Statistical Inference-I: Lecture Notes

Arindam Chatterjee
ISI Delhi

December 9, 2021

Contents

Abstract

These are lecture notes for the course Statistical Inference-I, for M.Stat first year students, taught during Fall 2021. The material has been drawn up from several resources, which will be pointed out in the lectures. There will be errors, notational inconsistencies and probably other issues, and I will appreciate any feedback for improvement. These class notes should not be shared outside the course.

1 Introduction

Inference is the main goal in the study of Statistics. We observe data and try to draw conclusions about the data-generating mechanism. The focus in this course would be on some basic theory for statistical inference. Before we discuss about inference, the primary step is to obtain data from an experiment or an observational study. There are many steps before we reach the stage of statistical inference using an data-set and these include data collection, data-cleaning, exploratory data analysis, etc. These are important steps but we ignore these issues during this course.

The collected data (either univariate or multivariate) are in the form of observations $\{x_1, \ldots, x_n\}$ and we assume that these are realized values of some random variables $\{X_1, \ldots, X_n\}$. We are interested in knowing the unknown distribution of the X_i's using the observed sample $\{x_1, \ldots, x_n\}$. In order to get an answer, one needs to make certain assumptions about the underlying distribution of the X_i's. These assumptions constitute a *model* for our data. For example, we may assume certain observations are from a Normal distribution (with an unknown mean and variance), but in truth, they may be arising from some other random mechanism. In practice, we can never know for sure, if our model is indeed correct or not. Choice of a correct model (using the observed data) is a challenging problem.

Write $\mathbf{X} = (X_1, \ldots, X_n)$ and assume that $\mathbf{X} \sim \mathbf{P}$, where $\mathbf{P}$ denotes the joint distribution and it is unknown. As the true underlying $\mathbf{P}$ is unknown, typically we assume that $\mathbf{P}$ is a member of a family of distributions $\mathcal{P}$. We may characterise $\mathcal{P}$ by a parameter θ (which may be vector valued) and then write $\mathcal{P} = \{\mathbf{P}_\theta : \theta \in \Theta\}$, where Θ denotes the entire collection of possible parameter values. Since $\mathcal{P}$ denotes the entire class of models, the true underlying distribution $\mathbf{P}$ will be of the form $\mathbf{P} = \mathbf{P}_{\theta_0}$, for some $\theta_0 \in \Theta$. A simple example would be $\mathcal{P} = \{N(\mu, \sigma^2) : \mu \in \mathbb{R},\ \sigma > 0\}$, with $\theta = (\mu, \sigma) \in \Theta = \mathbb{R} \times (0, \infty)$. The parameter θ is assumed to be finite dimensional and the parameter space Θ is assumed to be subset of the Euclidean space.

Another option is to use a nonparametric model, where $\mathcal{P}$ is not characterized by any Euclidean parameter. This gives us more flexibility in creating a model for our dataset, but inference becomes more difficult. An example would be $\mathcal{P} = \{\mathbf{P} : \mathbf{E}_{\mathbf{P}}(X^2) < \infty\}$, the class of all distributions with finite second moments. Another example can be, $\mathcal{P} = \{\mathbf{P} : \mathbf{P} \text{ has a pdf } p \text{ with } p(x) = p(-x) \text{ for each } x > 0\}$. This is the class of continuous distributions having a symmetric (about zero) pdf. In these two cases, the families can not be characterised by a finite dimensional parameter. Thus they are known as nonparametric families of distributions. The *parameter* in this case will be each $\mathbf{P}$ with finite second moments (or symmetric pdfs p), which are *infinite* dimensional quantities. There is another type of models, known as semiparametric models, where $\mathbf{P}$ can be characterized by a finite dimensional parameter and as well as an infinite dimensional parameter. The usual linear regression model with regression coefficient (α, β) and an unknown error distribution is a typical example. The study of such models is beyond the scope of this course.

In this course, the main focus is on parametric models. There are three main inference problems that we consider: obtain a *single* value for the unknown θ (so that it is near θ), known as *point estimation*; obtain a range of values for the undelying θ, known as *interval estimation*; and check a hypothesis about the underlying θ, known as *hypothesis testing*. Further on, we will explore decision theory which unifies and generalizes these three ideas. We will also look into the Bayesian approach towards inference, where the underlying θ is assumed to be random quantity. If time permits, we will try to go through some modern topics that are connected to statistical inference.

The majority of topics in this course are classical and were developed in early 1950's. Although the theory is elegant, but it has limited application especially in modern contexts. However, these topics provide a groundwork to study more modern and advanced statistical methods.

2 Sufficiency, exponential families, completeness and ancillarity

In this section we study the topics of sufficiency, completeness and ancillarity of a statistic. This section is required to develop the technical tools for constructing estimators and tests.

The random observables are denoted by $\mathbf{X} = (X_1, \ldots, X_n)$ and the observed realization of $\mathbf{X}$ is denoted by $\mathbf{x} = (x_1, \ldots, x_n)$. We assume that $\mathbf{X} \sim \mathbf{P}_\theta$, where $\theta \in \Theta$. We also assume that the joint distribution $\mathbf{P}_\theta$ has a pdf (or pmf) $p(\mathbf{x} : \theta)$. So, the underlying family of distributions can also be written as $\mathcal{P} = \{\mathbf{P}_\theta : \theta \in \Theta\} = \{p(\mathbf{x} : \theta) : \theta \in \Theta\}$. The parameter space $\Theta \subseteq \mathbb{R}^k$, (for some $k \geqslant 1$). The random vector $\mathbf{X} \in \mathcal{X}$ (the sample space). The notion of families will play an important role, so some examples are provided below.

Example 2.1. Assume $\{X_1, \ldots, X_n\}$ are i.i.d. Bernoulli(θ), with $\theta \in (0, 1)$. Here, the unknown parameter of interest is θ, the family of distributions associated with $\mathbf{X} = (X_1, \ldots, X_n)$ is $\mathcal{P} = \{p(\mathbf{x} : \theta) : \theta \in (0, 1)\}$ where $p(\mathbf{x} : \theta)$ is the joint pmf of $\mathbf{X}$.

Example 2.2. Assume $\{\mathbf{X}_1, \ldots, \mathbf{X}_n\}$ are i.i.d $N_k(\mathbf{0}, \Sigma)$. Here, the unknown parameter is Σ and $\Sigma \in$ the class of $k \times k$ symmetric p.d. matrices. The underlying family is $\mathcal{P} = \{\mathbf{P}_\Sigma : \Sigma \in \Theta\}$ where $\mathbf{P}_\Sigma$ denotes the multivariate Normal distribution with mean zero and varaince Σ. Here, Θ denotes set of all $k \times k$ symmetric p.d. matrices.

Example 2.3. Suppose $X \sim p(x : \theta)$ where $\theta \in \Theta = \{1, 2\}$, where $p(x : 1) =$ pmf of Binomial $(4, 1/2)$ and $p(x : 2) =$ pmf of Poisson(1). Then the underlying family is $\mathcal{P} = \{p(x : \theta) : \theta \in \Theta\}$.

Example 2.4. We illustrate the case of location and location-scale families of distributions.

(i) $\mathcal{P} = \{p(x : \mu) = h(x - \mu) : \mu \in \mathbb{R},\ h \text{ is a fixed pdf}\}$. This is an example of a *location* family. Consider $h(x) = \phi(x)$, the pdf of standard normal. This family will contain all normal distribution with some mean μ and unit variance.

(ii) $\mathcal{P} = \{p(x : \mu, \sigma) = \sigma^{-1} h((x - \mu)/\sigma) : \mu \in \mathbb{R},\ \sigma \in (0, \infty),\ h \text{ is a fixed pdf}\}$. This is an example of a *location-scale* family. Consider again the same choice of $h(x)$ as above. This family will contain all possible normal distributions.

(iii) The *scale* family is obtained by replacing $\mu = 0$ in the location-scale family. If $h(x) = \phi(x)$, then this will constitute the class of all normal distributions with mean zero and some finite variance.

The choice of h is fixed, instead of Normal distribution, we can use any other pmf of pdf to generate the location and scale families.

Our main goal is to know about the underlying θ. Once the data is collected, the statistician is interested in *data comprsession*, which leads to the idea of sufficiency. As the name suggests, we intend to retain only that part of the data which is relevant for our purposes. The following example makes this clear.

Example 2.5. Consider a coin tossing experiment with n coin tosses with outcomes denoted by 0 and 1. The observed values are $\{X_1, \ldots, X_n\}$. This data vector contains two pieces of information: (a) the number of 1's in the n tosses and, (b) the position of zero's and one's in the sequence of tosses.

In order to estimate the unknown success probability p of this coin, it is only relavent for us to know the information in part (a). Part (b) is completely irrelavent to us as far as estimation of p is concerned. If we are supplied with the information $T = \sum_{i=1}^{n} X_i$, then on the basis of this we can reconstruct a sequence $\{X_1', \ldots, X_n'\}$, such that the unconditional distributions of $\{X_1', \ldots, X_n'\}$ will be the same as of $\{X_1, \ldots, X_n\}$.

If two persons toss this same coin and obtain the same value of $T = t$, then both will have the same information regarding p, irrespective of how the zero's and one's were ordered. This leads to the idea of partitioning a sample space. Consider the sample space $\mathcal{X}$ of all possible outcomes, containing 2^n sample points (denoted by $\mathbf{x} = \{x_1, \ldots, x_n\}$). Given a value of $T = t \in \{0, 1, \ldots, n\}$, we can partition $\mathcal{X}$ into $(n+1)$ disjoint sets $\mathcal{A}_t = \{\mathbf{x} : T(\mathbf{x}) = t\}$. In this particular example, every element of $\mathcal{A}_t$ will be a permutation of another element of $\mathcal{A}_t$.

Now we describe the *Sufficiency Principle* and a *Sufficient Statistic.*

Definition 1 (Sufficiency Principle)**.** Consider a family of distributions $\mathcal{P} = \{\mathbf{P}_\theta : \theta \in \Theta\}$. If $T(\mathbf{X}) = T(X_1, \ldots, X_n)$ is sufficient for θ (or for the family of distributions $\mathcal{P}$), then any inference about θ should be based on the sample $\{X_1, \ldots, X_n\}$, only through the value $T(\mathbf{X})$, *i.e.* if $\mathbf{x}'$ and $\mathbf{x}''$ are two sample points, such that $T(\mathbf{x}') = T(\mathbf{x}'')$, then the inference about θ should be the same whether $\mathbf{X} = \mathbf{x}'$ or $\mathbf{X} = \mathbf{x}''$ is observed.

Now, we define a *Sufficient Statistic.*

Definition 2 (Sufficient Statistic)**.** A statistic $T(\mathbf{X})$ is a sufficient statistic for the family $\mathcal{P}$ (or sufficient for θ), if the conditional distribution of the sample $\mathbf{X}$, given the value of $T(\mathbf{X})$, does not depend on θ. Mathematically, $\mathbf{P}_\theta(\mathbf{X} \in A \mid T = t)$ is independent of θ for all sets A and all values of t.

Remark 1. When $\mathbf{P}_\theta(T = t) = 0$, for some t, then a more technical treatment can be used to define sufficiency. See (Lehmann and Romano (2005), Ch. 2) for more details if you are interested. Sometimes we say T `is sufficient for` θ, and that inherently means T is sufficient for the family of distributions (which is indexed by θ).

As an example we consider how a sufficient statistic leads to obtaining equal information, as compared to the full data-set $\{X_1, \ldots, X_n\}$. Suppose there are two persons, A and B. Person A observes $\mathbf{X} = \mathbf{x}$ and then computes $T(\mathbf{X}) = T(\mathbf{x}) = t$. While, person B observes only $T(\mathbf{X}) = t$. Since T is sufficient, the conditional probability distribution of $[\mathbf{X} \mid T = t]$ can be computed without knowing θ. Recall that

$$\mathbf{P}(\mathbf{X} = \mathbf{x} \mid T = t) = \begin{cases} 0 & \text{if } T(\mathbf{x}) \neq t, \\ \mathbf{P}(\mathbf{X} = \mathbf{x} \mid T = t) & \text{o.w.} \end{cases}$$

The support of the conditional distribution of $[\mathbf{X} \mid T = t]$ is the set $\mathcal{A}_t = \{\mathbf{y} : T(\mathbf{y}) = t\}$. Consider a new r.v. $\mathbf{X}'$, such that the conditional distribution of $\mathbf{X}'$ given each value of $T = t$ is specified in the following way:

$$\mathbf{P}(\mathbf{X}' = \mathbf{x} \mid T = t) = \mathbf{P}(\mathbf{X} = \mathbf{x} \mid T = t).$$

Since person B does not know the true value of $\mathbf{X}$, he can use a random number table to draw a new value of $\mathbf{X}'$, whose conditional distribution given T is the same as that of $\mathbf{X}$ given T. Then, the unconditional distribution of $\mathbf{X}'$ will be

$$\mathbf{P}_\theta(\mathbf{X}' = \mathbf{x}) = \sum_t \mathbf{P}(\mathbf{X}' = \mathbf{x} \mid T = t) \cdot \mathbf{P}_\theta(T = t) = \sum_t \mathbf{P}(\mathbf{X} = \mathbf{x} \mid T = t) \cdot \mathbf{P}_\theta(T = t) = \mathbf{P}_\theta(\mathbf{X} = \mathbf{x}).$$

This tells you that person B can obtain a r.v. $\mathbf{X}'$ which has the same unconditional distribution as that of the unknown $\mathbf{X}$. Hence, he loses no information, even without knowing the true value of $\mathbf{X}$. Once we know the value of the sufficient statistic T, the sample $\{X_1, \ldots, X_n\}$ will have no more information to convey about θ.

More simply, a sufficient statistic T splits the sample space of $\mathbf{X}$ into disjoint regions, depending the value of $T(\mathbf{x}) = t$, and within each such region, we do not need to keep track of the individual $\mathbf{x}$, that gave rise to $T(\mathbf{x}) = t$. All the information about the underlying parameters can be extracted by simply keeping track of values of T.

Although we have the definition of a sufficient statistic, it does not help us to find one. We will need to guess T and then verify the condition for sufficiency (in Definition 2) for all choices of A and t, which is not a very convenient method. Hence, we need a direct method to find a sufficient statistic.

Theorem 1 (Factorization Theorem)**.** *A necessary and sufficient condition for a statistic T to be sufficient for a family $\mathcal{P} = \{\mathbf{P}_\theta : \theta \in \Theta\}$ (of distributions dominated by a σ-finite measure μ) is that there exists a non-negative g_θ and h such that, the densities*[1] *p_θ of $\mathbf{P}_\theta$ satisfy*

$$p_\theta(\mathbf{x}) = g_\theta(T(\mathbf{x}))h(\mathbf{x}), \quad \textit{for all } \mathbf{x} \in \mathcal{X} \textit{ and each } \theta \in \Theta. \tag{2.1}$$

All we will need to do is to verify (2.1) to check if T is sufficient. In this theorem, we require $\mathbf{P}_\theta$ to have either a joint pdf or a joint pmf (referred above as *density*), in order to obtain the above factorization.

Proof of Theorem 1. We provide the proof when T and $\mathbf{X}$ are discrete random variables. Define the sets $A = [\mathbf{X} = \mathbf{x}]$ and $B = [T(\mathbf{X}) = T(\mathbf{x})]$. Then, $A \subseteq B$. If T is sufficient then the distribution of $[\mathbf{X} \mid T]$ is free of θ. Then,

$$\begin{aligned} p_\theta(\mathbf{x}) = \mathbf{P}_\theta(\mathbf{X} = \mathbf{x}) &= \mathbf{P}_\theta(\mathbf{X} = \mathbf{x},\ T(\mathbf{X}) = T(\mathbf{x})) \\ &= \mathbf{P}_\theta(\mathbf{X} = \mathbf{x} \mid T(\mathbf{X}) = T(\mathbf{x})) \cdot \mathbf{P}_\theta(T(\mathbf{X}) = T(\mathbf{x})) \\ &= h(\mathbf{x}) \cdot g_\theta(T(\mathbf{x})), \end{aligned}$$

as the first term will be free of θ by definition of a sufficient statistic and the second term will depend on θ and $\mathbf{x}$ through $T(\mathbf{x})$. This proves the necessity part of the theorem.

[1] We write $p_\theta(\mathbf{x})$ instead of $p(\mathbf{x} : \theta)$, although both are same.

Now assume that the factorization result in (2.1) holds. Then, for any t in the range of T, such that $T(\mathbf{x}) = t$,

$$\begin{aligned}\mathbf{P}_\theta(\mathbf{X} = \mathbf{x} \mid T(\mathbf{X}) = t) &= \frac{\mathbf{P}_\theta(\mathbf{X} = \mathbf{x})}{\mathbf{P}_\theta(T(\mathbf{X}) = t)} \\ &= \frac{g_\theta(T(\mathbf{x}))h(\mathbf{x})}{\sum_{\mathbf{y}:T(\mathbf{y})=t} \mathbf{P}_\theta(\mathbf{X} = \mathbf{y})} \\ &= \frac{g_\theta(T(\mathbf{x}))h(\mathbf{x})}{\sum_{\mathbf{y}:T(\mathbf{y})=t} g_\theta(T(\mathbf{y}))h(\mathbf{y})} \\ &= \frac{g_\theta(t)h(\mathbf{x})}{g_\theta(t)\sum_{y:T(\mathbf{y})=t} h(\mathbf{y})} = \frac{h(\mathbf{x})}{\sum_{y:T(\mathbf{y})=t} h(\mathbf{y})},\end{aligned}$$

which is free of θ. The quantity $g_\theta(t)$ has to be positive, otherwise the conditioning event will be a zero-probability event. In case we have t such that $T(\mathbf{x}) \neq t$, then the conditional probability will be zero and hence free of θ. This proves the sufficiency part.

In case $\mathbf{X}$ and T are continuous random variables, we replace the pmf $p(\mathbf{x} : \theta)$ by the pdf, and the theorem continues to hold. The proof of the continuous case requires the concept of conditional expectation in a measure-theoretic approach. The details are available in Lehmann and Romano (2005). □

Now, fix $\mathbf{x}$ and consider $p_\theta(\mathbf{x})$ as a function of θ. This function is known as the *likelihood* function and is denoted by,

$$L(\theta : \mathbf{x}) = p_\theta(\mathbf{x}), \quad \text{for all } \theta \in \Theta, \text{ for each fixed } \mathbf{x} \in \mathcal{X}. \tag{2.2}$$

The function,

$$\Lambda_{\mathbf{x}}(\theta_1, \theta_2) \equiv \frac{L(\theta_1 : \mathbf{x})}{L(\theta_2 : \mathbf{x})} = \frac{p_{\theta_1}(\mathbf{x})}{p_{\theta_2}(\mathbf{x})}, \quad \text{for all } \theta_1, \theta_2 \in \Theta, \tag{2.3}$$

is known as the *likelihood ratio function.* The next result shows a connection between the factorization result and the likelihood function.

Lemma 2. *Let $T(\mathbf{x})$ denote a statistic. Then the statements in (a) and (b) are equivalent.*

(a) There exists functions $h(\mathbf{x})$ and $g_\theta(T(\mathbf{x}))$, such that the factorization result in (2.1) holds.

(b) For any $\mathbf{x}, \mathbf{y} \in \mathcal{X}$, such that $T(\mathbf{x}) = T(\mathbf{y})$,

$$\Lambda_{\mathbf{x}}(\theta_1, \theta_2) = \Lambda_{\mathbf{y}}(\theta_1, \theta_2), \quad \textit{for all } \theta_1, \theta_2.$$

Proof of Lemma 2. If (2.1) holds and $T(\mathbf{x}) = T(\mathbf{y})$, then

$$\Lambda_{\mathbf{x}}(\theta_1, \theta_2) = \frac{p_{\theta_1}(\mathbf{x})}{p_{\theta_2}(\mathbf{x})} = \frac{g_{\theta_1}(T(\mathbf{x}))}{g_{\theta_2}(T(\mathbf{x}))} = \frac{g_{\theta_1}(T(\mathbf{y}))}{g_{\theta_2}(T(\mathbf{y}))} = \Lambda_{\mathbf{y}}(\theta_1, \theta_2),$$

for any pair of θ_1, θ_2, thereby proving part (b).

Conversely, suppose (b) holds. Fix any $\theta_0 \in \Theta$ and consider any other $\theta \in \Theta$. Then, for any $\mathbf{x}, \mathbf{y}$ satisfying $T(\mathbf{x}) = T(\mathbf{y})$, we can write

$$\frac{p_\theta(\mathbf{x})}{p_{\theta_0}(\mathbf{x})} = \Lambda_{\mathbf{x}}(\theta, \theta_0) = \Lambda_{\mathbf{y}}(\theta, \theta_0) = \frac{p_\theta(\mathbf{y})}{p_{\theta_0}(\mathbf{y})} \equiv g_{\theta,\theta_0}(T(\mathbf{x})), \tag{2.4}$$

for some function $g_{\theta,\theta_0}(\cdot)$, where the dependence on $\mathbf{x}$ (or $\mathbf{y}$), is only through the value of $T(\mathbf{x})$ (or $T(\mathbf{y})$). As θ_0 is fixed, we can write

$$g_{\theta,\theta_0}(T(\mathbf{x})) \equiv g_\theta(T(\mathbf{x})), \quad \text{for all } \theta \in \Theta \text{ and all } \mathbf{x} \in \mathcal{X}.$$

From (2.4), we can write

$$p_\theta(\mathbf{x}) = g_\theta(T(\mathbf{x})) \cdot p_{\theta_0}(\mathbf{x}) \equiv g_\theta(T(\mathbf{x})) \cdot h(\mathbf{x}).$$

Hence, we have shown that (b) is equivalent to (a). □

Example 2.6. Suppose $\{X_1, \ldots, X_n\}$ are an iid sample from $\mathcal{P} = \{f : f \text{ is a pdf }\}$. Then the order statistics $\big(X_{(1)}, \ldots, X_{(n)}\big)$ is sufficient for $\mathcal{P}$. Here the joint pdf of $\mathbf{X} = (X_1, \ldots, X_n)$ will be

$$f(\mathbf{x}) = \prod_{i=1}^n f(x_i) = \prod_{i=1}^n f(x_{(i)}).$$

So, we have $h(\mathbf{x}) = 1$ and $g_f(T(\mathbf{x})) = \prod_{i=1}^n f(x_{(i)})$, where $T(\mathbf{x}) = (x_{(1)}, \ldots, x_{(n)})$. Also, as the X_i are continuous, we can claim that $\mathbf{P}(X_i = X_j) = 0$, for each $i \neq j$. Hence, with probability 1, the order statistics satisfy the strict inequalities, $x_{(1)} < x_{(2)} < \cdots < x_{(n)}$. We note that, with $T(\mathbf{X}) = (X_{(1)}, \ldots, X_{(n)})$,

$$\mathbf{P}(\mathbf{X} = \mathbf{x} \mid T(\mathbf{X}) = (x_{(1)}, \ldots, x_{(n)})) = \begin{cases} 0 & \text{if } T(\mathbf{x}) \neq (x_{(1)}, \ldots, x_{(n)}), \\ 1/n! & \text{o.w.} \end{cases}$$

This is because, once the ordered values are supplied, the original sample can be only one of the possible $n!$ realizations which gave rise the supplied order statistics. The resulting conditional probability is free of the original distribution. The only requirement is to have a continuous distribution to prevent tied observations.

Example 2.7. Assume $\{X_1, \ldots, X_n\}$ are i.i.d. $\mathrm{N}(0, \sigma^2)$. Then the factorization theorem implies that $T_1(\mathbf{X}) = (X_1, \ldots, X_n)$, $T_2(\mathbf{X}) = (X_1^2, \ldots, X_n^2)$, $T_3 = (X_1^2 + \ldots + X_m^2, X_{m+1}^2 + \ldots + X_n^2)$ and $T_4 = X_1^2 + \ldots + X_n^2$ are all sufficient statistics. Each of them provide increasing reduction of data.

2.1 Minimal sufficiency

If T is a sufficient statistic for a family $\mathcal{P}$ and U is another statistic, such that $T = h(U)$ for some map h, then U will also be sufficient for the family $\mathcal{P}$. This follows directly from the factorization theorem. In Example 2.7, the statistic T_2 is a function of T_1 (but not the other way), similarly T_3 is a function of T_2 and T_4 is a function of T_3. The mapping connecting these statistics are not one-to-one maps. Hence, as i increases, T_i provides increasing reduction of data and provides a more coarser partition of the underlying sample space. The question then arises is there are sufficient statistic which provides the maximum possible reduction of data, or in other words, provides the coarsest possible partition of the sample space.

Definition 3 (Minimal sufficient statistic)**.** A statistic $T(\mathbf{X})$ is a minimal sufficient statistic for the family $\mathcal{P}$ if and only if, for any other sufficient statistic U (for the family $\mathcal{P}$), there exists some map h, such that $T = h(U)$. In other words, T is minimal sufficient, if it can expressed as a function of any other sufficient statistic.

This definition does not help us to find a minimal sufficient statistic. The theory for minimal sufficient statistics requires some technical tools that are beyond the scope of this course. However, the next result provides a necessary and sufficient condition for a statistic T to be minimal sufficient.

Lemma 3. *Recall the likelihood ratio defined in (2.3). A necessary and sufficient condition for a statistic $T(\mathbf{X})$ to be minimal sufficient is that,*

$$T(\mathbf{x}) = T(\mathbf{y}) \text{ if and only if } \Lambda_{\mathbf{x}}(\theta_1, \theta_2) = \Lambda_{\mathbf{y}}(\theta_1, \theta_2), \text{ for all } \theta_1, \theta_2. \tag{2.5}$$

Proof of Lemma 3. Suppose $T(\mathbf{x})$ satisfies (2.5) and $W(\mathbf{x})$ is a sufficient statistic. If T is not a function of W, then there exists two values $\mathbf{x}, \mathbf{y}$, for which $W(\mathbf{x}) = W(\mathbf{y})$, but $T(\mathbf{x}) \neq T(\mathbf{y})$. Using Lemma 2(b) on W, we can claim that

$$\Lambda_{\mathbf{x}}(\theta_1, \theta_2) = \Lambda_{\mathbf{y}}(\theta_1, \theta_2), \quad \text{for all } \theta_1, \theta_2.$$

But, using (2.5) it follows that $T(\mathbf{x}) = T(\mathbf{y})$, which is a contradiction. Therefore, T is a function of W. Since W was any arbitrary sufficient statistic, hence T is a function of every sufficient statistic and hence T is minimal sufficient.

We skip the proof for the reverse implication, *viz.* showing that T is minimal sufficient implies that (2.5) holds. For details refer to Theorem 6.1 of Young and Smith (2005). □

Remark 2. In Lemma 3 we provided a sufficient condition for a statistic to be minimal sufficient. For this, we only needed the following one sided implication in condition provided in (2.5),

$$T(\mathbf{x}) = T(\mathbf{y}) \quad \Leftarrow \quad \Lambda_{\mathbf{x}}(\theta_1, \theta_2) = \Lambda_{\mathbf{y}}(\theta_1, \theta_2), \quad \text{for all } \theta_1, \theta_2.$$

Usually this takes some effort.

There is a nice alternative way to find minimal sufficient statistics using the likelihood ratio.

Lemma 4. *The result contains two parts.*

(a) Suppose $\mathcal{P} = \{\mathbf{P}_0, \mathbf{P}_1, \ldots, \mathbf{P}_k\}$ is family containing $(k+1)$ distributions (a finite family). Assume $p_i(\mathbf{x})$ denotes the pdf (or pmf) of $\mathbf{P}_i$. Also assume that the family has common support. Then,

$$T(\mathbf{X}) = \left(\frac{p_1(\mathbf{X})}{p_0(\mathbf{X})}, \ldots, \frac{p_k(\mathbf{X})}{p_0(\mathbf{X})}\right)$$

is a minimal sufficient statistic for $\mathcal{P}$.

(b) Suppose $\mathcal{P}$ is a family of distributions (not necessarily a finite family) and $\mathcal{P}_0 \subset \mathcal{P}$ is a finite sub-family. Assume $\mathcal{P}_0$ and $\mathcal{P}$ have common support. Also assume that T is minimal sufficient for $\mathcal{P}_0$ and sufficient for $\mathcal{P}$. Then, T is minimal sufficient for $\mathcal{P}$.

Proof of Lemma 4. In part (a), firstly we need to show that T is sufficient. Write $h(\mathbf{x}) = p_0(\mathbf{x})$ and define the map, $g_i : \mathbb{R}^k \mapsto \mathbb{R}$ as,

$$g_i(t_1, \ldots, t_k) = \begin{cases} t_i & \text{if } 1 \leqslant i \leqslant k, \\ 1 & \text{if } i = 0. \end{cases}, \quad \text{for each } i = 0, 1, \ldots, k.$$

Then, we can write for each $i = 0, 1, \dots, k$,

$$p_i(\mathbf{x}) = \frac{p_i(\mathbf{x})}{p_0(\mathbf{x})} \cdot p_0(\mathbf{x}) = g_i\left(T(\mathbf{x})\right) \cdot h(\mathbf{x}), \quad \text{for each } \mathbf{x}.$$

Thus the factorization theorem can be now used to claim that T is sufficient. Now assume U is another sufficient statistic for $\mathcal{P}$. We need to show that $T = h(U)$ for some map h. As U is sufficient, the factorization theorem implies there are maps $\tilde{g}_i$ and $\tilde{h}$ such that,

$$p_i(\mathbf{x}) = \tilde{g}_i(U(\mathbf{x})) \cdot \tilde{h}(\mathbf{x}), \quad \text{for each } i = 0, 1, \dots, k,$$

Using this factorization in the expression for $T(\mathbf{x})$ we get,

$$T(\mathbf{x}) = \left(\frac{\tilde{g}_1(U(\mathbf{x}))}{\tilde{g}_0(U(\mathbf{x}))}, \dots, \frac{\tilde{g}_k(U(\mathbf{x}))}{\tilde{g}_0(U(\mathbf{x}))} \right) \equiv h(U(\mathbf{x})),$$

where $h : \mathbb{R} \mapsto \mathbb{R}^k$ is defined as

$$h(a) = \left(\frac{\tilde{g}_1(a)}{\tilde{g}_0(a)}, \dots, \frac{\tilde{g}_k(a)}{\tilde{g}_0(a)} \right),$$

which completes the proof.

If U is any other sufficient statistic for $\mathcal{P}$, then U is also sufficient for $\mathcal{P}_0$ and by definition of minimal sufficiency, $T = h(U)$ (with respect to all distributions in $\mathcal{P}_0$). Let $A = [T = h(U)]$ be a set in the (common) underlying sample space for all $\mathbf{P}(\in \mathcal{P})$). Then, as per the above information,

$$\mathbf{P}(A) = 1, \quad \text{for all } \mathbf{P} \in \mathcal{P}_0.$$

As the above statement is true for $\mathbf{P} \in \mathcal{P}_0$, and each $\mathbf{P} \in \mathcal{P}$ has common support, this implies, $\mathbf{P}(A^c) = 0$ for each $\mathbf{P} \in \mathcal{P}$. Hence, we can claim: $\mathbf{P}(T = h(U)) = 1$, for each $\mathbf{P} \in \mathcal{P}$, which proves minimal sufficiency. □

Example 2.8. Assume $\{X_1, \dots, X_n\}$ are i.i.d. Poisson (θ) where $\theta > 0$. In order to find a minimal sufficient statistic we use Lemma 3. Equating the likelihood ratio at two samples $\mathbf{x}$ and $\mathbf{y}$, we get

$$\Lambda_{\mathbf{x}}(\theta_1, \theta_2) = e^{-n(\theta_1 - \theta_2)} \cdot \left(\frac{\theta_1}{\theta_2} \right)^{\sum_{i=1}^n x_i} = e^{-n(\theta_1 - \theta_2)} \cdot \left(\frac{\theta_1}{\theta_2} \right)^{\sum_{i=1}^n y_i} = \Lambda_{\mathbf{y}}(\theta_1, \theta_2), \text{ for all } \theta_1, \theta_2 > 0.$$

Hence, we obtain the following relation: as $\theta_1, \theta_2 > 0$, the ratio $\theta_1/\theta_2 \in (0, \infty)$. Hence we get, the equality of two maps,

$$a^{\sum_{i=1}^n x_i} = a^{\sum_{i=1}^n y_i}, \quad \text{for each } a > 0.$$

This implies, $T(\mathbf{x}) \equiv \sum_{i=1}^n x_i = \sum_{i=1}^n y_i = T(\mathbf{y})$. Thus, we obtain the sufficient condition, as per Lemma 3 and we can claim that $T(\mathbf{X}) = \sum_{i=1}^n X_i$ is a sufficient statistic.

Example 2.9. Assume $\{X_1, \dots, X_n\}$ are i.i.d. Uniform(θ) where $\theta > 0$. Assume that $0 < \theta_1 < \theta_2$. Then, the likelihood ratio at a sample $\mathbf{x}$ is,

$$\Lambda_{\mathbf{x}}(\theta_1, \theta_2) = \frac{f(\mathbf{x} : \theta_1)}{f(\mathbf{x} : \theta_2)} = \begin{cases} (\theta_2/\theta_1)^n & \text{if } x_{(n)} \leqslant \theta_1, \\ 0 & \text{o.w.} \end{cases}$$

where $x_{(n)}$ denotes the largest order statistic. The case where $x_{(n)} > \theta_2$ is not considered as then the likelihood ratio is of the 0/0 form and remains undefined. Even otherwise, the likelihood ratio at two parameters θ_1, θ_2 makes sense only when the sample is such that $f(\mathbf{x} : \theta_1) + f(\mathbf{x} : \theta_2) > 0$, *i.e.*, the sample is within the support of at least one of the distributions.

Now, let $\mathbf{y}$ be another sample of size n and we are provided with the information (ref. Lemma 3), $L_{\mathbf{x}}(\theta_1, \theta_2) = L_{\mathbf{y}}(\theta_1, \theta_2)$, for each $0 < \theta_1 < \theta_2 < \infty$. Assume, **if possible** that $x_{(n)} < y_{(n)}$. Remember, as we are working with likelihood functions, the samples $\mathbf{x}$ and $\mathbf{y}$ are *fixed* and hence $\mathbf{x}_{(n)}$ and $\mathbf{y}_{(n)}$ are provided. Now choose θ_i, $i = 1, 2$, such that, $x_{(n)} < \theta_1 < y_{(n)} < \theta_2$. For this particular choice of (θ_1, θ_2),

$$L_{\mathbf{x}}(\theta_1, \theta_2) = \left(\frac{\theta_2}{\theta_2}\right)^n, \quad \text{and} \quad L_{\mathbf{y}}(\theta_1, \theta_2) = \left(\frac{\theta_2}{\theta_2}\right)^n \cdot \frac{\mathbf{1}(y_{(n)} \leqslant \theta_1)}{\mathbf{1}(y_{(n)} \leqslant \theta_2)} = 0.$$

So, at this (θ_1, θ_2), the likelihood ratios do not match, thereby leading to a contradiction to our original assumption (of equality of likelihood ratios). Thus, we can not have $x_{(n)} < y_{(n)}$ and similarly we can prove $x_{(n)} > y_{(n)}$ is also not possible, leading to the only possible conclusion, $T(\mathbf{x}) = x_{(n)} = y_{(n)} = T(\mathbf{y})$, thereby showing that $T(\mathbf{X}) = X_{(n)}$ is minimal sufficient.

Example 2.10. Assume $\{X_1, \ldots, X_n\}$ are i.i.d. $N(\mu, \sigma^2)$, where $(\mu, \sigma) \in \mathbb{R} \times (0, \infty)$. Also assume that $\{Y_1, \ldots, Y_n\}$ denotes another sample from the same distribution. Write the vector valued parameters $\boldsymbol{\theta}_i = (\mu_i, \sigma_i)$, for $i = 1, 2$, and assume that

$$L_{\mathbf{x}}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = L_{\mathbf{y}}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2), \quad \text{for each } \boldsymbol{\theta}_1, \boldsymbol{\theta}_2.$$

In order to show minimal sufficiency of $T(\mathbf{x}) = \left(\sum_{i=1}^n x_i, \sum_{i=1}^n x_i^2\right)^T$, we need to show that the above equality condition implies $T(\mathbf{x}) = T(\mathbf{y})$.

