with joint distribution

$$\begin{aligned} \mathbb{P}_{p,q}(X = x, Y = y) &= \binom{m}{x}p^x(1-p)^{m-x}\binom{n}{y}q^y(1-q)^{n-y} \\ &= \exp\left(x\log\frac{p}{1-p} + y\log\frac{q}{1-q}\right)(1-p)^m(1-q)^n\binom{m}{x}\binom{n}{y} \\ &= \exp\left(y\log\frac{q/(1-q)}{p/(1-p)} + (x+y)\log\frac{p}{1-p}\right)(1-p)^m(1-q)^n\binom{m}{x}\binom{n}{y} \\ &= \exp\left(u\theta + t\xi\right)(1-p)^m(1-q)^n\binom{m}{x}\binom{n}{y} \end{aligned}$$

Let $U = Y$, $\theta = \log\frac{q/(1-q)}{p/(1-p)}$, $T = X + Y$, $\xi = \log\frac{p}{(1-p)}$. Define odds ratio (also called cross-product ratio)

$$\rho = \frac{q/(1-q)}{p/(1-p)}$$

Suppose that we want to test $H_1 : q \le p$ vs $K_1 : q > p$, which can be transform to $H_1 : \theta \le 0$ vs $K_1 : \theta > 0$. Consider the conditional probability under $p = q \Longleftrightarrow \theta = 0$, by independence of X and Y, and since $p = q$, $X + Y \sim Binomial(m + n, p)$

$$\mathbb{P}_{p,p}(Y = y|X + Y = t) = \frac{\mathbb{P}_{p,p}(Y = y, X + Y = t)}{\mathbb{P}_{p,p}(X + Y = t)} = \frac{\mathbb{P}_p(X = t - y)\mathbb{P}_p(Y = y)}{\mathbb{P}_{p,p}(X + Y = t)}$$

$$= \frac{\binom{m}{t-y} p^{t-y}(1-p)^{m-t+y}\binom{n}{y}p^y(1-p)^{n-y}}{\binom{m+n}{t}p^t(1-p)^{m+n-t}} = \frac{\binom{m}{t-y}\binom{n}{y}}{\binom{m+n}{t}}$$

which obeys the Hyper-geometric distribution $(t, m, m + n)$.

The UMP unbiased test of $\rho_1 = \rho_2$, which is based on the (conditional) hypergeometric distribution, requires randomization to obtain an exact conditional level α for each t of the sufficient statistic T. Since in practice randomization is usually unacceptable, the one-sided test is frequently performed as

$$\phi_1(u, t) = \begin{cases} 1, & u > C(t) \\ 0, & u < C(t) \end{cases}$$

the test rejects when $Y \geq C(t)$, where $C(t)$ is the smallest integer for which $\mathbb{P}[Y \geq C(t)|T = t] \leq \alpha$. This conservative test is called Fisher's exact test. ♠

7.6 Testing for Independence in a 2×2 Table

Example 62: (Example in TSH Page 127)

Two characteristics A and B, which each member of a population may or may not possess, are to be tested for independence, where the random variable (U, V) take values (u, v) with $u \in \{A, A^c\}$, $v \in \{B, B^c\}$ The probabilities or proportion of individuals possessing properties A and B are denoted $\mathbb{P}(A)$ and $\mathbb{P}(B)$.

1. Desigin 1:

   - We sample m individuals with $v = B$, and there are x in m with $u = A$
   - We sample n individuals with $v = B^c$, and there are y in m with $u = A$

   Therefore, $X \sim Binomial(m, p(A|B))$, $Y \sim Binomial(n, p(A|B^c))$, where $p(u|v) = \mathbb{P}[U = u|V = v]$ with $X \perp Y$. Thus $U \perp V \iff \mathbb{P}[u = A|v = B] = \mathbb{P}[u = A|v = B^c]$.

   Now we transform it to comparing two Binomial Population, which is the same as previous problem.

2. Design 2: We sample s individuals from the population, as shown in the table below

	A	A^c	Total
B	X	X'	$M = X + X'$
B^c	Y	Y'	$N = Y + Y'$
Total	$T = X + Y$	$T' = Y' + Y'$	s

where $X = \#\{i : u_i = A, v_i = B\}$. Observe that (X, X', Y, Y') is sufficient, which has multinomial distribution with parameter (s, p), where $p = (p_{AB}, p_{A^cB}, p_{AB^c}, p_{A^cB^c})$. Note that $p_{AB} + p_{A^cB} + p_{AB^c} + p_{A^cB^c} = 1$, $s = X + X' + Y + Y'$, while

$$\mathbb{P}_p(X = x, X = x', Y = y, Y = y') = \frac{s!}{[x!] \cdot [x'!] \cdot [y!] \cdot [y'!]}(p_{AB})^x (p_{A^cB})^{x'} (p_{AB^c})^y (p_{A^cB^c})^{y'}$$

Since

$$\begin{aligned} & (p_{AB})^x (p_{A^cB})^{x'} (p_{AB^c})^y (p_{A^cB^c})^{y'} \\ = \; & \exp\left(x \log p_{AB} + x' \log p_{A^cB} + y \log p_{AB^c} + y' \log p_{A^cB^c}\right) \\ = \; & p_{A^cB^c}^s \exp\left(x \log \frac{p_{AB}p_{A^cB^c}}{p_{A^cB}p_{AB^c}} + (x + x') \log \frac{p_{A^cB}}{p_{A^cB^c}} + (x + y) \log \frac{p_{AB^c}}{p_{A^cB^c}} \right) \end{aligned}$$

Let

$$\theta = \log \frac{p_{AB}p_{A^cB^c}}{p_{A^cB}p_{AB^c}}$$
$$\theta = 0 \Longleftrightarrow p(A|B) = p(A|B^c) \Longleftrightarrow U \perp V$$

Note that

$$\begin{aligned} p_{A\cdot} &= p_{AB} + p_{AB^c}, & p_{\cdot B} &= p_{AB} + p_{A^cB} \\ p_{A^c\cdot} &= p_{A^cB} + p_{A^cB^c}, & p_{\cdot B^c} &= p_{AB^c} + p_{A^cB^c} \\ p(A|B) &= \tfrac{p_{AB}}{p_{\cdot B}}, & p(A|B^c) &= \frac{p_{AB^c}}{p_{\cdot B^c}} \end{aligned}$$

Suppose that we want to test $H_1 : \theta \le 0$ vs $K_1 : \theta > 0$. Let $M = X + X'$, $T = X + Y$. Given $M = m$, $T = t$, note that $Y + Y' = n = s - m$

$$\begin{aligned} & \mathbb{P}_p(X = x | M = m, T = t) = \mathbb{P}_p(X = x | X + X' = m, X + Y = t) \\ = \; & C_t(\rho)\binom{m}{x}\binom{n}{t-x}\rho^{t-x}, \quad x = 0, \cdots, t, \quad \rho = \frac{p_{A^cB}p_{AB^c}}{p_{AB}p_{A^cB^c}} \end{aligned}$$

which indicates that $X|M = m, T = t \overset{\theta=0}{\sim} Hyper - geometric(m, n, t)$, which is similar to the problem two Binomial populations, the test is

$$\phi_1(y, t) = \begin{cases} 1, & y > C(t) \\ 0, & y < C(t) \end{cases}$$

the test rejects when $Y \ge C(t)$, where $C(t)$ is the smallest integer for which $\mathbb{P}\left[Y \ge C(t) | T = t\right] \le \alpha$.

Remark Design 1 is superior than design 2, for example when

- If samples of size m and $n(m + n = s)$ are taken from B and B^c or from A and A^c, the best choice of m and n is $m = n = s/2$, and
- It is better to select samples of equal size $s/2$ from B and B^c than from A and A^c provided $|p_B - \frac{1}{2}| > |p_A - \frac{1}{2}|$, and
- Selecting the sample at random from the population at large is worse than taking equal samples either from A and A^c or from B and B^c.