After simplification we can show that the above equality relation reduces to the following equivalent conditions,

$$\begin{aligned}
\frac{\sum_{i=1}^n (x_i^2 - y_i^2)}{2\sigma_2^2} + \frac{\mu_2 \sum_{i=1}^n (y_i - x_i)}{\sigma_2^2} &= \frac{\sum_{i=1}^n (x_i^2 - y_i^2)}{2\sigma_1^2} + \frac{\mu_1 \sum_{i=1}^n (y_i - x_i)}{\sigma_1^2}, \text{ for all } \mu_1, \mu_2 \in \mathbb{R},\ \sigma_1, \sigma_2 > 0, \\
\Leftrightarrow \frac{\sigma_1^2}{\sigma_2^2} \cdot \left(\frac{A}{2} + B\mu_2\right) &= \frac{A}{2} + B\mu_1, \quad \text{where } A = \sum_i (x_i^2 - y_i^2),\ B = \sum_i (y_i - x_i), \\
\Leftrightarrow \delta^2 \cdot \left(\frac{A}{2} + B\mu_2\right) &= \frac{A}{2} + B\mu_1, \quad \text{for all } \delta = \frac{\sigma_1}{\sigma_2} \in (0, \infty), \text{ and } \mu_1, \mu_2 \in \mathbb{R}, \\
\Leftrightarrow \frac{A(1 - \delta^2)}{2} + B(\mu_1 - \delta^2\mu_2) &= 0, \quad \text{for all } \delta > 0,\ \mu_1, \mu_2 \in \mathbb{R}, \\
\Leftrightarrow g(\delta, \mu_1, \mu_2) &= 0, \quad \text{for all } \delta > 0,\ \mu_1, \mu_2 \in \mathbb{R}, \text{ where } g \text{ is a map.}
\end{aligned}$$

By assumption, g is a constant map on its domain. It is also partially differentiable w.r.t. each of its three arguments. Hence, taking partial derivatives, we obtain

$$\begin{aligned}
0 &= \frac{\partial}{\partial \mu_1} g(\delta, \mu_1, \mu_2) = B, \\
0 &= \frac{\partial}{\partial \mu_2} g(\delta, \mu_1, \mu_2) = -\delta^2 B, \\
0 &= \frac{\partial}{\partial \delta} g(\delta, \mu_1, \mu_2) = -A\delta - 2B\mu_2\delta.
\end{aligned}$$

From the first two relations we get $B = 0$ and using it in the last relation, we obtain $A = 0$. Thus, we have shown that $\sum_{i=1}^n x_i = \sum_{i=1}^n y_i$ and $\sum_{i=1}^n x_i^2 = \sum_{i=1}^n y_i^2$. This shows $T(\mathbf{x})$ (defined above) will be minimal sufficient.

Example 2.11. Suppose $\{X_1, \dots, X_n\}$ is an i.i.d. sample from the following pdf

$$f(x:\theta) = \frac{1}{2\theta}\cdot\{\mathbf{1}(0 \leqslant x \leqslant \theta) + \mathbf{1}(2\theta \leqslant x \leqslant 3\theta)\}, \quad \text{where } \theta \in (0,\infty).$$

Then the joint distribution of $\mathbf{X}$ will be,

$$\begin{aligned} f(\mathbf{x}:\theta) &= \frac{1}{(2\theta)^n}\cdot\prod_{i=1}^{n}[\mathbf{1}(0 \leqslant x_i \leqslant \theta) + \mathbf{1}(2\theta \leqslant x_i \leqslant 3\theta)] \\ &= \frac{1}{(2\theta)^n}\cdot\left[\mathbf{1}(0 \leqslant x_{(n)} \leqslant \theta) + \mathbf{1}(0 \leqslant x_{(1)} \leqslant \theta)\cdot\mathbf{1}(2\theta \leqslant x_{(n)} \leqslant 3\theta) + \mathbf{1}(2\theta \leqslant x_{(1)} \leqslant 3\theta)\right]. \end{aligned}$$

The factorization theorem now implies that $T(\mathbf{X}) = (X_{(1)}, X_{(n)})^T$ are jointly sufficient for the family $\{f(\mathbf{x}:\theta) : \theta > 0\}$. In order to show minimal sufficiency, we need to do some extra work.

Example 2.12. Assume $\{X_1, \dots, X_n\}$ is an i.i.d. sample from the double exponential distribution with pdf

$$f(x:\theta) = \frac{e^{-|x-\theta|}}{2}, \quad \text{where } \theta \in \mathbb{R}.$$

Then, we can show using the factorization theorem, $T(\mathbf{X}) = (X_{(1)}, \dots, X_{(n)})^T$ is sufficient for this family. Now consider Lemma 3, and note that proving the sufficient condition in (2.5) is equivalent to showing that,

$$\frac{f(\mathbf{x}:\theta)}{f(\mathbf{y}:\theta)} \quad \text{is a constant function of } \theta \;\Rightarrow\; T(\mathbf{x}) = T(\mathbf{y}). \tag{2.6}$$

The condition in (2.6) is equally important in finding a minimal sufficient statistic.

Now, apply (2.6) to the double exponential joint density at two sample points $\mathbf{x}$ and $\mathbf{y}$, and note that,

$$\begin{aligned} \frac{f(\mathbf{x}:\theta)}{f(\mathbf{y}:\theta)} &= \exp\left\{-\sum_{i=1}^{n}|x_{(i)} - \theta| + \sum_{i=1}^{n}|y_{(i)} - \theta|\right\} = \text{constant function of } \theta \\ \Leftrightarrow \sum_{i=1}^{n}|x_{(i)} - \theta| &= \sum_{i=1}^{n}|y_{(i)} - \theta| + c(\mathbf{x},\mathbf{y}), \quad \text{for all } \theta \in \mathbb{R}, \end{aligned} \tag{2.7}$$

where $c(\mathbf{x},\mathbf{y})$ depends only on $\mathbf{x}$ and $\mathbf{y}$. In order to minimal sufficiency of the order statistics, we have to show that (2.7) implies $X_{(i)} = Y_{(i)}$ for each $i = 1, \dots, n$.

Firstly note that the relation in (2.7) is true for each $\theta \in \mathbb{R}$, which is a very strong condition. Also note that we can assume, $\mathbf{P}(X_i = X_j) = 0 = \mathbf{P}(Y_i = Y_j)$, for each $i, j \in \{1, \dots, n\}$. Now fix any $\theta_\star < \min\{x_{(1)}, y_{(1)}\}$. For this $\theta_\star$, (2.7) reduces to

$$\sum_{i=1}^{n}x_{(i)} - n\theta_\star = \sum_{i=1}^{n}y_{(i)} - n\theta_\star + c(\mathbf{x},\mathbf{y}) \;\Rightarrow\; c(\mathbf{x},\mathbf{y}) = \sum_{i=1}^{n}(x_{(i)} - y_{(i)}).$$

If possible, assume that $x_{(1)} < y_{(1)}$. This can be split into two sub-cases: (a) $x_{(1)} < y_{(1)} \leqslant x_{(2)}$ and (b) $x_{(1)} < x_{(2)} \leqslant y_{(1)}$. In case (a), there exists infinitely many reals in the interval $(x_{(1)}, y_{(1)})$. Fix any such real $\theta_\star \in (x_{(1)}, y_{(1)})$. For this $\theta_\star$, (2.7) reduces to,

$$\begin{aligned} \sum_{i=2}^{n}\{x_{(i)} - \theta_\star\} - \{x_{(1)} - \theta_\star\} &= \sum_{i=1}^{n}\{y_{(i)} - \theta_\star\} + \sum_{i=1}^{n}\{x_{(i)} - y_{(i)}\} \\ \Leftrightarrow \sum_{i=2}^{n}x_{(i)} - (n-1)\theta_\star - \{x_{(1)} - \theta_\star\} &= \sum_{i=1}^{n}y_{(i)} - n\theta_\star + \sum_{i=1}^{n}\{x_{(i)} - y_{(i)}\} \\ \Leftrightarrow -\{x_{(1)} - \theta_\star\} &= \{x_{(1)} - \theta_\star\} \\ \Leftrightarrow \theta_\star &= x_{(1)}, \end{aligned}$$

which is contradiction. This shows that case (a) is not possible. If case (b) is true, again fix a $\theta_\star \in (x_{(1)}, x_{(2)})$ and note that (2.7) reduces to,

$$\sum_{i=2}^{n}\{x_{(i)} - \theta_\star\} - \{x_{(1)} - \theta_\star\} = \sum_{i=1}^{n}\{y_{(i)} - \theta_\star\} + \sum_{i=1}^{n}\{x_{(i)} - y_{(i)}\}$$

$$\Leftrightarrow \;\; \theta_\star = x_{(1)},$$

using the same argument. This shows, case (b) is not possible. Combining both, we can claim that $x_{(1)} < y_{(1)}$ is not possible. Similarly we can argue that $x_{(1)} > y_{(1)}$ is also not possible and hence the only possibility is $x_{(1)} = y_{(1)}$. Now, using this finding in (2.7) we can reuse the same argument for showing $x_{(2)} = y_{(2)}$, by noting that $c(\mathbf{x}, \mathbf{y}) = \sum_{i=2}^{n}\{x_{(i)} - y_{(i)}\}$. The process can be continued consecutively to show $x_{(i)} = y_{(i)}$ for each $i = 1, \ldots, n$, thereby showing that the order statistics are minimal sufficient.

Similarly, in case of Logistic distribution (with a location parameter θ) the order statistics can be shown as minimal sufficient, but require the use of different techniques for proving these results (see Chapter 1 in Lehmann and Casella (1998)).

Example 2.13. Suppose $\{X_1, \ldots, X_n\}$ are iid Cauchy $(\theta, 1)$ with pdf

$$f(x:\theta) = \frac{1}{\pi\big\{1 + (x-\theta)^2\big\}}, \quad x \in \mathbb{R},\; \theta \in \mathbb{R},$$

the sufficient statistic will be $\mathbf{T} = \big(X_{(1)}, \ldots, X_{(n)}\big)$. Assume that

$$\frac{f(\mathbf{x};\theta)}{f(\mathbf{y};\theta)} = \frac{\prod_{i=1}^{n} 1 + (y_i - \theta)^2}{\prod_{i=1}^{n} 1 + (x_i - \theta)^2} = \psi(\mathbf{x}, \mathbf{y}), \quad \text{for all } \theta \in \mathbb{R}. \tag{2.8}$$

The numerator and denominator in (2.8) are polynomials in θ of degree $2n$, and their ratio is independent of θ. Note that comparing the coefficients of the highest degree terms in the numerator and denominator, we have $\psi(\mathbf{x}, \mathbf{y}) = 1$. This implies

$$h_1(\theta) \equiv \prod_{i=1}^{n}\big\{1 + (y_i - \theta)^2\big\} = \prod_{i=1}^{n}\big\{1 + (x_i - \theta)^2\big\} \equiv h_2(\theta), \quad \text{for all } \theta \in \mathbb{R}.$$

Since they are of degree $2n$, let us factorize

$$h_i(\theta) = \prod_{k=1}^{2n}(\theta - v_{i,k}), \quad \text{for each } i = 1, 2,$$

where $\{v_{i,k} : 1 \leqslant k \leqslant 2n\}$ are the roots of h_i. Then, substituting $\theta = v_{1,1}$, we have

$$h_1(v_{1,1}) = 0 = h_2(v_{1,1}) = \prod_{k=1}^{2n}(v_{1,1} - v_{2,k}).$$

Since, the product is zero, there must exist an $v_{2,k}$ such that $v_{2,k} = v_{1,1}$. Similarly, there exists some $v_{2,i}$ such that $v_{2,i} = v_{1,j}$ for each $j \in \{1, \ldots, 2n\}$. Hence, h_1 and h_2 have the same set of roots. Computing the roots explicitly, the roots of h_1 are

$$\{y_k \pm \imath : 1 \leqslant k \leqslant n\}, \quad \text{where } \imath = \sqrt{-1},$$

and similarly the roots of h_2 are $\{x_k \pm \imath : k = 1, \ldots, n\}$. Since the roots are a permutation of one another, we must have $x_i = y_j$ for some j for all $i = 1, \ldots, n$. Hence $x_{(i)} = y_{(i)}$. This completes the proof.

For more results on minimal sufficient statistics, you can refer to Shao (2003) and Lehmann and Casella (1998). Although minimal sufficiency is a technical challenging topic, but it has limited utility, atleast in the context of this course. We will revisit this topic when the need arises.

2.2 Exponential families of distributions

Many commonly used distributions belong to a larger class of distributions, known as the exponential family of distributions. The exponential family, as the name suggests uses the exponential form to express a pdf or pmf in a standard form. If we develop certain results for the exponential family, then it enables us to use them for a wide range of distributions. Typical examples of distributions which belong to the exponential family are, Binomial, Poisson, Negative Binomial, Gamma, Normal, Exponential, Multinomial, Multivariate normal and many more. We will encounter more examples as we proceed. However, distributions like Uniform, Cauchy, Logistic, Double exponential, Hypergeometric etc., do not belong the exponential family.

The topic of exponential families is discussed because many of the exact inference procedures (unbiased minimum variance estimation, uniformly most powerful tests, etc.) are usable in the exponential family setup and they utilise the various useful properties of exponential families.

Definition 4 (k-parameter exponential family)**.** A random vector $\mathbf{X}$ is said to have a distribution which belongs to the *k-parameter exponential family*, if it has a pdf or pmf $p_{\boldsymbol{\theta}}(\mathbf{x})$, which can be expressed as,

$$p_{\boldsymbol{\theta}}(\mathbf{x}) = \exp\left\{\sum_{i=1}^{k} \eta_i(\boldsymbol{\theta})T_i(\mathbf{x}) - \psi(\boldsymbol{\theta})\right\} \cdot h(\mathbf{x}), \quad \text{for all } \mathbf{x} \in \mathcal{X}, \text{ for all } \boldsymbol{\theta} \in \Theta, \tag{2.9}$$

where, $T_1(\mathbf{x}), \ldots, T_k(\mathbf{x})$ are sufficient statistics for the family, $\eta_i : \Theta \mapsto \mathbb{R}$ are real valued parametric functions (for each $i = 1, \ldots, k$), $\psi(\boldsymbol{\theta})$ is a parametric function and $h(\cdot)$ depends only on $\mathbf{x}$. The underlying parameter space for $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^m$, for some $m \geqslant 1$.

As a special case of the k-parameter exponential family, we have the *one parameter exponential family* distribution, which can be represented as (2.9) with $k = 1$, *viz.*,

$$p_{\boldsymbol{\theta}}(\mathbf{x}) = \exp\{\eta(\boldsymbol{\theta})T(\mathbf{x}) - \psi(\boldsymbol{\theta})\} \cdot h(\mathbf{x}), \quad \text{for all } \mathbf{x} \in \mathcal{X},$$

where, $T(\mathbf{x})$ is a sufficient statistic for the family and $\eta : \Theta \mapsto \mathbb{R}$ is a parametric function. Here $\boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^m$, for some $m \geqslant 1$. Note, in the *one*-parameter exponential family, typically $\boldsymbol{\theta}$ is a scalar parameter (with $m = 1$).

Also note that, the representation in (2.9) is not unique. We can write $\eta_i T_i = (c\eta_i)(T_i/c)$, for some $c > 0$, and obtain an alternative representation. Also, this representation suggests that it does not have a support that depends on $\boldsymbol{\theta}$. If so, then it would require $h(\mathbf{x})$ to involve $\boldsymbol{\theta}$, as the exponential term in (2.9) is always positive.

Example 2.14. Assume $\{X_1, \ldots, X_n\}$ are i.i.d. $\mathrm{N}(\mu, \sigma^2)$, where $\boldsymbol{\theta} = (\mu, \sigma) \in \Theta = \mathbb{R} \times (0, \infty)$. Then, we can write

$$
\begin{aligned}
p_{\boldsymbol{\theta}}(\mathbf{x}) &= \frac{1}{\left(\sqrt{2\pi\sigma^2}\right)^n} \cdot \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right\}, \quad \text{for all } \mathbf{x}, \\
&= \exp\left\{\left(-\frac{1}{2\sigma^2}\right)\cdot\sum_{i=1}^{n}x_i^2 + \left(\frac{\mu}{\sigma^2}\right)\cdot\sum_{i=1}^{n}x_i - \frac{n\mu^2}{2\sigma^2} - n\log(\sigma)\right\} \cdot \frac{\prod_{i=1}^{n}\mathbf{1}(x_i \in \mathbb{R})}{\left(\sqrt{2\pi}\right)^n} \\
&= \exp\left\{\eta_1(\boldsymbol{\theta})T_1(\mathbf{x}) + \eta_2(\boldsymbol{\theta})T_2(\mathbf{x}) - \psi(\boldsymbol{\theta})\right\}\cdot h(\mathbf{x}),
\end{aligned}
$$

where, $\eta_1(\boldsymbol{\theta}) = -1/(2\sigma^2)$, $\eta_2(\boldsymbol{\theta}) = \mu/\sigma^2$, $T_1(\mathbf{x}) = \sum_{i=1}^{n} x_i^2$, $T_2(\mathbf{x}) = \sum_{i=1}^{n} x_i$, $\psi(\boldsymbol{\theta}) = \frac{n\mu^2}{2\sigma^2} + n\log(\sigma)$ and $h(\mathbf{x})$ is the last term not depending on $\boldsymbol{\theta}$. Thus, the joint distribution of $\mathbf{X}$ belongs to a two-parameter exponential family.

Assume $\{X_1, \ldots, X_n\}$ are i.i.d. Binomial(k, θ), where $\theta \in (0,1)$ and k is known. Then the joint law of $\mathbf{X}$ is

$$
\begin{aligned}
p_{\boldsymbol{\theta}}(\mathbf{x}) &= \prod_{i=1}^{n}\binom{k}{x_i}\theta^{x_i}(1-\theta)^{k-x_i}\mathbf{1}(x_i \in \{0,1,\ldots,k\}) \\
&= \theta^{\sum_{i=1}^{n}x_i}\cdot(1-\theta)^{nk-\sum_{i=1}^{n}x_i}\prod_{i=1}^{n}\binom{k}{x_i}\mathbf{1}(x_i \in \{0,1,\ldots,k\}) \\
&= \left(\frac{\theta}{1-\theta}\right)^{\sum_{i=1}^{n}x_i}\cdot(1-\theta)^{nk}\prod_{i=1}^{n}\binom{k}{x_i}\mathbf{1}(x_i \in \{0,1,\ldots,k\}) \\
&= \exp\left\{\left(\sum_{i=1}^{n}x_i\right)\cdot\log\left(\frac{\theta}{1-\theta}\right) + nk\log(1-\theta)\right\}\cdot h(\mathbf{x}) \\
&= \exp\left\{\left(\sum_{i=1}^{n}x_i\right)\cdot\log\left(\frac{\theta}{1-\theta}\right) - nk\log\left(\frac{1}{1-\theta}\right)\right\}\cdot h(\mathbf{x}) \\
&= \exp\{T(\mathbf{x})\cdot\eta(\theta) - \psi(\theta)\}\cdot h(\mathbf{x}),
\end{aligned}
$$

which shows that the Binomial distribution belongs to a one-parameter exponential family.

Assume $(X_1, \ldots, X_k)^T$ is a random vector with a Multinomial$(n, \theta_1, \ldots, \theta_k)$ distribution, where $0 < \theta_i < 1$, for each $i = 1, \ldots, k$, and $\sum_{i=1}^{k}\theta_i = 1$. So the parameter space for $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_{k-1})^T$ is $\Theta = \{(\theta_1, \ldots, \theta_{k-1}) : 0 < \theta_i < 1,\ \sum_{i=1}^{k-1}\theta_i < 1\}$. Notice that the sum constraint on the θ_i's implies one of the parameters is redundant. Also define $\mathcal{X} = \{(x_1, \ldots, x_{k-1}) : 0 \leqslant x_i \leqslant n,\ \sum_{i=1}^{k-1}x_i \leqslant n\}$. The joint pmf of $\mathbf{X} = (X_1, \ldots, X_{k-1})$ is

$$
\begin{aligned}
p_{\boldsymbol{\theta}}(\mathbf{x}) &= \frac{n!}{\prod_{i=1}^{k}x_i!}\cdot\prod_{i=1}^{k-1}\left(\frac{\theta_i}{1-\theta_i}\right)^{x_i}\cdot\left(1-\sum_{i=1}^{k-1}\theta_i\right)^{n}\cdot\mathbf{1}(\mathbf{x}\in\mathcal{X}) \\
&= \exp\left\{\sum_{i=1}^{k-1}x_i\cdot\log\left(\frac{\theta_i}{1-\theta_i}\right) - n\log\left(\frac{1}{1-\sum_{i=1}^{k-1}\theta_i}\right)\right\}\cdot h(\mathbf{x}).
\end{aligned}
$$

This shows that the distribution of $\mathbf{X}$ belongs to a $(k-1)$ parameter exponential family.

Example 2.15 (Marginal distribution of T). Consider the one-parameter exponential family density of $\mathbf{X}$ of the form

$$
p_{\boldsymbol{\theta}}(\mathbf{x}) = \exp\{\eta(\theta)T(\mathbf{x}) - \psi(\boldsymbol{\theta})\}\,h(\mathbf{x}), \quad \text{for all } \mathbf{x}\in\mathcal{X},
$$

where $\boldsymbol{\theta} \in \Theta$. Then T has also has an exponential family distribution of the same form. For simplicity, consider the discrete case where $p_{\boldsymbol{\theta}}$ is a pmf and $\mathbf{X}$ and T are discrete r.v's. Then for any t (in the range of T),

$$\mathbf{P}_{\boldsymbol{\theta}}(T = t) = \sum_{\mathbf{x}:T(\mathbf{x})=t} \mathbf{P}_{\boldsymbol{\theta}}(\mathbf{X} = \mathbf{x}) = e^{\eta(\boldsymbol{\theta})t - \psi(\boldsymbol{\theta})} \sum_{\mathbf{x}:T(\mathbf{x})=t} h(\mathbf{x}) = \exp\{\eta(\boldsymbol{\theta})t - \psi(\boldsymbol{\theta})\}\widetilde{h}(t),$$

where $\widetilde{h}(t) = \sum_{\mathbf{x}:T(\mathbf{x})=t} h(\mathbf{x})$. A similar argument can be carried out for the case where $p_{\boldsymbol{\theta}}$ is a pdf corresponding to a continuous r.v. $\mathbf{X}$. Note that the parametric functions $\eta(\boldsymbol{\theta})$ and $\psi(\boldsymbol{\theta})$ remain unchanged in the exponential family representation for the marginal distribution of T.

Instead of using parametric function $\eta_i : \Theta \mapsto \mathbb{R}$, which differs for each family of distributions, we can use a reparametrization to write $\eta_i = \eta_i(\boldsymbol{\theta})$ and work directly with the reparametrized pdf/pmf. This form of reparametrization leads to the *canonical form* of an exponential family.

Definition 5 (Canonical form of an exponential family)**.** Consider the exponential family density provided in (2.9) and write $\eta_i = \eta_i(\boldsymbol{\theta})$, for each $i = 1, \ldots, k$, and $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_k)^T$. Then, the reparametrized density is

$$p_{\boldsymbol{\eta}}(\mathbf{x}) = \exp\left\{\sum_{i=1}^{k} \eta_i T_i(\mathbf{x}) - \phi(\boldsymbol{\eta})\right\} \cdot h(\mathbf{x}), \quad \text{for all } \mathbf{x} \in \mathcal{X}, \text{ and all } \boldsymbol{\eta}, \tag{2.10}$$

where we assume that $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_k)^T = \big(\eta_1(\boldsymbol{\theta}), \ldots, \eta_k(\boldsymbol{\theta})\big)^T \in \mathcal{T} = \{\boldsymbol{\eta} : \eta_i = \eta_i(\boldsymbol{\theta}),\ \boldsymbol{\theta} \in \Theta\}$. The set $\mathcal{T}$ is known as the *natural parameter space* and the parameters $\boldsymbol{\eta}$ are known as the *natural parameters* for the exponential family. If the natural parameter space $\mathcal{T}$ contains an open set (in $\mathbb{R}^k$), then the exponential family (2.10) is said to be of *full rank.*

Example 2.16. In Example 2.14, in case of the $\mathrm{N}(\mu, \sigma^2)$ distribution, the original parameter space for $\boldsymbol{\theta} = (\mu, \sigma)$ is $\Theta = \mathbb{R} \times (0, \infty)$. We had, $\eta_1 = \eta_1(\boldsymbol{\theta}) = -1/(2\sigma^2)$ and $\eta_2 = \eta_2(\boldsymbol{\theta}) = \mu/\sigma^2$. As $\boldsymbol{\theta} \in \Theta$, we obtain the natural parameter space $\mathcal{T} = (-\infty, 0) \times \mathbb{R}$. This is different from the original parameter space. Also, in this case check that we can rewrite,

$$\psi(\boldsymbol{\eta}) = \frac{n\mu^2}{2\sigma^2} + \frac{n}{2}\log(\sigma^2) = -\frac{n\eta_2^2}{4\eta_1} - \frac{n}{2}\log(-2\eta_1).$$

Note that $\mathcal{T} = (-\infty, 0) \times \mathbb{R}$ and hence we can draw an open ball within this set. So, this is a *full rank* exponential family.

In the binomial example, we have $\eta = \eta(\theta) = \log(\theta/(1-\theta)$. As $\theta \in (0,1)$, $\eta \in \mathcal{T} = \mathbb{R}$, which contains an open set. So this is full rank exponential family. If we had selected $\Theta = \{1/4, 1/2, 3/4\}$, then $\mathcal{T}$ would be a discrete set containing three elements, which would not contain an open ball (and hence would not be a full rank exponential family).

In the multinomial case we will obtain a full rank exponential family with $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_{k-1})$.

Consider the family of distributions with pdf $p_\theta(x) = C\exp\{-(x-\theta)^4\}$, with $x \in \mathbb{R}$ and $\theta \in \mathbb{R}$. In this case, the joint pdf of $\mathbf{X}$ will be,

$$p_\theta(\mathbf{x}) = C^n \exp\left\{-\sum_{i=1}^{n} (x_i - \theta)^4\right\}, \quad \text{for all } \theta \in \mathbb{R}.$$

Then the vector of sufficient statistics for this family will be $\mathbf{T}(\mathbf{X}) = \left(\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2, \sum_{i=1}^n X_i^3\right)^T$, while the corresponding parameters will be $\boldsymbol{\eta}(\theta) = (\theta^3, \theta^2, \theta)^T$. If we write them in terms of natural parameters, we obtain $\eta_1 = \theta^3$, $\eta_2 = \theta^2$ and $\eta_3 = \theta$ and the natural parameter space will be,

$$\mathcal{T} = \left\{(a^3, a^2, a) : a \in \mathbb{R}\right\} \subset \mathbb{R}^3.$$

As a set in $\mathbb{R}^3$, $\mathcal{T}$ contains a single curve, and it does not contain an open set. Hence, this exponential family of distributions is not of full rank. When the natural parameters satisfy a nonlinear relationship, then the family is called a *curved* exponential family.

It can be shown that if an exponential family has full rank, then the natural sufficient statistics $(T_1(\mathbf{X}), \ldots, T_k(\mathbf{X}))$ are also minimal sufficient (cf. Corollary 6.16 of Lehmann and Casella (1998)). However, there are examples where the natural sufficient statistics are minimal sufficient, even though the family is not a full rank exponential family (Example 6.17 of Lehmann and Casella (1998)). Note that the natural parameter space $\mathcal{T}$ can also be written as,

$$\mathcal{T} = \left\{\boldsymbol{\eta} : \int \exp\{\sum_{i=1}^k \eta_i T_i(\mathbf{x})\} h(\mathbf{x})\ d\mathbf{x} < \infty\right\},$$

where the integral is replaced by a sum if we have a discrete distribution.

Example 2.17. Consider $k = 1$, $T_1(x) = x$ and $h(x) = e^{-|x|}$. Then the natural parameter space for this exponential family pdf will be,

$$\mathcal{T} = \left\{\eta : \int_{x\in\mathbb{R}} e^{\eta x} \cdot e^{-|x|}\ dx < \infty\right\}.$$

Writing this integral, we get

$$\int_{x\in\mathbb{R}} \exp\{\eta x - |x|\}\ dx = \int_{x<0} \exp\{x(\eta+1)\}\ dx + \int_{x\geqslant 0} \exp\{x(\eta-1)\}\ dx = \frac{1}{\eta+1} + \frac{1}{1-\eta}.$$

However, this requires us to ensure that $\eta < 1$ and $\eta > -1$, which implies the valid range of η is $(-1, 1)$. Hence the following pdf will constitute a valid exponential family,

$$f(x : \eta) = \exp\left\{\eta x - \log\left(\frac{2}{1-\eta^2}\right)\right\} \cdot e^{-|x|}, \quad x \in \mathbb{R} \text{ and } \eta \in (-1, 1).$$

This example shows that we can obtain *non-standard* distributions using the exponential family framework.

An important property of exponential families is the following: if each X_i are i.i.d. with a k-parameter exponential family density (pdf/pmf) of the form (2.9) with sufficient statistics $(T_1(X_i), \ldots, T_k(X_i))$, then the random vector $\mathbf{X} = (X_1, \ldots, X_n)$ also has an k-parameter exponential family density of the same form, with sufficient statistic $(\sum_{i=1}^n T_1(X_i), \ldots, \sum_{i=1}^n T_k(X_i))$, and normalizing constant $n\psi(\boldsymbol{\theta})$.

2.2.1 Some features of canonical exponential families

There are some other important properties of exponential families which we state below. Consider the canonical form of the exponential family and for simplicity consider the one-parameter case, with a single sufficient statistic. Hence, we are considering the following pdf (the case for pmf can be handled similarly),

$$f(\mathbf{x} : \eta) = \exp\{\eta T(\mathbf{x}) - \phi(\eta)\} h(\mathbf{x}), \quad \eta \in \mathcal{T}(\subseteq \mathbb{R}),\ \mathbf{x} \in \mathcal{X}. \tag{2.11}$$

Then the natural parameter space $\mathcal{T}$ is convex. Following the definition of the natural parameter space, it is enough for us to show that for any $\alpha \in (0,1)$ and $\eta_1, \eta_2 \in \mathcal{T}$,

$$\int_{\mathbf{x}\in\mathcal{X}} \exp\left\{(\alpha\eta_1 + (1-\alpha)\eta_2)T(\mathbf{x})\right\} h(\mathbf{x})\ d\mathbf{x} < \infty.$$

But, we can write the above integral as,

$$\int_{\mathbf{x}\in\mathcal{X}} \exp\left\{(\alpha\eta_1 + (1-\alpha)\eta_2)T(\mathbf{x})\right\} h(\mathbf{x})\ d\mathbf{x} = \int_{\mathbf{x}\in\mathcal{X}} \left(e^{\eta_1 T(\mathbf{x})}\right)^{\alpha} \cdot \left(e^{\eta_2 T(\mathbf{x})}\right)^{1-\alpha} \cdot h(\mathbf{x})\ d\mathbf{x}$$
$$\leqslant \left[\int_{\mathbf{x}\in\mathcal{X}} e^{\eta_1 T(\mathbf{x})} h(\mathbf{x})\ d\mathbf{x}\right]^{\alpha} \cdot \left[\int_{\mathbf{x}\in\mathcal{X}} e^{\eta_2 T(\mathbf{x})} h(\mathbf{x})\ d\mathbf{x}\right]^{(1-\alpha)} < \infty,$$

because both $\eta_1, \eta_2 \in \mathcal{T}$ and the integrals on the right are finite. Here, we used Holder's inequality[2] and $p = 1/\alpha$ and $q = 1/(1-\alpha)$. However, this proof actually has shown that

$$e^{\phi(\alpha\eta_1 + (1-\alpha)\eta_2)} \leqslant e^{\alpha\phi(\eta_1) + (1-\alpha)\phi(\eta_2)}, \quad \text{(this follows from the definition of } \phi(\cdot)\text{)},$$

which itself implies $\phi(\alpha\eta_1 + (1-\alpha)\eta_2) \leqslant \alpha\phi(\eta_1) + (1-\alpha)\phi(\eta_2)$, thereby also showing that $\phi(\cdot)$ in the canonical exponential family form (2.11) is also a convex function (on a convex domain $\mathcal{T}$).

Also, the map $\phi : \mathcal{T} \mapsto \mathbb{R}$ in (2.11) provides expression for moments of the natural sufficient statistic $T(\mathbf{X})$.

Lemma 5 (Moments of $T(\mathbf{X})$, Theorem 18.3 of DasGupta (2011))**.** *Fix any $\eta \in \mathcal{T}^0$ (where A^0 denotes the interior of a set A). Then,*

(a) $\mathbf{E}_\eta\left(T(\mathbf{X})\right) = \phi^{(1)}(\eta)$ *and* $\mathbf{Var}_\eta(T(\mathbf{X})) = \phi^{(2)}(\eta)$.

(b) The coefficients of skewness and kurtosis of $T(X)$ equal to,

$$\textit{skewness of } T = \frac{\psi^{(3)}(\eta)}{\left(\psi^{(2)}(\eta)\right)^{3/2}}, \quad \textit{kurtosis of } T = \frac{\psi^{(4)}(\eta)}{\left(\psi^{(2)}(\eta)\right)^{2}}.$$

(c) For any t such that $\eta + t \in \mathcal{T}$, the mgf of T exists and equals

$$M_\eta(t) = \exp\left\{\phi(\eta + t) - \phi(\eta)\right\}.$$

(d) The map $\phi(\eta) = \int_{\mathbf{x}} \exp\{\eta T(\mathbf{x})\}h(\mathbf{x})\ d\mathbf{x}$ is infinitely differentiable at each $\eta \in \mathcal{T}^0$ and the differentiation can be performed within the integral sign. Infact,

$$\frac{d^k}{d\eta^k}\ \phi(\eta) = \int_{\mathbf{x}} (T(\mathbf{x})^k \exp\{\eta T(\mathbf{x})\}\ h(\mathbf{x})\ d\mathbf{x}, \quad \textit{for each } k \geqslant 1 \textit{ and at each } \eta \in \mathcal{T}^0.$$

In case of a pmf the integral is replaced by a sum.

Example 2.18. Consider the family $\mathcal{P} = \{\mathbf{P}_\theta = \text{Bin}(n,\theta) : \theta \in (0,1)\}$, where n is unknown and $X \sim \mathbf{P}_\theta$. If we write this pmf in the usual exponential family form, then we obtain

$$p_\theta(x) = \exp\left\{x \log\left\{\frac{\theta}{1-\theta}\right\} + n\log(1-\theta)\right\} \cdot \binom{n}{x} \mathbf{1}(x \in \{0,1,\ldots,n\}),$$

[2] If f and g are maps on $\mathbb{R}^m$, then $\int |f(\mathbf{x})g(\mathbf{x})|\ d\mathbf{x} \leqslant (\int |f(\mathbf{x})|^p\ d\mathbf{x})^{1/p} \cdot (\int |g(\mathbf{x})|^q\ d\mathbf{x})^{1/q}$, for any $p, q \geqslant 1$, such that $\frac{1}{p} + \frac{1}{q} = 1$.

which leads to,

$$T(x) = x, \;\; \eta(\theta) = \log\left\{\frac{\theta}{1-\theta}\right\} \text{ and } \psi(\theta) = -n\log(1-\theta).$$

If we write this family in the canonical exponential family format, then we get, $\eta \in \mathcal{T} = \mathbb{R}$, with $\theta = \{1+\exp\{-\eta\}\}^{-1}$ and

$$\phi(\eta) = -n\log(1-\theta) = -n\log\left\{\left(1-\frac{1}{1+e^{-\eta}}\right)\right\} = n\log(1+e^{\eta}).$$

If we differentiate $\phi(\eta)$, we get (check),

$$\phi^{(1)}(\eta) = \frac{n}{1+e^{-\eta}} = \mathbf{E}_{\eta}(T(X)) = \mathbf{E}_{\theta}(T(X)) = n\theta.$$

This matches with the moment result for canonical exponential families, which is described in part (a) of Lemma 5. However, if we directly differentiate $\psi(\theta)$, we get,

$$\psi^{(1)}(\theta) = \frac{n}{1-\theta} \neq \mathbf{E}_{\theta}(T(X)).$$

So, the reparametrization in terms of the natural canonical parameters is important, and otherwise the moment relations will not hold. Obviously, after finding the moments of T in terms of η using these relations, we can revert back to the original expression in terms of θ. Similarly we can check that,

$$\phi^{(2)}(\eta) = \frac{ne^{\eta}}{\left(1+e^{\eta}\right)^2} = \mathbf{Var}_{\eta}(T(X)) = \mathbf{Var}_{\theta}(T(X)) = n\theta(1-\theta).$$

As a remark, note that the reparametrized version of $\mathbf{Var}_{\eta}(X)$ is an unfamiliar expression (as it is in terms of η), but it reduces to the original variance expression for binomials if we revert back to θ.

In case of multiparameter exponential family in the canonical form (2.10), with natural parameter $\boldsymbol{\eta}$ and sufficient statistics $(T_1(\mathbf{X}),\ldots,T_k(\mathbf{X}))$, for any $\boldsymbol{\eta} \in \mathcal{T}^0$, the map $\exp\{\phi(\boldsymbol{\eta})\}$ is infinitely partially differentiable w.r.t. each η_j, and the partial derivatives can be obtained by differentiating under the integral sign. Moreover, we have the following moment relations,

$$\mathbf{E}_{\boldsymbol{\eta}}(T_i(\mathbf{X})) = \frac{\partial}{\partial\eta_i}\,\phi(\eta) \quad \text{and} \quad \mathbf{cov}_{\boldsymbol{\eta}}(T_i(\mathbf{X}),T_j(\mathbf{X})) = \frac{\partial^2}{\partial\eta_i\partial\eta_j}\,\phi(\boldsymbol{\eta}), \quad \text{for all } i,j \in \{1,\ldots,k\}. \tag{2.12}$$

Example 2.19. Consider the Binomial(k,θ) example and check if $\mathbf{E}_{\eta}(T(X)) = \phi^{(1)}(\eta)$. Consider the $N_2(0,0,1,1,\rho)$ example and write it as an exponential family. Consider the multinomial case and show that the moment relations hold.