♠

7.7 Alternative Models for 2×2 Tables

Example 63: (Matched Pair Cross-over Designs, McNemar's test, Page 138 in TSH)

Suppose that A and B are two treatments, either of which can be assigned to each subject, as shown in the table below

Treatment	A(Success)	A^c(Failure)	Total
B(Success)	X	X'	$M = X + X'$
B^c(Failure)	Y	Y'	$N = Y + Y'$
Total	$T = X + Y$	$T' = Y' + Y'$	s

The null hypothesis of marginal homogeneity states that the two marginal probabilities for each outcome are the same, *i.e.*

$$H : p_{A\cdot} = p_{AB} + p_{AB^c} = p_{AB} + p_{A^cB} = p_{\cdot B^c} \Longleftrightarrow H : p_{AB^c} = p_{A^cB}$$

Thus the null and alternative hypotheses are

$$\begin{aligned} H_1 &: p_{AB^c} = p_{A^cB} \\ K_1 &: p_{AB^c} \neq p_{A^cB} \end{aligned}$$

The likelihood

$$\begin{aligned} &\mathbb{P}_p(X = x, X = x', Y = y, Y = y') \\ = &\frac{s!}{[x!] \cdot [x'!] \cdot [y!] \cdot [y'!]} (p_{A^cB^c})^s \left(\frac{p_{AB}}{p_{A^cB^c}}\right)^x \left(\frac{p_{A^cB}}{p_{A^cB^c}}\right)^{x'} \left(\frac{p_{AB^c}}{p_{A^cB^c}}\right)^y \\ = &\frac{s!}{[x!] \cdot [x'!] \cdot [y!] \cdot [y'!]} (p_{A^cB^c})^s \exp\left(y \log\left(\frac{p_{AB^c}}{p_{A^cB}}\right) + (x' + y)\log\left(\frac{p_{A^cB}}{p_{A^cB^c}}\right) + x \log\left(\frac{p_{AB}}{p_{A^cB^c}}\right)\right) \end{aligned}$$

Let $D = X' + Y$, $\theta = \log\left(\frac{p_{ABc}}{p_{A^cB}}\right)$. There exists a UMP unbiased test, McNemar's test, which rejects H in favor of the alternatives $p_{AB^c} < p_{A^cB}$ when $Y > C(X' + Y, X)$, where the conditional probability of rejection given $X' + Y = d$ and $X = x$ is α for all d and x. Under this condition, the numbers of pairs (A^c, B^c) and (A, B) are fixed, and the only remaining variables are Y and $X' = d - Y$ which specify the division of the d cases with mixed response between the outcomes (A^c, B) and (A, B^c). Conditionally, one is dealing with d binomial trials with success probability $p = \frac{p_{ABc}}{p_{AB^c}+p_{A^cB}}$, H becomes $p = \frac{1}{2}$, and the UMP unbiased test reduces to the sign test, *i.e.*$Y|(D = d, X = x) = Y|(Y' = y' = s - d - x, X = x) \sim Binomial(d, \frac{p_{ABc}}{p_{AB^c}+p_{A^cB}})$. ♠

7.8 The Sign Test

Example 64: (Example in TSH Page 135)

Let $X_1, \cdots, X_n \overset{i.i.d.}{\sim} F$ on $\mathbb{R}$.

1. For $p = \mathbb{P}_F(X \leq 0)$, we test $H : p \leq p_H$ vs $K : p > p_H$, while p_H is given. By example 52 , the test, which rejects large values of $M = \#\{i : X_i \geq 0\}$ is UMP.

2. Let

$$\begin{cases} p_- = \mathbb{P}_F(X < 0) \\ p_+ = \mathbb{P}_F(X > 0) \\ p_0 = \mathbb{P}_F(X = 0) \end{cases} \quad \begin{cases} F_-(x) = \mathbb{P}_F(X < x | X < 0) \\ F_+(x) = \mathbb{P}_F(X < x | X > 0) \end{cases}$$

We are going to test

$$H : p_+ \leq p_-$$

Note that any distribution of X can be specified by $(p_0, p_-, p_+, F_-, F_+)$. Consider any fixed distributions F_-^* and F_+^* with relevant densities ρ_-^* and ρ_+^*, and denote the family

$$\mathcal{F}_0 = \{F \text{ on } \mathbb{R}, F_- = F_-^*, F_+ = F_+^*\} \subset \mathcal{F} = \{F \text{ on } \mathbb{R}, F_-, F_+ \text{unknown}\}$$

and arbitrary p_-, p_+, p_0. Any test that is unbiased for testing H in the original family of distributions $\mathcal{F}$ is also unbiased for testing H in the smaller family $\mathcal{F}_0$.

To determine the UMP unbiased test of H in $\mathcal{F}_0$, the joint density of the X's at a point $(x_1, \cdots, x_n)$ with

$$x_{i_1}, \cdots, x_{i_r} < 0 = x_{j_1} = \cdots = x_{j_s} < x_{k_1}, \cdots, x_{k_m}$$

is

$$p_-^r p_0^s p_+^m \rho_-^*(x_{i_1}) \cdots \rho_-^*(x_{i_r}) \rho_+^*(x_{k_1}) \cdots \rho_+^*(x_{k_m})$$

Let

$$\begin{cases} R := \#\{i : X_i < 0\} \\ S := \#\{i : X_i = 0\} \\ M := \#\{i : X_i > 0\} \end{cases}$$

By factorization lemma, the statistics (R, S, M) are jointly sufficient for (p_-, p_0, p_+), which are distributed according to the multinomial distribution $multinomial(n, p_-, p_0, p_+)$, with $r + s + m = n$, $p_- + p_0 + p_+ = 1$, *i.e.*, the density

$$\begin{aligned} \frac{n!}{r!s!m!} p_-^r p_0^s p_+^m &= \frac{n!}{r!s!m!} \left(\frac{p_0}{1 - p_0 - p_+} \right)^r \left(\frac{p_+}{1 - p_0 - p_+} \right)^m (1 - p_0 - p_+)^n \\ = & \frac{n!}{r!s!m!} \exp\left(r \log\left(\frac{p_0}{1 - p_0 - p_+} \right) + m \log\left(\frac{p_+}{1 - p_0 - p_+} \right) \right) (1 - p_0 - p_+)^n \end{aligned}$$

and is then seen to constitute an exponential family with $U = M$, $T = R$, $\theta = \log\left(\frac{p_+}{1-p_0-p_+}\right)$, $\nu = \log\left(\frac{p_0}{1-p_0-p_+}\right)$. Rewriting the hypothesis

$$H : p_+ \leq p_- \Longleftrightarrow H : p_+ \leq 1 - p_0 - p_+$$

it is seen to be equivalent to $\theta = 0$. There exists therefore a UMP unbiased test of H, which is obtained by considering s as fixed and determining the best unbiased conditional test of H given $S = s$. Since the conditional distribution of $M|S = s \sim Binomial(n - s, p)$ with $p = \frac{p_+}{p_+ + p_-}$, the problem reduces to that of testing the hypothesis $p = \frac{1}{2}$ in a binomial distribution with $n - s$ trials. By theorem 4.4.1, the UMP unbiased test for the class $\mathcal{F}_0$ is in the form of

$$\phi(s,m) = \begin{cases} 1, & m > C(s) \\ \gamma(s), & m = C(s) \\ 0, & \text{otherwise} \end{cases}$$

where $\gamma(s)$ and $C(s)$ are determined by $\mathbb{E}_{p_+=p_-}[\phi(M,S)|s] = \alpha$. Note that

- $\phi(s,m)$ is UMP unbiased test for the family $\mathcal{F}_0$.
- Any test that is unbiased for testing H in the original family of distributions $\mathcal{F}$ is also unbiased for testing H in the smaller family $\mathcal{F}_0$.
- It turns out that $\phi(s,m)$ is also unbiased for testing H in $\mathcal{F}$ and is independent of F_-^* , F_+^*.
- Let ψ be any other unbiased test of H in $\mathcal{F}$, and consider any fixed alternative, which without loss of generality can be assumed to be in $\mathcal{F}_0$. Since ψ is unbiased for $\mathcal{F}$, it is unbiased for testing $p_+ \leq p_-$ in $\mathcal{F}_0$; the power of ϕ against the particular alternative is therefore at least as good as that of ψ. Hence ϕ is UMP unbiased.

♠

Chapter 8

Unbiasedness: Applications to Normal Distributions

8.1 Statistics Independent of a Sufficient Statistic

Theorem 8.1.1. *(Theorem 5.1.1 in TSH)*
Suppose that the distribution of X is given by

$$d\mathbb{P}_{\theta,\nu}(x) = C(\theta,\nu)\exp\left(\theta U(x) + \sum_{i=1}^{n}\nu_i T_i(x)\right) d\mu(x)$$

1. *Assuming that $V = h(U,T)$ is independent of T at $\theta = \theta_0$ and h is increasing in u for each t, then the UMP unbiased test for $H_1 : \theta \le \theta_0$ vs $K_1 : \theta > \theta_0$, is in the form of*

$$\psi_1(v) = \begin{cases} 1, & v > C \\ \gamma, & v = C \\ 0, & \text{otherwise} \end{cases}$$

   where γ and C are determined by $\mathbb{E}_{\theta_0}[\psi_1(V)] = \alpha$

2. *Assuming that $V = h(U,T)$ is independent of T at $\theta = \theta_1$ and $\theta = \theta_2$, while h is increasing in u for each t, then the UMP unbiased test for $H_2 : \theta \le \theta_1$ or $\theta \ge \theta_2$ vs $K_2 : \theta_1 < \theta < \theta_2$, is in the form of*

$$\psi_2(v) = \begin{cases} 1, & C_1 < v < C_2 \\ \gamma_1, & v = C_1 \\ \gamma_2, & v = C_2 \\ 0, & \text{otherwise} \end{cases}$$

   where γ and C_1, C_2 are determined by $\mathbb{E}_{\theta_1}[\psi_2(V)] = \mathbb{E}_{\theta_2}[\psi_2(V)] = \alpha$.

3. *Assuming that $V = h(U,T)$ is independent of T at $\theta = \theta_1$ and $\theta = \theta_2$, while h is increasing in u for each t, then the UMP unbiased test for $H_3 : \theta_1 \le \theta \le \theta_2$ vs $K_3 : \theta < \theta_1$ or $\theta > \theta_2$, is in the form of*

$$\psi_3(v) = \begin{cases} 1, & v < C_1 \text{ or}, v > C_2 \\ \gamma_1, & v = C_1 \\ \gamma_2, & v = C_2 \\ 0, & \text{otherwise} \end{cases}$$

where γ and C_1, C_2 are determined by $\mathbb{E}_{\theta_1}[\psi_3(V)] = \mathbb{E}_{\theta_2}[\psi_3(V)] = \alpha$.

4. *Assuming that $V = h(U,T)$ is independent of T at $\theta = \theta_0$, while $h(u,t) = a(t)u + b(t)$ with $a(t) > 0$, then the UMP unbiased test for $H_4 : \theta = \theta_0$ vs $K_4 : \theta \neq \theta_0$, is in the form of*

$$\psi_3(v) = \begin{cases} 1, & v < C_1 \text{ or}, v > C_2 \\ \gamma_1, & v = C_1 \\ \gamma_2, & v = C_2 \\ 0, & \text{otherwise} \end{cases}$$

where γ and C_1, C_2 are determined by $\mathbb{E}_{\theta_0}[\psi_4(V)] = \alpha$, $\mathbb{E}_{\theta_0}[V\psi_4(V)] = \alpha\mathbb{E}_{\theta_0}[V]$.

8.2 Testing the Parameters of a Normal Distribution

Example 65: (Example in TSH Page 154)

Let $X_1, \cdots, X_n \overset{i.i.d.}{\sim} \mathsf{N}(\xi, \sigma^2)$. The four hypotheses $\sigma \le \sigma_0$, $\sigma \ge \sigma_0$, $\xi \le \xi_0$, $\xi \ge \xi_0$ concerning the variance σ^2 and mean ξ of a normal distribution were discussed in example 53, and it was pointed out that there is a UMP test for testing $H : \sigma \le \sigma_0$ vs $K : \sigma > \sigma_0$, which rejects large values of $\sum_{i=1}^n (x_i - \bar{x})^2$. Meanwhile, there is no UMP test for $H : \sigma \ge \sigma_0$ vs $K : \sigma < \sigma_0$. We shall now show that the standard (likelihood-ratio) tests are UMP unbiased for the above four hypotheses as well as for some of the corresponding two-sided problems.

For varying ξ and σ, the densities

$$(2\pi\sigma^2)^{-n/2} \exp\left(-\frac{n\xi^2}{2\sigma^2} - \frac{1}{2\sigma^2}\sum_{i=1}^n x_i^2 + \frac{n\xi}{\sigma^2}\frac{1}{n}\sum_{i=1}^n x_i\right)$$

1. Focus on σ: Consider the density as a two-parameter exponential family, which coincides

$$\theta = -\frac{1}{2\sigma^2}, \quad \nu = \frac{n\xi}{\sigma^2}, \quad U(x) = \sum_{i=1}^n x_i^2, \quad T(x) = \frac{1}{n}\sum_{i=1}^n x_i = \bar{x}$$

- By theorem 7.4.1 there exists a UMP unbiased test of the hypothesis $\theta \ge \theta_0$, which for $\theta_0 = -\frac{1}{2\sigma_0^2}$ is equivalent to $H : \sigma \ge \sigma_0$ vs $K : \sigma < \sigma_0$, in the form of

$$\phi(u,t) = \begin{cases} 1, & u \le C_0(t) \\ 0, & \text{otherwise} \end{cases}$$

Note the fact that $\sum_{i=1}^n (X_i - \bar{X})^2$ is independent of $\bar{X}$, then $C_0(t)$ is determined by

$$\mathbb{P}_{\sigma_0}(U \leq C_0(t)|T = t) = \mathbb{P}_{\sigma_0}(\sum_{i=1}^{n} X_i^2 \leq C_0(\bar{X})|\bar{X} = \bar{x})$$
$$= \mathbb{P}_{\sigma_0}(\sum_{i=1}^{n} X_i^2 - n\bar{X}^2 \leq C_0(\bar{X}) - n\bar{X}^2|\bar{X} = \bar{x})$$
$$= \mathbb{P}_{\sigma_0}(\sum_{i=1}^{n} (X_i - \bar{X})^2 \leq C_0(\bar{X}) - n\bar{X}^2|\bar{X} = \bar{x})$$
$$= \mathbb{P}_{\sigma_0}(\frac{1}{\sigma_0^2} \sum_{i=1}^{n} (X_i - \bar{X})^2 \leq C_1) = \alpha$$

since $\frac{1}{\sigma_0^2}\sum_{i=1}^{n}(X_i^2 - \bar{X})^2$ has a χ^2_{n-1} distribution with $n-1$ degrees of freedom, the determining condition for $C_1 = \sigma_0^2 \chi^2_{n-1,\alpha}$.
The same result can be obtained through Theorem 8.1.1. A statistic $V = h(U, T)$ of the kind required by the theorem, which independent of $\bar{X}$ for $\sigma = \sigma_0$ and all ξ, is in the form of

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

Since $V = h(u, t)$ is an increasing function of u for each t, it follows that the UMP unbiased test has a rejection region of the form $v < C_1$. Then the following are the same as before.