Example 2.20. Consider the Cauchy r.v. X with pdf

$$f(x:\theta) = \frac{1}{\pi\left\{1+(x-\theta)^2\right\}}, \quad \text{for all } x \in \mathbb{R},$$

and $\theta \in \mathbb{R}$. This distribution does not belong to an exponential family and we will prove this. Assume, if possible that X belonged to an one-parameter exponential family, such that

$$\frac{1}{\pi\left\{1+(x-\theta)^2\right\}} = a(\theta)h(x)e^{\eta(\theta)T(x)}, \quad \text{for all } x \in \mathbb{R}, \text{ and all } \theta \in \mathbb{R},$$

where $a(\theta) = e^{-\psi(\theta)}$, $\eta(\cdot)$ and $T(\cdot)$ denotes some suitable parametric function and a sufficient statistic. If we use $\theta = 0$ in this relation, then we obtain

$$\begin{aligned}\log(1/\pi) - \log(1+x^2) &= \log\{a(0)\} + \log\{h(x)\} + \eta(0)T(x) \\ \Leftrightarrow\ T(x) &= K_1 - K_2 \log\{h(x)(1+x^2)\},\end{aligned}$$

with constants K_1, K_2 free of x and θ. Using this expression for $T(x)$, we can write

$$\begin{aligned}\log(1/\pi) - \log\{1+(x-\theta)^2\} &= \log\{a(\theta)\} + \log\{h(x)\} + \eta(\theta)T(x) \\ \Leftrightarrow\ \eta(\theta) &= \frac{\log(1/\pi) - \log\{a(\theta)\} - \log\{1+(x-\theta)^2\} - \log\{h(x)\}}{K_1 - K_2\log\{h(x)(1+x^2)\}}.\end{aligned}$$

The l.h.s. above depends only on θ, while the expression on the r.h.s. depends on both x and θ. This indicates that the r.h.s. is a constant function of x. Suppose we fix $\theta = 1$, then it implies each $x \in \mathbb{R}$ must satisfy the equation,

$$\eta(1) = \frac{K_3 - \log\{h(x)\} - \log\{1+(x-1)^2\}}{K_1 - K_2\log\{h(x)(1+x^2)\}}$$

If we apply a change of measure, by absorbing the factor $h(x)$ into the dx, and defining a new measure $d\mu(x) = h(x)dx$, then we can assume that the exponential family density (w.r.t. the measure μ) will have $h(x) = 1$ and other factors remain unchanged. The earlier argument also carries on, and we will obtain

$$\eta(1) = \frac{K_3 - \log\{1+(x-1)^2\}}{K_1 - K_2\log\{(1+x^2)\}}, \quad \text{for all } x \in \mathbb{R}.$$

Clearly this is impossible as the quantity on the r.h.s. is non-constant function of x. So, the one-parameter exponential family representation for Cauchy$(\theta, 1)$ distribution is not feasible.

There are some futher facts which we state without details. Consider the exponential family representation in (2.10) with sufficient statistics $(T_1, \ldots, T_r, T_{r+1}, \ldots, T_k)$ and natural parameters $(\eta_1, \ldots, \eta_r, \eta_{r+1}, \ldots, \eta_k)$. Then, the marginal distribution of $(T_1, \ldots, T_r)$ and $(T_{r+1}, \ldots, T_k)$ also follows an exponential family distribution and the conditional distribution of $(T_1, \ldots, T_r)$ given $(T_{r+1}, \ldots, T_k)$ also follows an exponential family distribution. For details, refer to Section 2.7 of Lehmann and Romano (2005). Exponential families have been widely used in various areas of statistics, for more details refer to Sundberg (2019).

2.2.2 Interchanging differentiation and integration in exponential families

In Lemma 5 we found that $\phi(\eta)$ can be differentiated by interchanging the integration and differentiation operations. We provide an explanation for this using the Dominated Convergence Theorem (DCT) (see Theorem 16.4 of Billingsley (1995)). We state the DCT below (without bothering about measure-theoretic details).

Theorem 6 (DCT). *Assume $\{f_n : n \geqslant 1\}$ is a sequence of real valued maps defined on some common domain $\mathcal{X}$. Assume that:*

(a) there exists some map f, such that $f_n(x) \to f(x)$, at each $x \in \mathcal{X}$.

(a) there exists a map g, such that $|f_n(x)| \leqslant g(x)$ for each $x \in \mathcal{X}$, and $\int_{\mathcal{X}} |g(x)|\, dx < \infty$.

Then, both f_n and f are integrable and $\int_{\mathcal{X}} f_n(x)\ dx \to \int_{\mathcal{X}} f(x)\ dx$, as $n \to \infty$.

So, the DCT is an theorem which provides conditions under which the limit of an integral would equal the integral of the limiting function. As we see, it allows us to take the limit within the integral. As an example consider the functions $f_n(x) = n^2 \mathbf{1}(0 < x < n^{-1})$. Then, for each fixed $x \in (0, \infty)$, $f_n(x) \to f(x) = 0$, so that $\int_0^\infty f(x)\ dx = 0$. But we see that $\int_0^{1/n} n^2\ dx = n \to \infty$, as $n \to \infty$. So the limit of the integral does not converge to the integral of the limit function. The proof of the DCT is beyond the scope of this course.

We are interested in showing that[3],

$$\frac{d}{d\eta}\ \phi(\eta) = \int_{\mathcal{X}} T(\mathbf{x}) e^{\eta T(\mathbf{x})} h(\mathbf{x})\ d\mathbf{x}, \quad \text{at each } \eta \in \mathcal{T}^0.$$

We consider an $\eta \in \mathcal{T}^0$ to ensure that we have no issues with taking derivatives from the right and left. For simplicity we consider an univariate η.

Recall that the exponential family density in the canonical form is

$$p(\mathbf{x} : \eta) = \exp\{\eta T(\mathbf{x}) - \phi(\eta)\} h(\mathbf{x}), \quad \text{for all } \mathbf{x} \text{ and all } \eta \in \mathcal{T}.$$

As a result,

$$e^{\phi(\eta)} = \int_{\mathcal{X}} e^{\eta T(\mathbf{x})} \cdot h(\mathbf{x})\ d\mathbf{x} \equiv g(\eta), \quad \text{(say)}.$$

Using this and taking derivatives on both sides we get,

$$e^{\phi(\eta)} \cdot \phi^{(1)}(\eta) = \frac{d}{d\eta}\ g(\eta).$$

The r.h.s. above is a derivative of an integral. From the definition of a derivative, we obtain

$$\begin{aligned} \frac{d}{d\eta} g(\eta) &\equiv \lim_{n\to\infty} \frac{g(\eta + a_n) - g(\eta)}{a_n}, \quad \text{where } a_n \to 0 \text{ as } n \to \infty, \\ &= \lim_{n\to\infty} \frac{1}{a_n} \cdot \int_{\mathcal{X}} \left\{ e^{(\eta + a_n)T(\mathbf{x})} - e^{\eta T(\mathbf{x})} \right\} \cdot h(\mathbf{x})\ d\mathbf{x} \\ &= \lim_{n\to\infty} \int_{\mathcal{X}} h(\mathbf{x}) \cdot e^{\eta T(\mathbf{x})} \cdot \frac{\left(e^{a_n T(\mathbf{x})} - 1\right)}{a_n}\ d\mathbf{x} \equiv \lim_{n\to\infty} \int_{\mathcal{X}} f_n(\mathbf{x})\ d\mathbf{x}, \quad \text{(say)}, \end{aligned}$$

where, we define the sequence of maps

$$f_n(\mathbf{x}) = h(\mathbf{x}) \cdot e^{\eta T(\mathbf{x})} \cdot \frac{\left(e^{a_n T(\mathbf{x})} - 1\right)}{a_n}, \quad \text{for each } \mathbf{x} \in \mathcal{X} \text{ and each } n \geqslant 1.$$

Our goal is to verify the required conditions for the DCT, for this sequence of maps $\{f_n : n \geqslant 1\}$. Note that for any fixed $\mathbf{x} \in \mathcal{X}$,

$$\begin{aligned} \lim_{n\to\infty} f_n(\mathbf{x}) = h(\mathbf{x}) \cdot e^{\eta T(\mathbf{x})} \cdot \lim_{n\to\infty} \frac{e^{a_n T(\mathbf{x})} - 1}{a_n} &= h(\mathbf{x}) \cdot e^{\eta T(\mathbf{x})} \cdot \lim_{a\to 0} \frac{e^{a T(\mathbf{x})} - 1}{a} \\ &= h(\mathbf{x}) \cdot e^{\eta T(\mathbf{x})} \cdot T(\mathbf{x}) \equiv f(\mathbf{x}), \quad \text{(say)}. \end{aligned}$$

So we have identified the limiting map f. We will use the inequalities, $|e^t - 1| \leqslant |t| e^{|t|}$, and $|t| \leqslant e^{|t|}$ (both are true for all $t \in \mathbb{R}$). Also set

$$a_n = \frac{\epsilon}{n}, \quad \text{for each } n \geqslant 1,$$

[3] A similar approach can be used to handle interchanging of differentiation and a (possibly infinite) summation.

where $\epsilon > 0$, is a fixed positive number which will be decided later. Then, with this choice of a_n,

$$\begin{aligned}
|f_n(\mathbf{x})| &\leqslant h(\mathbf{x})e^{\eta T(\mathbf{x})} \cdot \left| \frac{e^{a_n T(\mathbf{x})} - 1}{a_n} \right| \\
&\leqslant h(\mathbf{x})e^{\eta T(\mathbf{x})} \cdot \frac{1}{|a_n|} \cdot |a_n T(\mathbf{x})| \cdot e^{|a_n T(\mathbf{x})|}, \quad \text{(using the first inequality)}, \\
&= h(\mathbf{x})e^{\eta T(\mathbf{x})} \cdot \frac{|(\epsilon/n) \cdot T(\mathbf{x})|}{|\epsilon/n|} \cdot \exp\{|(\epsilon/n) \cdot T(\mathbf{x})|\} \\
&\leqslant h(\mathbf{x})e^{\eta T(\mathbf{x})} \cdot \frac{|\epsilon \cdot T(\mathbf{x})|}{|\epsilon|} \cdot \exp\{|\epsilon \cdot T(\mathbf{x})|\} \\
&\leqslant \frac{1}{\epsilon} \cdot h(\mathbf{x})e^{\eta T(\mathbf{x})} \cdot \exp\{2\epsilon \cdot |T(\mathbf{x})|\} \quad \text{(using second inequality)} \\
&\leqslant \frac{1}{\epsilon} \cdot h(\mathbf{x}) \cdot \exp\{\eta T(\mathbf{x}) + 2\epsilon \cdot |T(\mathbf{x})|\} \\
&\leqslant \frac{1}{\epsilon} \cdot \left[h(\mathbf{x}) \cdot e^{(\eta+2\epsilon)T(\mathbf{x})} + h(\mathbf{x}) \cdot e^{(\eta-2\epsilon)T(\mathbf{x})} \right] \equiv \widetilde{g}(\mathbf{x}), \quad \text{(say)}.
\end{aligned}$$

The choice of $\epsilon > 0$ is made in such a way that $\eta \pm 2\epsilon \in \mathcal{T}$, which will ensure that

$$\int_{\mathcal{X}} \exp\{(\eta \pm 2\epsilon)T(\mathbf{x})\}h(\mathbf{x})\ d\mathbf{x} < \infty.$$

This is possible because $\eta \in \mathcal{T}^0$ and as such we can find an $\epsilon > 0$ small enough, so that $(\eta - 2\epsilon, \eta + 2\epsilon) \subset \mathcal{T}$. By definition of $\mathcal{T}$, the above integral will be finite. Note, $\epsilon > 0$, may be small, but a fixed quantity, and hence $1/\epsilon$ will be finite (in the definition of $\widetilde{g}$). As $|f_n(\mathbf{x})| \leqslant \widetilde{g}(\mathbf{x})$ for all $\mathbf{x}$ and all $n \geqslant 1$, and $\widetilde{g}$ is integrable, this satisfies the second requirement of the DCT. Now, the interchange of limit and integration can be performed to obtain,

$$\begin{aligned}
& e^{\phi(\eta)} \cdot \phi^{(1)}(\eta) = \lim_{n\to\infty} \int_{\mathcal{X}} f_n(\mathbf{x})\ d\mathbf{x} = \int_{\mathcal{X}} \left[\lim_{n\to\infty} f_n(\mathbf{x})\right]\ d\mathbf{x} = \int_{\mathcal{X}} f(\mathbf{x})\ d\mathbf{x} = \int_{\mathcal{X}} T(\mathbf{x})e^{\eta T(\mathbf{x})}h(\mathbf{x})\ d\mathbf{x} \\
\Leftrightarrow\ & \phi^{(1)}(\eta) = \mathbf{E}_\eta T(\mathbf{X}).
\end{aligned}$$

The same argument can be repeated for higher order derivatives to obtain the expression in Lemma 5(d). The choice of a_n is crucial as it allows us to use how much *inside* the parameter η is within the parameter space, through the ϵ term[4].

2.3 Completeness and Ancillarity

A sufficient statistic contains *all* available information about the underlying parameters. The distribution of a sufficient statistic depends on the parameters, but the distribution of the data given the sufficient statistic is free of parameters. A statistic $V(\mathbf{X})$ is called *ancillary* if its distribution does not depend on the underlying parameter $\boldsymbol{\theta}$. This implies, if V is ancillary, it does not contain any information about the parameters. A statistic is called *first order ancillary*, if $\mathbf{E}_{\boldsymbol{\theta}}(V(\mathbf{X}))$ is a constant that does not depend on $\boldsymbol{\theta}$.

For example, if $\{X_1, \ldots, X_n\}$ are i.i.d. $N(0, \sigma^2)$, then $\sum_{i=1}^n X_i^2/X_1^2$ is ancillary. If $\{X_1, \ldots, X_n\}$ are i.i.d. $N(\mu, 1)$, then $(X_i - X_j)$ is ancillary. If $\{X_1, X_2\}$ are i.i.d. Uniform$(0, \theta)$, then $V = (X_1 - X_2)/(X_1 + X_2)$

[4] So far, any other alternative approach I tried has not been successful.

is ancillary. The property of ancillarity in the above examples is closely related to the location, scale, or location-scale structure of the underlying family of distributions. In the first of the above examples, we have

$$\frac{\sum_{i=1}^n X_i^2}{X_1^2} = \frac{\sum_{i=1}^n (X_i/\sigma)^2}{(X_1/\sigma)^2} = \frac{\sum_{i=1}^n Y_i^2}{Y_1^2},$$

whose distribution is free of σ. This is because each Y_i is i.i.d. $N(0,1)$. In the $N(\mu,1)$ case, it is a location family and we can write $X_i - X_j = (X_i - \mu) - (X_j - \mu)$ which is a difference of standard normals. In the Uniform$(0,\theta)$ case, X_i/θ is an Uniform$(0,1)$ r.v. and hence the ratio will have a distribution that will be free of θ.

Completeness is a purely technical concept that was created for obtaining desired results. Assume $\mathbf{X} \sim \mathbf{P}$ where $\mathbf{P} \in \mathcal{P}$ and let $T(\mathbf{X})$ be a real valued statistic. Then, $T(\mathbf{X})$ is called *complete* if

$$\mathbf{E_P} h(T) = 0, \quad \text{for all } \mathbf{P} \in \mathcal{P} \text{ implies } \quad \mathbf{P}(h(T) = 0) = 1, \quad \text{for all } \mathbf{P} \in \mathcal{P}. \tag{2.13}$$

If (2.13) holds only for bounded maps h, then T is called *boundedly complete*. Note, in order to show completeness of T, we must show that for any such h for which

$$\mathbf{E_P} h(T) = 0, \quad \text{for all } \mathbf{P} \in \mathcal{P}$$

holds, then for each such h we must have

$$\mathbf{P}(h(T) = 0) = 1, \quad \text{for all } \mathbf{P} \in \mathcal{P}.$$

Even if this fails to be true for some particular choice of map h_1 (say), then it will imply that T is not a complete statistic. So, a proof of completeness requires us to show (2.13) holds for an arbitrary choice of h.

Example 2.21. Assume $X \sim$ Binomial(n,θ) with $\theta \in (0,1)$. Assume h is a map which satisfies $\mathbf{E}_\theta h(X) = 0$, for each $\theta \in (0,1)$. This implies,

$$\sum_{i=0}^n h(i)\binom{n}{i}\theta^i(1-\theta)^{n-i} = 0, \quad \text{for all } \theta \in (0,1),$$
$$\Leftrightarrow \sum_{i=0}^n h(i)\binom{n}{i}\rho^i = 0, \quad \text{for all } \rho = \frac{\theta}{1-\theta} \in (0,\infty).$$

The lhs is a map in ρ and differentiable. Differentiating the lhs n times w.r.t ρ and equating to zero, we obtain

$$h(n)\binom{n}{n} = 0 \;\Rightarrow\; h(n) = 0.$$

Similarly again differentiating $(n-1)$ times and using $h(n) = 0$, we obtain $h(n-1) = 0$, and continuing this argument we get

$$h(0) = h(1) = \ldots = h(n) = 0.$$

This shows, for any $\theta \in (0,1)$, $\mathbf{P}_\theta(h(X) = 0) = 1$. Thus, X is complete statistic (under this family of distributions). Note that h may be a map with a larger domain, however we are not interested in values of $h(k)$ for $k > n$, as $\mathbf{P}_\theta(X > n) = 0$ for all θ, and hence for our purposes, $h(X) = 0$ w.p. 1, a.e. θ.

Lemma 7. *If T is a complete statistic and also a sufficient statistic for the family $\mathcal{P}$, and if a minimal sufficient statistic for $\mathcal{P}$ exists, then T is minimal sufficient.*

Proof of Lemma 7. If S is minimal sufficient, then by definition of minimal sufficient statistic, $S = h(T)$ for some map h. Write

$$g(T) = T - \mathbf{E}(T \mid S) = T - \mathbf{E}(T \mid h(T)).$$

Now, $\mathbf{E}_{\mathbf{P}} g(T) = \mathbf{E}_{\mathbf{P}}(T) - \mathbf{E}_{\mathbf{P}}\left[\mathbf{E}(T \mid h(T))\right] = 0$, for all $\mathbf{P} \in \mathcal{P}$, using the property of conditional expectations. So, by completeness property of T, the map g must satisfy,

$$1 = \mathbf{P}(g(T) = 0) = \mathbf{P}(T = \mathbf{E}(T \mid S)), \quad \text{for all } \mathbf{P} \in \mathcal{P}.$$

As $\mathbf{E}(T \mid S)$ is a function of S, this shows T is a function of S. Hence, T must be also a minimal sufficient statistic. □

The main reason for studying exponential families and completeness is the following fundamental result.

Theorem 8 (Completeness of full-rank exponential families)**.** *Assume $\mathbf{X}$ has a k-parameter canonical exponential family distribution of the form (2.10) where $\mathcal{T}$ is the natural parameter space. Assume that $\mathcal{T}$ contains an open set, or in other words, the exponential family has full rank. Then $\mathbf{T}(\mathbf{X}) = (T_1(\mathbf{X}), \ldots, T_k(\mathbf{X}))^T$ is complete and sufficient for $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_k)^T$ (or for the family $\mathcal{P} = \{p_{\boldsymbol{\eta}}(\mathbf{x}) : \boldsymbol{\eta} \in \mathcal{T}\}$).*

The proof of Theorem 8 is not difficult but requires some concepts that are beyond this course. If you are interested, a complete proof is provided in Theorem 4.4 of Bhattacharya et al. (2016). We are primarily interested in the application of this theorem, specially for unbiased estimation. The full rank condition is easy to verify once the exponential family is written in its canonical form. On the other hand, direct verification of completeness is not straightforward.

Example 2.22. Continuing from Example 2.21, consider family $\mathcal{P}_1 = \mathcal{P} \cup Q$, where $\mathcal{P}_0 = \{\text{Binomial}(n, \theta) : 0 < \theta < 1\}$ and $Q = \text{Poisson }(1)$. So, we have included the Poisson r.v. as an additional member and constructed a larger family. We have seen that X is a complete (and sufficient) statistic for $\mathcal{P}_0$. However, if we apply the argument for the larger family $\mathcal{P}_1$, then if h satisfies

$$\mathbf{E}_{\mathbf{P}} h(X) = 0, \quad \text{for all } \mathbf{P} \in \mathcal{P}_1,$$

then it implies, for all $\theta \in (0, 1)$,

$$\sum_{x=0}^{n} h(x) \cdot \binom{n}{x} \theta^x (1-\theta)^{n-x} = 0, \quad \text{for all } \theta \in (0,1), \text{ and} \quad \sum_{x=0}^{\infty} h(x) \cdot \frac{e^{-1}}{x!} = 0.$$

By earlier arguments, we obtain $h(0) = h(1) = \cdots = h(n) = 0$. Using this in the Poisson r.v. based constraint, we get

$$\sum_{x=n+1}^{\infty} \frac{h(x)}{x!} = 0. \tag{2.14}$$

This constraint can be satisfied by taking $h(n+3) = h(n+4) = \cdots = 0$, and setting $h(n+1)$ and $h(n+2)$ to satisfy

$$\frac{h(n+1)}{(n+1)!} = \frac{h(n+2)}{(n+2)!} \Leftrightarrow h(n+2) = -(n+2)h(n+1).$$

We can set $h(n+1) = a(\neq 0)$ and $h(n+2) = -(n+2)a(\neq 0)$. Similarly, (2.14) can be satisfied by infinitely many non-zero choices of h. This shows

$$\mathbf{P}(h(X) = 0) < 1, \quad \text{for some } \mathbf{P} \in \mathcal{P}_1.$$

As a result, X is not a complete statistic for this family. This shows that if a statistic is complete for a smaller family, then it need not be complete in a larger family.

Think: what will happen if $\mathcal{P}_1 = \mathcal{P}_0 \cup \{\text{Poisson}(\lambda) : \lambda > 1\}$.

If the smaller family *dominates* the larger family, then we can extend completeness for the larger family. We say that a family $\mathcal{P}_0$ *dominates* another family $\mathcal{P}_1$ (with $\mathcal{P}_0 \subseteq \mathcal{P}_1$) if, for any set A satisfying

$$\mathbf{P}(\mathbf{X} \in A) = 0, \quad \text{for all } \mathbf{P} \in \mathcal{P}_0 \quad \Rightarrow \quad \mathbf{P}(\mathbf{X} \in A) = 0, \quad \text{for all } \mathbf{P} \in \mathcal{P}_1.$$

In other words, if an event A has zero probability under all $\mathbf{P} \in \mathcal{P}_0$, then it must have zero probability under each $\mathbf{P} \in \mathcal{P}_1$. Definitely, in Example 2.22, the family of Binomial r.v.'s does not dominate the larger family (containing the Poisson r.v.), as the Binomials were supported on $\{0, 1, \ldots, n\}$.

Example 2.23. Consider the families $\mathcal{P}_0 = \{\text{Uniform}(0,\theta) : \theta \in (0,\infty) \cap \mathbb{Q}\}$ and $\mathcal{P} = \{\text{Uniform}(0,\theta) : \theta > 0\}$. Then $\mathcal{P}_0$ dominates $\mathcal{P}$. Assume A is an event such that, $\mathbf{P}_\theta(X \in A) = 0$, for all $\theta \in \mathcal{P}_0$. We wish to show the probability will be also zero for any irrational θ.

Consider any irrational $\theta_0 > 0$. Then, there exists a sequence of rational $\{\theta_m : m \geqslant 1\}$, such that $\theta_m \downarrow \theta$. Firstly note that, for any *fixed* $x > 0$, $\mathbf{1}(0 \leqslant x \leqslant \theta_m) \downarrow \mathbf{1}(0 \leqslant x \leqslant \theta)$, as $m \to \infty$.

If $x > \theta$, then $\mathbf{1}(0 \leqslant x \leqslant \theta) = 0$ and for some large enough m_0, we will have $\theta < \theta_m < x$, for all $m \geqslant m_0$, leading to $\mathbf{1}(0 \leqslant x \leqslant \theta_m) = 0$, for all $m \geqslant m_0$. Note this choice of m_0 will depend on x. If $x < \theta$, then both indicators will be equal to one. At $x = \theta$, definitely $\mathbf{1}(0 \leqslant \theta \leqslant \theta) = 1$. And, $\mathbf{1}(0 \leqslant \theta \leqslant \theta_m) = 1$, for all $m \geqslant 1$, leading to the convergence at $x = \theta$.

For any θ_m, define the map $f_m(x) = \theta_m^{-1} \cdot \mathbf{1}(x \in A \cap [0, \theta_m])$, for all $x \in (0, \infty)$. Then, as per information

$$0 = \mathbf{P}_{\theta_m}(X \in A) = \int_0^\infty f_m(x)\, dx, \quad \text{for all } m \geqslant 1.$$

But, for any fixed $x > 0$,

$$f_m(x) = \frac{\mathbf{1}(x \in A) \cdot \mathbf{1}(0 \leqslant x \leqslant \theta_m)}{\theta_m} \to \frac{\mathbf{1}(x \in A) \cdot \mathbf{1}(0 \leqslant x \leqslant \theta)}{\theta} \equiv f(x), \quad \text{(say)}.$$

Also for any $\epsilon > 0$, there exists $m_0 \in \mathbb{N}$, such that $\theta < \theta_m < \theta + \epsilon$, for all $m \geqslant m_0$. It is important to note that the choice of m_0 does not depend on x, and only depends on ϵ and θ. Thus, for all $m \geqslant m_0$,

$$0 \leqslant f_m(x) = \frac{\mathbf{1}(x \in A \cap [0, \theta_m])}{\theta_m} \leqslant \frac{\mathbf{1}(x \in A \cap [0, \theta + \epsilon])}{\theta} \equiv g(x), \quad \text{(say), for all } x > 0.$$

Then[5],

$$\int_0^\infty g(x)\ dx = \frac{1}{\theta}\int_0^{\theta+\epsilon} \mathbf{1}(x \in A)\ dx \leqslant \frac{\theta+\epsilon}{\theta} < \infty.$$

Hence both conditions of the DCT are satisfied. Hence, we can claim that

$$0 = \mathbf{P}_{\theta_m}(X \in A) = \int_0^\infty f_m(x)\ dx \to \int_0^\infty f(x)\ dx = \mathbf{P}_\theta(X \in A),$$

proving the desired result[6].

Lemma 9 (Completeness for dominated families)**.** *Assume $\mathcal{P}_0$ and $\mathcal{P}_1$ are two families of distributions such that $\mathcal{P}_0 \subset \mathcal{P}_1$ and $\mathcal{P}_0$ dominates $\mathcal{P}_1$. Then, if T is complete for $\mathcal{P}_0$, it will be complete for $\mathcal{P}_1$.*

Proof of Lemma 9. Since T is complete for $\mathcal{P}_0$, for any h satisfying

$$\begin{aligned} \mathbf{E}_{\mathbf{P}} h(T) = 0 &\quad \text{for each } \mathbf{P} \in \mathcal{P}_0, \\ \Rightarrow\ \mathbf{P}(h(T) = 0) = 1 &\quad \text{for each } \mathbf{P} \in \mathcal{P}_0, \\ \Rightarrow\ \mathbf{P}(h(T) \neq 0) = 0 &\quad \text{for each } \mathbf{P} \in \mathcal{P}_0, \\ \Rightarrow\ \mathbf{P}(h(T) \neq 0) = 0 &\quad \text{for each } \mathbf{P} \in \mathcal{P}_1. \end{aligned}$$

This proves the result. □

Theorem 10. *Consider the family $\mathcal{P}$ of all continuous distributions, i.e., $\mathcal{P} = \{f : f \text{ is a pdf}\}$. Assume $\{X_1, \ldots, X_n\}$ is an i.i.d. sample from this family. Then the order statistics $U(\mathbf{X}) = (X_{(1)}, \ldots, X_{(n)})$ are complete and sufficient for the family $\mathcal{P}$.*

Proof of Theorem 10. The sufficiency of the order statistics has already been proved. It remains to show the completeness part. Consider a subfamily $\mathcal{P}_0$ of $\mathcal{P}$ consisting of the following pdf's indexed by a parameter vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n) \in \mathbb{R}^n$,

$$\mathcal{P}_0 = \left\{ f(x : \boldsymbol{\theta}) = C(\boldsymbol{\theta}) \exp\left\{ -x^{2n} + \sum_{j=1}^n \theta_j x^j \right\} \mathbf{1}(x \in \mathbb{R}) : \theta_1, \ldots, \theta_n \in \mathbb{R} \right\}, \tag{2.15}$$

where $C(\boldsymbol{\theta})$ is a normalizing constant. Definitely, one needs to verify if the density provided is indeed a pdf.

Firstly note that for $\boldsymbol{\theta} \in \mathbb{R}^n$, and $|x| \geqslant 2$,

$$|\sum_{j=1}^n \theta_j x^j| \leqslant \sum_{j=1}^n |\theta_j||x|^j \leqslant \max_{1 \leqslant j \leqslant n} |\theta_j| \cdot \frac{|x| \cdot \{|x|^n - 1\}}{|x| - 1} \leqslant \theta^\star \cdot |x|^n, \quad \text{where } \theta^\star = 2 \max_{1 \leqslant j \leqslant n} |\theta_j|.$$

This implies,

$$\exp\{\sum_{j=1}^n \theta_j x^j - x^{2n}\} \leqslant \exp\left\{\theta^\star |x|^n - (|x|^n)^2\right\}.\text{‘}$$

[5] In particular, if we choose $\epsilon = \theta$, the upper bound will be equal to two. We could have used the upper bound function $\tilde{g}(x) = \theta^{-1}\mathbf{1}(x \in A)$, but we can not directly claim that $\int_0^\infty \mathbf{1}(x \in A)\ dx < \infty$.

[6] Essentially, $A \cap (0, \theta)$ has zero length if we consider any rational $\theta > 0$. So, can $A \cap (0, \theta)$ have positive length for any irrational $\theta > 0$? Intuitively, the answer is a clear 'No'. But the entire apparatus is needed to provide a rigorous proof.

Write, $a = |x|^n$, and consider the map: $g(a) = \theta^\star a - a^2$, for $a > 0$, where $\theta^\star > 0$. We claim that there exists a $M > 1$, such that for all $a > M$, $g(a) < -a^2/2$. Write $h(a) = a^2/2 - \theta^\star a$. Then, $h(0) = 0$ and $h'(a) = a - \theta^\star > 0$, when $a > M = \theta^\star$, or equivalently for all $|x| > (\theta^\star)^{1/n}$. Thus, we can claim,

$$\exp\{\sum_{j=1}^{n} \theta_j x^j - x^{2n}\} \leqslant \exp\left\{\theta^\star |x|^n - (|x|^n)^2\right\} \leqslant \exp\left\{-\frac{x^{2n}}{2}\right\}, \quad \text{for all } |x| > (\theta^\star)^{1/n}.$$

At the same time, $-a^{2n} < -a^2$, for all $a > 1$. Hence, we can claim that

$$\exp\{\sum_{j=1}^{n} \theta_j x^j - x^{2n}\} \leqslant \exp\left\{\theta^\star |x|^n - (|x|^n)^2\right\} \leqslant \exp\left\{-\frac{x^2}{2}\right\}, \quad \text{for all } |x| > \max\left\{1, (\theta^\star)^{1/n}\right\}.$$

This shows that the exponent term in the function $f(x : \boldsymbol{\theta})$ is dominated by an integrable map over all $|x| > \max\left\{1, (\theta^\star)^{1/n}\right\}$. Thus each $f(x : \boldsymbol{\theta})$ is integrable over $\mathbb{R}$ and hence a proper pdf. The exact form of $C(\boldsymbol{\theta})$ will be hard to calculate, but that is unnecessary for this proof.

Assume $\mathbf{X} = (X_1, \ldots, X_n)$ is an i.i.d. sample from $f(x : \boldsymbol{\theta})$ (in the family $\mathcal{P}_0$). The pdf's in the family $\mathcal{P}_0$ are in the natural canonical form with sufficient statistic

$$T(\mathbf{X}) = \left(\sum_{i=1}^{n} X_i, \sum_{i=1}^{2} X_i^2, \ldots, \sum_{i=1}^{n} X_i^n\right)^T, \tag{2.16}$$

and natural parameter space $\mathcal{T} = \mathbb{R}^n$. Hence, by our earlier result on completeness for exponential families, we can claim that $T(\mathbf{X})$ will be complete for $\mathcal{P}_0$. But, the family $\mathcal{P}_0$ consists of distributions which are supported on the entire real line ($\mathbb{R}$). So, $\mathcal{P}_0$ dominates the family $\mathcal{P}$ and as a result $T(\mathbf{X})$ will be complete for $\mathcal{P}$, using Lemma 9.

Now, note that if T is complete for the family $\mathcal{P}$ and has 1-1 correspondence with $U(\mathbf{X})$, then U will also be complete for $\mathcal{P}$. This step is achieved by introducing the statistic, $V(\mathbf{X}) = (V_1, \ldots, V_n)^T$, where

$$V_1(\mathbf{X}) = \sum_i X_i,\ V_2(\mathbf{X}) = \sum_{i<j} X_i X_j,\ V_3(\mathbf{X}) = \sum_{i<j<k} X_i X_j X_k,\ \ldots\ , V_n = X_1 X_2 \cdots X_n.$$

The target will be to show that there is a 1-1 correspondence between the statistics $U(\mathbf{X})$, $T(\mathbf{X})$ and $V(\mathbf{X})$.

The first step is to note that the statistics $T(\mathbf{X})$ and $V(\mathbf{X})$ satisfy the following identities which are also known as Newton-Girard identities,

$$T_k - V_1 T_{k-1} + V_2 T_{k-2} - \cdots + (-1)^{k-1} V_{k-1} T_1 + (-1)^k k V_k = 0, \quad \text{for all } k \geqslant 1. \tag{2.17}$$

For $k = 1$, (2.17) reduces to $T_1 - 1 \times V_1 = \sum_{i=1}^n X_i - \sum_{i=1}^n V_i = 0$. For $k = 2$, (2.17) reduces to,

$$T_2 - V_1 T_1 + 2V_2 = \sum_i X_i^2 - \left(\sum_i X_i\right)\left(\sum_i X_i\right) + 2\sum_{i<j} X_i X_j = 0.$$

The proof of these identities for general k is not difficult, but a complete discussion on these is beyond the scope of this course[7]. Proofs of these identities are available in in these lecture notes by Mosse (2019) (which also provides a combinatorial proof), and also an elementary discussion is available in this article on brilliant.org (best for a beginners introduction), more in-depth details are available in this wikipedia article. Most importantly, given the set of values $\{T_1, \ldots, T_n\}$ it is possible to uniquely obtain $\{V_1, \ldots, V_n\}$

[7] Personally, I have become aware of these identities recently.

and vice-versa (solving each equation consecutively), hence there is a 1-1 correspondence between these statistics.

Now, let P be a polynomial with roots $\{x_1, \ldots, x_n\}$. Then we can write,

$$P(t) = \prod_{i=1}^{n}(t - x_i) = t^n - V_1(\mathbf{x})t^{n-1} + V_2(\mathbf{x})t^{n-2} - \cdots + (-1)^n V_n(\mathbf{x}) = \prod_{i=1}^{n}\big(t - x_{(i)}\big),$$

where $V_i(\mathbf{x})$ denotes that the fact that V_i's are based on x_i's. So $P(t)$ is a n-th degree polynomial with coefficients $\{-V_1(\mathbf{x}), V_2(\mathbf{x}), -V_3(\mathbf{x}), \ldots, (-1)^n V_n(\mathbf{x})\}$. Now, if coefficients of two n-degree polynomials match, then their roots must match and vice-versa.

Hence if $V_i(\mathbf{x}) = V_i(\mathbf{y})$ for all $i = 1, \ldots, n$, it implies $\{x_1, \ldots, x_n\}$ is a permutation of $\{y_1, \ldots, y_n\}$ and that implies these sets must have the same ordered statistics, $x_{(i)} = y_{(i)}$ for each $i = 1, \ldots, n$. The reverse correspondence is also clear.

Hence there is a 1-1 correspondence between the statistics $V(\mathbf{X}) = (V_1(\mathbf{X}), \ldots, V_n(\mathbf{X}))^T$ and $(X_{(1)}, \ldots, X_{(n)})^T$. This shows that all three sets of statistics are complete and sufficient for the family $\mathcal{P}$. □

Theorem 10 is important when it comes to nonparametric families and minimum variance unbiased estimation. It allows us to use the order statistics to construct unbiased estimators, which would automatically become minimum variance estimators (more on this later).

Suppose T is complete for $\mathcal{P}$ and U is another statistic such that $U = h(T)$ where h is a (measurable) map. Then if g is a map such that $\mathbf{E}_{\mathbf{P}} g(U) = 0$, for all $\mathbf{P} \in \mathcal{P}$, then for each $\mathbf{P}$,

$$\mathbf{E}_{\mathbf{P}} g\big(h(T)\big) = 0 \;\Leftrightarrow\; \mathbf{E}_{\mathbf{P}} f(T) = 0, \quad \text{where } f = g \circ h \;\Rightarrow\; \mathbf{P}\big(f(T) = g(U) = 0\big) = 1.$$

However, if U is complete and $U = h(T)$, then there is no assurance that T will be complete. If h is a one-to-one function, then if $u_1 = h(t_1) = h(t_2) = u_2$, it implies $t_1 = t_2$. This implies there exists an inverse map h^{-1} such that, $t = h^{-1}(u)$. Let g be any map such that $\mathbf{E}_{\mathbf{P}} g(T) = 0$ for each $\mathbf{P} \in \mathcal{P}$. Then, $\mathbf{E}_{\mathbf{P}} g(T) = \mathbf{E}_{\mathbf{P}} g\big(h^{-1}(U)\big) = 0$. This implies $g \circ h^{-1}(U) = 0$ w.p. 1, and as a result T will be also complete.