- For $H : \sigma \leq \sigma_1$ or $\sigma \geq \sigma_2$ vs $K : \sigma_1 < \sigma < \sigma_2$, previous derivation also shows that the rejection region if UMP unbiased test is

$$C_1 < \sum_{i=1}^{n} (x_i - \bar{x})^2 < C_2$$

with
$$\int_{C_1/\sigma_1^2}^{C_2/\sigma_1^2} \chi^2_{n-1}(y)dy = \int_{C_1/\sigma_2^2}^{C_2/\sigma_2^2} \chi^2_{n-1}(y)dy = \alpha$$

- For $H : \sigma = \sigma_0$ vs $K : \sigma \neq \sigma_0$, since $V = h(u, t)$ is linear in u, then the UMP unbiased test is in the form of

$$\phi(v) = \begin{cases} 1, & v \leq C_1 \text{ or } v \geq C_2 \\ 0, & \text{otherwise} \end{cases}$$

with

$$\mathbb{P}_{\sigma_0}(\{V \leq C_1\} \cup \{V > C_2\}) = \alpha$$
$$\mathbb{E}_{\sigma_0}\left[V \mathbf{1}_{\{V \leq C_1\} \cup \{V > C_2\}\})\right] = \alpha \mathbb{E}_{\sigma_0}[V]$$

2. Focus on ξ: Consider the density as a two-parameter exponential family, which coincides

$$\theta = \frac{n\xi}{\sigma^2}, \quad \nu = -\frac{1}{2\sigma^2}, \quad U(x) = \bar{x}, \quad T(x) = \sum_{i=1}^{n} x_i^2$$

Without loss of generality, assuming $\xi_0 = 0$.

- For $H : \xi \leq 0$ vs $K : \xi > 0 \Longleftrightarrow H : \theta \leq 0$ vs $K : \theta > 0$, since

$$V = \frac{\bar{X}}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}} = \frac{U}{\sqrt{T - nU^2}}$$

is independent of $T = \sum_{i=1}^{n} X_i^2$ when $\xi = 0$, by theorem 8.1.1, the UMP unbiased rejection region is $V > C_0'$, or equivalently

$$t(x) > C_0, \quad t = \frac{\sqrt{n}\bar{x}}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

♠

8.3 Comparing the Means and Variances of Two Normal Distributions

Example 66: (Example in TSH Page 158)

We consider first the comparison of two variances σ^2 and τ^2, which occurs for example when one is concerned with the variability of analyses made by two different laboratories or by two different methods, and specifically the hypotheses $H : \sigma^2/\tau^2 \leq \Delta_0$ and $H' : \sigma^2/\tau^2 = \Delta_0$.

The proof is too long. Readers who are interested in this proof can check the solution in Page 158 Testing Statistical Hypothesis. ♠

Chapter 9

Invariant Test

9.1 Symmetry and Invariance

Lemma 9.1.1. *(Lemma 6.1.1 in TSH)*

Let g, g' be two transformations preserving Ω. Then the transformations $g'g$ and g^{-1} defined by

$$(g'g)x = g'(gx), \quad g(g^{-1}x) = x \quad \forall x \in \mathcal{X}$$

also preserve Ω and satisfy

$$\overline{g'g} = \overline{g'}\bar{g} \quad \overline{(g^{-1})} = (\bar{g})^{-1}$$

Proof If $X \sim \mathbb{P}_\theta$, then $gX \sim \mathbb{P}_{\bar{g}\theta}$, therefore $(g'g)X = g'(gX) \sim \mathbb{P}_{\bar{g'}\bar{g}\theta}$. Since $(g'g)X \sim \mathbb{P}_{\overline{g'g}}$, then $\overline{g'g} = \overline{g'}\bar{g}$. The other part is the same. □

Let C be a class of transformations satisfying these two conditions, and let G be the smallest class of transformations containing C such that $g, g' \in G$ implies that gg' and g^{-1} belong to G. Then G is a group of transformations, all of which by Lemma 9.1.1 preserve both Ω and Ω_H. Any class C of transformations leaving the problem invariant can therefore be extended to a group G. It follows further from Lemma 9.1.1 that the class of induced transformations $\bar{g}$ form a group $\bar{G}$. The two equations in Lemma 9.1.1 express the fact that $\bar{G}$ is a homomorphism of G.

Definition 9.1.1. *(Invariant Problem)*

*The problem of testing $H : \theta \in \Omega_H$ against $K : \theta \in \Omega_K$ remains **invariant** under a transformation g if $\bar{g}$ preserves both Ω_H and Ω_K, i.e., the parameter set Ω_H, Ω_K remains invariant under g if*

$$\begin{aligned} \bar{g}\theta &\in \Omega_H, \quad && \forall \theta \in \Omega_H \\ \bar{g}\theta &\in \Omega_K, \quad && \forall \theta \in \Omega_K \end{aligned}$$

In addition, $\bar{g}\theta \in \Omega$ for all $\theta \in \Omega$.

In the presence of symmetries in both sample and parameter space represented by the groups G and $\bar{G}$, it is natural to restrict attention to tests ϕ which are also symmetric, *i.e.*invariant under G, that is, which satisfy

$$\phi(gx) = \phi(x) \quad \forall x \in \mathcal{X} \text{ and } g \in G$$

9.2 Maximal Invariants

Definition 9.2.1. *(Maximal Invariant)*
A function M is said to be maximal invariant if it is invariant and if

$$M(x_1) = M(x_2) \implies x_2 = gx_1 \text{ for some } g \in G$$

that is, if it is constant on the orbits but for each orbit takes on a different value. All maximal invariants are equivalent in the sense that their sets of constancy coincide.

Theorem 9.2.1. *(Theorem 6.2.1 in TSH)*
Let $M(x)$ be a maximal invariant with respect to G. Then, a necessary and sufficient condition for ϕ to be invariant is that it depends on x only through $M(x)$; that is, that there exists a function h for which $\phi(x) = h[M(x)]$ for all x.

Proof

- $\Longleftarrow$: If $\phi(x) = h[M(x)]$ for all x, then $\phi(gx) = h[M(gx)] = h[M(x)] = \phi(x)$ so that ϕ is invariant and if $M(x_1) = M(x_2)$ for some $x_2 = gx_1$ for some $g \in G$.
- $\Longrightarrow$: If ϕ is invariant, define h by $\phi(x) = h[M(x)]$. If $M(x_1) = M(x_2)$, then $\phi(x_2) = \phi(x_1)$ which implies that $x_2 = gx_1$ for some $g \in G$ and therefore $M(x)$ is a maximal invariant with respect to G.

$\square$

Example 67:

- For transformation $x = (x_1, \cdots, x_n) \mapsto x + c = (x_1 + c, \cdots, x_n + c)$

$$\begin{cases} M_1(x) = (x_1 - x_n, \cdots, x_{n-1} - x_n) \\ M_2(x) = (x_1 - \bar{x}, \cdots, x_n - \bar{x}) \end{cases} \quad \text{are maximally invariant}$$

- For transformation $x = (x_1, \cdots, x_n) \mapsto cx = (cx_1, \cdots, cx_n)$

$$\begin{cases} M_1(x) = (\frac{x_1}{x_n}, \cdots, \frac{x_{n-1}}{x_n}) \\ M_2(x) = (\frac{x_1}{\bar{x}}, \cdots, \frac{x_{n-1}}{\bar{x}}) \\ M_3(x) = (\frac{x_1}{y}, \cdots, \frac{x_{n-1}}{y}), y = \left(\prod_{i=1}^{n} x_i\right)^{1/n} \end{cases} \quad \text{are maximally invariant}$$

- For transformation $x = (x_1, \cdots, x_n) \mapsto ax + b = (ax_1 + b, \cdots, ax_n + b)$

$$\left\{ M(x) = \left(\frac{x_1 - x_n}{x_{n-1} - x_n}, \cdots, \frac{x_{n-2} - x_n}{x_{n-1} - x_n}\right) \right. \quad \text{is maximally invariant}$$

- Group of orthogonal transformation: $M(x) = \|x\|_p$ is maximally invariant.
- Group of permutations: $M(x) = (x_{(1)}, \cdots, x_{(n)})$ is maximally invariant

- For transformation $g \in G$: $x = (x_1, \cdots, x_n) \mapsto \lambda(x) = (\lambda(x_1), \cdots, \lambda(x_n))$, where λ is strictly increasing and continuous on $\mathbb{R}$, let r_i be rank of x_i

$$M(x) = (r_1, \cdots, r_n) \quad \text{is maximally invariant}$$

♠

Theorem 9.2.2. *(Theorem 6.2.2 in TSH)*

Let G be a group of transformations, and let D and E be two subgroups generating G. Suppose that $y = s(x)$ is maximal invariant with respect to D, and that for any $e \in E$

$$s(x_1) = s(x_2) \implies s(ex_1) = s(ex_2)$$

If $z = t(y)$ is maximal invariant under the group E^* of transformations e^* defined by

$$e^* y = s(ex) \quad when y = s(x)$$

then $z = t[s(x)]$ is maximal invariant with respect to G.

9.3 Most Powerful Invariant Tests

Theorem 9.3.1. *(Theorem 6.3.1 in TSH)*

Suppose the problem of testing Ω_0 against Ω_1 remains invariant under a finite group $G = g_1, \cdots, g_N$ and that $\bar{G}$ is transitive over Ω_0 and Ω_1. Then there exists a UMP invariant test of Ω_0 against Ω_1, and it rejects Ω_0 when

$$\frac{\sum_{k=1}^{N} f_{\bar{g}_k \theta_1(x)}/N}{\sum_{k=1}^{N} f_{\bar{g}_k \theta_0(x)}/N}$$

is sufficiently large, where θ_0 and θ_1 are any elements of Ω_0 and Ω_1, respectively.

Example 68: (Example 6.3.2 in TSH)

Let $X = (X_1, \cdots, X_n)$, and suppose that the density of X is $f_k(x_1 - \xi, \cdots, x_n - \xi)$ under $H_i (i = 0, 1)$, where ξ ranges from $-\infty$ to ∞. The problem of testing H_0 against H_1 is invariant under the group G of transformations

$$gx = (x_1 + c, \cdots, x_n + c), \quad -\infty < c < \infty$$

where $\bar{g}\xi = \xi + c$. A maximal invariant under G is $Y = (X_1 - X_n, \cdots, X_{n-1} - X_n)$. The distribution of Y is independent of ξ and under H_k has the density

$$h_k(y) = \int_{-\infty}^{\infty} f_k(y_1 + z, y_2 + z, \cdots, z) dz$$

When referred to Y, the problem of testing H_0 against H_1 therefore becomes one of testing a simple hypothesis against a simple alternative. By theorem 6.2.1, the test, which is independent of ξ, that rejects large values of $\frac{h_1(y)}{h_0(y)}$ is most powerful among tests that are functions of y, which is invariant under G. Therefore, it is UMP among all invariant tests. ♠

Theorem 9.3.2. *(Theorem 6.3.2 in TSH)*
If $M(x)$ is invariant under g, and if $\nu(\theta)$ maximal invariant under the induced group $\bar{G}$, then the distribution of $M(X)$ depends only on $\nu(\theta)$.

Proof Let $\nu(\theta_1) = \nu(\theta_2)$. Then $\theta_2 = \bar{g}\theta_1$, and hence

$$\mathbb{P}_{\theta_2}(M(X) \in B) = \mathbb{P}_{\bar{g}\theta_1}(M(X) \in B) = \mathbb{P}_{\theta_1}(M(gX) \in B) = \mathbb{P}_{\theta_1}(M(X) \in B)$$

This result can be paraphrased by saying that the principle of invariance identifies all parameter points that are equivalent with respect to $\bar{G}$. □

Example 69: (Example 6.3.3 in TSH)
If $X_1, \cdots, X_n$ is a sample from $\mathsf{N}(\xi, \sigma^2)$, the hypothesis $H : \sigma \geq \sigma_0$ remains invariant under the transformations $X_i' = X_i + c, c \in \mathbb{R}$. In terms of the sufficient statistics

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

these transformations become $Y' = Y + c$, $(S^2)' = S^2$, and a maximal invariant is S^2. The class of invariant tests is therefore the class of tests depending on S^2. Notice that $S^2 \sim \chi^2_{n-1}$.

To test $H : \sigma \geq \sigma_0$ against $K : \sigma < \sigma_0$ It follows from Theorem 6.3.1 that there exists a UMP invariant test, with rejection region $S^2 \leq C = \sigma_0^2\chi^2_{n-1,\alpha}$. ♠

Example 70: (Example 6.3.4 in TSH) Suppose that

$$X_1, \cdots, X_m \sim \mathsf{N}(\xi, \sigma^2) \perp Y_1, \cdots, Y_n \sim \mathsf{N}(\eta, \tau^2)$$

To test $H : \tau^2 \leq \sigma^2$ against $K : \tau^2 > \sigma^2$. Consider the group G of transformation

$$(x_1, \cdots, x_n, y_1, \cdots, y_n) \mapsto (ax_1 + b, \cdots, ax_n + b, ay_1 + c, \cdots, ay_n + c), a > 0, b \in \mathbb{R}$$

$(T_1, T_2, T_3, T_4) = (\bar{X}, \bar{Y}, S^2, T^2)$ are jointly sufficient for this problem, where

$$\bar{X} = \tfrac{1}{n}\textstyle\sum_{i=1}^{n} X_i, \qquad S^2 = \sum_{i=1}^{n}(X_i - \bar{X})^2$$
$$\bar{Y} = \tfrac{1}{n}\textstyle\sum_{i=1}^{n} Y_i, \qquad T^2 = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$$

Under transformation g, $(t_1, t_2, t_3, t_4) \mapsto (at_1 + b, at_2 + c, a^2t_3, a^2t_4)$, thus (T_1, T_2, T_3, T_4) is equivariant. Define

$$M(t_1, t_2, t_3, t_4) = \frac{t_4}{t_3} \text{ is maximally invariant}$$

The statistic

$$Z = \frac{T_4/(n-1)}{T_3/(m-1)} \sim \frac{\tau^2}{\sigma^2} F_{n-1,m-1}$$

Let $\theta = \frac{\tau^2}{\sigma^2}$, $\theta_0 = 1$. Then we are going to test

$$H : \theta \leq \theta_0 \quad vs \quad K : \theta > \theta_0$$

Notice the facts that Z has MLR with respect to density $F_{n-1,m-1}(z)$, by theorem 6.3.1, there exists a UMP test among these invariant tests, which rejects $\{z > f_{n-1,m-1}^{(1-\alpha)}\}$ ♠

Remark In most situations of interests, the reduction by sufficiency is without loss of generality, by Theorem 6.5.1, Theorem 6.5.3 in TSH.

Example 71: (Problem 6.19 in TSH, Testing a correlation coefficient)

Let $(X_1, Y_1), \cdots, (X_n, Y_n)$ be a sample from a bivariate normal distribution, *i.e.*, $(X_1, Y_1), \cdots, (X_n, Y_n) \sim \mathsf{N}_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where

$$\boldsymbol{\mu} = \begin{bmatrix} \xi \\ \eta \end{bmatrix}, \quad \boldsymbol{\Sigma} = \begin{bmatrix} \sigma^2 & \rho\sigma\tau \\ \rho\sigma\tau & \tau^2 \end{bmatrix}$$

1. For testing $H : \rho \leq \rho_0$ against $K : \rho > \rho_0$, there exists a UMPI (UMP invariant) test with respect to the group of all transformations $X_i' = aX_i + c$, $Y_i' = bY_i + d$, where $a, b > 0$. This test rejects when the sample correlation coefficient R is too large.

   Proof $(\bar{X}, \bar{Y}, S^2, T^2, V)$ are jointly sufficient for this problem, where

   $$\bar{X} = \tfrac{1}{n}\textstyle\sum_{i=1}^n X_i, \qquad S^2 = \sum_{i=1}^n (X_i - \bar{X})^2, \quad V = \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})$$
   $$\bar{Y} = \tfrac{1}{n}\textstyle\sum_{i=1}^n Y_i, \qquad T^2 = \sum_{i=1}^n (Y_i - \bar{Y})^2$$

   The group of transformation $G = \{g : g(x_i, y_i) = (ax_i + b, cy_i + d)\}$, where $a, b > 0$ and $c, d \in \mathbb{R}$. Under the transformation g

   $$g(\bar{X}, \bar{Y}, S^2, T^2, V) = (a\bar{X} + c, b\bar{Y} + d, a^2S^2, b^2T^2, abV)$$

   Therefore, define sample correlation $R = \frac{V}{ST}$, which is maximally invariant with respect to group G. Notice the fact that THE family of distributions(parametrized by ρ) has monotone likelihood ratio with respect to R(to be proved). By Theorem 6.3.1, the test rejects large values of R is UMPI. □

2. The problem of testing $H : \rho = \rho_0$ against $K : \rho \neq \rho_0$ remains invariant in addition under the transformation $Y_i' = -Y_i$,$X_i' = X_i$. With respect to the group generated by this transformation and those of (1) there exists a UMP invariant test, with rejection region $|R| > C$.