Example 2.24. In Example 2.21 we found that $T(X) = X$ is complete for the family $\mathcal{P} = \{\text{Binomial}(n, \theta) : 0 < \theta < 1\}$. Consider the sub-family $\mathcal{P}_0 = \{\text{Binomial}(n, \theta) : \theta \in \{1/4, 3/4\}\}$. For simplicity assume $n = 2$ (other cases can also be considered). Then, if $\mathbf{E}_\theta h(X) = 0$ for all $\theta \in \{1/4, 3/4\}$, it implies

$$h(0)(3/4)^2 + 2h(1)(1/4)(3/4) + h(2)(1/4)^2 = 0 \quad \text{and} \quad h(0)(1/4)^2 + 2h(1)(1/4)(3/4) + h(2)(3/4)^2 \quad = 0.$$

After some calculations we can find infinitely many non-zero choices of $\{h(0), h(1), h(2)\}$ such that these two equations hold. Note that family $\mathcal{P}_0$ still belongs to an exponential family, however it will no longer be a full-rank family (as $\mathcal{T}$ would be a three point set and would not be able to contain an open interval). However that would not imply that the natural sufficient statistic family is not-complete (as full-rank is a sufficient condition).

Remark 3. Even if a statistic is complete for a larger family, it need not be complete for a sub-family. This is in contrast to the result on sufficiency, where a statistic which is sufficient for a larger family is also sufficient for a sub-family. The primary reason is because a larger family imposes a larger number of constraints on the *behavior* of h, while a sub-family imposes fewer constraints, thereby allowing h to take non-zero values, even though it expectation becomes zero. Similarly as seen in Example 2.22, adding a new member to a family may also result in breakdown of completeness, even though that introduces more constraints!

Remark 4. Completeness is a purely technical concept and usually we aim to find a statistic which is both sufficient and complete. A sufficient statistic T is most successful in reducing data if no nonconstant function of T is ancillary or even first order ancillary. A statistic T is said to be *first order ancillary* if $\mathbf{E}_\theta(T)$ is a constant independent of θ. So first order ancillary means, $\mathbf{E}_\theta g(T) = c$, for all $\theta \in \Theta$, implies $g(T) = c$, a.e. $\mathcal{P}$. If we subtract c, this reduces to the completeness condition described earlier.

We will need a result on differentiation of integrals to handle some particular problems regarding completeness. Assume $h_1, h_2 : [a, b] \mapsto \mathbb{R}$ are continuously differentiable and $f : [a, b] \times \mathbb{R} \mapsto \mathbb{R}$. Then,

$$\frac{d}{dx} \int_{h_1(x)}^{h_2(x)} f(x, y)\, dy = \int_{h_1(x)}^{h_2(x)} \frac{\partial}{\partial x} f(x, y)\, dy + f(x, h_2(x)) \cdot h_2'(x) - f(x, h_1(x)) \cdot h_1'(x). \tag{2.18}$$

Example 2.25. Let $\{X_1, \ldots, X_n\}$ be i.i.d. Uniform $(0, \theta)$, and $\theta > 0$. We know $T = X_{(n)}$ is sufficient, and we want to show that T is also complete. Then, assume that h satisfies $\mathbf{E}_\theta h(X_{(n)}) = 0$, for all $\theta > 0$. The pdf of $T = X_{(n)}$ is,

$$f(t : \theta) = \frac{nt^{n-1}}{\theta^n} \mathbf{1}(0 < t < \theta).$$

This implies

$$\int_0^\theta h(t) \frac{nt^{n-1}}{\theta^n}\, dt = 0 \;\Leftrightarrow\; \int_0^\theta h(t) t^{n-1}\, dt = 0, \quad \text{for all } \theta > 0.$$

Applying formula (2.18) and differentiating both sides we obtain

$$0 = \int_0^\theta \frac{\partial}{\partial \theta}\, h(t) t^{n-1}\, dt + h(\theta) \frac{d}{d\theta}\,(\theta) - h(0) \frac{d}{d\theta}\,(0) = h(\theta), \quad \text{for all } \theta > 0.$$

Since $\mathbf{P}_\theta(X_{(n)} > 0) = 1$ for each $\theta > 0$, this implies $\mathbf{P}_\theta(h(X_{(n)}) = 0) = 1$, for each $\theta > 0$, showing that $X_{(n)}$ is complete.

We can also approach this question without using differentiation. For any real valued function h we can always decompose it as follows,

$$h(x) = h^+(x) - h^-(x), \quad \text{where} \quad h^+(x) = \max\{h(x), 0\} \text{ and } h^-(x) = -\min\{h(x), 0\}.$$

This means,

$$h^+(x) = \begin{cases} h(x) & \text{if } h(x) \geqslant 0, \\ 0 & \text{o.w.} \end{cases} \quad \text{and} \quad h^-(x) = \begin{cases} -h(x) & \text{if } h(x) < 0, \\ 0 & \text{o.w.} \end{cases}$$

This shows, both h^+ and h^- are non-negative maps, even though h could take any real value. These maps, h^+ and h^- are called positive and negative parts of a function h (and any map can be written in this manner). Also note that, $|h(x)| = h^+(x) + h^-(x)$. The advantage of working with non-negative maps is that we can always define integrals/sums (over x) of non-negative maps without encountering a situation where the expression is of the form $\infty - \infty$ (you will learn more about integration in other courses). Also non-negative maps (if their integrals/sums are finite) are like probability densities (except the integral/sum may be greater than 1).

Now assume, for some map h, $\mathbf{E}_\theta h(T) = 0$ for each $\theta > 0$ and split $h(x) = h^+(x) - h^-(x)$. Then, we obtain

$$\int_0^\theta h(t)t^{n-1}\, dt = 0 \;\Leftrightarrow\; \int_0^\theta h^+(t)t^{n-1}\, dt = \int_0^\theta h^-(t)t^{n-1}\, dt, \text{ for each } \theta > 0.$$

Subtracting one from another, we get for any $\theta_1 < \theta_2$,

$$\int_{\theta_1}^{\theta_2} h^+(t)t^{n-1}\, dt = \int_{\theta_1}^{\theta_2} h^-(t)t^{n-1}\, dt.$$

This implies, the integrals on both sides agree over all intervals within $(0, \infty)$. Using this fact along with Dynkin's π-λ theorem (see Theorem 3.3 of Billingsley (1995)) one can then show that

$$\int_A h^+(t)t^{n-1}\, dt = \int_A h^-(t)t^{n-1}\, dt,$$

for all Borel sets A in $((0, \infty), \mathcal{B}(0, \infty))$, where $\mathcal{B}(0, \infty)$ denotes the Borel σ-field on $(0, \infty)$. Now one can apply Theorem 16.10 of Billingsley (1995) to claim that $h^+(t) = h^-(t)$, a.e. λ (where λ denotes the Lebesgue measure on $\mathbb{R}$). This means h^+ and h^- coincide except possibly on a set whose length (Lebesgue measure) is zero. The exact details are beyond this course at this point. As a result we will have $h(t) = 0$, a.e. λ. As $\mathbf{P}_\theta(T > 0) = 1$, for each θ, this shows that $\mathbf{P}_\theta(h(T) = 0) = 1$, for each $\theta > 0$.

This method avoids the differentiation approach, and is usable when certain maps are non-differentiable.

Example 2.26. Assume $\{X_1, \ldots, X_n\}$ is i.i.d. $N(\mu, \sigma^2)$, where $\mu \in \mathbb{R}$ and $\sigma > 1$. If we write this distribution in the natural exponential form, then it would form a two-parameter canonical exponential family with the natural parameter space $\mathcal{T} = \{(\eta_1, \eta_2) : \eta_1 = -1/(2\sigma^2) > -1/2,\ \eta_2 = \mu/\sigma^2 \in \mathbb{R}\}$. Hence it will be possible to inscribe a two-dimensional rectangle within $\mathcal{T}$, implying that the natural sufficient statistics $(T_1 = \sum_i X_i,\ T_2 = \sum_i X_i^2)$ will be complete.

Example 2.27. Assume $\{X_1, \ldots, X_n\}$ is an i.i.d. sample from $N(\theta, \theta^2)$, where $\theta \in \mathbb{R}$. Then if write the joint distribution in the canonical exponential family form, it will be two-parameter canonical exponential family with natural parameters $\eta_1 = -1/2\theta^2$ and $\eta_2 = 1/\theta$ and corresponding sufficient statistics $T_1 = \sum_{i=1}^n X_i^2$ and $T_2 = \sum_{i=1}^n X_i$. Also note that the natural paraameter space does not contain a two-dimensional rectangle and hence the family is not full rank.

But, $\mathbf{E}_\theta(T_1) = n\mathbf{E}_\theta(X_1^2) = 2n\theta^2$ and $\mathbf{E}_\theta(T_2^2) = n\theta^2 + (n\theta)^2 = (n + n^2)\theta^2 = n(n+1)\mathbf{E}_\theta(T_1/(2n))$. So, $\mathbf{E}_\theta h(T_1, T_2) = 0$, for all θ, where $h(t_1, t_2) = t_1/(2n) - t_2^2/\{n(n+1)\}$, however $\mathbf{P}_\theta(h(T_1, T_2) = 0) < 1$, for each θ. This implies the family is not complete. Note, in absence of full rank property, we had to do a direct verification of completeness (or lack of it).

Example 2.28. Assume $\{X_1, \ldots, X_n\}$ is i.i.d. with pdf $f(x : \theta) = \exp\{-(x-\theta)\}\mathbf{1}(x > \theta)$, where $\theta \in \mathbb{R}$. This is the shifted exponential family of distributions, which does not belong to the exponential family framework. However $T = X_{(1)}$ can be shown to be sufficient for this family and using a similar approach as in the Uniform $(0, \theta)$ case, it can be shown that T is complete. The range of θ values in this family is extremely important.

Similarly if we consider the exponential distribution with pdf, $f(x:\theta) = \theta\exp\{-\theta x\}\mathbf{1}(x>0)$, then the family will have sufficient statistic $T = 1/\sum_{i=1}^n X_i$ and this will have 1-1 correspondence with $\sum_{i=1}^n X_i$, which can be shown to be complete, using the exponential family framework for showing completeness.

Example 2.29. Suppose X and Y are independent Binomial(k,θ) and Binomial(k,θ^2) random variables, where $\theta \in (0,1)$. Assume k is known. Then, the joint distribution of (X,Y) is in the two-parameter canonical exponential family framework and it can be shown that the natural parameter space does not contain a two-dimensional rectangle (this means that the **natural sufficient statistics** may not be complete). However,

$$\mathbf{E}_\theta(X^2) = k\theta(1-\theta) = k\theta - k\theta^2 = \mathbf{E}_\theta(X) - \mathbf{E}_\theta(Y) \;\Rightarrow\; \mathbf{E}_\theta(X^2 - X + Y) = 0, \quad \text{for all } \theta \in (0,1).$$

However the map $h(x,y) = x^2 - x + y$ is not identically zero for all $x, y \in \{0,1,\ldots,k\}$. This shows that (X,Y) will not be complete.

Example 2.30. We revisit the Uniform$(0,\theta)$ problem in Example 2.25. Assume in this case $\theta \in (1,\infty)$ (instead of the usual parameter space $\theta > 0$). Then, the previous argument for showing completeness of $X_{(n)}$ will fail, as no information on the values of $h(x)$ for $x \in (0,1)$ can be obtained. In fact, if we choose

$$h(x) = \begin{cases} 1 & \text{if } 0 < t < 1/2, \\ -\frac{1}{2^n - 1} & \text{if } 1/2 \leqslant t < 1, \\ 0 & \text{if } t \geqslant 1. \end{cases}$$

Then, we can check

$$\mathbf{E}_\theta h(X_{(n)}) = \int_0^{1/2} \frac{nt^{n-1}}{\theta^n}\, dt - \frac{1}{2^n - 1} \cdot \int_{1/2}^1 \frac{nt^{n-1}}{\theta^n}\, dt = 0, \quad \text{for all } \theta > 1.$$

So, $X_{(n)}$ is infact not complete[8] if we consider the underlying family with $\theta > 1$. The similar argument shows that $X_{(n)}$ will not be complete if we restrict $\theta \in [c,\infty)$, for any $c > 0$.

Example 2.31 ($\star$). Assume $\{X_1,\ldots,X_n\}$ is an i.i.d. sample from Uniform(θ_1,θ_2), where $-\infty < \theta_1 < \theta_2 < \infty$. Then, factorization theorem shows that $(T_1 = X_{(1)}, T_2 = X_{(n)})$ will be jointly sufficient for this family. Assume $h(x,y)$ is a map such that

$$\mathbf{E}_{\theta_1,\theta_2} h(T_1,T_2) = 0, \quad \text{for each } \theta_1 < \theta_2.$$

The joint pdf of (T_1,T_2) will be (see Section 4.7 of Rohatgi and Saleh (2015)),

$$f(t_1,t_2:\theta_1,\theta_2) = \frac{n(n-1)}{(\theta_2-\theta_1)^n} \cdot (t_2-t_1)^{n-2} \cdot \mathbf{1}(\theta_1 < t_1 < t_2 < \theta_2).$$

Using the expectation condition we obtain for each $\theta_1 < \theta_2$,

$$\int_{\theta_1}^{\theta_2}\int_{t_1}^{\theta_2} h(t_1,t_2) \cdot \frac{n(n-1)}{(\theta_2-\theta_1)^n} \cdot (t_2-t_1)^{n-2}\, dt_2\, dt_1 = 0 \quad \Leftrightarrow \quad \int_{\theta_1}^{\theta_2}\int_{t_1}^{\theta_2} h(t_1,t_2) \cdot (t_2-t_1)^{n-2}\, dt_2\, dt_1 = 0,$$

$$\Leftrightarrow \int_{A_{\theta_1,\theta_2}} h^+(t_1,t_2)(t_2-t_1)^{n-2}\, dt_2 dt_1 = \int_{A_{\theta_1,\theta_2}} h^-(t_1,t_2)(t_2-t_1)^{n-2}\, dt_2 dt_1, \tag{2.19}$$

[8] Think why this argument will fail to extend in case we allow $\theta > 0$.

where

$$A_{\theta_1,\theta_2} = \left\{(t_1,t_2) \in \mathbb{R}^2 : \theta_1 < t_1 < t_2 < \theta_2\right\}$$

is a triangular region. Note, on A_{θ_1,θ_2}, we have $(t_2 - t_1)^{n-2} > 0$, for all $\theta_1 < \theta_2$. The question is, if these integrals match on such triangular sets, will they match on arbitrary open sets? The brief answer is yes. The detailed discussion on how to show this requires knowledge of measure theory and we skip the details. The end result is that we can show $h^+(t_1,t_2) = h^-(t_1,t_2)$ for all $t_1 < t_2$, except possibly on a set with area zero. Eventually this would imply completeness.

Example 2.32. Let $\{X_1,\ldots,X_n\}$ denote an i.i.d. sample from the density,

$$f(x:\mu,\sigma) = \frac{1}{\sigma}\exp\left\{-\frac{(x-\mu)}{\sigma}\right\}\cdot \mathbf{1}(x \geqslant \mu), \quad \text{with } \mu \in \mathbb{R},\ \sigma > 0.$$

It is clear that the statistic $\mathbf{T} = (T_1,T_2)^T = \left(X_{(1)}, \sum_{i=1}^n (X_i - X_{(1)})\right)^T$ is sufficient. We aim to show that $\mathbf{T}$ is complete. Note, the joint pdf can not be written in an exponential family format. Suppose $h(T_1,T_2)$ satisfies, $\mathbf{E}_{\mu,\sigma} h(T_1,T_2) = 0$, for all μ and σ. This implies,

$$\int_0^\infty \int_\mu^\infty h(t_1,t_2)\ f_{\mu,\sigma}(t_1,t_2)\ dt_1 dt_2 = 0, \quad \text{for all } (\mu,\sigma). \tag{2.20}$$

From this equation, we can not obtain a relation of the form,

$$\int_a^b \int_b^d h(t_1,t_2)\ f_{\mu,\sigma}(t_1,t_2)\ dt_1 dt_2 = 0, \quad \text{for all } a < b \text{ and } c < d, \text{ and all } (\mu,\sigma).$$

Hence, we need to work with marginal densities of T_1 and T_2. It can be shown that T_1 and T_2 are independent with densities,

$$f_{\mu,\sigma}(t_1) = \frac{1}{\sigma}\exp\left\{-\frac{n(t_1-\mu)}{\sigma}\right\}\mathbf{1}(t_1 \geqslant \mu), \quad \text{and} \quad f_\sigma(t_2) = \frac{1}{\sigma}\exp\left\{-\frac{(n-1)t_2}{\sigma}\right\}\mathbf{1}(t_2 \geqslant 0).$$

One can show by using some subtle arguments that T_1 and T_2 will be complete for this family. The details require some delicate measure-theoretic arguments (see Theorem 7.1 of Lehmann and Scheffé (2012)), that are beyond the scope of this course.

Example 2.33. Consider the r.v. X with pmf

$$\mathbf{P}_\theta(X = x) = \begin{cases} \theta & \text{if } x = -1, \\ (1-\theta)^2\theta^x & \text{if } x = 0,1,2,\ldots. \end{cases}$$

where $\theta \in (0,1)$. Assume h is a map such that $\mathbf{E}_\theta h(X) = 0$ for all θ. This implies, for each $\theta \in (0,1)$,

$$\begin{aligned} &\sum_{x=-1}^\infty h(x)\mathbf{P}_\theta(X = x) = 0, \\ \Leftrightarrow\ & h(-1)\theta + (1-\theta)^2 \sum_{x=0}^\infty h(x)\theta^x = 0, \\ &\qquad \Leftrightarrow \sum_{x=0}^\infty h(x)\theta^x = -\frac{\theta h(-1)}{(1-\theta)^2}, \\ &\qquad \Leftrightarrow \sum_{x=0}^\infty h(x)\cdot\theta^x = -h(-1)\cdot\theta\left\{1 + 2\theta + 3\theta^2 + \ldots\right\} = -h(-1)\sum_{x=0}^\infty x\cdot\theta^x. \end{aligned} \tag{2.21}$$

Both sides of the above relation are true at any $\theta > 0$ and both sides are finite sums. We will use the following fact about power series.

Suppose $f(x) = \sum_{n \geqslant 0} a_n (x - x_0)^n$ and $g(x) = \sum_{n \geqslant 0} b_n (x - x_0)^n$ are two power series which converge absolutely on the region $B(x_0 : R) = \{x : |x - x_0| < R\}$ for some radius $R > 0$. Also assume that $f(x) = g(x)$ for all $x \in B(x_0 : R)$. Then, $a_n = b_n$ for all $n \geqslant 0$. A proof of this result can be found in this lecture notes by Wehler (2019) (cf. Proposition 1.11).

If we write the left and right sides of (2.21) as $f(\theta)$ and $g(\theta)$ respectively, with the assumption that they converge absolutely[9] for all $\theta \in (-R, R) \subseteq (-1, 1)$. Then applying the above result, we would obtain $h(x) = -xh(-1)$, for all $x \geqslant 0$.

We have the option of deciding the value of $h(-1)$. If $a = h(-1) \neq 0$, then $h(\cdot)$ becomes unbounded and non-zero, and yet satisfies the completeness condition. If $a = h(-1) = 0$, then $h(x) = 0$ for all $x \in \{-1, 0, 1, \ldots\}$ and we obtain $\mathbf{P}_\theta(h(X) = 0) = 1$, for each θ, ensuring the completeness condition. So, X is boundedly complete (completeness is satisfied only for bounded maps) and yet X is not complete.

2.4 Basu's theorem

At the other extreme of a sufficient statistic is an *ancillary* statistic. A statistic $T(\mathbf{X})$ is called ancillary if its distribution does not depend on θ. This means T does not contain any information about θ, even though T is a map based on $\mathbf{X}$. We look into ancillary statistics with the goal of developing Basu's theorem, which states that a complete sufficient statistic is independent of an ancillary statistic. For example if $\{X_1, \ldots, X_n\}$ are i.i.d. $N(\mu, 1)$, then $\bar{X}_n$ is sufficient while the sample variance is ancillary, and they turn out to be independent. Eventually Basu's theorem becomes useful in many estimation problems. Basu's theorem is elegant and is considered to be one of the most simple yet profound results in Statistics and has been applied to a wide range of problems. For excellent commentaries, insight and more information one can consult the excellent papers by Lehmann (1981) and Ghosh (2002) and as well as the book by DasGupta (2011).

Theorem 11 (Basu's Theorem)**.** *Assume $\mathbf{X} \sim \mathbf{P}_\theta$ where $\theta \in \Theta$. Suppose $V(\mathbf{X})$ is ancillary and $T(\mathbf{X})$ is complete and sufficient for the family. Then V and T are independent.*

Proof of Theorem 11. For any set A, write $p_A = \mathbf{P}_\theta(V \in A)$, since the probability must be free of θ. Also write $\eta(t) = \mathbf{P}_\theta(V(\mathbf{X}) \in A \mid T(\mathbf{X}) = t)$, which is also independent of θ, as T is sufficient. Then, for each $\theta \in \Theta$,

$$\mathbf{E}_\theta \eta(T) = \mathbf{P}_\theta(V \in A) = p_A \;\Rightarrow\; \mathbf{E}_\theta(\eta(T) - p_A) = 0.$$

As T is also complete, and the map $h(t) \equiv (\eta(t) - p_A)$ has zero expectation, then

$$\mathbf{P}_\theta(\eta(T) = p_A) = 1, \quad \text{for each } \theta \in \Theta.$$

[9] We are not told what will happen to both terms on each side of (2.21) if $\theta < 0$. However, we need to assume there exists a common interval of convergence for each series. For $\theta > 0$, the term on the l.h.s. of (2.21) will be finite, as because the expectation of $h(X)$ exists. But, a reasonable way to show this would be to ensure this would be to show that, $\sum_{k \geqslant -1} |h(k)||\theta|^k < \infty$. If the latter holds, then indeed the series would converge also for any $-1 < \theta < 0$.

This is true for any arbitrary set A. This shows that the conditional distribution of V given T, is same as the unconditional distribution of V, which implies independence of V and T. $\square$

We will go through some interesting examples where Basu's theorem is used to obtain difficult results in an easy manner.

Example 2.34. Assume $\{X_1, \ldots, X_n\}$ are i.i.d. $N(\mu, \sigma^2)$, where $(\mu, \sigma) \in \mathbb{R} \times (0, \infty)$. It is well known (and proved using cumbersome sampling distribution calculations) that $\bar{X}_n$ and $\sum_{i=1}^n (X_i - \bar{X}_n)^2$ are independently distributed. However, we can apply the location-scale structure of Normal distributions and Basu's theorem to obtain the same result with less effort.

Let $\{Z_1, \ldots, Z_n\}$ be i.i.d. $N(0,1)$ random variables. Then we can express, $X_i = \mu + \sigma Z_i$ for each $i = 1, \ldots, n$. Hence it will be enough to show that the statistics $\bar{Z}_n$ and $\sum_{i=1}^n (Z_i - \bar{Z}_n)^2$ are independent, because

$$\big(\bar{X}_n,\ \sum_{i=1}^n (X_i - \bar{X}_n)^2\big) \overset{d}{=} \big(\mu + \sigma \bar{Z}_n,\ \sigma^2 \sum_{i=1}^n (Z_i - \bar{Z}_n)^2\big).$$

Now consider the smaller family $\mathcal{P}_0 = \{N(\theta, 1) : \theta \in \mathbb{R}\}$ and assume $\{Y_1, \ldots, Y_n\}$ are i.i.d. $N(\theta, 1)$. Then the joint distribution of $\mathbf{Y}$ forms a full-rank one-parameter exponential family with natural sufficient statistic $T(\mathbf{Y}) = \bar{Y}_n$, which will be also complete, while the sum of squares term $\sum_{i=1}^n (Y_i - \bar{Y}_n)^2$ will be free of θ and hence ancillary. Applying Basu's theorem, $\bar{Y}_n$ and $\sum_{i=1}^n (Y_i - \bar{Y}_n)^2$ will be independent for any underlying θ, and in particular for $\theta = 0$. This shows that $\bar{Z}_n$ and $\sum_{i=1}^n (Z_i - \bar{Z}_n)^2$ will be independent. This is exactly what we wanted to prove!

Example 2.35. Assume $\{X_1, \ldots, X_n\}$ are i.i.d $N(0,1)$ random variables and write $M_n(\mathbf{X})$ = median of X_i's. We want to find $\mathbf{cov}(M_n, \bar{X}_n)$. A direct calculation will be difficult.

To find this covariance, we embed the $N(0,1)$ distribution into a larger family, $\mathcal{P} = \{N(\theta, 1) : \theta \in \mathbb{R}\}$ and let $\{Y_1, \ldots, Y_n\}$ be i.i.d. $N(\theta, 1)$. Let $M_n(\mathbf{Y})$ denote the median of Y_i's. Then, we note that[10],

$$\bar{Y}_n - M_n(\mathbf{Y}) = (\bar{X}_n + \theta) - (M_n(\mathbf{X}) + \theta) = \bar{X}_n - M_n(\mathbf{X}),$$

which is ancillary. Also $\bar{Y}_n$ is complete and sufficient for the family $\mathcal{P}$. Thus $\bar{Y}_n$ and $(\bar{Y}_n - M_n(\mathbf{Y}))$ are independent, by Basu's theorem. Hence they have zero covariance (for any choice of θ) and

$$0 = \mathbf{cov}_{\theta=0}\left(\bar{Y}_n,\ \bar{Y}_n - M_n(\mathbf{Y})\right) = \mathbf{cov}\left(\bar{X}_n,\ \bar{X}_n - M_n(\mathbf{X})\right) = \mathbf{Var}(\bar{X}_n) - \mathbf{cov}\big(\bar{X}_n,\ M_n(\mathbf{X})\big).$$

This shows that $\mathbf{cov}\big(\bar{X}_n,\ M_n(\mathbf{X})\big) = 1/n$.

Other interesting examples will arise as we explore the theory of unbiased estimation.

3 Unbiased minimum variance estimation

The problem of point estimation involves a real (or vector) valued function $\tau(\theta) : \Theta \mapsto \mathbb{R}^m$, which is to be estimated. This *parametric function* $\tau(\theta)$ is called an *estimand.* An *estimator* $T(\mathbf{X}) : \mathcal{X} \mapsto \mathbb{R}^m$ is a real

[10] Median of $\{a_1, \ldots, a_n\} = c +$ Median of $\{b_1, \ldots, b_n\}$, where $a_i = c + b_i$.

valued (or vector) valued function defined on the sample space. It is used to estimate $\tau(\theta)$. The value $T(\mathbf{x})$, taken on by $T(\mathbf{X})$, when the observed data is $\mathbf{x}$ is known as the *estimate* of $\tau(\theta)$.

The family of distributions $\{\mathbf{P}_\theta : \theta \in \Theta\}$ (or in terms of pdf/pmf $\{p_\theta : \theta \in \Theta\}$) is said to be *identifiable* if, for any $\theta_1, \theta_2 \in \Theta$,

$$\theta_1 = \theta_2 \quad \Leftrightarrow \quad \mathbf{P}_{\theta_1} = \mathbf{P}_{\theta_2}.$$

Hence, the distributions differ at least a single point (in case of discrete distributions) or they differ over an interval (in case of continuous distributions). If the family fails to satisfy this criteria, then it is called a non-identifiable family of distributions. For example, if $X \sim N(\theta, 1)$ and $Y = |X|$, then the family of distributions of Y is non-identifiable (over $\theta \in \mathbb{R}$).

There are different methods for estimation, *viz.*, maximum likelihood method and other variations of likelihood based approach, method of moments and estimating equations approach, empirical likelihood, minimum variance estimation, resampling based methods, etc. All approaches have their strengths and weaknesses and we apply a particular approach depending on the scenario. However, a natural and early approach was to consider unbiased estimators (UE's) and find the best estimators among such UE's. We will now discuss this topic in detail.

3.1 Accuracy of an estimator and unbiased estimators

Irrespective of the choice of any method for estimation, we require an estimator $T(\mathbf{X})$ should be close to the estimand $\tau(\theta)$. But $T(\mathbf{X})$ is random and we can define different notions of closeness or accuracy of T. This should take into account the randomness of $\mathbf{X}$. For example, we can define accuracy measures like

$$\mathbf{P}_\theta\left(\|T(\mathbf{X}) - \tau(\theta)\| \geqslant \epsilon\right), \quad \text{for any fixed } \epsilon > 0, \text{ or} \quad \mathbf{E}_\theta\|T(\mathbf{X}) - \tau(\theta)\|^p, \quad \text{for some } p > 0,$$

or one may use some other measure of accuracy. The first approach tries to find how closely T is concentrated around the target $\tau(\theta)$, when $\mathbf{P}_\theta$ is the true distribution. The second approach uses moments (provided they are assumed to exist). In both cases, a *small* value of these two quantities will suggest that T is a *good* estimator. However, both these accuracy measures are defined for each θ. As such, we can compare two estimators T_1 and T_2 (using the same measure of accuracy), pointwise at each θ, but it may become difficult to compare them over the entire parameter space Θ (with T_1 performing better than T_2 over certain part of the parameter space). There can be global measures of accuracy over the whole parameter space, like

$$\sup_{\theta \in \Theta} \mathbf{P}_\theta\left(\|T(\mathbf{X}) - \tau(\theta)\| \geqslant \epsilon\right) \quad \text{or} \quad \sup_{\theta} \mathbf{E}_\theta\|T(\mathbf{X}) - \tau(\theta)\|^2, \quad \text{or} \quad \int_\theta \omega(\theta)\mathbf{E}_\theta\|T(\mathbf{X}) - \tau(\theta)\|^p \, d\theta,$$

or some other variation, which takes into account the behavior of the estimators over the entire Θ and produces a single measure of accuracy. Here, $\omega(\theta)$ is a pre-chosen non-negative weight function which describes the *importance* of each $\theta \in \Theta$. More generally everything stated above can be discussed under the framework of decision theory, which will be studied later in the course.

In case moments are assumed to exist, the most commonly used pointwise measure of accuracy for an estimator is the mean-squared error (MSE),

$$\mathrm{MSE}_\theta(T) = \mathbf{E}_\theta\|T(\mathbf{X}) - \tau(\theta)\|^2, \quad \text{for each } \theta \in \Theta.$$

In case $T(\mathbf{X})$ is unbiased for $\tau(\theta)$, then MSE reduces to variance.

Remark 5. Our search for the *best* estimator of a parametric function $\tau(\theta)$, at **every** value of θ is futile. For any fixed $\theta = \theta_0$, the best estimator of $\tau(\theta_0)$ will be

$$T_0(\mathbf{x}) = \tau(\theta_0), \quad \text{for all } \mathbf{x}.$$

No estimator will be able to perform better at θ_0, in terms of estimating $\tau(\theta_0)$, as there is absolutely no error in estimation. At any other value of θ, $T_0(\mathbf{X})$ will be a very bad estimator, unless $\tau(\theta)$ is constant function (which itself does not sound like an interesting estimation problem). Hence, if we are attempting to find an estimator (say T), that is the best at each θ, then T should be better than T_0 at θ_0 and T_1 at θ_1 and so on. This is clearly impossible, and obviously if $\tau(\theta)$ is unknown (because the underlying θ is unknown), estimators like T_0 will make no sense.

As we see in Remark 5, we try to be less ambitious and construct an estimator that is not the uniformly best at every θ, but will be *reasonably well* at all θ values. There are different ways to accomplish this goal by restricting to a smaller class of estimators, like unbiased estimators, equivariant estimators and using global criteria like minimax or Bayes estimators (to be discussed later). Our discussion will focus on unbiased estimators (UE). An estimator $T(\mathbf{X})$ of a parametric function $\tau(\theta)$ is *unbiased*, if $\mathbf{E}_\theta(T(\mathbf{X})) = \tau(\theta)$ for all θ. If the estimator is not unbiased, we call it biased. For each θ, the bias of T will be, $\text{bias}_\theta(T) = \mathbf{E}_\theta(T) - \tau(\theta)$.

For the moment, we focus on estimation of univariate parameters $\tau(\theta) : \Theta \mapsto \mathbb{R}$. To choose among the class of all UE's of $\tau(\theta)$, we will use the point-wise accuracy measure of $\mathbf{Var}_\theta(T) = \mathbf{E}_\theta\big(T(\mathbf{X}) - \tau(\theta)\big)^2$. We will say that an estimator T will be the *best* estimator $\tau(\theta)$ in this sense if,

(i) $\mathbf{E}_\theta(T) = \tau(\theta)$ for all θ.

(ii) For any other estimator W satisfying $\mathbf{E}_\theta(W) = \tau(\theta)$ for all θ, we will have

$$\mathbf{Var}_\theta(T) \leqslant \mathbf{Var}_\theta(W), \quad \text{for all } \theta.$$

Such an estimator, if it exists will be known as the UMVUE (uniformly minimum variance unbiased estimator) of $\tau(\theta)$. The definition suggests that we should check the variance condition against all possible UE's of $\tau(\theta)$, which will be formidable task. To circumvent this problem, one of the initial approaches was to try to find a lower bound on the variance of an UE of $\tau(\theta)$. Suppose we have a value $b(\theta)$, such that $\mathbf{Var}_\theta(W) \geqslant b(\theta)$, for all θ and all UE's W. Assume that we are able to obtain an UE T, such that $\mathbf{Var}_\theta(T) = b(\theta)$ for all θ. This would be equivalent to saying that T has the minimum possible variance among all UE's of $\tau(\theta)$, which means T is an UMVUE of $\tau(\theta)$. The Cramer-Rao lower bound (CRLB) is such a lower bound, available when the underlying distribution P_θ satisfies certain regularity assumptions.

Similar to an UMVUE, there is a notion of LMVUE (locally minimum variance unbiased estimator), in which case the estimator T is LMVUE if it is superior over a subset of Θ. We will not pursue this topic further.

3.2 Uniformly minimum variance unbiased estimators

In this section we study methods to construct UMVUE's. We first state a result which states that the search for UMVUE's should utilize sufficient statistics. A sufficient statistic improves the performance of any UE.

Theorem 12 (Rao-Blackwell)**.** *Suppose W is any UE of $\tau(\theta)$ and T is a sufficient statistic for θ. Define $\phi(T) = \mathbf{E}_\theta(W|T)$. Then we can improve upon W using $\phi(T)$:*

(a) $\phi(T)$ is an UE of $\tau(\theta)$.

(b) $\mathbf{Var}_\theta(\phi(T)) \leqslant \mathbf{Var}_\theta(W)$, for all θ.

Proof. We have $\mathbf{E}_\theta\left(\phi(T)\right) = \mathbf{E}_\theta\left(\mathbf{E}\left(W|T\right)\right) = \mathbf{E}_\theta(W) = \tau(\theta)$, for any θ. Hence $\phi(T)$ is an UE of $\tau(\theta)$. For any θ,

$$\begin{aligned}\mathbf{Var}_\theta\left(W\right) &= \mathbf{Var}_\theta\left(\mathbf{E}(W|T)\right) + \mathbf{E}_\theta\left(\mathbf{Var}(W|T)\right)\\ &= \mathbf{Var}_\theta\left(\phi(T)\right) + \mathbf{E}_\theta\left(\text{a non negative r.v.}\right), \quad (\text{variance is} \geqslant 0 \text{ for any r.v.})\\ &\geqslant \mathbf{Var}_\theta\left(\phi(T)\right), \quad (\text{for any r.v. } Z \geqslant 0,\ \mathbf{E}(Z) \geqslant 0).\end{aligned}$$

It should be noted that since T is sufficient, the distribution of $\phi(T)$ is independent of θ, and hence $\phi(T)$ will be an estimator. □

This result does not provide a technique to identify an UMVUE, but shows that our search for UMVUE's can be constrained to estimators which are based on sufficient statistics. The next result states a method of finding UMVUE's, and can be used to show that an estimator is not an UMVUE. An estimator U is said to be an estimator of 0, if $\mathbf{E}_\theta(U) = 0$ for all θ. Let

$$\mathcal{U} = \left\{U : \mathbf{E}_\theta(U) = 0,\ \mathbf{E}_\theta(U^2) < \infty, \quad \text{for all } \theta\right\}.$$

Thus $\mathcal{U}$ represents the class of all UE's of 0, with finite variance. Suppose T is an UE of $\tau(\theta)$. Then

$$W = T - U,$$

is also an UE of $\tau(\theta)$. This is true for any choice of $U \in \mathcal{U}$. Thus the totality of all UE's of $\tau(\theta)$ can be generated by starting from any fixed UE (say, T) and then generating W, by going over all possible choices of U.

Theorem 13. *A necessary and sufficient condition for T to be an UMVUE of $\tau(\theta)$ is that:*

$$\mathbf{E}_\theta(TU) = 0, \quad \textit{for all } U \in \mathcal{U} \quad \textit{and all} \quad \theta \in \Theta. \tag{3.22}$$

Proof of Theorem 13. Note that since $\mathbf{E}_\theta(U) = 0$, the condition (3.22) is equivalent to saying that $\mathbf{cov}_\theta(T, U) = 0$ for all U and for all θ.

We start with the necessity part. Suppose T is UMVUE for $\tau(\theta)$. Fix $U \in \mathcal{U}$ and a $\theta \in \Theta$, and for arbitrary real λ, let $T' = T + \lambda U$. Then T' is an UE of $\tau(\theta)$, and since T is UMVUE,

$$\mathbf{Var}_\theta\left(T + \lambda U\right) \geqslant \mathbf{Var}_\theta(T), \quad \text{for all } \lambda.$$

This is equiavlent to

$$\lambda^2 \mathbf{Var}_\theta(U) + 2\lambda \mathbf{cov}_\theta(T,U) \geqslant 0, \quad \text{for all } \lambda.$$

This is quadratic function in λ, and since it takes positive values for all $\lambda \in \mathbb{R}$, it can only be possible if[11] $\mathbf{cov}_\theta(T,U) = 0$. Since U was arbitrary, we have the necessity part.

For the sufficiency part: assume $\mathbf{E}_\theta(TU) = 0$ for all $U \in \mathcal{U}$. Let T' be another UE of $\tau(\theta)$. Then, $T - T' \in \mathcal{U}$, since it is an UE of 0. Thus, $\mathbf{E}_\theta\left(T(T-T')\right) = 0$, for all θ. Thus, $\mathbf{E}_\theta(T^2) = \mathbf{E}_\theta(TT')$, and since T, T' have the same expectations,

$$\begin{aligned} \mathbf{Var}_\theta(T) &= \mathbf{cov}_\theta(T,T') \leqslant \left(\mathbf{Var}_\theta(T)\cdot \mathbf{Var}_\theta(T')\right)^{1/2} \\ \Rightarrow \mathbf{Var}_\theta(T) &\leqslant \mathbf{Var}_\theta(T'), \quad \text{for all } \theta. \end{aligned}$$

Since T' was arbitrary, we have the proof. □

This method is fairly useful in showing that an unbiased estimator is not an UMVUE, as the theorem provides an *iff* condition. A direct use of this theorem is difficult, as finding out (or characterizing) all UE's of 0 is a difficult task.

It could be made much simpler if the only unbiased estimator of 0 is 0 itself, which means $\mathcal{U}$ consists of only one estimator: the trivial estimator which takes the value 0 everywhere. In such a situation, our task of verifying (3.22) will be trivially easy, since $\mathbf{E}_\theta(T\cdot 0) = 0$, for any T. And we will be able to say that T is uncorrelated with all UE's of 0 (they only one is 0 itself). The next result makes this more precise.

Theorem 14 (Lehmann-Scheffe)**.** *Suppose $f_\theta(\mathbf{x})$ is the joint pdf/pmf of $\{X_1,\ldots,X_n\}$. Let T be complete and sufficient statistic for θ. Let $W^* = g(T)$ be an UE of $\tau(\theta)$, where g is some function of T. Then W^* is an UMVUE of $\tau(\theta)$.*

Proof of Theorem 14. Let W be any other UE of $\tau(\theta)$. We need to show that $\mathbf{Var}_\theta(W^*) \leqslant \mathbf{Var}_\theta(W)$ for all θ. Let $W_1 = \mathbf{E}(W|T)$. Then by Rao-Blackwell theorem,

(a) W_1 is an UE of $\tau(\theta)$.

(b) $\mathbf{Var}_\theta(W_1) \leqslant \mathbf{Var}_\theta(W)$ for all θ.

Note that both W^* and W_1 are functions of T, and hence their difference is also a function (denoted as h) of T. This implies

$$\begin{aligned} \mathbf{E}_\theta(W^* - W_1) = 0 = \mathbf{E}_\theta(h(T)), &\quad \text{for all } \theta, \\ \Leftrightarrow\ h(T) = 0, &\quad \text{with probability 1, for all } \theta, \\ \Leftrightarrow\ W^* = W_1, &\quad \text{with probability 1, for all } \theta. \end{aligned}$$

Hence, $\mathbf{Var}_\theta(W^*) = \mathbf{Var}_\theta(W_1) \leqslant \mathbf{Var}_\theta(W)$, (by part (b) above) for all $\theta \in \Theta$. □

[11] try to draw a figure for this parabolic curve: $a\lambda^2 + 2b\lambda$, with $a \geqslant 0$, and see that the function will take negative values on some set of λ values, unless $b = 0$.

This theorem provides us two methods to construct UMVUE's of a parametric function $\tau(\theta)$:

(i) We find a complete sufficient statistic T for θ and construct a function $g(T)$ such that $\mathbf{E}_\theta(g(T)) = \tau(\theta)$, for all θ. Then, $g(T)$ (and in the above proof, this is W^*) will be an UMVUE of $\tau(\theta)$. This is the direct method.

(ii) We find an UE of $\tau(\theta)$, which is denoted as W. Then we find the conditional expectation of W given T:

$$\phi(T) = \mathbf{E}(W|T),$$

which is function of T only (and independent of θ). Then by Rao-Blackwell $\phi(T)$ will be an UE of $\tau(\theta)$ and by the proof of the above theorem $\phi(T)$ (denoted above by W_1) will have the minimum variance. This is an indirect method, with an additional step of calculating the conditional expectation.

The choice of which method to apply depends on the situation. It is easy to check that if T_i is the UMVUE of $\tau_i(\theta)$, then $\sum_{i=1}^k T_i$ is the UMVUE of $\sum_{i=1}^k \tau_i(\theta)$. The next result shows that an UMVUE is unique.

Theorem 15. *If T_1 is an UMVUE of $\tau(\theta)$, then T_1 is unique.*

Proof. Suppose T_2 be any other UMVUE of $\tau(\theta)$. Then $\mathbf{Var}_\theta(T_1) = \mathbf{Var}_\theta(T_2)$ for all θ. Note that $\mathbf{E}_\theta(T_1 - T_2) = 0$ for all θ. Hence from Theorem 14 (see above), we have for all θ,

$$\begin{aligned} &\mathbf{E}_\theta(T_1(T_1 - T_2)) = 0, \; \Leftrightarrow \; \mathbf{E}_\theta(T_1^2) = \mathbf{E}_\theta(T_1 T_2), \\ \Leftrightarrow \; &\mathbf{Var}_\theta(T_1) = \mathbf{cov}_\theta(T_1, T_2) = \mathbf{Var}_\theta(T_2), \quad (*) \\ \Leftrightarrow \; &\frac{\mathbf{cov}_\theta(T_1, T_2)}{\sqrt{\mathbf{Var}_\theta(T_1)}\sqrt{\mathbf{Var}_\theta(T_2)}} = 1 \; \Leftrightarrow \; \mathbf{corr}_\theta(T_1, T_2) = 1. \end{aligned}$$

This implies that there is a linear relationship between T_1 and T_2. Hence $T_1 = a(\theta)T_2 + b(\theta)$ for some functions a and b, with probability 1 for all θ. This will lead to:

$$\mathbf{cov}_\theta(T_1, T_2) = a(\theta)\mathbf{Var}_\theta(T_2) = \mathbf{Var}_\theta(T_2), \quad \text{(by earlier relation (*))},$$

and this implies that[12] $a(\theta) = 1$. And since $\mathbf{E}_\theta(T_1) = \mathbf{E}_\theta(T_2) = \tau(\theta)$, this leads to $b(\theta) = 0$. Thus, for all θ, $T_1 = T_2$, with probability 1. Hence the UMVUE is unique. □

We summarize some methods of finding UMVUE's:

(a) Based on the Lehmann-Scheffe theorem: directly solve for the function g, such that $g(T)$ is an UE of $\tau(\theta)$. For this method we need the distribution of the complete sufficient statistic T.

(b) We can find any UE of $\tau(\theta)$, say W. Then find $\mathbf{E}(W|T)$ as an UMVUE of $\tau(\theta)$. For this method, we do not need the distribution of T, but we need to find the conditional expectation. Since the UMVUE found, will be unique, it will not matter which W is selected initially. We should try to choose a W which makes evaluation of the conditional expectation easier.

[12] the trivial case where $\mathbf{Var}_\theta(T_1) = \mathbf{Var}_\theta(T_2) = 0$, will imply $T_1 = T_2 = \tau(\theta)$, w.p. 1

(c) In the absence of a complete sufficient statistic, finding an UMVUE can be difficult. In such cases we can try to use Theorem 13, but this will usually involve characterizing the class $\mathcal{U}$ of all estimators of 0.

(d) A completely different approach is the use of CRLB, which provides a lower bound to the variance of all UE's of $\tau(\theta)$. If we manage to obtain an UE, T, which has the same variance as the CRLB, we can conclude that T is an UMVUE. This method works only if all CRLB regularity assumptions are satistfied, but still provides us a method to assess the performance of UMVUE's (or other UE's).

3.2.1 Examples of UMVU estimators

Example 3.1. Assume $T \sim \text{Binomial}(n, \theta)$ where $\theta \in (0,1)$. Then T belongs to an exponential family of full rank and it is complete and sufficient. Consider the parametric function $\tau(\theta) = \theta(1-\theta)$. Suppose $g(T)$ is an UE of $\tau(\theta)$. Then, for each $\theta \in (0,1)$,

$$\begin{aligned}
\mathbf{E}_\theta g(T) &= \theta(1-\theta) \\
\Rightarrow \sum_{t=0}^{n} g(t)\binom{n}{t}\theta^t(1-\theta)^{n-t} &= \theta(1-\theta) \\
\Rightarrow \sum_{t=0}^{n} g(t)\binom{n}{t}\left(\frac{\theta}{1-\theta}\right)^t &= \frac{\theta}{(1-\theta)^{n-1}} \\
\Rightarrow \sum_{t=0}^{n} g(t)\binom{n}{t}\rho^t &= \rho(1+\rho)^{n-2}, \quad \text{with } \rho = \theta/(1-\theta), \\
\Rightarrow \sum_{t=0}^{n} g(t)\binom{n}{t}\rho^t &= \sum_{t=1}^{n-1}\binom{n-2}{t-1}\rho^t, \quad \text{for all } \rho > 0.
\end{aligned}$$

Equating coefficients[13] on both sides we obtain

$$g(T) = \begin{cases} \frac{T(n-T)}{n(n-1)} & \text{if } 1 \leqslant T \leqslant (n-1), \\ 0 & \text{o.w.} \end{cases}$$

This is an UMVUE of $\theta(1-\theta)$. This method is based on writing the unbiasedness condition in terms of a statistic based on T and then directly solving for values of $g(T)$. Another alternative is to use conditioning on an UE of $\tau(\theta)$.

Example 3.2. Assume $\{X_1, \ldots, X_n\}$ are i.i.d. Uniform $(0, \theta)$, with $\theta > 0$. We are interested in estimating $\tau(\theta) = \theta$. Then $T = X_{(n)}$ is complete and sufficient for this family and $U = 2X_1$ is an UE of $\tau(\theta)$. We can directly find an UE of $\tau(\theta)$ using T and that will be the UMVUE. However we will be using the conditioning approach. The estimator $g(T) = \mathbf{E}(U \mid T) = 2\mathbf{E}(X_1 \mid T)$ will be an UMVUE of $\tau(\theta)$.

It can be shown that

$$[X_1 \mid X_{(n)} = t] \overset{d}{=} \begin{cases} t & \text{w.p. } 1/n, \\ \text{Uniform}(0,t) & \text{w.p. } (n-1)/n. \end{cases}$$

[13] We are interested in obtaining a *single* choice of $\{g(t) : 0 \leqslant t \leqslant n\}$, so that the sums match. If we use the same coefficients on both sides, then definitely the sums will match. We do not results on equality of power series on a certain radius of convergence, to find or establish uniqueness of $g(t)$'s. However, the uniqueness of UMVUE's ensures this particular choice of $g(T)$ is unique.

The above distributional result is intuitive, but the proof involves tedious calculations. Using this result, we can write

$$g(t) = \mathbf{E}[2X_1 \mid X_{(n)} = t] = \frac{2t}{n} + \frac{2t}{2} \cdot \frac{n-1}{n} = \frac{(n+1)t}{n}.$$

Thus $g(T) = (n+1)T/n$ is an UMVUE of $\tau(\theta) = \theta$.

Example 3.3. Assume $\{X_1, \ldots, X_n\}$ is an i.i.d. sample from Uniform $(0, \theta)$, where $\theta > 0$. Let $\tau(\theta) = g(\theta)$, where g is a differentiable map on $(0, \infty)$. We aim to find an UMVUE of $\tau(\theta)$. As $T = X_{(n)}$ is complete and sufficient for this family, assume that $h(X_{(n)})$ is possible UMVUE (if it exists). Then it satisfies the unbiasedness condition,

$$\mathbf{E}_\theta h(X_{(n)}) = g(\theta), \quad \text{for all } \theta > 0, \quad \Leftrightarrow \int_0^\infty h(u) \cdot \frac{nu^{n-1}}{\theta^n}\, du = g(\theta)$$

$$\Leftrightarrow\ n \int_0^\infty h(u) u^{n-1}\, du = \theta^n g(\theta) \quad \Leftrightarrow\ n \cdot h(\theta)\theta^{n-1} = \theta^n g^{(1)}(\theta) + n\theta^{n-1} g(\theta) \quad \text{(taking derivatives on both sides)}$$

$$\Leftrightarrow\ h(\theta) = \frac{\theta^n g^{(1)}(\theta) + n\theta^{n-1} g(\theta)}{n\theta^{n-1}}.$$

Since the above relation is true for each $\theta > 0$, and $\mathbf{P}_\theta(X_{(n)} > 0) = 1$, for all $\theta > 0$, we can claim that the UMVUE of $g(\theta)$ will be,

$$h(X_{(n)}) = g(X_{(n)}) + \frac{X_{(n)} g^{(1)}(X_{(n)})}{n},$$

where $g^{(1)}(\cdot)$ denotes the first derivative of g.

Example 3.4. Assume $\{X_1, \ldots, X_n\}$ is an i.i.d. sample from Poisson(θ), where $\theta > 0$. Suppose we are interested in finding an UMVUE of

$$\tau(\theta) = \sum_{j=0}^\infty a_j \theta^j,$$

where $\{a_j : j \geqslant 0\}$ is a sequence of constants, so that the infinite sum converges absolutely for all $\theta > 0$. Then, $T = \sum_{i=1}^n X_i$ is a complete and sufficient statistic for this family of distributions. If $h(T)$ is an UMVUE of $\tau(\theta)$ then,

$$\mathbf{E}_\theta h(T) = \sum_{t=0}^\infty h(t) \cdot e^{-n\theta} \frac{(n\theta)^t}{t!} = \tau(\theta)$$

$$\Rightarrow \sum_{t=0}^\infty \frac{h(t) n^t}{t!} \cdot \theta^t = e^{n\theta} \sum_{j=0}^\infty a_j \theta^j = \left(\sum_{r=0}^\infty \frac{(n\theta)^r}{r!} \right) \cdot \left(\sum_{j=0}^\infty a_j \theta^j \right) = \sum_{t=0}^\infty \left(\sum_{(j,k):j+k=t} \frac{n^k a_j}{k!} \right) \theta^t, \quad \text{for all } \theta > 0.$$

Then, equating coefficients we obtain

$$h(t) = \frac{t!}{n^t} \cdot \left(\sum_{(j,k):j+k=t} \frac{n^k a_j}{k!} \right), \quad \text{for all } t = 0, 1, 2, \ldots.$$

As a simple example, we can choose $\tau(\theta) = \theta^m$, then we set $a_m = 1$ and $a_j = 0$ for all $j \neq m$. With this choice, we get

$$h(t) = \begin{cases} 0 & \text{if } t < m, \\ \frac{t!}{n^m (t-m)!} & \text{if } t \geqslant m. \end{cases}$$

Example 3.5. Assume $\{X_1, \dots, X_n\}$ are i.i.d. $N(\mu, \sigma^2)$, where μ is known and $\sigma > 0$, is unknown. Then, $S_n^2 = \sum_{i=1}^n \left(X_i - \bar{X}_n\right)^2$ is complete and sufficient for this family. Then $\mathbf{E}_\sigma(S_n^2/\sigma^2) = n$, and hence S_n^2/n is UMVUE for σ^2. Using properties for chi-square r.v.'s, we can show that,

$$\mathbf{E}\left(\sigma^{-r} S_n^r\right) = k_{n,r}^{-1}, \quad \text{where } k_{n,r}^{-1} = \frac{2^{r/2}\Gamma(n+r)/2}{\Gamma(n/2)}, \text{ if } n > -r.$$

So, $k_{n,r} S_n^r$ is an UMVUE of σ^r. In case both μ and σ are unknown, then the UMVUE for σ^r remains same.

Example 3.6. Assume $\{X_1, \dots, X_n\}$ are i.i.d. $N(\mu, 1)$, where $\mu \in \mathbb{R}$. We are interested in finding an UMVUE of $\tau_u(\mu) = \mathbf{P}_\mu(X_1 \leqslant u)$, where u is a fixed real number. Then, $U = \mathbf{1}(X_1 \leqslant u)$ is an UE of $\tau(\mu)$. Also $T = \bar{X}_n$ is complete and sufficient for this family. And

$$\begin{aligned} h(\bar{x}_n) &= \mathbf{E}(U \mid \bar{X}_n = \bar{x}_n) = \mathbf{P}(X_1 \leqslant u \mid \bar{X}_n = \bar{x}_n) \\ &= \mathbf{P}(X_1 - \bar{X}_n \leqslant u - \bar{x}_n \mid \bar{X}_n = \bar{x}_n) \\ &= \mathbf{P}(X_1 - \bar{X}_n \leqslant u - \bar{x}_n) = \Phi\left(\sqrt{\frac{n}{n-1}} \cdot (u - \bar{x}_n)\right). \end{aligned}$$

So, $h(\bar{X}_n)$ will be the UMVUE of $\tau_u(\mu)$. The conditioning can be removed because $X_1 - \bar{X}_n$ is ancillary (for any underlying μ) and $\bar{X}_n$ is complete and sufficient. So, Basu's theorem is applicable, making these r.v.'s independent.

Now assume $\{X_1, \dots, X_n\}$ is an i.i.d. sequence of r.v.'s following $N(\mu, \sigma^2)$, where $\mu \in \mathbb{R}$ and $\sigma > 0$ are both unknown parameters. Then, $\mathbf{T}(\mathbf{X}) = (\bar{X}_n,\ S_n^2 = \sum_{i=1}^n \left(X_i - \bar{X}_n\right)^2)$ are complete and sufficient for this family. The estimand is $\tau_u(\mu, \sigma) = \mathbf{P}_{\mu,\sigma}(X_1 \leqslant u) = \Phi((u - \mu)/\sigma)$. As earlier $U = \mathbf{1}(X_1 \leqslant u)$ will be an UE and the UMVUE will be,

$$\begin{aligned} h(\mathbf{T}) &= \mathbf{P}\big(X_1 \leqslant u \mid \bar{X}_n = \bar{x}_n,\ S_n^2 = s_n^2\big) \\ &= \mathbf{P}\left(\frac{X_1 - \bar{X}_n}{S_n} \leqslant \frac{u - \bar{x}_n}{s_n} \,\middle|\, \bar{X}_n = \bar{x}_n,\ S_n^2 = s_n^2\right) = \mathbf{P}\left(\frac{X_1 - \bar{X}_n}{S_n} \leqslant \frac{u - \bar{x}_n}{s_n}\right). \end{aligned}$$

One needs to derive the density of $(X_1 - \bar{X}_n)/S_n$ to find the expression for the abvove probability and it will be in terms of $\bar{x}_n$ and s_n. This will be the UMVUE of $\tau_u(\mu, \sigma)$.

Example 3.7. Assume $\{X_1, \dots, X_n\}$ are i.i.d. $N(\mu, \sigma^2)$ and $\{Y_1, \dots, Y_m\}$ are i.i.d. $N(\xi, \tau^2)$, where the sets of random variables are independent. Assume $\mu, \eta \in \mathbb{R}$ and $\sigma, \tau \in (0, \infty)$. Then $\mathbf{T} = (\bar{X}_n, \bar{Y}_m, S_X^2, S_Y^2)$ will be complete and sufficient. Based on this one can find UMVUE's of various parametric functions, *viz.*, $(\mu - \eta)$, or σ/τ,

Example 3.8. Assume $\{X_1, \dots, X_n\}$ are i.i.d. with pdf

$$f(x : \mu, \sigma) = \sigma^{-1} \exp\left\{-\sigma^{-1}(x - \mu)\right\} \mathbf{1}(x > \mu), \quad \text{where } \mu \in \mathbb{R} \text{ and } \sigma > 0.$$

If σ is known, then $T = X_{(1)}$ is complete and sufficient for the family and one can check that $n\sigma^{-1}(X_{(1)} - \mu) \sim \exp(1)$. Hence, the UMVUE of μ would be, $(X_{(1)} - \sigma/n)$.

If μ is known and σ is unknown, then the family reduces to a one-parameter exponential family with complete and sufficient statistic $T = \sum_{i=1}^n (X_i - \mu)$ and one can find an UMVUE of σ using this statistic.

If μ and σ are both unknown, then $\mathbf{T} = (X_{(1)}, \sum_{i=1}^{n}(X_i - X_{(1)}))$ will be complete and sufficient and they will be independently distributed (see Example 2.32). One can show that in this case,

$$\frac{n(X_{(1)} - \mu)}{\sigma} \sim \exp(1) \quad \text{and} \quad \frac{2\sum_{i=1}^{n}(X_i - X_{(1)})}{\sigma} \sim \chi^2_{2(n-1)}.$$

Using these relations we can find UMVUE's of μ and σ as,

$$X_{(1)} - \frac{\sum_{i=1}^{n}(X_i - X_{(1)})}{n(n-1)} \quad \text{and} \quad \frac{\sum_{i=1}^{n}(X_i - X_{(1)})}{n-1},$$

respectively.

Example 3.9 (UMVUE in a nonparametric family). Assume $\{X_1, \ldots, X_n\}$ is an i.i.d. sample from the family $\mathcal{P} = \{f : f \text{ is a pdf}\}$. Then $\mathbf{T} = (X_{(1)}, \ldots, X_{(n)})$ will be sufficient and complete for this family (see Theorem 10).

An important and useful fact is the following: *an estimator $\phi(X_1, \ldots, X_n)$ will be a function of $\mathbf{T}$ if and only if ϕ is symmetric in its arguments.*

Suppose ϕ is symmetric. To show it is a function of the ordered statistic $\mathbf{T}(\mathbf{X})$ we have to show that if $\mathbf{T}(\mathbf{x}) = \mathbf{T}(\mathbf{y})$ (for some other n-tuple $\mathbf{y} \in \mathbb{R}^n$), then we must have $\phi(\mathbf{x}) = \phi(\mathbf{y})$. Now, $\mathbf{T}(\mathbf{x}) = \mathbf{T}(\mathbf{y})$ implies $x_{(i)} = y_{(i)}$ for each $i = 1, \ldots, n$. On the other hand,

$$\phi(\mathbf{x}) = \phi(\mathbf{T}(\mathbf{x})) = \phi(\mathbf{T}(\mathbf{y})) = \phi(\mathbf{y}),$$

which completes the proof. We have used symmetricity property of ϕ for first and third equality relations in the above argument.

Conversely, assume $\phi(\mathbf{x})$ is a function of $\mathbf{T}(\mathbf{x})$. So, $\phi(\mathbf{x}) = h(\mathbf{T}(\mathbf{x}))$, for some map h. Let $\{i_1, \ldots, i_n\}$ be a permutation of $\{1, \ldots, n\}$. Then,

$$\phi\left(x_{i_1}, \ldots, x_{i_n}\right) = h\left(\mathbf{T}\left(x_{i_1}, \ldots, x_{i_n}\right)\right) = h(\mathbf{T}(\mathbf{x})) = \phi(\mathbf{x}),$$

because ordered statistic of a n-tuple remain invariant under permutation of its components. This proves that ϕ is symmetric in its arguments.

The implication of this result is that, we need to make sure any UE of a parametric function $\tau(f)$ in this family is a symmetric function of the X_i's.

For example, if we consider the parametric function $\tau(f) = \mathbf{P}_f(X_1 \leqslant u)$, where u is a fixed real number and $\mathbf{P}_f$ indicates that f is the underlying pdf. Then, an UE of $\tau(f)$ would be,

$$\phi(X_1, \ldots, X_n) \equiv \frac{1}{n}\sum_{i=1}^{n}\mathbf{1}(X_i \leqslant u) = \frac{1}{n}\sum_{i=1}^{n}\mathbf{1}(X_{(i)} \leqslant u) = F_n(u),$$

the empirical cdf at u, which is a symmetric function of its arguments (and alternatively it is also expressible as a function of the ordered statistics, thus making it symmetric). So $F_n(u)$ will always be an UMVUE in this family.

Example 3.10 (Continuation of Example 3.9). This is a continuation of Example 3.9. Next, consider the parametric function $\tau(f) = \int x f(x)\, dx = \mathbf{E}_f(X_1)$, the mean of the distribution (with pdf f). The underlying family will be $\mathcal{P}_1 = \{f : f \text{ is a pdf, and } \mathbf{E}_f|X| < \infty\}$, the family of continuous distributions with finite mean.

Firstly note that $\mathcal{P}_1 \subset \mathcal{P}$. Although sufficiency of ordered statistics is ensured, but we need to verify if completeness of ordered statistics continues to hold? The method is to re-trace the proof of completeness in Theorem 10 for this sub-family and see if it continues to hold if $\mathcal{P}$ is replaced by $\mathcal{P}_1$. As it turns out, the family $\mathcal{P}_0$ (see (2.15)) is a subset of $\mathcal{P}_1$ and further any set with probability zero under $\mathcal{P}_0$ will be a zero probability set under any member of $\mathcal{P}_1$. Hence, the argument used in Theorem 10 will be applicable and the ordered statistics will be complete for $\mathcal{P}_1$.

It remains to show that for any $f \in \mathcal{P}_0$ (see (2.15)) satisfies $\mathbf{E}_f|X| < \infty$. But, from our earlier discussion in the proof of Theorem 10, we see that for any $f \in \mathcal{P}_0$,

$$\int_{x \in \mathbb{R}} |x| f(x : \boldsymbol{\theta})\, dx \leqslant K_1 + C(\boldsymbol{\theta}) \int_{|x| > M} |x| \exp\{-x^2/2\}\, dx,$$

where K_1 denotes the integral within the range $|x| \leqslant M$ (and K_1 will be finite) and $M = \max\{1, (\theta^\star)^{1/n}\}$. The last integral on the r.h.s. will be finite and in fact, the integral

$$\int_{|x| > M} |x|^r \exp\{-x^2/2\}\, dx < \infty, \quad \text{for any } r \geqslant 1.$$

This shows that if we restricted ourselves to consecutive sub-families

$$\mathcal{P}_r = \{f \in \mathcal{P} : \mathbf{E}_f|X|^r < \infty\}, \quad r \geqslant 1,$$

then also the ordered statistics would continue to be complete. This allows us to use the ordered statistics for finding UMVUE's in families where higher moments exist. So, in the family $\mathcal{P}_r$, the UMVUE of $\mathbf{E}(X^r)$ would be $n^{-1} \sum_{i=1}^n X_i^r$.

Example 3.11 (Non-existence of UMVUE of the mean in a family). As we saw in Example 3.10 above, the UMVUE of the mean in the family $\mathcal{P}_1$ is the sample mean. However there is a particular sub-family of distributions, where the UMVUE of the mean fails to exist!

We say that a continuous r.v. X with pdf f is symmetric if there exists an $a \in \mathbb{R}$, such that

$$f(a - t) = f(a + t), \quad \text{for all } t > 0. \tag{3.23}$$

The Normal distribution, Double exponential distribution, Cauchy distribution, etc., are members of the family of symmetric continuous distributions. We consider the sub-family of symmetric pdf's where the mean exists. Define the class of distributions,

$$\mathcal{H} = \{f : f \text{ is a pdf, } \mathbf{E}_f|X| < \infty, \text{ and } f(a-t) = f(a+t) \text{ for all } t > 0 \text{ and for some } a \in \mathbb{R}.\} \tag{3.24}$$

Consider any particular $f \in \mathcal{H}$ and let a be its center of symmetry and X denote the underlying r.v. with pdf f. We are interested in finding an UMVUE of

$$\mu(f) = \mathbf{E}_f(X), \quad \text{for all } f \in \mathcal{H}.$$

Note that

$$\begin{aligned}\int xf(x)\ dx &= \int_{-\infty}^{a} xf(x)\ dx + \int_{a}^{\infty} xf(x)\ dx\\ &= -\int_{\infty}^{0} (a-t)f(a-t)\ dt + \int_{0}^{\infty} (a+t)f(a+t)\ dt\\ &= a\int_{0}^{\infty} f(a-t)\ dt + a\int_{0}^{\infty} f(a+t)\ dt - \int_{0}^{\infty} tf(a-t)\ dt + \int_{0}^{\infty} tf(a+t)\ dt\\ &= a\int_{\infty}^{\infty} f(x)\ dx = a.\end{aligned}$$

In the second line we substituted $x = a - t$ and $x = a + t$, in the first and second integrals on the right side and in the penultimate step we used $f(a + t) = f(a - t)$ for each $t > 0$, to cancel the last two terms. Also note that if X is the underlying r.v., then $Y = (X - a) \overset{d}{=} (a - X) = Z$, because[14]

$$P(Y \leqslant t) = P(X \leqslant t + a) = P(X \geqslant a - t) = P(Z \leqslant t), \quad \text{for all } t \in \mathbb{R},$$

and this follows from (3.23). If $\{X_1, \ldots, X_n\}$ are i.i.d. f, then define $Y_i = (X_i - a)$ and $Z_i = (a - X_i)$ for all $i = 1, \ldots, n$. Since Y and Z have the same distribution, $Y_{(1)} \overset{d}{=} Z_{(1)}$. And note that

$$X_{(1)} = a + Y_{(1)} \quad \text{and} \quad X_{(n)} = a - Z_{(1)}.$$

Hence, at $f = f_a$,

$$\mathbf{E}_f\left(\frac{X_{(1)} + X_{(n)}}{2}\right) = \frac{1}{2}\left[2a + \mathbf{E}(Y_{(1)}) - \mathbf{E}(Z_{(1)})\right] = a = \mu(f),\ \text{for all } f \in \mathcal{H}. \tag{3.25}$$

This fact, although very simple, would play an important role in our arguments.

Now we return to our original objective. We aim to show that no UMVUE of $\mu(f)$ exists for all $f \in \mathcal{H}$. Assume the contrary: suppose T is the required UMVUE. Then

(a) $\mathbf{E}_f(T) = \mu(f)$ for all $f \in \mathcal{H}$.

(b) for any T_1 satisfying (a) above, $\mathbf{Var}_f(T) \leqslant \mathbf{Var}_f(T_1)$ for all $f \in \mathcal{H}$.

Now consider the sub-family

$$\mathcal{H}_1 = \{f : f \text{ is the pdf of } N(\mu, 1),\ \mu \in \mathbb{R}\}.$$

In this case for any $f \in \mathcal{H}_1$, $\mu(f) = \mu$. Using the completeness and sufficiency of $\bar{X}_n$, we can say that $\bar{X}_n$ will be the UMVUE of $\mu(f)$ in the family $\mathcal{H}_1$.

By part (a) above, $\mathbf{E}_f(T) = \mu(f)$ for all $f \in \mathcal{H}_1$. Since $\bar{X}_n$ is the UMVUE of $\mu(f)$ in $\mathcal{H}_1$, we must have

$$\mathbf{Var}_f(\bar{X}_n) \leqslant \mathbf{Var}_f(T) \quad \text{for all } f \in \mathcal{H}_1.$$

Also note that for any $f \in \mathcal{H}$, $\mathbf{E}_f(\bar{X}_n) = \mu(f)$. Hence by part (b) above,

$$\mathbf{Var}_f(T) \leqslant \mathbf{Var}_f(\bar{X}_n) \quad \text{for all } f \in \mathcal{H}.$$

[14] For example, $X \sim N(a, 1)$ has center of symmetry at a. Then $Y = X - a \sim N(0, 1)$ and $Z = a - X \sim N(0, 1)$.

Combining the above we obtain

$$\mathbf{Var}_f(T) = \mathbf{Var}_f(\bar{X}_n) \quad \text{for all } f \in \mathcal{H}_1.$$

Since $\bar{X}_n$ is the UMVUE of $\mu(f)$ in $\mathcal{H}_1$, by the uniqueness of UMVUE's, we must have

$$\mathbf{P}_f(\bar{X}_n = T) = 1, \quad \text{for all } f \in \mathcal{H}_1.$$

Now, note that if any set A satisfies $\mathbf{P}_f(X \in A) = 0$ for all $f \in \mathcal{H}_1$, then it must satisfy $\mathbf{P}_f(X \in A) = 0$ for all $f \in \mathcal{H}$. This simply states that the family $\mathcal{H}_1$ 'dominates' the family $\mathcal{H}$. Hence we can claim that

$$\mathbf{P}_f(\bar{X}_n = T) = 1, \quad \text{for all } f \in \mathcal{H}. \tag{3.26}$$

Now consider another sub-family of $\mathcal{H}$,

$$\mathcal{H}_2 = \{f : f \text{ is the pdf of } U(\theta_1 - \theta_2, \theta_1 + \theta_2) \text{ for } \theta_1 \in \mathbb{R} \text{ and } \theta_2 > 0\}.$$

Define $W = (X_{(1)} + X_{(n)})/2$. Then W is known (by usual completeness and sufficiency arguments) to be the UMVUE of $\mu(f)$ for all $f \in \mathcal{H}_2$. By part (a) above, once again

$$\mathbf{Var}_f(W) \leqslant \mathbf{Var}_f(T) = \mathbf{Var}_f(\bar{X}_n), \quad \text{for all } f \in \mathcal{H}_2.$$

The equality above follows from (3.26). Now, from (3.25), we have

$$\mathbf{E}_f(W) = \mu(f), \quad \text{for any } f \in \mathcal{H}.$$

Hence by part (b) above, we must have

$$\mathbf{Var}_f(T) = \mathbf{Var}_f(\bar{X}_n) \leqslant \mathbf{Var}_f(W), \quad \text{for all } f \in \mathcal{H} \supset \mathcal{H}_2.$$

Following earlier arguments, we can say

$$\mathbf{Var}_f(T) = \mathbf{Var}_f(\bar{X}_n) = \mathbf{Var}_f(W) \quad \text{for all } f \in \mathcal{H}_2.$$

Once again using the uniqueness of UMVUE's in a family, we must have

$$\mathbf{P}_f(\bar{X}_n = W) = 1, \quad \text{for all } f \in \mathcal{H}_2. \tag{3.27}$$

It means, in the Uniform$(\theta_1 - \theta_2,\ \theta_1 + \theta_2)$ family, the sample mean is equal to the average value of the extreme order statistics w.p. 1. If $n > 2$, then the above is impossible. This leads to a contradiction to the original assumption about the existence of T. Hence, there does not exist any UMVUE of $\mu(f)$ for the family $\mathcal{H}$.

Infact, if $n > 2$, then by our earlier calculations,

$$\mathbf{E}_f\left(\frac{X_{(1)} + X_{(n)}}{2} - \frac{1}{n}\sum_{i=1}^{n} X_{(i)}\right) = 0, \quad \text{for all } f \in \mathcal{H}.$$

But that does not imply

$$\mathbf{P}_f\left(\frac{X_{(1)} + X_{(n)}}{2} - \bar{X}_n = 0\right) = 1, \quad \text{for all } f \in \mathcal{H}.$$

This also implies that the order statistics $(X_{(1)}, \ldots, X_{(n)})$ are not complete for the family $\mathcal{H}$.

The only possibility arises when $n = 2$ and $\bar{X}_n$ can be used as the UMVUE of $\mu(f)$. For $n = 1$, X_1 is complete for the family $\mathcal{H}_1$ and $\mathcal{H}_1$ dominates $\mathcal{H}$. This implies, X_1 is complete (and trivially sufficient) in $\mathcal{H}$ and $\mathbf{E}_f(X_1) = \mu(f)$ for all $f \in \mathcal{H}$. So the UMVUE of $\mu(f)$ would be X_1.

For the larger family of all continous random variables, $(X_{(1)}, \ldots, X_{(n)})$ are complete and sufficient, and no such problem arises.

In the next result we provide a method to obtain an UMVUE, when no complete statistic is available.

Theorem 16. *Consider the family $\mathcal{P} = \{\mathbf{P}_\theta : \theta \in \Theta\}$. Suppose,*

$$\mathcal{U} = \{U : \mathbf{E}_\theta U = 0 \ \text{ and } \ \mathbf{E}_\theta U^2 < \infty, \ \text{for all } \theta\}$$

denote the class of unbiased estimators of zero (with finite variance). Suppose, $\widetilde{T}$ is a sufficient statistic for the family. Also define the class of unbiased estimators of zero, based on $\widetilde{T}$ as

$$\mathcal{U}_{\widetilde{T}} = \{U : U = g(\widetilde{T}), \ \mathbf{E}_\theta U = 0 \ \text{ and } \ \mathbf{E}_\theta U^2 < \infty, \ \text{for all } \theta, \text{ where } g \text{ is some function of } \widetilde{T}\}.$$

Consider an unbiased estimator T of $\tau(\theta)$, such that $T = h(\widetilde{T})$. Then, T will be an UMVUE of $\tau(\theta)$ iff

$$\mathbf{E}_\theta(TU) = 0, \quad \text{for all } \theta \text{ and all } U \in \mathcal{U}_{\widetilde{T}}. \tag{3.28}$$

This theorem provides a way to obtain an UMVUE without the existence of a complete sufficient statistic, as long as a sufficient statistic exists for the family. The only requirement is zero covariance with all UE's of zero which are based on the sufficient statistic.