   Proof Under combined transformations of $Y_i' = -Y_i$,$X_i' = X_i$ and $X_i' = aX_i + c$, $Y_i' = bY_i + d$, where $a, b > 0$, $|R| = \frac{|V|}{ST}$ is maximally invariant. The distribution of R is symmetric about 0, which yields the test rejecting large value of $|R|$. □

♠

9.4 Unbiasedness and Invariance

Theorem 9.4.1. *(Theorem 6.6.1 in TSH) Suppose that for a given testing problem there exists a UMP unbiased test ϕ^* which is unique (up to sets of measure zero), and that there also exists a UMP almost invariant test ϕ' with respect to some group G. Then the latter is also unique (up to sets of measure zero), and the two tests coincide a.e.*

Proof If $U(\alpha)$ is the class of unbiased level-α tests, then $\phi(x) \in U(\alpha)$ if and only if $\phi(gx) \in U(\alpha)$ for all $g \in G$. Denoting the power function of the test ϕ by $\beta_\phi(\theta)$, we thus have

$$\begin{aligned}
\beta_{\phi^* g}(\theta) &= \beta_{\phi^*}(\bar{g}\theta) = \sup_{\phi \in U(\alpha)} \beta_\phi(\bar{g}\theta) = \sup_{\phi \in U(\alpha)} \beta_{\phi g}(\theta) \\
&= \sup_{\phi g \in U(\alpha)} \beta_{\phi g}(\theta) = \sup_{\phi \in U(\alpha)} \beta_\phi(\theta) = \beta_{\phi^*}(\theta)
\end{aligned}$$

It follows that ϕ^* and $\phi^* g$ have the same power function, and, because of the uniqueness assumption, that ϕ^* is almost invariant. Therefore, if ϕ' is UMP almost invariant, we have $\beta_{\phi'}(\theta) \geq \beta_{\phi^*}(\theta)$ for all θ. On the other hand, ϕ' is unbiased, as is seen by comparing it with the invariant test $\phi(x) \equiv \alpha$, and hence $\beta_{\phi'}(\theta) \leq \beta_{\phi^*}(\theta)$ for all θ. Since ϕ' and ϕ^* therefore have the same power function, they are equal a.e. because of the uniqueness of ϕ^*, as was to be proved. $\square$

9.5 Rank Test

Example 72: (Example in TSH Page 240, Two Sample Problem)

$X_1, \cdots, X_m \overset{i.i.d.}{\sim} F \perp Y_1, \cdots, Y_n \overset{i.i.d.}{\sim} G$, where F and G are continuous cumulative distributions on $\mathbb{R}$. We want to test

$$H_1 : G = F \quad vs \quad K_1 : G(z) \leq F(z), \forall z \text{ and } G \neq F$$

$i.e. K_1 : Y$'s are stochastically larger than the X's.

The two-sample problem of testing H_1 against K_1 remains invariant under the group T of all transformations

$$(x_1, x_2, \cdots, x_m, y_1, \cdots, y_n) \mapsto (\rho(x_1), \rho(x_2), \cdots, \rho(x_m), \rho(y_1), \cdots, \rho(y_n))$$

where ρ is continuous and strictly increasing, and ρ is one-to-one on $\mathbb{R}$ with a finite number of discontinuities. Let

$$\begin{aligned}
R_i &= \text{Rank of } X_i \text{ in } \{X_1, X_2, \cdots, X_m, Y_1, \cdots, Y_n\} \\
S_j &= \text{Rank of } Y_j \text{ in } \{X_1, X_2, \cdots, X_m, Y_1, \cdots, Y_n\}
\end{aligned}$$

Note the fact that

$$(R; S) = (R_1, R_2, \cdots, R_m, S_1, \cdots, S_n)$$

is Maximally invariant under group transformation T. Meanwhile, order statistics

$$\left.\begin{array}{r} X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(m)} \\ Y_{(1)} \leq Y_{(2)} \leq \cdots \leq Y_{(n)} \end{array}\right\} \text{jointly sufficient and equivariant for } (R; S)$$

since the distribution of $(R_1, R_2, \cdots, R_m, S_1, \cdots, S_n)$ is symmetric in the first m and in the last n variables for all distributions F and G.

Therefore, by one of these sets alone since each of them determines the other. Any invariant test is thus a rank test, that is, it depends only on the ranks of the observations, for example on $(S_1, \cdots, S_n)$.

1. Wilcoxon Rank-sum test: rejects large values of $\sum_{j=1}^n S_j$. To obtain the p-value from this test, randomly pick n observations from total $n+m$ variables. Draw the histogram to get p-value.

2. Norm Scores tets: rejects large $\sum_{j=1}^n \eta(S_j)$, where $\eta(S) = \mathbb{E}\left[Z_{(S)}\right]$, $Z_1, \cdots, Z_{m+n} \overset{i.i.d.}{\sim} \mathsf{N}(0,1)$.

3. Kolmogorov-Smirnov test: rejects large value $\sup_{t\in\mathbb{R}}(\hat{F}(t) - \hat{G}(t))$, where

$$\hat{F}(t) = \frac{1}{m}\sum_{i=1}^m \mathbf{1}_{\{X_i \le t\}}, \quad \hat{G}(t) = \frac{1}{n}\sum_{i=1}^n \mathbf{1}_{\{Y_i \le t\}}$$

♠

Example 73: (Example in TSH Page 240, Testing for Symmetry)

$(X_1, Y_1), \cdots, (X_N, Y_N)$ are a sample from a continuous bivariate distribution F on $\mathbb{R}^2$. The hypothesis of no effect is then equivalent to the assumption that F is symmetric with respect to the line $y = x$, *i.e.*

$$H_2 : F(x,y) = F(y,x) \quad vs \quad K_2 : X \overset{stochastic}{\le} Y$$

Make the transformation

$$(X_i, Y_i) \mapsto (Y_i - X_i, X_i + Y_i) = (Z_i, W_i)$$

Under the hypothesis this distribution is symmetric with respect to the w-axis, while under the alternatives the distribution is shifted in the direction of the positive z-axis The problem is unchanged if all the w's are subjected to the same transformation $w_i' = \lambda(w_i)$, where λ is $1:1$ and has at most a finite number of discontinuities, and $(Z_1, \cdots, Z_n)$ constitutes a maximal invariant under this group, *i.e.*

$$M\left((Z_1, W_1), \cdots, (Z_N, W_N)\right) := (Z_1, \cdots, Z_N) \quad \text{is maximally invariant}$$

The Z's are a sample from a continuous univariate distribution D, for which the hypothesis of symmetry with respect to the origin

$$\begin{aligned} H_2' : Z \sim -Z \iff & \quad D(-z) = 1 - D(z) \quad \forall z \in \mathbb{R} \\ vs \quad K_2' : Z \overset{stochastic}{\ge} -Z \iff & \quad D(-z) \ge 1 - D(z) \quad \forall z \in \mathbb{R} \end{aligned}$$

This problem is invariant under the group G of all transformations $z_i' = \rho(z_i)$, for $i = 1, \cdots, N$, such that ρ is

- one-to-one continuous on $\mathbb{R}$;

- odd(so it respects H_2')
- strictly increasing(so it respects K_2')