Proof of Theorem 16. We will show that if T satisfies condition (3.28), then it must satisfy

$$\mathbf{E}_\theta(TU) = 0, \quad \text{for all } \theta \text{ and all } U \in \mathcal{U}.$$

Suppose $U \in \mathcal{U}$. Then, $\mathbf{E}_\theta(U \mid \widetilde{T}) \in \mathcal{U}_{\widetilde{T}}$. And

$$\begin{aligned}
\mathbf{E}_\theta(TU) &= \mathbf{E}_\theta\left[\mathbf{E}_\theta\left(TU|\widetilde{T}\right)\right] \\
&= \mathbf{E}_\theta\left[\mathbf{E}_\theta\left(h(\widetilde{T}) \cdot U|\widetilde{T}\right)\right] \quad (\text{by definition of } \widetilde{T}), \\
&= \mathbf{E}_\theta\left[h(\widetilde{T}) \cdot \mathbf{E}_\theta\left(U|\widetilde{T}\right)\right] \\
&= \mathbf{E}_\theta\left[T \cdot V\right], \quad (\text{where } V = \mathbf{E}_\theta(U|\widetilde{T}) \in \mathcal{U}_{\widetilde{T}}) \\
&= 0 \quad (\text{due to (3.28).})
\end{aligned}$$

This completes the proof of the sufficient condition. Now, assume that T is the UMVUE of $\tau(\theta)$. Then it must satisfy

$$\mathbf{E}_\theta(TU) = 0 \quad \text{for all } \theta \text{ and all } U \in \mathcal{U}.$$

Since, $\mathcal{U}_{\widetilde{T}} \subset \mathcal{U}$, the necessity part follows. □

Example 3.12 (Finding UMVUE in absence of a complete statistic). Suppose $\{X_1, \ldots, X_n\}$ are i.i.d. Uniform $(0, \theta)$, with $\theta \in \Theta = [1, \infty)$. In this case, $T = X_{(n)}$ is sufficient for the family, but T is not complete. To obtain an UMVUE, we use Theorem 16.

Our target is $\tau(\theta) = \theta$ for the above mentioned family. In this case $\widetilde{T} = X_{(n)}$ and we will need to characterize the class $\mathcal{U}_{X_{(n)}}$. Suppose, $U(X_{(n)})$ is an UE of zero. Then, for all $\theta \geqslant 1$,

$$\begin{aligned} 0 = \mathbf{E}_\theta U(X_{(n)}) = \int_0^\theta U(t) \frac{nt^{n-1}}{\theta^n}\, dt \ &\Leftrightarrow \ 0 = \int_0^\theta U(t) t^{n-1}\, dt \\ \Leftrightarrow \ 0 &= \int_0^1 U(t)t^{n-1}\, dt + \int_1^\theta U(t)t^{n-1}\, dt \\ \Leftrightarrow \ 0 &= c + \int_1^\theta U(t)t^{n-1}\, dt \\ \Leftrightarrow \ 0 &= c + \psi(\theta), \quad \text{(say)}. \end{aligned}$$

This implies, $\psi(\theta) = -c$ for all $\theta \geqslant 1$. Thus, since $\psi(\theta)$ is differentiable on $(1, \infty)$, we can say, $\psi'(\theta) = 0$ for all $\theta > 1$. Also,

$$0 = \psi'(\theta) = U(\theta)\theta^{n-1},$$

implies, $U(x) = 0$ for all $x > 1$. Since for all θ, $\mathbf{P}_\theta(X = 1) = 0$, we can say,

$$U(t) = 0 \quad \text{for all } t \geqslant 1, \text{ w.p. } 1 \text{ for all } \theta \in [1, \infty). \tag{3.29}$$

Now, we need to find the UE of θ, based on $X_{(n)}$. Suppose, $T = h(X_{(n)})$. In order to satisfy (3.28) we must have

$$\int_0^\theta h(t)U(t)t^{n-1}\, dt = 0 = \int_0^1 h(t)U(t)t^{n-1}\, dt, \tag{3.30}$$

due to (3.29). Consider an $h(t)$ of the following form:

$$h(t) = \begin{cases} c & \text{if } 0 \leqslant t \leqslant 1, \\ bt & \text{if } t > 1, \end{cases}$$

where, c and b are some constants. The remark discusses this choice of the form of $h(\cdot)$. Now due to unbiasedness, for all $\theta \geqslant 1$,

$$\begin{aligned} \mathbf{E}_\theta h(X_{(n)}) = \theta &= \int_0^\theta h(t) \cdot \frac{nt^{n-1}}{\theta^n}\, dt \\ &\Rightarrow \theta = cP_\theta(X_{(n)} \leqslant 1) + b\int_1^\theta \frac{nt^n}{\theta^n}\, dt \\ &\Rightarrow \theta = \frac{c}{\theta^n} + \frac{nb}{\theta^n} \cdot \frac{\theta^{n+1} - 1}{n+1} \\ &\Rightarrow \theta = \frac{nb\theta}{n+1} + \frac{1}{\theta^n} \cdot \left(c - \frac{nb}{n+1}\right). \end{aligned}$$

Now equating coeffienct of θ on both sides, we have

$$b = \frac{n+1}{n}, \quad \text{and} \quad c = 1.$$

Hence,

$$T = h(X_{(n)}) = \begin{cases} 1 & \text{if } 0 \leqslant X_{(n)} \leqslant 1, \\ \frac{(n+1)X_{(n)}}{n} & \text{if } X_{(n)} > 1, \end{cases} \tag{3.31}$$

is the required UMVUE of θ.

Remark: Note that we already know $\int_0^1 U(t)t^{n-1}\ dt = 0$. So, in order to satisfy (3.30) we can choose $h(t) = c$ for all $t \in [0,1]$. This will simplify our calculations. Note that we are at liberty to choose $h(\cdot)$ freely as long as $T = h(X_{(n)})$ is an UE of θ and satisfies (3.28). Hence, this choice of $h(t)$ on $[0,1]$ is valid. We are concerned about the values of h on $(1,\infty)$.

Regarding the choice of $h(t)$ on $(1,\infty)$, we can argue as follows. Suppose, $h(t) = \psi(t)$ for all $t > 1$, where $\psi(t)$ is some arbitrary function. Then, due to the unbiasedness condition, for all $\theta \geqslant 1$,

$$\begin{aligned} \theta^{n+1} &= c + \int_1^\theta \psi(t)nt^{n-1}\ dt \\ \Rightarrow \int_0^1 1\ dt + \int_1^\theta (n+1)t^n\ dt &= \int_0^1 c\ dt + \int_1^\theta \psi(t)nt^{n-1}\ dt \\ \Rightarrow \int_0^\theta h_1(t)\ dt &= \int_0^\theta h_2(t)\ dt \\ \Rightarrow \int_0^\theta (h_1(t) - h_2(t))\ dt &= 0. \end{aligned}$$

Differentiating wrt θ, we get,

$$h_1(\theta) = h_2(\theta) \quad \text{for all } \theta > 1.$$

This leads to $\psi(t) = (n+1)t/n$ for all $t > 1$. This is exactly the form of $h(\cdot)$ which we had selected.

Remark: Note that $h(X_{(n)})$ is also sufficient for θ. It is enough to show that

$$g(X_{(n)}) = 1 \cdot \mathbf{1}(X_{(n)} \in [0,1]) + X_{(n)} \cdot \mathbf{1}(X_{(n)} > 1),$$

is sufficient, since g and h are 1-1 functions. We can write

$$f(\mathbf{x} : \theta) = \theta^{-n}\mathbf{1}(x_{(n)} \leqslant \theta) = \theta^{-n}\mathbf{1}(g(x_{(n)}) \leqslant \theta),$$

which shows $g(X_{(n)})$ is sufficient. Also it is complete. Because, if for all $\theta \geqslant 1$, $\mathbf{E}_\theta(\phi(g(X_{(n)}))) = 0$, then

$$0 = \int_0^\theta \phi(g(u))u^{n-1}\ du = \int_0^1 \phi(1)u^{n-1}\ du + \int_1^\theta \phi(u)u^{n-1}\ du.$$

Let, $\theta \to 1^+$, then $\phi(1) = 0$ and this implies, for all $\theta > 1$,

$$\int_1^\theta \phi(u)u^{n-1}\ du = 0.$$

This concludes that $\phi(u) = 0$ for all $u \geqslant 1$ and w.p. 1 under all $\theta \geqslant 1$.

3.3 The Cramer-Rao lower bound on the variance of unbiased estimators

We explored direct methods for constructing UMVUE's. The Cramer-Rao Lower Bound (CRLB) provides an indirect method, by finding the lowest possible variance of an UE of a parametric function. Any unbiased estimator with a variance matching this lower bound automatically becomes an UMVUE.

Suppose $\mathbf{X}$ has a pdf or pmf $p(\mathbf{x} : \theta)$ for each θ. The CRLB states that the variance of any unbiased estimator $W = W(\mathbf{X})$ of $\tau(\theta)$, will satisfy

$$\mathbf{Var}_\theta(W) \geqslant \frac{(\tau'(\theta))^2}{\mathbf{E}_\theta\left(\frac{\partial}{\partial\theta}\log p(\mathbf{x} : \theta)\right)^2}, \quad \text{for all } \theta. \tag{3.32}$$

The quantity on the right side of (3.32) is called the CRLB for the parametric function $\tau(\theta)$. The expected value in the denominator is denoted as $I(\theta)$, and known as the Information number (or Fisher information). Note that the bound has nothing to do with the estimator W, but depends on the choice of the estimand $\tau(\theta)$. This bound is universal for any UE of $\tau(\theta)$. The idea is that the larger the number $I(\theta)$, the lower will be the CRLB. If any UE attains this lower variance bound, it will definitely be an UMVUE of $\tau(\theta)$.

However, finding the CRLB is not always enough to find an UMVUE. In some situations, an UMVUE $T(\mathbf{X})$ of $\tau(\theta)$ might exist and yet it may not attain the CRLB. On the other hand, if certain regularity conditions are not satisfied (like common support for $p(\mathbf{x} : \theta)$), the CRLB result is not even applicable. The lower bound in (3.32) is also valid in case $\tau(\theta)$ is a vector valued estimand. We state and prove the general result for vector valued $\tau(\theta)$.

Consider a family $\mathcal{P} = \{\mathbf{P}_\theta : \theta \in \Theta\}$ of distributions. In all cases A^T denotes the transpose of a matrix A. We describe the basic framework.

(a) $\mathbf{X} = \{X_1, \ldots, X_n\} \sim \mathbf{P}_\theta$, with $\theta \in \Theta \subseteq \mathbb{R}^k$. Thus, we are considering a parameter vector θ of the form $\theta = (\theta_1, \ldots, \theta_k)$.

(b) P_θ has a pdf (or pmf), which is denoted by $p(\mathbf{x} : \theta)$.

(c) $T(\mathbf{X}) = (T_1, \ldots, T_s)$ is an unbiased estimator of $\tau(\theta) = (\tau_1(\theta), \ldots, \tau_s(\theta))$, that is $\mathbf{E}_\theta T_j(\mathbf{X}) = \tau_j(\theta)$ for all θ. We assume $\tau : \Theta \mapsto \mathbb{R}^s$ or for simplicity $\tau : \mathbb{R}^k \mapsto \mathbb{R}^s$.

(d) The Jacobian matrix of the map τ is the $s \times k$ matrix

$$J_\tau(\theta) = \frac{\partial\tau(\theta)}{\partial\theta} = \begin{pmatrix} \frac{\partial\tau_1(\theta)}{\partial\theta_1} & \cdots & \frac{\partial\tau_1(\theta)}{\partial\theta_k} \\ . & \cdots & . \\ . & \cdots & . \\ \frac{\partial\tau_s(\theta)}{\partial\theta_1} & \cdots & \frac{\partial\tau_s(\theta)}{\partial\theta_k} \end{pmatrix}.$$

The transpose $J'_\tau(\theta)$ will be a $k \times s$ matrix and is called the **derivative matrix** of the map τ. We will denote this by

$$\dot\tau(\theta) = J'_\tau(\theta). \tag{3.33}$$

We obviously assume that all partial derivatives exist. The ith column of $\dot\tau(\theta)$ will be of length k and will contain all partial derivatives of the function τ_i.

Theorem (CRLB Inequality)**.** *Suppose the following conditions are true:*

(A1) Θ *is an open subset of* $\mathbb{R}^k$*.*

(A2) The set, $\mathcal{C} = \{x : p(x : \theta) = 0\}$ *does not depend on* θ*. For all* θ *and all* $i = 1, \dots, k$*,*

$$\frac{\partial p(\mathbf{x} : \theta)}{\partial \theta_i}$$

exists for almost all $\mathbf{x}$*. In other words, the set of* $\mathbf{x}$ *values, where this partial derivative will not exist has probability zero under* P_θ *for all* $\theta \in \Theta$*.*

(A3) The $k \times k$ *matrix with* (i, j)*th element*

$$I_{i,j}(\theta) = \mathbf{E}_\theta \left[\frac{\partial}{\partial \theta_i} \log p(\mathbf{X} : \theta) \cdot \frac{\partial}{\partial \theta_j} \log p(\mathbf{X} : \theta) \right],$$

will be positive definite. This matrix $I(\theta)$ *is called the Fisher information matrix for* θ*.*

(A4) $\int_{\mathbf{x}} p(\mathbf{x} : \theta)\ d\mathbf{x}$ *and* $\int_{\mathbf{x}} T_j(\mathbf{x}) p(\mathbf{x} : \theta)\ d\mathbf{x}$ *can both be differentiated with respect to* θ *under the integral sign, for all* $1 \leqslant j \leqslant s$*.*

(A5) $\int_{\mathbf{x}} p(\mathbf{x} : \theta)\ d\mathbf{x}$ *can be twice differentiated under the integral sign.*

If assumptions (A1)-(A4) are true, then

$$\mathbf{Var}_\theta\left(T(\mathbf{X})\right) \geqslant (\dot{\tau}(\theta))'[I(\theta)]^{-1}\dot{\tau}_\theta, \tag{3.34}$$

where, for two matrices A, B*, the inequality* $A \geqslant B$ *implies* $(A - B)$ *is positive semi definite. If, in addition (A5) holds then,*

$$I(\theta) = -\left(\left(\mathbf{E}_\theta \left\{ \frac{\partial^2}{\partial \theta_i \partial \theta_j} \log p(\mathbf{X} : \theta) \right\}\right)\right)_{i,j=1,\dots,k}. \tag{3.35}$$

Proof. Note that for all $1 \leqslant j \leqslant s$,

$$\tau_j(\theta) = \mathbf{E}_\theta T_j(\mathbf{X}) = \int_{\mathbf{x}} T_j(\mathbf{x}) p(\mathbf{x} : \theta)\ d\mathbf{x}. \tag{3.36}$$

Note that the (i, j)-th element of the $(s \times k)$ Jacobian matrix $J_\tau(\theta)$ is

$$\frac{\partial}{\partial \theta_j} \tau_i(\theta).$$

Hence, the (i, j)-th element of the $(k \times s)$ matrix $\dot{\tau}_\theta$ is the (j, i)th element of $J_\tau(\theta)$ and

$$\begin{aligned}
\dot{\tau}_\theta(i, j) = \frac{\partial}{\partial \theta_i} \tau_j(\theta) &= \frac{\partial}{\partial \theta_i} \int_{\mathbf{x}} T_j(\mathbf{x}) p(\mathbf{x} : \theta)\ d\mathbf{x} \\
&= \int_{\mathbf{x}} T_j(\mathbf{x}) \cdot \left(\frac{1}{p(\mathbf{x} : \theta)} \cdot \frac{\partial}{\partial \theta_i} p(\mathbf{x} : \theta) \right) \cdot p(\mathbf{x} : \theta)\ d\mathbf{x} \\
&= \mathbf{E}_\theta \left[T_j(\mathbf{X}) \cdot \frac{\partial}{\partial \theta_i} \log p(\mathbf{X} : \theta) \right] \\
&= \mathbf{E}_\theta \left[\{T_j(\mathbf{X}) - \mathbf{E}_\theta T_j(\mathbf{X})\} \cdot \frac{\partial}{\partial \theta_i} \log p(\mathbf{X} : \theta) \right] \quad \text{(why!)} \\
&= \mathbf{cov}_\theta \left(T_j(\mathbf{X}), \frac{\partial}{\partial \theta_i} \log p(\mathbf{X} : \theta) \right).
\end{aligned}$$

Hence,

$$\dot{\tau}(\theta) = \mathbf{E}_\theta \left[\begin{pmatrix} \frac{\partial}{\partial\theta_1} \log p(\mathbf{X}:\theta) \\ \vdots \\ \frac{\partial}{\partial\theta_k} \log p(\mathbf{X}:\theta) \end{pmatrix} \begin{pmatrix} T_1(\mathbf{X}) & \cdots & T_s(\mathbf{X}) \end{pmatrix} \right] = \mathbf{E}_\theta \left[\dot{L}(\theta) T' \right], \quad \text{(say)},$$

where $\dot{L}(\theta)$ is the $k \times 1$ vector of partial derivatives of $\log p(\mathbf{X}:\theta)$. Note that $\mathbf{E}_\theta \dot{L}(\theta) = \mathbf{0}_{k\times 1}$, because,

$$\mathbf{E}_\theta \left(\frac{\partial}{\partial\theta_i} \log p(\mathbf{X}:\theta) \right) = \int_{\mathbf{x}} \left(\frac{\partial}{\partial\theta_i} \log p(\mathbf{x}:\theta) \right) p(\mathbf{x}:\theta)\, d\mathbf{x} = \frac{\partial}{\partial\theta_i} \int_{\mathbf{x}} p(\mathbf{x}:\theta)\, d\mathbf{x} = 0.$$

Now, for every $\mathbf{u} \in \mathbb{R}^s$ and $\mathbf{v} \in \mathbb{R}^k$,

$$\begin{aligned} \mathbf{cov}_\theta(\mathbf{u}'T, \dot{L}(\theta)'\mathbf{v}) &= \mathbf{E}_\theta \left(\mathbf{u}'T \cdot \dot{L}(\theta)'\mathbf{v} \right) \\ &= \mathbf{E}_\theta \left[\mathbf{u}' (T - \tau) (\dot{L}(\theta))' \mathbf{v} \right] = \mathbf{u}' \left(\mathbf{E}_\theta T (\dot{L}(\theta))' - \tau \mathbf{E}_\theta (\dot{L}(\theta))' \right) \mathbf{v} \\ &= \mathbf{u}'(\dot{\tau}(\theta))'\mathbf{v}, \end{aligned} \tag{3.37}$$

due to the above relations. Note that the right side of (3.37) is a scalar. Now applying C-S inequality,

$$\left(\mathbf{u}'(\dot{\tau}(\theta))'\mathbf{v} \right)^2 = \left(\mathbf{cov}_\theta(\mathbf{u}'T, \dot{L}(\theta)'\mathbf{v}) \right)^2 \leqslant \mathbf{Var}_\theta \left(\mathbf{u}'T \right) \cdot \mathbf{Var}_\theta \left(\dot{L}(\theta)'\mathbf{v} \right) = \left(\mathbf{u}'\mathbf{Var}_\theta(T)\mathbf{u} \right) \cdot \left(\mathbf{v}'I(\theta)\mathbf{v} \right).$$

Now, use $\mathbf{v} = (I(\theta))^{-1}\dot{\tau}(\theta)\mathbf{u}$. We then obtain,

$$\begin{aligned} \left(\mathbf{u}'(\dot{\tau}(\theta))'\mathbf{v} \right)^2 &= \left(\mathbf{u}'(\dot{\tau}(\theta))'(I(\theta))^{-1}\dot{\tau}(\theta)\mathbf{u} \right)^2 \\ &\leqslant \left(\mathbf{u}'\mathbf{Var}_\theta(T)\mathbf{u} \right) \cdot \left(\mathbf{u}'(\dot{\tau}_\theta)'(I(\theta))^{-1}I(\theta)(I(\theta))^{-1}\dot{\tau}(\theta)\mathbf{u} \right) \\ &= \left(\mathbf{u}'\mathbf{Var}_\theta(T)\mathbf{u} \right) \cdot \left(\mathbf{u}'(\dot{\tau}_\theta)'(I(\theta))^{-1}\dot{\tau}(\theta)\mathbf{u} \right), \end{aligned}$$

which implies, for all $\mathbf{u} \in \mathbb{R}^s$,

$$\mathbf{u}' \left[\mathbf{Var}_\theta(T) - (\dot{\tau}(\theta))'(I(\theta))^{-1}\dot{\tau}(\theta) \right] \mathbf{u} \geqslant 0.$$

This completes the required proof of (3.34). The proof of (3.35) can be completed by following similar techniques as the univariate case.

□

3.3.1 Some more discussion on the CRLB

If $\mathbf{X}$ and $\mathbf{Y}$ are independent, with pdf (or pmf) $f(\mathbf{x}:\theta)$ and $g(\mathbf{y}:\theta)$ and with Fisher information $I_{\mathbf{X}}(\theta)$ and $I_{\mathbf{Y}}(\theta)$, then the Fisher information available from $(\mathbf{X}, \mathbf{Y})$ will be

$$I_{(\mathbf{X},\mathbf{Y})}(\theta) = I_{\mathbf{X}}(\theta) + I_{\mathbf{Y}}(\theta).$$

Suppose, $\mathbf{X} = \{X_1, \ldots, X_n\}$, where the X_i's are *iid* with pdf (or pmf) $p(x:\theta)$, with $I_{X_1}(\theta)$ being the Fisher information from X_1, then $I_{\mathbf{X}}(\theta) = nI_{X_1}(\theta)$. If we apply this on the right side of (3.32), we would obtain

$$\mathbf{Var}_\theta(W) \geqslant \frac{(\tau'(\theta))^2}{n\mathbf{E}_\theta \left(\frac{\partial}{\partial\theta} \log p(x:\theta) \right)^2}, \quad \text{for all } \theta.$$

In the special case where X_i are *iid* with joint pdf/pmf $p(x:\theta) = \prod_{i=1}^{n} p(x_i:\theta)$, an UE $W(\mathbf{X})$ of $\tau(\theta)$ will attain the CRLB *if and only if*,

$$a(\theta)\,(W(\mathbf{x}) - \tau(\theta)) = \frac{\partial}{\partial\theta}\log\prod_{i=1}^{n} p(x_i:\theta),$$

where $a(\theta)$ is some function of θ (for a proof, see Corollary 7.3.15 in Casella and Berger (1990)). This condition typically holds in one-parameter exponential families (see page 120-121 of Lehmann and Casella (1998)) and is an useful tool to check if an UE of $\tau(\theta)$ will attain the CRLB. The inequality in (3.32) or (3.34) is known as the *Information inequality.*

Usually, for applying the CRLB, we need to verify all the stated regularity conditions. Specially the proof of the CRLB depends heavily on the interchange of the integration and differentiation operations. This can be a difficult thing to check in every situation. In two common situations, this interchange can be carried out without any problems:

(a) The support of $p(x:\theta)$ is finite and does not depend on θ.

(b) If the density belongs to an exponential family. This takes care of many cases, when the density has infinite support, but still the range does not depend on θ.

In other cases, when neither of the above two assumptions are satisfied, we need to check separately if the interchange operation is still valid. The CRLB does not provide us a method to explicitly find UMVUE's, if they exist. It turns out that when a sufficient and complete statistic exists for the family, finding UMVUE's can be easier.

The main idea behind the CRLB inequality is to use the Cauchy-Schwarz inequality cleverly. The CS inequality states that for any two random variables, Z_1 and Z_2, we have, $\big(\mathbf{cov}(Z_1,Z_2)\big)^2 \leqslant \mathbf{Var}(Z_1)\cdot\mathbf{Var}(Z_2)$. In case of the CRLB, one uses

$$Z_1 \equiv W(\mathbf{X}) \quad \text{and} \quad Z_2 \equiv \frac{\partial}{\partial\theta}\ \log f(\mathbf{X}:\theta).$$

Then,

$$\begin{aligned}\tau(\theta) = \frac{\partial}{\partial\theta}\ \mathbf{E}_\theta W(\mathbf{X}) &= \int W(\mathbf{x})\cdot\frac{\partial}{\partial\theta}\ p(\mathbf{x}:\theta)\ d\mathbf{x} \quad \text{(interchanging derivative and integral)},\\ &= \mathbf{E}_\theta\left[W(\mathbf{X})\cdot\frac{\partial}{\partial\theta}\ \log p(\mathbf{x}:\theta)\right] = \mathbf{E}_\theta(Z_1Z_2).\end{aligned}$$

But, again interchanging integral and derivative it is easy to check that $\mathbf{E}_\theta(Z_2) = 0$, for all θ, since $p(\mathbf{x}:\theta)$ is a pdf for all θ. Hence, $\mathbf{cov}_\theta(Z_1,Z_2) = \mathbf{E}_\theta(Z_1Z_2) \leqslant \sqrt{\mathbf{Var}_\theta(Z_1)\cdot\mathbf{Var}_\theta(Z_2)}$. Thus, using CS inequality we obtain,

$$\begin{aligned}\tau^2(\theta) &\leqslant \mathbf{Var}_\theta(W(\mathbf{X}))\cdot\mathbf{Var}_\theta\left(\frac{\partial}{\partial\theta}\ \log p(\mathbf{X}:\theta)\right)\\ \Rightarrow\ \tau^2(\theta) &\leqslant \mathbf{Var}_\theta(W(\mathbf{X}))\cdot\mathbf{E}_\theta\left(\frac{\partial}{\partial\theta}\ \log p(\mathbf{X}:\theta)\right)^2,\\ \Rightarrow\ \frac{\tau^2(\theta)}{\mathbf{E}_\theta\big(\frac{\partial}{\partial\theta}\ \log p(\mathbf{X}:\theta)\big)^2} &\leqslant \mathbf{Var}_\theta(W(\mathbf{X})).\end{aligned}$$

This immediately allows us to find when equality will be attained in the CRLB inequality. If the second derivative of $\log p(\mathbf{x} : \theta)$ exists for all $\mathbf{x}$ and θ and interchange of differentiation and integration is allowed then

$$I_{\mathbf{X}}(\theta) = -\mathbf{E}_\theta\left[\frac{\partial^2}{\partial\theta^2}\,\log p(\mathbf{X} : \theta)\right].$$

The following result holds for exponential families. It is given in Theorem 5.4 of Lehmann and Casella (1998). You can try the proof.

Lemma 17. *Suppose X is from an exponential family $\{p(x : \theta) : \theta \in \Theta\}$, with pdf (or pmf),*

$$p(x : \theta) = h(x)\exp\left[\eta(\theta)T(x) - \psi(\theta)\right],$$

where Θ is an open subset of $\mathbb{R}$. Then,

(a) The CRLB regularity conditions are satisfied.

(b) Let, $\tau(\theta) = \mathbf{E}_\theta(T)$. Then, $\mathbf{Var}_\theta(T) = 1/I_{\mathbf{X}}(\tau(\theta))$.

In this context it is important to note that the Fisher information $I(\theta)$ is the information that X contains about θ. But, if we choose a different parametrization, with $\theta = h(\xi)$, where h is differentiable, then the information that X contains about ξ is

$$I_*(\xi) = I(h(\xi))\big(h'(\xi)\big)^2, \tag{3.38}$$

where $I(\cdot)$ is the Fisher information function in terms of θ and $I_*(\cdot)$ is the Fisher information in terms of ξ. Similarly if θ is vector valued and $\theta = h(\xi)$, and h is differentiable, then the Fisher information that $\mathbf{X}$ contains about ξ will be,

$$\left(\frac{\partial}{\partial\xi}\,h(\xi)\right)I(h(\xi))\left(\frac{\partial}{\partial\xi}\,h(\xi)\right)^{T}.$$

Example 3.13. Let $X \sim$ Poisson (λ), $\lambda > 0$. Then the Fisher information in terms of λ will be ,

$$I(\lambda) = \mathbf{E}_\lambda\left(\frac{X}{\lambda} - 1\right)^2 = \frac{1}{\lambda}. \tag{3.39}$$

Consider $\xi = \sqrt{\lambda}$, a different parametrization. Then, $\lambda = h(\xi) = \xi^2$. In terms of this parameter, the Poisson pmf will be:

$$\mathbf{P}_\xi(X = x) = \exp(-\xi^2)\frac{\xi^{2x}}{x!}, \quad x = 0, 1, 2, \ldots, \quad \xi > 0.$$

Then we can directly compute the Fisher information in terms of ξ, by reparametrizing the Poisson r.v. in terms of ξ:

$$\begin{aligned} I_*(\xi) &= \mathbf{E}_\xi\left(\frac{\partial}{\partial\xi}\left(-\eta^2 + X\log\eta^2 - \log(X!)\right)\right)^2, \quad \text{(complete the details)},\\ &= 4\mathbf{E}_\xi\left(\frac{X}{\eta} - \eta\right)^2,\\ &= 4, \quad (\text{using } \mathbf{E}_\xi(X^2) = \lambda + \lambda^2 = \eta^2 + \eta^4 \text{ and } \mathbf{E}_\xi(X) = \lambda = \eta^2). \end{aligned}$$

On the other hand, if we use (3.38) and (3.39), we find

$$I_*(\xi) = \frac{1}{h(\xi)}\big(h'(\xi)\big)^2 = \frac{1}{\xi^2}(4\xi^2) = 4,$$

which matches with the direct calculation done above by reparametrizing the Poisson distribution in terms of ξ. This shows how the information contained about different parameters changes, with a constant information $(= 4)$ available for $\sqrt{\lambda}$, compared to $1/\lambda$, available for λ.

Example 3.14. Suppose $\{X_1, \ldots, X_n\}$ are i.i.d. Poisson(λ) with $\lambda > 0$. We are interested in estimating $\tau(\lambda) = \lambda$. Find the CRLB for any UE of this estimand, and then show that $\bar{X}_n$ is the UMVUE.

4 Decision Theory

We have investigated unbiased minimum variance estimators, where the performance of an estimator is assessed by the MSE criteria. Moreover, we require that such (unbiased) estimators attain the minimum variance at each underlying parameter point. Both of these requirements can be changed and a much broader and abstract viewpoint can be constructed to handle various statistical inference problems. The subject of statistical decision theory provides results on these lines. The initial discussion will be abstract, but we will provide examples.

We assume that a random vector $\mathbf{X} \sim f(\mathbf{x} : \boldsymbol{\theta})$, where $\boldsymbol{\theta} \in \Theta$ is the parameter space and $f(\mathbf{x} : \boldsymbol{\theta})$ denotes the underlying pdf of pmf. The true $\boldsymbol{\theta}$ which generates the random vector will be unknown to the statistician, and we say $\boldsymbol{\theta}$ is the underlying state of *nature* and Θ is the set of all possible states of *nature.* A statistician observes the data $\mathbf{X}$ and has to decide on an *action* (an abstract term which will be explained soon) $d(\mathbf{X})$, where d is a map from the underlying sample space $\mathcal{X}$ to the *action space* $\mathcal{A}$. The *action* is chosen in such a manner, so that a pre-defined *loss* is minimized. The *loss* is measured through a *loss function* $L(\boldsymbol{\theta}, \mathbf{a})$, where $\boldsymbol{\theta} \in \Theta$ and $\mathbf{a} \in \mathcal{A}$. So, $L : \Theta \times \mathcal{A} \mapsto [0, \infty)$ is loss function, taking non-negative values. Different choices of loss functions lead to different ways of choosing optimum actions.

As the statistician observes a random $\mathbf{X}$, his action will be a random quantity $d(\mathbf{X}) \in \mathcal{A}$, and as a result the loss will be $L(\boldsymbol{\theta}, d(\mathbf{X}))$, which will again be a random quantity. Over repeated realizations of $\mathbf{X}$, the *expected loss* of the statistician (if he continues to use the action (mapping) $d : \mathcal{X} \mapsto \mathcal{A}$) will be

$$R(d, \boldsymbol{\theta}) = \mathbf{E}_{\boldsymbol{\theta}} L(\boldsymbol{\theta}, d(\mathbf{X})),$$

which is known as the *risk* of the decision rule d at the parameter $\boldsymbol{\theta}$.

Example 4.1. Assume $\Theta = \mathbb{R}$ and θ is a real-valued parameter. We have a real-valued estimator $d(\mathbf{X})$ (the range of d is the entire real line) for estimating θ and we measure the performance (loss) of d by the *squared-error* loss,

$$L(\theta, a) = (\theta - a)^2, \quad \text{for all } \theta \in \mathbb{R} \text{ and } a \in \mathbb{R}.$$

Note, as the estimator is real-valued, the *action space* $\mathcal{A} = \mathbb{R}$. So, in this case the loss of d will be $L(\theta, d(\mathbf{X})) = (\theta - d(\mathbf{X}))^2$, which will be a random quantity. The risk of d under squared error loss will be $R(d, \theta) = \mathbf{E}_\theta L(\theta, d(\mathbf{X})) = \mathbf{E}_\theta(\theta - d(\mathbf{X}))^2 = \mathrm{MSE}_\theta(d(\mathbf{X}))$.

If we used the absolute error loss, then $L(\theta, a) = |\theta - a|$, and we will obtain $R(d, \theta) = \mathbf{E}_\theta |d(\mathbf{X}) - \theta|$. Similarly one can use the truncated squared error loss function, $L(\theta, a) = \min\{K, (\theta - a)^2\}$, where $K > 0$ is a pre-chosen value.

If we decided to use the loss, $L(\theta, a) = \mathbf{1}(|\theta - a| \geqslant K)$, where $K > 0$ is pre-chosen, then we obtain the risk,

$$R(d, \theta) = \mathbf{E}_\theta \mathbf{1}(|\theta - d(\mathbf{X})| \geqslant K) = \mathbf{P}_\theta(|d(\mathbf{X}) - \theta| \geqslant K),$$

which measures the probability of $|d(\mathbf{X}) - \theta|$ exceeding K. Minimizing this loss function will ensure control of the tail probability of $|d(\mathbf{X}) - \theta|$, which is a very different goal from minimizing the MSE. One can also use an assymmetric loss function, $L(\theta, a) = \mathbf{1}(\theta \leqslant a)$, which will provide control over $R(d, \theta) = \mathbf{P}_\theta(d(\mathbf{X}) - \theta \geqslant 0)$.

Many other examples can be constructed, but one needs to have a clear reason for choosing a particular loss function.

Example 4.2. In case $\boldsymbol{\theta}$ and $\mathbf{a}$ are k-dimensional parameter and action, then we can think of various loss functions: $L(\boldsymbol{\theta}, \mathbf{a}) = \|\boldsymbol{\theta} - \mathbf{a}\|^2$ (usual Euclidean k-norm), or $L(\boldsymbol{\theta}, \mathbf{a}) = \sum_{i=1}^k |\theta_i - a_i| = \|\boldsymbol{\theta} - \mathbf{a}\|_1$ (ℓ_1-norm in $\mathbb{R}^k$), or $L(\boldsymbol{\theta}, \mathbf{a}) = \max_{1 \leqslant i \leqslant k} |\theta_i - a_i|$ (max-norm), and probably many other choices are possible.

Example 4.3. We are interested in estimating an unknown pdf f over its entire support $\mathcal{X}$. Then, $\theta = f(\cdot)$ is a function-valued parameter. In this case we can consider the integrated squared-error loss, $L(\theta, a) = \int_{\mathcal{X}} (f(u) - a(u))^2 \, du$, where a is a function-valued action, as this will measure the overall accuracy over the entire support. The risk of any estimator $\widehat{f}_n(\cdot)$ under this loss will be,

$$R(\widehat{f}_n, f) = \mathbf{E}_f L(\widehat{f}_n, f) = \mathbf{E}_f \int_{\mathcal{X}} \left(\widehat{f}_n(u) - f(u)\right)^2 du.$$

Example 4.4 (Hypothesis testing). In hypothesis testing, we are interested in comparing two statements about the underlying parameter, instead of estimating it. Suppose the null hypothesis is $H_0 : \theta \in \Theta_0$ and the alternative hypothesis is $H_1 : \theta \in \Theta_1$, where Θ_i, $i = 0, 1$, are subsets of the parameter space (and $\Theta_1 = \Theta \setminus \Theta_0$). The action space is $\mathcal{A} = \{a_0, a_1\}$, where a_i denotes the accepting H_i, for $i = 0, 1$. We can introduce a 0 - 1 loss function of the form,

$$L(\theta, a_0) = \begin{cases} 0 & \text{if } \theta \in \Theta_0, \\ 1 & \text{if } \theta \in \Theta_1. \end{cases} \quad \text{and} \quad L(\theta, a_1) = \begin{cases} 1 & \text{if } \theta \in \Theta_0, \\ 0 & \text{if } \theta \in \Theta_1. \end{cases}$$

In this case, we are assuming that there are two underlying states of *nature* (the parameter θ), *viz.*, either $\theta \in \Theta_0$ or $\theta \in \Theta_1$. Correspondingly, there are two possible actions: accept (choose) either of these two states on the basis of data. Any decision rule d will be a map, $d : \mathcal{X} \mapsto \{a_0, a_1\}$. A *critical region* R for a hypothesis testing problem is the sub-part of the sample space where d maps to a_1 (hence rejecting H_0), *i.e.*,

$$d(\mathbf{x}) = \begin{cases} a_0 & \text{if } \mathbf{x} \in R^c, \\ a_1 & \text{if } \mathbf{x} \in R. \end{cases} \quad \text{for all } \mathbf{x} \in \mathcal{X}.$$

The risk of d under the zero-one loss will be,

$$\mathbf{E}_\theta L(\theta, d(\mathbf{X})) = \begin{cases} \mathbf{P}_\theta(\mathbf{X} \in R) & \text{if } \theta \in \Theta_0, \\ \mathbf{P}_\theta(\mathbf{X} \in R^c) & \text{if } \theta \in \Theta_1. \end{cases}$$
$$= \begin{cases} \mathbf{P}_\theta(\text{type I error}) & \text{if } \theta \in \Theta_0, \\ \mathbf{P}_\theta(\text{type II error}) & \text{if } \theta \in \Theta_1. \end{cases}$$

One can possibly use other loss functions, but it will result in a different risk function.

Example 4.5 (Interval estimation). Assume $\theta \in \mathbb{R}$ is a parameter and we are interested in an a two-sided confidence interval (CI) for θ. A two-sided CI is a random set $d(\mathbf{X}) = (L(\mathbf{X}), U(\mathbf{X}))$, so that it contains θ with a pre-specified probability. The action space $\mathcal{A} = \{(a, b) : -\infty < a < b < \infty\}$. If we have a loss function $L(\theta, (a, b))$, then the risk will be $\mathbf{E}_\theta L(d(\mathbf{X}), \theta)$. For example, we can consider the loss function,

$$L(\theta, (a, b)) = L_0(a - \theta) + L_1(\theta - b) + k(b - a),$$

where L_0 and L_1 are monotone non-decreasing functions with $L_0(u) = L_1(u) = 0$, for all $u \leqslant 0$, and k is a non-negative constant. Good choices of L_0 and L_1 include convex maps (or even linear maps). Using this approach one can control various optimality properties (length, coverage) of the desired optimal CI. See Winkler (1972) for detailed discussion.

It should be noted that a strictly decision theoretical approach is usually avoided for interval estimation, but the above discussion suggests that we can frame interval estimation in a decision theoretic framework.

Decision theory can be very flexible, and can handle several other types of problems, including *no-data* problems (we will see examples) where an action is chosen, purely on the basis of the given loss function. The next example does not fit any of the classical inference setup's discussed so far.

Example 4.6. Assume that $\Theta = \{\theta_1, \theta_2\}$ denotes the possible states of *nature* and the decision maker can choose an action from $\mathcal{A} = \{a_1, a_2, a_3\}$ (the action space). The following loss matrix (in Table 1) is provided:

Table 1: Values of $L(\theta, a)$

	Action		
Parameter	a_1	a_2	a_3
θ_1	0	1	2
θ_2	2	0	1

Also, we are told that X is a random variable with p.m.f. $f(x : \theta)$, whose values are given in Table 2 The first question is: how many possible decision rules $d(x)$ exist? Recall, d is a map from the underlying sample space of X to the action space, *i.e.*, $d : \mathcal{X} \mapsto \mathcal{A}$. Hence, there exists $3 \times 3 = 9$ possible decision rules.

Table 2: Values of $f(x:\theta)$

Parameter	$x = 0$	$x = 1$
θ_1	p_1	$1 - p_1$
θ_2	p_2	$1 - p_2$

They are:

$$\begin{array}{lll} d_1:\ d_1(0) = a_1,\ d_1(1) = a_1, & d_2:\ d_2(0) = a_1,\ d_2(1) = a_2, & d_3:\ d_3(0) = a_1,\ d_3(1) = a_3, \\ d_4:\ d_4(0) = a_2,\ d_4(1) = a_1, & d_5:\ d_5(0) = a_2,\ d_5(1) = a_2, & d_6:\ d_6(0) = a_2,\ d_6(1) = a_3, \\ d_7:\ d_7(0) = a_3,\ d_7(1) = a_1, & d_8:\ d_8(0) = a_3,\ d_8(1) = a_2, & d_9:\ d_9(0) = a_3,\ d_8(1) = a_3. \end{array}$$

In general, if $\mathcal{X}$ has m elements and $\mathcal{A}$ has k elements, then there will be k^m possible decision rules. The risk of d_1 will be,

$$\begin{aligned} R(d_1, \theta) = \mathbf{E}_\theta L(\theta, d_1(X)) &= \sum_x L(\theta, d_1(x)) \cdot \mathbf{P}_\theta(X = x) \\ &= L(\theta, d_1(0)) \cdot \mathbf{P}_\theta(X = 0) + L(\theta, d_1(1)) \cdot \mathbf{P}_\theta(X = 1) \\ &= L(\theta, a_1) \cdot \mathbf{P}_\theta(X = 0) + L(\theta, a_1) \cdot \mathbf{P}_\theta(X = 1) \\ &= \begin{cases} L(\theta_1, a_1) \cdot \mathbf{P}_{\theta_1}(X = 0) + L(\theta_1, a_1) \cdot \mathbf{P}_{\theta_1}(X = 1) & \text{if } \theta = \theta_1, \\ L(\theta_2, a_1) \cdot \mathbf{P}_{\theta_2}(X = 0) + L(\theta_2, a_1) \cdot \mathbf{P}_{\theta_2}(X = 1) & \text{if } \theta = \theta_2, \end{cases} \\ &= \begin{cases} 0 \cdot p_1 + 0 \cdot (1 - p_1) & \text{if } \theta = \theta_1, \\ 2 \cdot p_2 + 2 \cdot (1 - p_2) & \text{if } \theta = \theta_2, \end{cases} \\ &= \begin{cases} 0 & \text{if } \theta = \theta_1, \\ 2 & \text{if } \theta = \theta_2. \end{cases} \end{aligned}$$

The risk of d_1 can be depicted as a vector, $(R(d_1, \theta_1), R(d_1, \theta_2)) = (0, 2)$. Similarly, the risk of all other decision rules can be found.

Example 4.7. If we consider Example 4.6 and for simplicity of calculation, we change the action space to $\mathcal{A} = \{a_1, a_2\}$. Then, we can have only four decision rules: d_1, d_2, d_3 and d_4 (given above). Now, let us fix $p_1 = 1/4$ and $p_2 = 1/2$. The risk of these four rules will be,

$$(R(d, \theta_1), R(d, \theta_2)) = \begin{cases} (0,\ 2) & \text{if } d = d_1, \\ (1 - p_1,\ 2p_2) = (3/4,\ 1) & \text{if } d = d_2, \\ (p_1,\ 2(1 - p_2)) = (1/4,\ 1) & \text{if } d = d_3, \\ (1,\ 0) & \text{if } d = d_4. \end{cases}$$

The *risk-set* for this decision problem is given in Figure 1. The four vertices of the shaded polygon represent the risk points of the four decision rules. The shaded area represents risks of *randomized* decision rules, which can be constructed by applying a probability distribution on the *non-randomized* rules $\{d_1, d_2, d_3, d_4\}$.

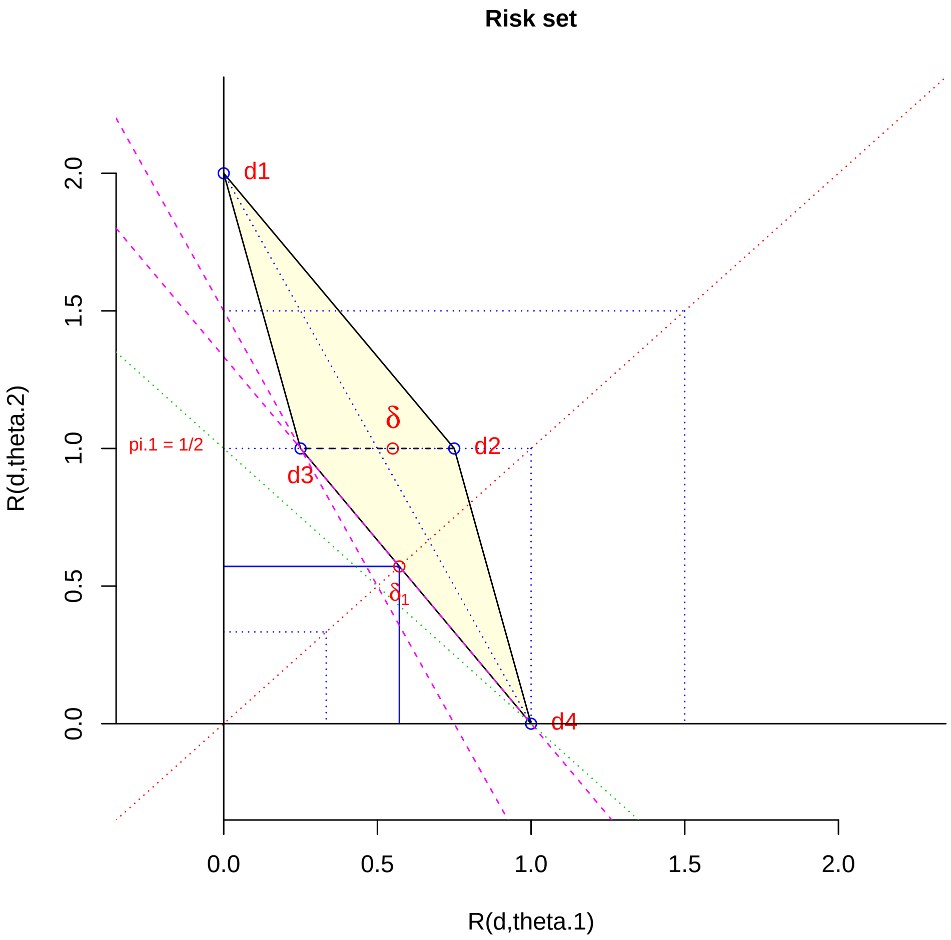

Figure 1: Risk set of the decision problem in Example 4.7.

A decision rule d is a map from $\mathcal{X}$ to $\mathcal{A}$. In this approach, once a decision rule d is chosen, and any realization $\mathbf{x}$ is observed, it will map to a fixed action a (depending on what is the value of $d(\mathbf{x})$). However, instead of always choosing a fixed action, one can decide to randomly pick an action, even if the same $\mathbf{x}$ is observed. Such a decision rule is called a *randomized* decision rule (and typically the notation $\delta(\mathbf{x})$ is used to denote them). The collection of all randomized decision rules is based on a probability distribution on the set of non-randomized rules. In case of Example 4.7, we have four non-randomized rules. Hence any randomized rule δ can be written as

$$\delta(\mathbf{x}) = \sum_{i=1}^{4} \pi_i \cdot d_i(\mathbf{x}), \quad \text{where } 0 \leqslant \pi_i \leqslant 1 \text{ and } \sum_{i=1}^{4} \pi_i = 1.$$

The above relation implies, $\delta = d_i$, w.p. π_i, for each $i = 1, \ldots, 4$. Correspondingly the risk of a randomized decision rule will be,

$$R(\delta,\theta) = \mathbf{E}_\theta L(\delta(\mathbf{X}),\theta) = \sum_{\mathbf{x}} L(\delta(\mathbf{x}),\theta)\cdot \mathbf{P}_\theta(\mathbf{X}=\mathbf{x}) = \sum_{\mathbf{x}} \left[\sum_i \pi_i \cdot L(d_i(\mathbf{x}),\theta)\right]\cdot \mathbf{P}_\theta(\mathbf{X}=\mathbf{x}) = \sum_i \pi_i \cdot R(d_i,\theta).$$

As we go over all possible probability distributions $\{\pi_i\}$, the risk points of all (randomized) decision rules will be same as the convex hull of the risk points of all non-randomized rules. This argument is clearly seen in Figure 1, where the shaded yellow region represents the convex hull of the four risk points for non-randomized rules. In particular, any non-randomized rule can also be considered as a special case of a randomized rule.

In Figure 1, the point $(11/20, 1)$ (drawn with a red-circle) corresponds to the risk of a randomized decision rule $\delta(\mathbf{x})$, where

$$\delta = (1/5)\cdot d_1 + (2/5)\cdot d_2 + (1/5)\cdot d_3 + (1/5)\cdot d_4.$$

However, this risk point also lies on the line joining the risk points of d_3 and d_2, implying that δ could have an alternate representation in terms of d_2 and d_3. It should be noted that Figure 1 shows the risk at θ_1 and θ_2, on each of its two axes. When the parameter space contains more points, the risk set will be a higher-dimensional region, and in case the parameter space is infinite, then such plots can not be drawn.

Example 4.8 (Comparing decision rules). Assume $X \sim N(\theta,1)$ and $L(\theta,a) = (\theta-a)^2$, where $a \in \mathbb{R}$. Consider the following class of decision rules for estimating θ,

$$\mathcal{D} = \{d_c(x) = cx : c \geqslant 0\}.$$

The main question is, which rule from $\mathcal{D}$ should be preferred? Note,

$$R(d_c,\theta) = \mathbf{E}_\theta(cX-\theta)^2 = \mathbf{E}_\theta[cX - c\theta + c\theta - \theta]^2 = c^2 + (1-c)^2\theta^2, \text{ for all } \theta.$$

Note, $R(d_1,\theta) = 1$, for all θ, and has a constant risk over the entire parameter space. This implies, if $c > 1$, then

$$R(d_c,\theta) \geqslant c^2 > 1, \quad \text{for all } \theta.$$

So, any rule from $\mathcal{D}$, with $c > 1$ can be discarded from our consideration, as d_c will be *uniformly* worse than d_1. However, if we compare the risks of d_1, $d_{1/2}$ and $d_{1/4}$, as seen in Figure 2, then none of three rules are uniformly better than another on the entire parameter space. Hence, comparison of decision rules can be done *pointwise* (at a fixed parameter point), or over the entire parameter space. The UMVUE's studied earlier, had uniformly smaller risk over the entire parameter space, when restricted to the class of all unbiased estimators.

This brings to us the question of how to compare decision rules over the entire parameter space? If d_1 and d_2 are two decision rules, such that

$$R(d_1,\theta) \leqslant R(d_2,\theta) \quad \text{for all } \theta \in \Theta, \text{ and} \quad R(d_1,\theta_1) < R(d_2,\theta_1), \quad \text{for some } \theta_1 \in \Theta,$$

then we say d_1 is better than d_2. A decision rule d is called *admissible* if there does not exist any other rule that is better than d. If there exists a rule that is better than d, then d is called *inadmissible*. This means, a rule d is inadmissible if there exists another rule which has lower risk than d at all parameter points. On

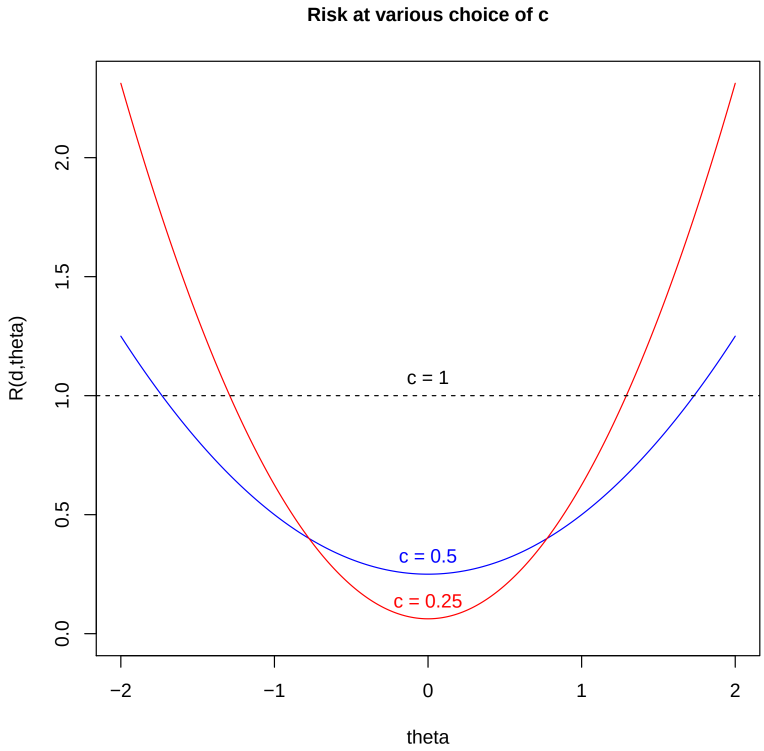

Figure 2: Risk plots for the rule d_c at various choices of c.

the other hand, if a rule is admissible, that *does not* imply it has the lowest risk in comparison to all other rules.

If we see the risk plots in Figure 2, then the rule d_c (for any $c > 1$) will be inadmissible, as there exists another rule (d_1), which is better than d_c. But, the each of the rules d_c for $0 < c \leqslant 1$, will be admissible (verify this). Checking the admissibility of a rule involves comparing the risks over the entire parameter space. If we consider the risk set in Figure 1, the entire shaded area (including the boundary) contains the risk points for all possible decision rules for that specific problem in Example 4.7. Which are the risk points, for which the corresponding decision rules are admissible?

4.1 Bayes and Minimax rules

There are various ways of selecting a decision rule. As in the case of Example 4.8, we may wish restrict to rules which are unbiased. In that case, the only choice is $c = 1$. Or, we may use criteria like invariance or equivariance (see Chapter 3 of Lehmann and Casella (1998)). In case of hypothesis testing or interval estimation, we may want to restrict to unbiased tests or confidence intervals. The task then reduces to finding the best rule within that sub-class. The alternate approach is to compare rules by studying their overall behavior on the entire parameter space.

The Bayes and Minimax approaches are two methods that achieve an overall comparison of decision rules. Assume $\{\pi(\boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta\}$ is a probability distribution on the parameter space Θ. This can be also be viewed as a weight function or a *priority* function, as set by the statistician. It can also signify the underlying *belief* of the statistician about the unknown parameter or also describe any *prior* information about $\boldsymbol{\theta}$, which is available from other sources. This distribution is also known as a *prior* distribution on the underlying parameter space. The choice of π is very crucial and involves a lot of details (see Robert (2007)), which we can not cover in this course.

The *Bayes risk* of d with respect to the prior π is defined as,

$$r(\pi, d) = \mathbf{E}_{\boldsymbol{\theta}\sim\pi} R(d, \boldsymbol{\theta}) = \mathbf{E}_{\boldsymbol{\theta}\sim\pi}\left[\mathbf{E}_{\mathbf{X}|\boldsymbol{\theta}} L(\boldsymbol{\theta}, d(\mathbf{X}))\right] = \begin{cases} \int_{\boldsymbol{\theta}} R(d, \boldsymbol{\theta}) \cdot \pi(\boldsymbol{\theta})\ d\boldsymbol{\theta} & \text{if } \pi \text{ is a p.d.f.} \\ \sum_{\boldsymbol{\theta}\in\Theta} R(d, \boldsymbol{\theta}) \cdot \pi(\boldsymbol{\theta}) & \text{if } \pi \text{ is a p.m.f.} \end{cases}$$

Example 4.9. Continuing from Example 4.8, assume we have the prio $\pi(\theta) \sim N(0, \tau^2)$, where $\tau > 0$ is a known[15] variance. Note θ is a real valued parameter, and $\pi(\theta)$ is a Normal distribution (supported on $\mathbb{R}$). Then, for $d_c(x) = cx$, we had found $R(d_c, \theta) = \theta^2 + c^2(1-\theta)^2$. Thus, the Bayes risk under this Normal prior will be

$$r(\pi, d_c) = \mathbf{E}_{\pi(\theta)}\left[c^2 + c^2(1-\theta)^2\right] = c^2 + (1-c)^2\tau^2.$$

As we see, the Bayes risk is free of θ and provides a summary measure of performance of d_c. Then, the Bayes risk is minimized at $c = \tau^2/(1+\tau^2)$, implying $d(x) = \left(\tau^2/(1+\tau^2)\right) \cdot x$ is the Bayes rule under this prior. The choice of c shows that $c \to 0$, when $\tau \downarrow 0$, and when $\tau \uparrow \infty$, then $c \to 1$, indicating that $d_1(x) = x$ is the preferred rule, when all values of θ are equally important. On the other hand, when values of θ close to zero are preferred (as $\tau \downarrow 0$) then a rule with c close to zero will be important. The choice will change if we consider any other prior. Hence, the choice of prior is extremely important from a Bayesian perspective.

At this point, it is important to introduce the *posterior distribution* of a parameter $\boldsymbol{\theta}$. As we are imposing a prior probability on $\boldsymbol{\theta}$, we can consider $\boldsymbol{\theta}$ as a *random vector* and *view* the distribution of $\mathbf{X}$ as the conditional distribution of $[\mathbf{X} \mid \boldsymbol{\theta}]$. Hence

$$\begin{aligned} \boldsymbol{\theta} &\sim \pi(\boldsymbol{\theta}) \quad \text{(prior distribution)} \\ [\mathbf{X} \mid \boldsymbol{\theta}] &\sim f(\mathbf{x} \mid \boldsymbol{\theta}) \quad \text{(viewed as conditional pdf/pmf of } \mathbf{X} \text{ given } \boldsymbol{\theta}), \\ (\mathbf{X}, \boldsymbol{\theta}) &\sim f(\mathbf{x}, \boldsymbol{\theta}) = f(\mathbf{x} \mid \boldsymbol{\theta}) \cdot \pi(\boldsymbol{\theta}) \quad \text{(as a joint pdf/pmf)}, \\ [\boldsymbol{\theta} \mid \mathbf{X}] &\sim \frac{f(\mathbf{x}, \boldsymbol{\theta})}{m(\mathbf{x})} = \frac{f(\mathbf{x} \mid \boldsymbol{\theta}) \cdot \pi(\boldsymbol{\theta})}{\int_\Theta f(\mathbf{x}, \mathbf{u})\ d\mathbf{u}} = \frac{f(\mathbf{x} \mid \boldsymbol{\theta}) \cdot \pi(\boldsymbol{\theta})}{\int_\Theta f(\mathbf{x} \mid \mathbf{u}) \cdot \pi(\mathbf{u})\ d\mathbf{u}} \quad \text{(using Bayes theorem)}, \end{aligned}$$

where the last line displays the *posterior density* (pdf/pmf) of $\boldsymbol{\theta}$ conditional on $\mathbf{X}$. The posterior density depicts the updated *belief* about $\boldsymbol{\theta}$, once the data has been observed.

The Bayesian perspective differs in terms of treating the underlying parameter as quantity which is not *fixed*, but allows the flexibility of incorporating any prior knowledge (from earlier datasets, expert opinion, etc.), and also allows development of very complex models, where $\boldsymbol{\theta}$ itself may depend on many other factors.

[15] This is itself another parameter, that is used to specify the prior distribution of the original parameter θ. Such parameters are known as *hyper-parameters* and they can also be unknown in practice.

The viewpoint (fixed $\boldsymbol{\theta}$ or flexible $\boldsymbol{\theta}$) can change in many cases depending on the situation and data in hand. In many instances, a Bayesian approach is more practicable.

Example 4.10. Continuing with Example 4.7, assume we have the prior: $\pi(\theta_1) = 1/3$ and $\pi(\theta_2) = 2/3$. Then, for the four non-randomized rules, the Bayes risk will be (check!)

$$r(\pi, d_1) = 16/12,\ r(\pi, d_2) = 11/12,\ r(\pi, d_3) = 9/12,\ \text{and } r(\pi, d_4) = 4/12.$$

Thus, with this prior, d_4 will be the Bayes rule. But, what if we consider all randomized decision rules? Will d_4 still be the (best) Bayes rule among all rules?

Now we focus on *minimax* rules. The minimax criteria selects a decision rule which minimizes the worst (max) possible risk. The approach is very conservative, assuming that the statistician will encounter the worst possible risk and protects against this worst possible scenario. If we compare in terms of the worst risk, then d_1 is better than d_2 if $\sup_{\boldsymbol{\theta}} R(\boldsymbol{\theta}, d_1) < \sup_{\boldsymbol{\theta}} R(\boldsymbol{\theta}, d_2)$. If there exists a non-randomized rule $d_\star$ such that

$$\sup_{\boldsymbol{\theta}} R(\boldsymbol{\theta}, \delta_\star) = \inf_{d \in \mathcal{D}} \sup_{\boldsymbol{\theta}} R(\boldsymbol{\theta}, d),$$

where $\mathcal{D}$ is the class of all non-randomized rules for a decision problem, then $d_\star$ is called a minimax rule.

If we consider Example 4.8 then, $R(d_c, \theta) = c^2 + (1-c)^2\theta^2$. Hence,

$$\sup_{\theta \in \mathbb{R}} R(\theta, d_c) = \begin{cases} 1 & \text{if } c = 1, \\ \infty & \text{if } c \neq 1. \end{cases}$$

So d_1 turns out to be the minimax rule in this case.

On the other hand, in Example 4.7, we obtain

$$\sup_{\theta \in \{\theta_1, \theta_2\}} R(\theta, d_i) = \begin{cases} 2 & \text{if } i = 1, \\ 1 & \text{if } i = 2, \\ 1 & \text{if } i = 3, \\ 1 & \text{if } i = 4. \end{cases}$$

Hence, each of the three non-randomized rules d_2, d_3 and d_4 minimize the maximum risk, and all three rules are minimax rules.

4.1.1 Randomized Bayes and minimax rules

Assume that Θ is a finite parameter space containing k parameter points $\{\theta_1, \ldots, \theta_k\}$. The risk set will be a convex subset of $\mathbb{R}^k$ and each point within this risk set will correspond to the risk value of some randomized decision rule. Once a prior distribution π with probabilities $\pi(\theta_i) = \pi_i$ is supplied, then the Bayes risk of any randomized decision rule δ will be, $r(\pi, \delta) = \sum_{i=1}^k \pi_i R(\delta, \theta_i)$.

The lowest Bayes risk is achieved, if there exists a rule δ_π such that

$$r(\pi, \delta_\pi) = \inf_{\delta \in \mathcal{D}} r(\pi, \delta),$$

where $\mathcal{D}$ denotes the class of all randomized rules for the decision problem. In some cases, there may be an unique Bayes rule. In some cases, if the risk set is not closed, then the infimum may not be attained at any rule. In case there are more than one Bayes rules, then there will exist infinitely many of them, expressible as the convex combination of these two rules.

We avoid a general discussion of Bayes rules in general parameter spaces. For more details one can refer to Berger (1993) or Ferguson (2014).

More material to be added.

5 Testing of hypothesis

In hypothesis testing, the primary goal is to assess and select among two competing statements about the underlying parameter. These statements are called the null and alternative hypothesis. Usually the null hypothesis is chosen in such a way such that the resulting testing procedure has nice properties or perhaps it makes technical calculations easier. As seen earlier, the hypothesis testing problem can be formulated in a decision theoretic framework.

Assume $\mathbf{X} \sim f(\mathbf{x} : \boldsymbol{\theta})$ where f denotes the underlying pdf/pmf of $\mathbf{X}$ and $\boldsymbol{\theta}$ denotes the underlying parameter. This parameter can also be infinite dimensional, but we will focus on the finite dimensional case. The null hypothesis is written as $H_0 : \boldsymbol{\theta} \in \Theta_0$ where $\Theta_0 \subset \Theta$ and Θ is the parameter space. The alternative hypothesis is denoted by $H_1 : \boldsymbol{\theta} \in \Theta_1$ where $\Theta_1 = \Theta \backslash \Theta_0$. Obviously, in order to avoid ambiguity we will assume that $\Theta_0 \cap \Theta_1 = \varnothing$. We will describe the testing problem as, test of $H_0 : \boldsymbol{\theta} \in \Theta_0$ vs $H_1 : \boldsymbol{\theta} \in \Theta_1$. In case Θ_i $(i = 0, 1)$ specifies a *single* distribution, then H_i is called a *simple* hypothesis, and if it specifies more than distributions (of $\mathbf{X}$), then it is called a *composite* hypothesis.

For example consider $X \sim N(\theta, 1)$ and θ is real valued. Consider the null hypothesis $H_0 : \theta = 0$ against the alternative hypothesis $H_1 : \theta = 1$. In this case, although the usual parameter space for a Normal mean is $\mathbb{R}$, but for purposes of the testing problem we will consider $\Theta = \Theta_0 \cup \Theta_1 = \{0, 1\}$. In this cases, both H_0 and H_1 are simple hypothesis. If we consider $H_1 : \theta > 0$, then it will be a composite hypothesis. Both null and alternative can be either simple or composite hypothesis.

Based on the data $\mathbf{x}$, we construct a rule which specifies a critical region $S \subset \mathcal{X}$, such that if $\mathbf{x} \in S$, then we reject H_0, and if $\mathbf{x} \notin S$, then either we accept H_0, or make a randomized decision. The decision to accept/reject H_0 can be error-prone, as the underlying true state is unknown[16]. There are two types of errors that are possible: (a) Type-I error: we reject H_0 when H_0 is true, and (b) Type-II error: we reject H_1 when H_1 is true. As a statistician we would like devise a testing rule (critical region) such both error probabilities are minimized. However, it can be shown that it is not possible to minimize both error probabilities simultaneously.

A test function $\phi : \mathcal{X} \mapsto [0, 1]$ is a (measurable) map. Infact $\phi(\mathbf{x})$ denotes the probability of rejecting H_0 when $\mathbf{x}$ is observed. Typical examples would be, $\phi(\mathbf{x}) = \mathbf{1}(\mathbf{x} \in S)$, which corresponds to always rejecting H_0 when $\mathbf{x} \in S$ and always accepting H_0 if $\mathbf{x} \notin S$. We can also have randomized tests of the form,

$$\phi(\mathbf{x}) = \begin{cases} 1 & \text{if } \mathbf{x} \in S_1, \\ \gamma(\mathbf{x}) & \text{if } \mathbf{x} \in S_2, \\ 0 & \text{o.w.,} \end{cases}$$

where $\mathcal{X} = S_1 \cup S_2 \cup S_3$ is disjoint partition. If $\mathbf{x} \in S_1$ (or S_3) we always reject (or accept), and if $\mathbf{x} \in S_2$ then we reject H_0 with changing probabilities $\gamma(\mathbf{x}) \in [0, 1]$. If $S_2 = \mathcal{X}$, then the above test reduces to a purely randomized test. The randomized test is visualized in this manner: assume for simplicity that $\gamma(\mathbf{x}) = \gamma \in (0, 1)$, then once $\mathbf{x}$ is observed a coin with probability of heads equal to γ is tossed, and if head

[16] We will assume that always either one of these two states H_0 or H_1 is true and there is no third possibility.

appears, then H_0 is rejected. Corresponding to any test function ϕ, we define its *power function* as

$$\beta_\phi(\boldsymbol{\theta}) = \mathbf{E}_{\boldsymbol{\theta}}\phi(\mathbf{X}), \quad \text{for any } \boldsymbol{\theta} \in \Theta. \tag{5.40}$$

The power function represents the expected probability of rejecting the null hypothesis when the true parameter is $\boldsymbol{\theta}$.

As it is not possible to minimize both types of error probabilities simultaneously, the standard approach is to control one of the error probabilities upto a maximum permissible level, typically the type-I error probability is controlled, and minimize the type-II error probability. Consider any testing problem $H_0 : \boldsymbol{\theta} \in \Theta_0$ against $H_1 : \boldsymbol{\theta} \in \Theta_1$ and any test function ϕ for this testing problem. If $\alpha \in [0, 1]$ is any pre-specified value (decided before the testing procedure), then in order to control the type-I error probability, we require ϕ to satisfy

$$\sup_{\boldsymbol{\theta} \in \Theta_0} \mathbf{E}_{\boldsymbol{\theta}}\phi(\mathbf{X}) \leqslant \alpha. \tag{5.41}$$

This ensures that the probability of type-I error never exceeds the pre-specified level α. Any test ϕ satisfying the upper bound in (5.41) is called a *level* α test. In some cases, the supremum may equal the upper bound α, while in some cases, the supremum may be strictly smaller. The value of the supremum on the left side of (5.41) is called the *size* of the test. Definitely any size α' test is a level α test, for any $\alpha' \leqslant \alpha$. When $\Theta_0 = \{\boldsymbol{\theta}_0\}$, consisting of a single value, then the size will be $\mathbf{E}_{\boldsymbol{\theta}_0}\phi(\mathbf{X})$.

The choice of α is very crucial, as it decides how much power we can achieve at the alternative. Typically small choices like $\alpha = 0.05$ or $\alpha = 0.01$ are used, but there is no reason to always use such choices. If α is too small, the power of the test will reduced, and if α is too large, it is desirable to decrease α as long as there is appreciable loss in power. If the experimenter has very strong conviction about the validity of H_0, then he might want to obtain extremely strong evidence against the null, before rejecting it. This can be achieved by setting a very low value of α, making the test very conservative. Setting a small α implies we are making the critical region a very low probability set (under the null) and outcomes from that set will be highly unlikely to occur.

Before discussing more results, we provide some important definitions.

Definition 6 (Most Powerful (MP) test)**.** Assume $\mathbf{X} \sim \mathbf{P}_{\boldsymbol{\theta}}$. Consider a simple versus simple testing problem $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ against $H_1 : \boldsymbol{\theta} = \boldsymbol{\theta}_1$, where $\boldsymbol{\theta}_0 \neq \boldsymbol{\theta}_1$. A test function ϕ is called the Most Powerful (MP) test of level α ($\in [0, 1]$) if:

(a) $\mathbf{E}_{\boldsymbol{\theta}_0}\phi(\mathbf{X}) \leqslant \alpha$.

(b) For any other test ϕ_1 satisfying (a) above, we have $\mathbf{E}_{\boldsymbol{\theta}_1}\phi(\mathbf{X}) \geqslant \mathbf{E}_{\boldsymbol{\theta}_1}\phi_1(\mathbf{X})$.

Hence, among all level α tests, the test ϕ has the highest power.

The notion of MP test is very important. Typically if a test ϕ is MP of level α, then the choice of ϕ will depend on both $\boldsymbol{\theta}_0$ and $\boldsymbol{\theta}_1$. At another alternative parameter, this same test may not remain MP. However, in case we are lucky, the same test may continue to remain MP at different alternative values. This leads to the definition of an UMP test.

Definition 7 (UMP test)**.** Assume $\mathbf{X} \sim \mathbf{P}_{\boldsymbol{\theta}}$ and we consider testing the composite hypotheses, $H_0 : \boldsymbol{\theta} \in \Theta_0$ against $H_1 : \boldsymbol{\theta} \in \Theta_1$. A test ϕ is said to be Uniformly Most Powerful (UMP) of level α for this testing problem, if

(a) $\sup_{\boldsymbol{\theta}\in\Theta_0} \mathbf{E}_{\boldsymbol{\theta}}\phi(\mathbf{X}) \leqslant \alpha$.

(b) For any other test ϕ satisfying (a) above, we have $\mathbf{E}_{\boldsymbol{\theta}}\phi(\mathbf{X}) \geqslant \mathbf{E}_{\boldsymbol{\theta}}\phi_1(\mathbf{X})$, for all $\boldsymbol{\theta} \in \Theta_1$.

A UMP level α test controls the type-I error on the null parameter space and maximizes power at all parameter points, among all other level α tests. It is uniformly MP, because the same test is MP at different alternative points.

This immediately implies that if ϕ is a level α UMP test for testing $H_0 : \boldsymbol{\theta} \in \Theta_0$ against $H_1 : \boldsymbol{\theta} \in \Theta_1$, then ϕ will be MP[17] for testing $H_0 : \boldsymbol{\theta} \in \Theta_0$ against $H_1' : \boldsymbol{\theta} = \boldsymbol{\theta}_1$, at any $\boldsymbol{\theta}_1 \in \Theta_1$.

5.1 The size-power diagram for a simple against simple testing problem

Consider any simple against simple testing problem. Suppose, we want to know what are the possible values of size $(= \alpha)$ and power $(= \beta)$, that are achievable for this testing problem. This is same as constructing the risk set in a decision problem.

More specifically, say $\mathbf{X} \sim \mathbf{P}$ and we wish to test $H_0 : \mathbf{P} = \mathbf{P}_0$ against $H_1 : \mathbf{P} = \mathbf{P}_1$. Definitely we can consider any constant test function $\phi_\alpha(\mathbf{x}) = \alpha$ (for all $\mathbf{x}$), which leads to (size, power) $= (\alpha, \alpha)$. At any fixed size α, the highest possible value of power will be attained by the MP test of size α. Also if ϕ is any test with[18] $(\mathbf{E}_0\phi, \mathbf{E}_1\phi) = (\alpha, \beta)$, then the test $\psi = 1 - \phi$, will lead us to $(\mathbf{E}_0\psi, \mathbf{E}_1\psi) = (1-\alpha, 1-\beta)$. So, the reflection of any point (α, β) w.r.t. the central point $(1/2, 1/2)$, will also belong to the size-power set. Moreover, if ϕ_1 and ϕ_2 are two tests with corresponding size-power values (α_i, β_i), $i = 1, 2$, then the test function

$$\xi \equiv c\phi_1 + (1-c)\phi_2,$$

will be a convex combination of these two, and all corresponding size-power values lying on the line joining (α_1, β_1) and (α_2, β_2) will also be inside this size-power set. This implies, this collection of points will be a convex subset of $[0,1] \times [0,1]$.

Example 5.1. If we consider the specific problem: $\{X_1, \ldots, X_n\}$ are i.i.d. $N(\mu, 1)$ and we want to test $H_0 : \mu = 0$ against $H_1 : \mu = \theta_1 (> 0)$. Then it can be shown that the MP test of size α will be, $\phi_\alpha(\mathbf{x}) = \mathbf{1}(\sqrt{n}\bar{x}_n > \tau_{1-\alpha})$, where $\tau_{1-\alpha}$ is the $(1-\alpha)$ quantile of standard normal distribution. The power of this MP test will be,

$$\mathbf{E}_{\theta_1}\phi_\alpha(\mathbf{X}) = \mathbf{P}_{\theta_1}(\sqrt{n}\bar{X}_n > \tau_{1-\alpha}) = 1 - \Phi(\tau_{1-\alpha} - \sqrt{n} \cdot \theta_1).$$

Figure 3 shows the size-power diagram for this testing problem at $\theta_1 = 0.5$, at various sample sizes n. The figure shows that for a fixed n, as size increases, the maximum power increases. However, if we allow n to

[17] Note, in this case H_0 is a composite null, but H_1' is a simple alternative and hence we can call it a MP test (instead of an UMP test). The uniform term is used because of the composite alternative hypothesis. So the MP test terminology is also applicable for a composite null against a simple alternative, with suitable modification for the definition of the level of the test.

[18] Here $\mathbf{E}_i$ is used to denote expectation w.r.t. $\mathbf{P}_i$, $i = 0, 1$.

increase, then for sufficiently large n, one can attain power close to 1, and make the size equal to zero (at least in this specific testing problem).