If $z_{i_1}, \cdots, z_{i_m} < 0 < z_{j_1}, \cdots, z_{j_n}$, where $i_1 < \cdots < i_m$ and $j_1 < \cdots < j_n$, let

$$\begin{aligned}(r_1, \ldots, r_m) = & \text{ rank of } (z_{i_1}, \cdots, z_{i_m}) \text{ among } \{|z_1|, \cdots, |z_N|\} \\ (s_1, \ldots, s_n) = & \text{ rank of } (z_{j_1}, \cdots, z_{j_n}) \text{ among } \{|z_1|, \cdots, |z_N|\} \\ (\epsilon_1, \cdots, \epsilon_N) = & \text{ sign of } (z_1, \cdots, z_N)\end{aligned}$$

The transformations ρ preserve the sign of each observation, and hence in particular also the numbers m and n. Since ρ is a continuous, strictly increasing function of $|z|$, it leaves the order of the absolute values invariant and therefore the ranks r_i and s_j. Note the following facts

- Fact: under null $H_2 : Z \sim -Z$, $(r_1, \ldots, r_m, s_1, \ldots, r_n) \perp (\epsilon_1, \ldots, \epsilon_N)$, where $N = m + n$.
- Fact: under null $H_2 : Z \sim -Z$, $|Z| \perp sign(Z)$ and $\mathbb{P}[sign(Z) = 1] = \frac{1}{2}$.
- Fact: under null $H_2 : Z \sim -Z$, $T = \sum_{i \in I} S_i - \sum_{i \notin I} R_i \sim \sum_{i=1}^N i\xi_i$, where $\{\xi_i\}_{i=1}^N$ is Rademacher process, such test are distribution free.

Under the null hypothesis, the distribution D is symmetric about 0, we can focus on the tests which are the functions of $\{s_1, \ldots, s_n\}$, $i.e. I = \{i : \epsilon_i = 1, z_i > 0\}$, for example

- Wilcoxon signed-rank test: rejects large value of $\sum_i S_i$.
- Smirnov test for symmetry: rejects large values of

$$\sup_{z \in \mathbb{R}} [\hat{D}(z) + \hat{D}(-z) - 1], \quad \hat{D}(t) = \frac{1}{n} \sum_{i=1}^n \mathbf{1}_{\{z_i \leq t\}}$$

♠

Example 74: (Example in TSH Page 240, Testing for Independence)

$(X_1, Y_1), \cdots, (X_N, Y_N)$ is a sample from a bivariate distribution F, one will be interested in the hypothesis

$$\begin{aligned} & H_3 : F(x, y) = G_1(x)G_2(y) \quad \Longleftrightarrow \quad X \text{ and } Y \text{ are independent} \\ vs \quad & K_3 : X \text{ and } Y \text{ are positive associated}\end{aligned}$$

The problem remains invariant with respect to group of transformations

$$(x_i, y_i) \mapsto (\gamma(x_i), \lambda(y_i))$$

where γ and λ are strictly increasing and continuous and one-to-one on $\mathbb{R}$. Let

$$\begin{aligned}(R_1, \ldots, R_N) = & \text{ rank of } (X_1, \cdots, X_N) \text{ among } \{X_1, \cdots, X_N\} \\ (S_1, \ldots, S_N) = & \text{ rank of } (Y_1, \cdots, Y_N) \text{ among } \{Y_1, \cdots, Y_N\}\end{aligned}$$

Therefore

$$((R_1, S_1), \cdots, (R_N, S_N)) \quad \text{is maximally invariant}$$

Here is an example test, Spearman rank correlation test

- The statistic is proportional to

$$\begin{aligned} & \sum_{i=1}^{N}(R_i-\bar{R})(S_i-\bar{S}), \quad \text{where } \bar{S}=\bar{R}=\frac{N+1}{2} \\ = & \sum_{i=1}^{N} R_i S_i - N\left(\frac{N+1}{2}\right)^2 \end{aligned}$$

  The test rejects large values of $\sum_{i=1}^{N} R_i S_i$.

- Similarly, write

$$\sum_{i=1}^{N}(R_i-S_i)^2=\sum_{i=1}^{N}(R_i^2+S_i^2)-2\sum_{i=1}^{N} R_i S_i=\frac{N(N+1)(2N+1)}{3}-2\sum_{i=1}^{N} R_i S_i$$

  this test rejects small values of $\sum_{i=1}^{N}(R_i-S_i)^2$.

- Fact: Spearman rank correlation is consistent for

$$\rho = 3\mathbb{E}\left[sign[(X_1-X_2)(Y_1-Y_3)]\right], \quad (X_1,Y_1),(X_2,Y_2),(X_3,Y_3) \overset{i.i.d.}{\sim} F$$

Here is an example test, Kendall's coefficient

- The statistic

$$\frac{1}{\binom{n}{2}} \sum \sum_{1\le i<j\le n} sign[(X_i-Y_i)(X_j-Y_j)]$$

- Kendall's coefficient is consistent for

$$\tau := \mathbb{E}\left[sign[(X_1-X_2)(Y_1-Y_2)]\right], \quad (X_1,Y_1),(X_2,Y_2) \overset{i.i.d.}{\sim} F$$

Here is an example test, Hoeffding's test

- The statistic

$$\begin{aligned} \hat{F}(x,y) &= \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\{X_i\le x, Y_i\le y\}} \\ \hat{F}_X(x) &= \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\{X_i\le x\}}, \quad \hat{F}_Y(y)=\frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\{Y_i\le y\}} \end{aligned}$$

- The test rejects large values of

$$\sup_{x,y\in\mathbb{R}} \left(\hat{F}(x,y)-\hat{F}_X(x)\hat{F}_Y(y)\right)$$

The distribution of $(R_1,S_1),\cdots,(R_N,S_N)$ is symmetric in these N pairs for all distributions of (X,Y). It follows that a sufficient statistic is $(S_1,\ldots,S_N)$ where $(1,S_1),\cdots,(N,S_N)$ is a permutation of $(R_1,S_1),\cdots,(R_N,S_N)$ and where therefore S_i is the rank of the variable Y associated with the ith smallest X.

When calculating p values for tests all above, we could take permutation of $(S_1, \cdots, S_N)$, compute each statistic and then get the distribution. ♠

Example 75: (Example in TSH Page 240, Testing for Monotoncity)

$(X_1, Y_1), \cdots, (X_N, Y_N)$ is a sample. If the chosen values are $x_1 < \cdots < x_N$ and F_i denotes the distribution of Y given xi, the Y's are independently distributed with continuous cumulative distribution functions $F_1, \cdots, F_N$. The hypothesis of independence of Y from x becomes

$$H_4 : F_1 = \cdots = F_N \Longleftrightarrow Y_1 \sim Y_2 \sim \cdots \sim Y_N$$
$$K_4 : F_1 \geq \cdots \geq F_N \Longleftrightarrow Y_1 \overset{stochastic}{\leq} Y_2 \overset{stochastic}{\leq} \cdots \overset{stochastic}{\leq} Y_N$$

The problem is invariant under transformation

$$(y_1, \cdots, y_n) \mapsto (\lambda(y_1), \cdots, \lambda(y_n))$$

where λ is strictly increasing and continuous and one-to-one on $\mathbb{R}$. Let

$$(S_1, \ldots, S_N) = \text{ rank of } (Y_1, \cdots, Y_N) \text{ among } \{Y_1, \cdots, Y_N\}$$

Therefore

$$(S_1, \cdots, S_N) \quad \text{ is maximally invariant}$$

Example of tests are Spearman and Kendall's test. ♠

Chapter 10

Solutions

Example 76: Let $X_1, \cdots, X_n$ be i.i.d. $\mathsf{N}(a, \sigma^2)$, and independently, let $Y_1, \cdots, Y_n$ be i.i.d. $\mathsf{N}(b, \sigma^2)$.

1. Assume σ^2 is known. Consider testing $a \geq b$ versus $a < b$. Derive a UMP test at level α. [There is a point mass prior that is least favorable.]

   Proof First, we fix an alternative (a_1, b_1) where $a_1 < b_1$. Choose a point mass prior at $(\frac{a_1+b_1}{2}, \frac{a_1+b_1}{2})$, then we are going to test H_Λ versus K. The likelihood ratio is

$$\begin{aligned} LR &= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^{2n} \exp\Big\{ -\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i - a_1)^2 - \frac{1}{2\sigma^2}\sum_{i=1}^n (y_i - b_1)^2 \\ &+ \frac{1}{2\sigma^2}\sum_{i=1}^n (x_i - \frac{a_1+b_1}{2})^2 + \frac{1}{2\sigma^2}\sum_{i=1}^n (y_i - \frac{a_1+b_1}{2})^2 \Big\} \\ &\propto \exp\left\{ \frac{n\bar{x}(a_1 - b_1)}{2\sigma^2} + \frac{n\bar{y}(b_1 - a_1)}{2\sigma^2} - \frac{n(a_1 - b_1)^2}{4\sigma^2} \right\} \\ &= \exp\left\{ (a_1 - b_1)\frac{n(\bar{x} - \bar{y})}{2\sigma^2} - \frac{n(a_1 - b_1)^2}{4\sigma^2} \right\} \end{aligned}$$

   By theorem 6.2.1, the most powerful test is in the form of

$$\phi = \begin{cases} 1 & , (\bar{x} - \bar{y}) > C \\ \gamma & , (\bar{x} - \bar{y}) > C \\ 0 & , (\bar{x} - \bar{y}) < C \end{cases}$$

   The test ϕ does not depend on (a_1, b_1), then we can go through all alternatives. Meanwhile, the test phi is level α under the original null. By theorem 6.5.3, ϕ is UMP test. $\square$

2. Assume σ^2 is known. Consider testing $a = b$ versus $a \neq b$. Derive a UMP test at level α.

   Proof We write the likelihood as exponential family, let $\theta = a_1 - b_1$, $U = \frac{n(\bar{x}-\bar{y})}{2\sigma^2}$, $\theta = 0 \iff a_1 = b_1$. By theorem 7.4.1 item 4, there exists UMPU test $\phi_4(u, t)$. $\square$

3. Assume σ^2 is unknown. Consider testing $a = b$ versus $a \neq b$. Is there a UMPU test at level α?

   Proof We write the likelihood as exponential family, let $\theta = \frac{a_1 - b_1}{2\sigma^2}$, $U = \bar{x} - \bar{y}$,$\theta = 0 \iff a_1 = b_1$. By theorem 7.4.1 item 4, there exists UMPU test $\phi_4(u, t)$. □

♠

Example 77: Let $X_1, \cdots, X_n$ be independent with $X_i \sim Poisson(\lambda_i)$. Consider testing $\sum_{i=1}^n \lambda_i \leq 1$ versus $\sum_{i=1}^n \lambda_i > 1$at some level $\alpha \in (0, 1)$. Find a UMP test. (Specify the test as much as possible.) [HINT: there is a point mass prior that is least favorable.] ♠

Proof Fix an alternative $\lambda^{(1)} = (\lambda_1^{(1)}, \cdots, \lambda_n^{(1)})$. Define $S_n^{(1)} = \sum_{i=1}^n \lambda_i^{(1)}$ and

$$\lambda^{(0)} = (\lambda_1^{(0)}, \cdots, \lambda_n^{(0)}) = (\frac{\lambda_1^{(1)}}{S_n^{(1)}}, \cdots, \frac{\lambda_n^{(1)}}{S_n^{(1)}})$$

We choose a point mass prior at $\lambda^{(0)}$, i.e. $\Lambda = \delta(\lambda_{(0)})$. Then the test is transformed into $H_\Lambda : \sum_{i=1}^n \lambda_i^{(0)} = 1$ versus $K : \sum_{i=1}^n \lambda_i^{(0)} > 1$. The likelihood ration

$$L = \frac{f_{\lambda^{(1)}}}{f_{\lambda^{(0)}}} = [S_n^{(1)}]^{\sum_{i=1}^n x_i} e^{1 - S_n^{(1)}} = g(T)$$

where $T = \sum_{i=1}^n x_i$ and $g(T)$ is monotonely increasing with respect to T. By theorem 6.2.1, the most powerful test is in the form of

$$\phi = \begin{cases} 1 & , \quad T > C \\ \gamma, & , \quad T = C \\ 0 & , \quad T < C \end{cases}$$

while C is determined by $\mathbb{E}_{\lambda^{(0)}}[\phi] = \alpha$. Under H_Λ, we have $T = \sum_{i=1}^n x_i \sim Poi(\sum_{i=1}^n \lambda_i^{(0)})$. For the remaining part of the original null, $\sum_{i=1}^n \lambda_i \leq 1$,

$$\mathbb{E}_\lambda[\phi] = \mathbb{P}_\lambda(T > C) + \gamma \mathbb{P}_\lambda(T = C) \leq \mathbb{P}_{\lambda^{(0)}}(T > C) + \gamma \mathbb{P}_{\lambda^{(0)}}(T = C) = \mathbb{E}_{\lambda^{(0)}}[\phi] = \alpha$$

Meanwhile, the test does not depend on the choice of $\lambda^{(1)}$, thus we can go through the whole alternative. By theorem 6.5.3, ϕ is UMP test for H versus K. □

Example 78: Consider $X_1, \cdots, X_n$ that are iid normal with mean ξ and variance σ^2, and independently, $Y_1, \cdots, Y_n$ that are i.i.d. normal with mean γ and variance τ^2. Consider testing $\sigma = \tau$ versus $\sigma \neq \tau$. Let $x = (x_1, \cdots, x_n)$ and $x = (y_1, \cdots, y_n)$ be generic notation for observations.

1. Show that the problem is invariant w.r.t.$(x, y) \to (ax + b, ay + c)$, where $a > 0$ and $b, c \in \mathbb{R}$. Also, show that the problem is invariant w.r.t. $(x, y) \to (y, x)$.

Proof Since $\mathbb{V}\text{ar}(ax+b) = a^2\sigma^2$, $\mathbb{V}\text{ar}(ax+c) = a^2\tau^2$, and $a > 0$, then we have $a^2\sigma^2 = a^2\tau^2 \iff \sigma^2 = \tau^2$. Let $g : (x, y) \to (ax+b, ay+c)$, $\bar{g} : (\sigma, \tau) \to (a\sigma, a\tau)$, then

$$\begin{aligned} \forall \quad (\sigma, \tau) \in \Omega_H, &\quad \bar{g}(\sigma, \tau) \in \Omega_H \\ \forall \quad (\sigma, \tau) \in \Omega_K, &\quad \bar{g}(\sigma, \tau) \in \Omega_K \end{aligned}$$

Thus the transformation g leaves the problem invariant. Also,

$$\begin{aligned} g' &: (x, y) \to (y, x) \\ \bar{g}' &: (\sigma, \tau) \to (\tau, \sigma) \end{aligned}$$

Thus the problem is invariant w.r.t. transformation $g' : (x, y) \to (y, x)$. □

2. Let G denote the group of such transformations. Derive a maximally invariant (MI) function for G. **Proof** Define $T_1 = \bar{x}$, $T_2 = \bar{y}$, $T_3 = \sum_{i=1}^{m}(y_i - \bar{y})$, (T_1, T_2, T_3, T_4) are jointly sufficient. Under transformation $g \in G$

$$(t_1, t_2, t_3, t_4) \to (at_1 + b, at_2 + c, a^2 t_3, a^2 t_4)$$

Define $\max\{\frac{t_3}{t_4}, \frac{t_4}{t_3}\}$, which is maximally invariant function for G. □

3. Derive a UMPI test. **Proof** Let $\theta = \max\{\frac{t_3}{t_4}, \frac{t_4}{t_3}\}$. The test is transformed into $H : \theta = \theta_0 = 1$ versus $K : \theta > 1$. Define $Z = \frac{T_4/(n-1)}{T_3/(m-1)} \sim \theta\theta F_{n-1,m-1}$. The likelihood ratio is monotone with respect to z. By theorem 6.3.1, the UMPI test is in the form of

$$\phi = \begin{cases} 1 & , z > C \\ \beta & , z > C \\ 0 & , z < C \end{cases}$$

where C and β are determined by $\mathbb{E}_{\theta_0}[\phi(Z)] = \alpha$. □

♠