In Figure 4, the same plot is shown for the alternative value $\theta = 0.15$, which is much closer to the null parameter value. As seen in this plot, the power curve increases much slowly, as compared to Figure 3.

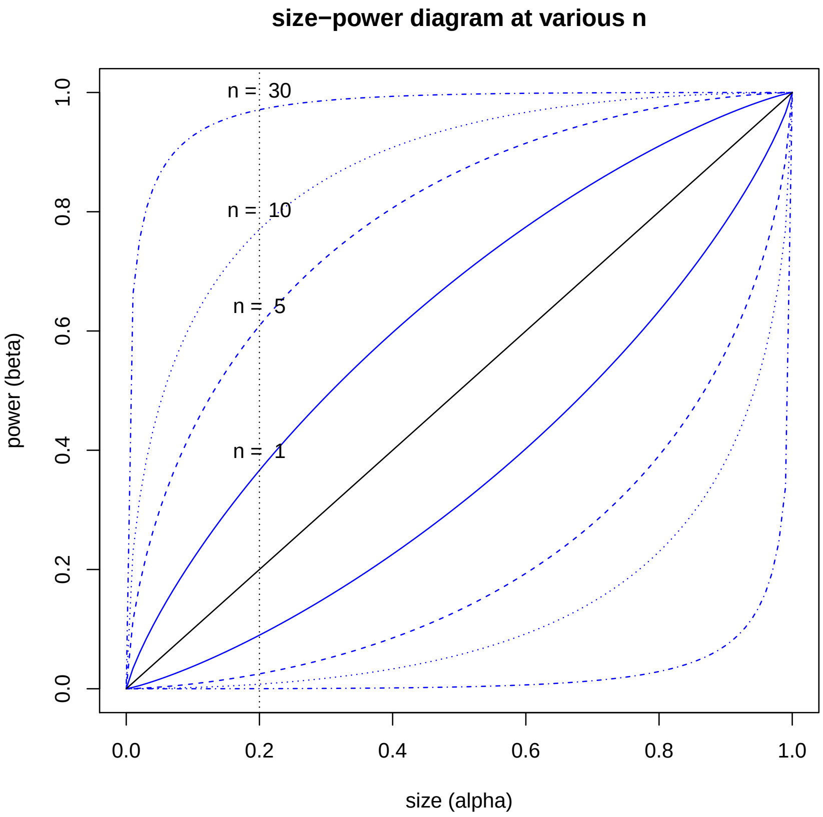

Figure 3: Size-Power diagram for the specific testing problem in Example 5.1, at alternative $H_1 : \theta = 0.5$ and at various sample sizes (n). Vertical line at size $= 0.2$, shows increasing power of MP tests as n increases.

At the other extreme, if $\mathbf{P}_0$ and $\mathbf{P}_1$ have disjoint supports, then for any sample size n, the size-power set will be equal to $[0,1]^2$. This can be shown as follows. Let f_i denote the pdf (or pmf) of $\mathbf{P}_i$, $i = 0, 1$. Naturally it is enough for us to show that there is a test ϕ such that $\mathbf{E}_0\phi(\mathbf{X}) = 0$ and $\mathbf{E}_1\phi(X) = 1$. As the distributions have disjoint support, hence[19] $\{\mathbf{x} : f_0(\mathbf{x}) > 0\} \cap \{\mathbf{x} : f_1(\mathbf{x}) > 0\} = \varnothing$. We set $\phi(\mathbf{x}) = \mathbf{1}(f_1(\mathbf{x}) > 0)$. Hence we reject the null whenever $\mathbf{x}$ is in the support of the alternative distribution. This will ensure the required size and power.

Conversely, if $(0, 1)$ lies within the size-power diagram, then $\mathbf{P}_0$ and $\mathbf{P}_1$ will have disjoint supports. This is because, there exists a ϕ such that (we drop the $d\mathbf{x}$ term from the notation for ease of writing),

$$\mathbf{E}_0\phi(\mathbf{X}) = \int \phi \ f_0 = 0 \quad \text{and} \quad \mathbf{E}_1\phi(\mathbf{X}) = \int \phi \ f_1 = 1,$$

[19] Instead of empty set, we can use a set with length zero, but the argument will be same.

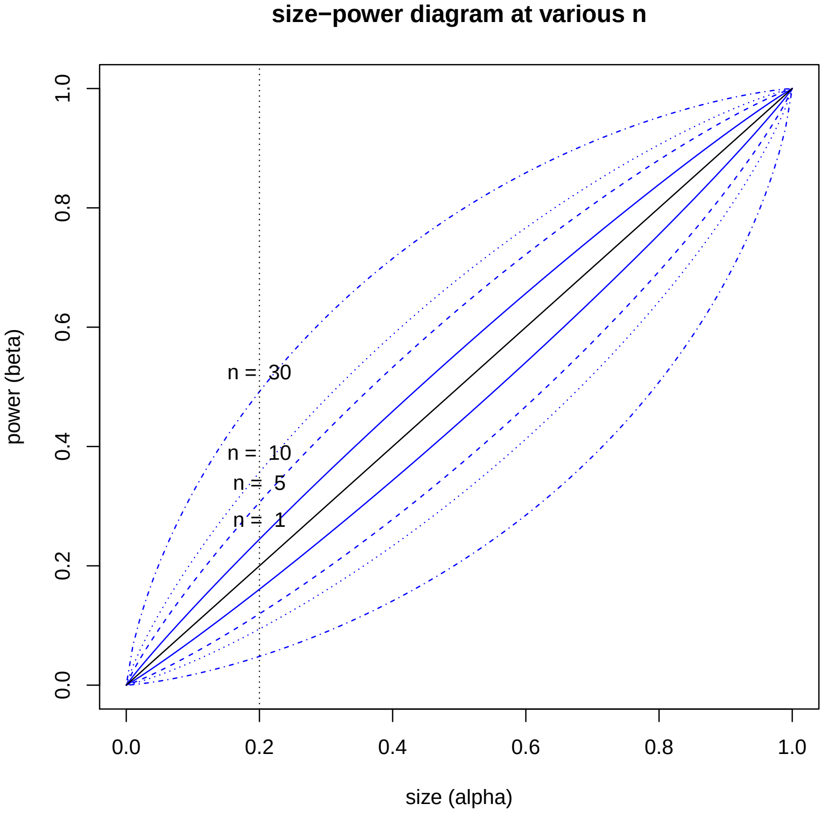

Figure 4: Size-Power diagram for the specific testing problem in Example 5.1, at alternative $H_1 : \theta = 0.15$ and at various sample sizes (n). Vertical line at size $= 0.2$, shows increasing power of MP tests as n increases.

$$\begin{aligned}
&\Leftrightarrow 0 = \int_{\phi>0, f_0>0} \phi \; f_0 \quad \text{and} \quad 0 = \int_{1-\phi>0, f_1>0} (1-\phi) \; f_1, \\
&\Leftrightarrow \{\phi > 0\} \cap \{f_0 > 0\} = \varnothing \quad \text{and} \quad \{\phi < 1\} \cap \{f_1 > 0\} = \varnothing, \quad (\star), \\
&\Leftrightarrow \{f_0 > 0\} \subseteq \{\phi = 0\} \quad \text{and} \quad \{f_1 > 0\} \subseteq \{\phi = 1\}.
\end{aligned}$$

Since $\{\phi = 0\} \cap \{\phi = 1\}$ are necessarily disjoint, this implies $\{f_0 > 0\} \cap \{f_1 > 0\} = \varnothing$, thereby completing the proof.

Another important case arises when the size-power diagram is equal to the set $\{(\alpha, \alpha) : \alpha \in [0, 1]\}$. Then, both the null and alternative distributions are identical. In fact, even if there is *any* $\alpha \in (0, 1)$, for which the MP test has power $= \alpha$, then it will imply that $\mathbf{P}_0 = \mathbf{P}_1$. The argument can be carried out as follows.

Consider the size-power diagram for any simple versus simple testing problem ($\mathbf{P}_0$ against $\mathbf{P}_1$). Assume there exists some $\alpha_1 \in (0, 1)$, such that MP test of size α_1 has power $= \alpha_1$. If possible, assume there is an α_2, such that the MP test of size α_2 has power $\beta_2 (> \alpha_2)$. If $\alpha_2 > \alpha_1$, then we join the points $(0, 0)$ and (α_2, β_2) with a straight line. Convexity implies that this straight line should be included inside the size-power set. However at $x = \alpha_1$, the straight line passes above the point (α_1, α_1) and this implies there exists a test of size α_1 with power $> \alpha_1$, contradicting the fact that the MP test of size α_1 has power $= \alpha_1$. Hence, there cannot exist any $\beta_2 > \alpha_2$, within this size-power set. Similarly if $\alpha_2 < \alpha_1$, we do the same process by joining with the point $(1,1)$. As a result, there exists no points above the straight line $\{y = x\}$ which are inside the size-power set. Symmetry implies, there exists no points below the straight line $\{y = x\}$ which are inside the size-power set. So for **any** test function ϕ for this testing problem, the only possibility is,

$$\mathbf{E}_0\phi(\mathbf{X}) = \int \phi \ f_0 = \int \phi \ f_1 \mathbf{E}_1\phi(\mathbf{X}).$$

Now choose $\phi(\mathbf{x}) = \mathbf{1}(\mathbf{x} \in C)$, where C any Borel-set (on the union of their supports). Then as $\phi(\mathbf{x})$ is a valid test function, it implies

$$\mathbf{P}_0(\mathbf{X} \in C) = \mathbf{P}_1(\mathbf{X} \in C), \quad \text{for any arbitrary Borel-set } C.$$

This shows that the two probability distributions are same.

5.2 Neyman Pearson Lemma

We begin with some basic notation. For any r.v. X with pdf/pmf $f(x)$, write

$$\int g(x)f(x)\ d\mu(x) = \begin{cases} \sum_x g(x)f(x) & \text{if } X \text{ has a pmf } f, \\ \int g(x)f(x)\ dx & \text{if } X \text{ has a pdf } f. \end{cases}$$

Although we are not going into details: but you can think of $\mu(A)$ = length of the set A, in case X is a continuous r.v., and in case of a discrete r.v. X, $\mu(A)$ is the number of elements in the set A (further details on these will be provided in your Probability course). Also it is useful for us to note that if

$$\int g(x)f(x)\ d\mu(x) = 0, \text{for any } g > 0,$$

then the set $\{x : g(x) > 0\}$ must have probability zero under the r.v. X. Note that X can be vector valued. Now we state the NP Lemma:

Theorem 18 (NP Lemma)**.** *Consider the testing problem $H_0 : X \sim P_0$ against $H_1 : X \sim P_1$, where both H_0 and H_1 are simple hypotheses. Suppose f_i denotes the pdf*[20] *(or the pmf) of P_i, and $\mathbf{E}_i$ denotes the expectation under P_i, $i = 0, 1$.*

(a) Sufficiency: *Any test function $\phi : \mathcal{X} \mapsto [0,1]$ of the form*

$$\phi(x) = \begin{cases} 1 & \text{if } f_1(x) > kf_0(x), \\ \gamma(x) & \text{if } f_1(x) = kf_0(x), \\ 0 & \text{if } f_1(x) < kf_0(x), \end{cases} \tag{5.42}$$

for some $\gamma : \mathcal{X} \mapsto [0,1]$ and $k \in [0, \infty)$ is a MP test of its size $= \alpha = \mathbf{E}_0\phi(X)$. For $k = +\infty$, the test

$$\phi(x) = \begin{cases} 1 & \text{if } f_0(x) = 0, \\ 0 & \text{if } f_0(x) > 0, \end{cases} \tag{5.43}$$

is a MP test of size $\alpha = 0$.

[20] in this case, $f_i(x) = \frac{dF_i(x)}{dx}$.

(b) EXISTENCE: *For every $\alpha \in [0,1]$, a MP test of size α exists for some choice of k, and is of the form* (5.42) *with $\gamma(x) = \gamma_0$ (a constant), or is of the form* (5.43).

(c) UNIQUENESS/NECESSITY: *If ϕ_1 is another MP test of size $\alpha \in (0,1)$, then it is of the form* (5.42) *except on the set $\{x : f_1(x) = kf_0(x)\}$, i.e., $\phi_1(x) = \phi(x)$ for all $x \in \{f_1(x) \neq kf_0(x)\}$ with probability 1 under both P_i, $i = 0,1$. If ϕ is another MP test of size $\alpha = 0$, then ϕ_1 must be of the form* (5.43), *except on a set of probability zero, under both P_i, $i = 0,1$.*

Remark 6. The sufficiency part gives the precise form of an MP test (except up to the knowledge of k and γ). For $\alpha = 1$, the test

$$\phi(x) = 1 \quad \text{for all } x. \tag{5.44}$$

will be the MP test. This test will have power $\beta = 1$. Size $\alpha = 1$ tests are not too interesting as it does not put any restrictions on the type I error probability. Also note that the tests described in (5.43) and (5.44) can be described as special cases of the test in (5.42). For the test in (5.43), we can write

$$\{f_1 \gtreqless kf_0\} = \left\{\frac{f_1}{k} \gtreqless f_0\right\},$$

and then use $k = +\infty$ and $\gamma = 1$, to derive the MP test in (5.43). For the test in (5.44), we can use $k = 0$ and $\gamma = 1$.

Proof of NP-Lemma. We begin with the **sufficiency** part. This states that any test of the form (5.42) or (5.43) will be MP of its size $= \alpha \in [0,1)$. We need to show the following: if ϕ_1 is another test with $\mathbf{E}_0\phi_1(X) \leqslant \alpha$, then $\mathbf{E}_1\phi(X) \geqslant \mathbf{E}_1\phi_1(X)$. For the test ϕ in (5.42), define the sets $A = \{x : f_1(x) > kf_0(x)\}$ and $B = \{x : f_1(x) < kf_0(x)\}$. On the set A, $\phi = 1$ and we must have $(\phi - \phi_1) \geqslant 0$. Similarly, on the set B, $\phi = 0$ and hence $(\phi - \phi_1) \leqslant 0$. Using this we can write

$$\begin{aligned}\int (\phi - \phi_1)(f_1 - kf_0)\, d\mu &= \int_A \underbrace{(\phi - \phi_1)(f_1 - kf_0)}_{\geqslant 0}\, d\mu + \int_B \underbrace{(\phi - \phi_1)(f_1 - kf_0)}_{\geqslant 0}\, d\mu \\ &\geqslant 0.\end{aligned}$$

This implies,

$$\mathbf{E}_1\phi(X) - \mathbf{E}_1\phi_1(X) \geqslant k\left[\mathbf{E}_0\phi(X) - \underbrace{\mathbf{E}_0\phi_1(X)}_{\leqslant \alpha}\right] \geqslant 0.$$

Hence proving that ϕ is MP of size $\alpha \in (0,1)$.

Now consider the case for $\alpha = 0$ and the test in (5.43). Let ϕ_1 be any other test of size $\alpha = 0$. Then

$$\mathbf{E}_0\phi_1(X) = 0 = \int \phi_1 f_0\, d\mu = \int_{\{f_0 > 0\}} \phi_1 f_0\, d\mu.$$

Hence, $\phi_1(x) = 0$ on the set $\{x : f_0(x) > 0\}$. And

$$\begin{aligned}\mathbf{E}_1\phi(X) - \mathbf{E}_1\phi_1(X) = \int (\phi - \phi_1) f_1 \, d\mu &= \int_{\{f_0>0\}} (\phi - \phi_1) f_1 \, d\mu + \int_{\{f_0=0\}} (\phi - \phi_1) f_1 \, d\mu \\ &= \int_{\{f_0>0\}} (0 - 0) f_1 \, d\mu + \int_{\{f_0=0\}} \underbrace{(1 - \phi_1)}_{\geqslant 0} f_1 \, d\mu \geqslant 0.\end{aligned}$$

For the size $\alpha = 1$ test defined in (5.44), the power is 1, and hence it will be MP.

Now we study the **existence** part. In this case, we will need choices of k and γ for the test described in (5.42). For the MP tests of size 0 and 1, the choices of k and γ have been discussed in Remark 6. So we look into the case of $\alpha \in (0, 1)$. The only condition that dictates the choices of these values, is the size condition. The fact that the tests will be MP (with appropriate choices of the constants has already been proved). We will show that in (5.42), it will be enough to choose $\gamma(x) = \gamma_0$, a constant.

We know that k and γ_0 have to be chosen in such a way that

$$P_0\left(f_1(X) > k f_0(X)\right) + \gamma_0 P_0\left(f_1(X) = k f_0(X)\right) = \alpha. \tag{5.45}$$

Define the r.v.

$$Y = \frac{f_1(X)}{f_0(X)} \cdot \mathbf{1}\left(f_0(X) > 0\right).$$

Then, $P_0(Y \in [0, \infty)) = 1$, which implies that Y is a real valued r.v. under P_0. This will be an important fact in our calculations[21] Rewriting (5.45) we have

$$P_0\left(Y \leqslant k\right) - \gamma_0 \cdot P_0\left(Y = k\right) = 1 - \alpha. \tag{5.46}$$

Let, $k_\alpha = \inf\{y : P_0(Y \leqslant y) \geqslant (1 - \alpha)\}$, which is the $(1 - \alpha)$ quantile of Y. Since, $\alpha \in (0, 1)$, $(1 - \alpha) \in (0, 1)$ and since Y is a proper real valued r.v. under P_0, we must have $k_\alpha \in (0, \infty)$. Now, if $P_0(Y \leqslant k_\alpha) = 1 - \alpha$, then we set $k = k_\alpha$ and $\gamma_0 = 0$. Otherwise, set $k = k_\alpha$ and solve for γ_0 in (5.46), with[22]

$$\gamma_0 = \frac{P_0(Y \leqslant k_\alpha) - (1 - \alpha)}{P_0(Y = k_\alpha)} \in (0, 1). \tag{5.47}$$

Thus we obtain the choices of $\gamma(x) = \gamma_0$ and k in (5.42), which proves the existence of the MP test.

Now we study the **uniqueness/necessity** part: we aim to show that the MP tests in (5.42) and (5.43) are essentially unique. Consider the case for $\alpha = 0$. The MP test is given by (5.43). Suppose ϕ_1 is any other MP test of size $\alpha = 0$. Then,

$$\mathbf{E}_0\phi_1(X) = 0 \quad \text{and} \quad \mathbf{E}_1\phi_1(X) = \mathbf{E}_1\phi(X). \tag{5.48}$$

The size condition implies that

$$0 = \int \phi_1 f_0 \, d\mu = \int_{\{f_0>0\}} \phi_1 f_0 \, d\mu$$

[21] If $Y = +\infty$, w.p. $\delta > 0$, then we would have problems in defining the ηth (with $\eta > 1 - \delta$) quantile of Y.

[22] Check: for any r.v. Z with cdf G and $\xi_\eta = G^{-1}(\alpha)$, we have $G(G^{-1}(\alpha)) \geqslant \alpha$ for all $\alpha \in (0, 1)$. And for any $x \in \mathbb{R}$, $G^{-1}(G(x)) \leqslant x$.

Hence on the set $\{f_0 > 0\}$, we have $\phi_1 = 0$. And since both tests have same power, we have

$$\begin{aligned} 0 &= \int (\phi - \phi_1) f_1 \, d\mu \\ &= \int_{\{f_0=0\}} (1-\phi_1) f_1 \, d\mu + \int_{\{f_0>0\}} (0-0) f_1 \, d\mu \\ &= \int_{\{f_0=0\}\cap\{f_1>0\}} (1-\phi_1) f_1 \, d\mu. \end{aligned}$$

This can lead to two situations:

(i) $\phi_1 = 1$ on the set $\{f_0 = 0\} \cap \{f_1 > 0\}$. In that case we already know that $\phi = 0 = \phi_1$ on the set $\{f_0 > 0\}$. And $\phi = 1 = \phi_1$ on $\{f_0 = 0\} \cap \{f_1 > 0\}$. The only set where ϕ and ϕ_1 can mismatch is $\{f_0 = 0\} \cap \{f_1 = 0\}$, which implies that

$$P_i\left(\phi(X) = \phi_1(X)\right) = 1, \quad \text{for } i = 0, 1.$$

(ii) The other situation is that the set $\{f_0 = 0\} \cap \{f_1 > 0\}$ has zero probability, which can be expressed by

$$\mu\left(\{f_0 = 0\} \cap \{f_1 > 0\}\right) = 0.$$

In that case we must have $\{f_1 > 0\} \subseteq \{f_0 > 0\}$ with probability 1 (almost everywhere μ). On the set $\{f_0 > 0\}$, we already have $\phi = 0 = \phi_1$. On the set $\{f_0 = 0\}$ we have $\phi = 1$, but $\{f_0 = 0\} \subseteq \{f_1 = 0\}$ with probability 1, and hence it does not matter if ϕ_1 matches ϕ on this set (draw a simple figure depicting the set inclusions, and the argument will be clear).

So in both cases, we have $\phi = \phi_1$, except on a set which has probability 0 under both P_i. Hence the uniqueness of the MP test of size $\alpha = 0$ is proved.

Now we consider for $\alpha \in (0,1)$ and the MP test described in (5.42). Suppose ϕ_1 is also MP of size α, *i.e.*,

$$\mathbf{E}_i \phi(X) = \mathbf{E}_i \phi_1(X), \quad i = 0, 1.$$

Then

$$\begin{aligned} 0 &= \mathbf{E}_1(\phi(X) - \phi_1(X)) - k \cdot E_0(\phi(X) - \phi_1(X)) \\ &= \int_{\{f_1 \neq k f_0\}} (\phi - \phi_1)(f_1 - k f_0) \, d\mu \\ &= \int_{\{f_1 \neq k f_0\} \cap \{\phi \neq \phi_1\}} (\phi - \phi_1)(f_1 - k f_0) \, d\mu. \end{aligned} \tag{5.49}$$

Now note that after some tedious simplifications (and using arguments similar to that used in the sufficiency part) we have[23],

$$\{\phi \neq \phi_1\} \cap \{f_1 \neq k f_0\} = [\{\phi > \phi_1\} \cap \{f_1 > k f_0\}] \cap [\{\phi < \phi_1\} \cap \{f_1 < k f_0\}] = C \cap D \quad \text{(say)}.$$

The integral in (5.49) can then be written as (note that the integrands are strictly > 0)

$$0 = \int_{\{f_1 \neq k f_0\} \cap \{\phi \neq \phi_1\}} (\phi - \phi_1)(f_1 - k f_0) \, d\mu = \int_C \underbrace{(\phi - \phi_1)(f_1 - k f_0)}_{>0} \, d\mu + \int_D \underbrace{(\phi - \phi_1)(f_1 - k f_0)}_{>0} \, d\mu.$$

[23] You **must** verify this yourself.

This implies that $\mu(C) = 0 = \mu(D)$. The implications are as follows: on the set $\{f_1 > kf_0\}$, we already know that it is impossible to have $\{\phi < \phi_1\}$, and we have found $\mu(C) = 0$, which rules out $\{\phi > \phi_1\}$ on the set $\{f_1 > kf_0\}$. This leaves out the only possibility of $\{\phi = \phi_1\}$ on the set $\{f_1 > kf_0\}$. Similarly on the set $\{f_1 < kf_0\}$, we must have $\{\phi = \phi_1\}$. The only ambiguity remains on the set $\{f_1 = kf_0\}$, where ϕ_1 can be arbitrarily chosen, as long as it satisfies the size condition.

Hence, $\phi = \phi_1$ with probability 1 under both P_i, $i = 0, 1$, on the set $\{f_1 \neq kf_0\}$. This completes the proof. □

Remark 7. An intuitive explanation about the non-uniqueness of MP tests on the set $\{f_1 = kf_0\}$ can be given as follows: note the form of ϕ in (5.42) (with $\gamma(x) = \gamma_0$ in (5.47)). Now suppose we $\{f_1 = kf_0\}$ contains more than one point. The MP test obtained by the steps in the sufficiency and existence part, places a constant value (a constant randomization probability) on all points of this set. And the overall contribution to the size α, from this set is

$$\int_{\{f_1 = kf_0\}} \phi f_0 \ d\mu = \gamma_0 \int_{\{f_1 = kf_0\}} f_0 \ d\mu. \tag{5.50}$$

Now, instead of using a constant value γ_0 on this set, we can arbitrarily change ϕ so that the integral on the left side of (5.50) remains the same (drawing a simple figure helps). If we try to put different values of ϕ_1 on the set $\{f_1 > kf_0\}$, then we would need to tweak ϕ_1 to either a smaller (or higher) value than 1 on an interval (or a set of points), and as a result we would need to compensate on some other part of this set, by setting a value of ϕ_1 larger than (or smaller than) 1. This is impossible as ϕ_1 cannot exceed 1. The same argument works for the set $\{f_1 < kf_0\}$.

Remark 8. Note that the MP test ϕ of size α is the best test in the class of tests $\mathcal{C}_1 = \{\psi : \mathbf{E}_0\psi(X) \leqslant \alpha\}$. And ϕ is also the MP test in the class $\mathcal{C}_2 = \{\psi : \mathbf{E}_0\psi(X) = \alpha\}$. The condition $\mathbf{E}_0\phi(X) = \alpha$ implies $\phi \in \mathcal{C}_2 \subseteq \mathcal{C}_1$.

Remark 9. A MP test ϕ maximizes the power (and hence minimizes the probability of type II error) under the constraint that the size $\alpha = \alpha_0$, for some specified $\alpha_0 \in [0, 1]$. If β denotes the power of the test, then this is equivalent to minimizing

$$(1 - \mathbf{E}_1\phi(X)) + \lambda(\mathbf{E}_0\phi(X) - \alpha_0),$$

where λ is the Lagrange multiplier. This is equivalent to minimizing

$$1 - \lambda\alpha_0 + \int_{R_1} (\lambda f_0 - f_1) \ d\mu,$$

where $R_1 = R_1(\lambda)$ is the 'rejection region' (to be selected). This can be done if we define R_1 to be the set containing all sample points satisfying $\{f_1 > \lambda f_0\}$. And similarly place all points in $\{f_1 < \lambda f_0\}$ in a set R_3. And the points in $\{f_1 = kf_0\}$ can be placed in another set R_2. This idea can be carried on to show that a λ can be selected to give a MP test of size α_0.

References

Berger, J. O. (1993). *Statistical decision theory and Bayesian analysis.* Springer Series in Statistics. Springer-Verlag, New York. Corrected reprint of the second (1985) edition.

Bhattacharya, R., Lin, L., and Patrangenaru, V. (2016). *A course in Mathematical Statistics and Large Sample Theory.* Springer.

Billingsley, P. (1995). *Probability and Measure.* Wiley Series in Probability and Statistics. Wiley.

Casella, G. and Berger, R. L. (1990). *Statistical Inference.* The Wadsworth & Brooks/Cole Statistics/Probability Series. Wadsworth & Brooks/Cole Advanced Books & Software, Pacific Grove, CA.

DasGupta, A. (2011). *Probability for Statistics and Machine Learning: Fundamentals and Advanced topics.* Springer Science & Business Media.

Ferguson, T. S. (2014). *Mathematical statistics: A decision theoretic approach*, volume 1. Academic press.

Ghosh, M. (2002). Basu's theorem with applications: A personalistic review. *Sankhya: The Indian Journal of Statistics, Series A (1961-2002)*, 64(3):509–531.

Lehmann, E. L. (1981). An interpretation of completeness and basu's theorem. *Journal of the American Statistical Association*, 76(374):335–340.

Lehmann, E. L. and Casella, G. (1998). *Theory of Point Estimation.* Springer Texts in Statistics. Springer-Verlag, New York, Second edition.

Lehmann, E. L. and Romano, J. P. (2005). *Testing Statistical Hypotheses.* Springer Texts in Statistics. Springer, New York, Third edition.

Lehmann, E. L. and Scheffé, H. (2012). Completeness, similar regions, and unbiased estimation - part II. In *Selected Works of EL Lehmann*, pages 269–286. Springer.

Robert, C. P. (2007). *The Bayesian choice.* Springer Texts in Statistics. Springer, New York, second edition. From decision-theoretic foundations to computational implementation.

Rohatgi, V. K. and Saleh, A. M. E. (2015). *An Introduction to Probability and Statistics.* John Wiley & Sons.

Shao, J. (2003). *Mathematical Statistics.* Springer Texts in Statistics. Springer.

Sundberg, R. (2019). *Statistical Modelling by Exponential Families.* Institute of Mathematical Statistics Textbooks. Cambridge University Press.

Winkler, R. L. (1972). A decision-theoretic approach to interval estimation. *Journal of the American Statistical Association*, 67(337):187–191.

Young, G. A. and Smith, R. L. (2005). *Essentials of Statistical Inference.* Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

1 MLR based UMP tests

Let $\{X_1, \ldots, X_n\}$ be a sample with pdf/pmf $f(\mathbf{x} : \theta)$. Then the likelihood function is defined as $L(\theta) = f(\mathbf{x} : \theta)$. Consider a testing problem of the form $H_0 : \theta \in \Theta_0$ v/s $H_1 : \theta \in \Theta_1$. We want to find an UMP test for this problem, in general there will be no such test. However, under additional conditions such tests can be developed. We consider the following simplified definition.

Definition 1 (Monotone Likelihood Ratio)**.** Let $\Theta \subseteq \mathbb{R}$ and $\{f(\mathbf{x} : \theta) : \theta \in \Theta\}$ denote the family of pdf/pmf (densities) of $\mathbf{X} = \{X_1, \ldots, X_n\}$. Let $T : \mathcal{X} \mapsto \mathbb{R}$ be a statistic. Assume that the distributions under θ_1 and θ_2 are distinct, if $\theta_1 \neq \theta_2$. Then, the family of densities has a MLR (monotone likelihood ratio) in $T(\mathbf{X})$ if for any $\theta_1 < \theta_2$, with $\theta_1, \theta_2 \in \Theta$, there exists a non-decreasing function $g \equiv g_{\theta_1,\theta_2} : \mathbb{R} \mapsto [0, \infty]$, such that

$$\frac{f(\mathbf{x} : \theta_2)}{f(\mathbf{x} : \theta_1)} = g_{\theta_1,\theta_2}(T(\mathbf{x})), \quad \text{for all } \mathbf{x} \in A, \tag{1.1}$$

with $A = A_{\theta_1,\theta_2} = \{\mathbf{x} : f(\mathbf{x} : \theta_1) + f(\mathbf{x} : \theta_2) > 0\}$. We will define the ratio in (1.1) as $+\infty$, if $f(\mathbf{x} : \theta_2) > 0$ and $f(\mathbf{x} : \theta_1) = 0$.

Example 1 (One parameter exponential family)**.** Let $\mathbf{X}$ have the density,

$$f(\mathbf{x} : \theta) = \exp\{\eta(\theta)T(\mathbf{x}) - \psi(\theta)\} \cdot h(\mathbf{x}), \quad \mathbf{x} \in \mathcal{X}, \ \theta \in \Theta \subseteq \mathbb{R},$$

with

$$A_{\theta_1,\theta_2} = A = \{\mathbf{x} : h(\mathbf{x}) > 0\}, \quad \text{for all } \theta_1, \theta_2 \in \Theta.$$

Then, for $\theta_1 < \theta_2$ and $\mathbf{x} \in A$,

$$\frac{f(\mathbf{x} : \theta_2)}{f(\mathbf{x} : \theta_1)} = \exp\left\{(\eta(\theta_2) - \eta(\theta_1))\, T(\mathbf{x}) - \psi(\theta_2) + \psi(\theta_1)\right\},$$

will be non-decreasing in $T(\mathbf{x})$, if $\eta(\theta)$ is non-decreasing in θ.

Example 2. Let $\{X_1, \ldots, X_n\}$ be iid Uniform $(0, \theta)$, $\theta > 0$. Then the family has MLR in $T(\mathbf{X}) = X_{(n)}$.

Example 3. Let $\{X_1, \ldots, X_n\}$ be iid Cauchy $(\theta, 1)$, $\theta \in \mathbb{R}$. The family has no MLR, because the ratio $f(\mathbf{x} : \theta_2)/f(\mathbf{x} : \theta_1)$ is not monotone in any function of $\mathbf{x}$.

Theorem 1 (UMP tests for one-sided composite hypothesis)**.** *Assume that the family of distributions $\{f(\mathbf{x} : \theta) : \theta \in \Theta\}$, with $\Theta \subseteq \mathbb{R}$ has a MLR in $T(\mathbf{X})$.*

(i) For testing $H_0 : \theta \leqslant \theta_0$ against $H_1 : \theta > \theta_0$, a UMP size α test is given by,

$$\phi(\mathbf{x}) = \begin{cases} 1 & \text{if } T(\mathbf{x}) > c, \\ \gamma & \text{if } T(\mathbf{x}) = c, \\ 0 & \text{if } T(\mathbf{x}) < c, \end{cases} \quad \text{for some } c \in [-\infty, \infty] \text{ and } \gamma \in [0, 1], \tag{1.2}$$

such that, $\mathbf{E}_0\phi(\mathbf{X}) = \alpha$, provided $\alpha \in (0, 1]$.

(ii) The power function $\beta_\phi(\theta) = \mathbf{E}_\theta\phi(\mathbf{X})$ is non-decreasing in θ.

Proof of Theorem 1. Initially, let us consider the simple vs simple testing problem[1]: $H_0 : \theta = \theta_0$ against $H_1 : \theta = \theta_1$ $(> \theta_0)$. Consider the case where the size $\alpha = 1$. By NP-lemma, the MP size $\alpha = 1$ test is given by, $\psi(\mathbf{x}) = 1$, for all $\mathbf{x}$. Consider $\phi(\mathbf{x})$ in (1.2) with the following choice: $c = -\infty$. Since T is real valued, ϕ reduces to

$$\phi(\mathbf{x}) = \begin{cases} 1 & \text{if } T(\mathbf{x}) > -\infty, \\ \gamma & \text{if } T(\mathbf{x}) = -\infty, \\ 0 & \text{if } T(\mathbf{x}) < -\infty. \end{cases}$$

The only non-empty set is $\{T(\mathbf{x}) > -\infty\} = \mathcal{X}$. Hence, ϕ is equivalent ψ defined above.

Now consider the case of $\alpha \in (0, 1)$. We claim, $c \in \mathbb{R}$, due to the following reason. If $c = -\infty$, then we have $\alpha = 1$ case, which has been already considered. If $c = +\infty$, then ϕ reduces to, $\phi(\mathbf{x}) = 0$, if $T(\mathbf{x}) < +\infty$. This leads to, $\mathbf{E}_{\theta_0}\phi(\mathbf{X}) = 0$ (which is currently out of our consideration).

Going back to the case of $\alpha \in (0, 1)$, due to NP-lemma the MP-size α test for $H_0 : \theta = \theta_0$ against $H_1 : \theta = \theta_1$ $(> \theta_0)$, is of the form,

$$\psi(\mathbf{x}) = \begin{cases} 1 & \text{if } f_{\theta_1}(\mathbf{x}) > k f_{\theta_0}(\mathbf{x}), \\ \gamma(\mathbf{x}) & \text{if } f_{\theta_1}(\mathbf{x}) = k f_{\theta_0}(\mathbf{x}), \\ 0 & \text{if } f_{\theta_1}(\mathbf{x}) < k f_{\theta_0}(\mathbf{x}). \end{cases}$$

Note that, $g_{\theta_0,\theta_1}(T(\mathbf{x})) = f(\mathbf{x} : \theta_1)/f(\mathbf{x} : \theta_0)$ is non-decreasing in $T(\mathbf{x})$. So,

$$[T(\mathbf{x}) > c] \Rightarrow [g(T(\mathbf{x})) \geqslant g(c)].$$

Hence, we can write the test ϕ in (1.2) as,

$$\begin{aligned} \phi(\mathbf{x}) &= \begin{cases} 1 & \text{if } T(\mathbf{x}) > c \text{ and } g_{\theta_0,\theta_1}(T(\mathbf{x})) > g_{\theta_0,\theta_1}(c), \\ 1 & \text{if } T(\mathbf{x}) > c \text{ and } g_{\theta_0,\theta_1}(T(\mathbf{x})) = g_{\theta_0,\theta_1}(c), \\ \gamma & \text{if } g_{\theta_0,\theta_1}(T(\mathbf{x})) = g_{\theta_0,\theta_1}(c), \\ 0 & \text{if } T(\mathbf{x}) < c \text{ and } g_{\theta_0,\theta_1}(T(\mathbf{x})) = g_{\theta_0,\theta_1}(c), \\ 0 & \text{if } T(\mathbf{x}) < c \text{ and } g_{\theta_0,\theta_1}(T(\mathbf{x})) < g_{\theta_0,\theta_1}(c). \end{cases} \\ &= \begin{cases} 1 & \text{if } g_{\theta_0,\theta_1}(T(\mathbf{x})) = \frac{f(\mathbf{x}:\theta_1)}{f(\mathbf{x}:\theta_0)} > g_{\theta_0,\theta_1}(c), \\ \gamma(\mathbf{x}) & \text{if } g_{\theta_0,\theta_1}(T(\mathbf{x})) = \frac{f(\mathbf{x}:\theta_1)}{f(\mathbf{x}:\theta_0)} = g_{\theta_0,\theta_1}(c), \\ 0 & \text{if } g_{\theta_0,\theta_1}(T(\mathbf{x})) = \frac{f(\mathbf{x}:\theta_1)}{f(\mathbf{x}:\theta_0)} < g_{\theta_0,\theta_1}(c), \end{cases} \\ &= \begin{cases} 1 & \text{if } \frac{f(\mathbf{x}:\theta_1)}{f(\mathbf{x}:\theta_0)} > k, \\ \gamma_{\theta_0,\theta_1}(\mathbf{x}) & \text{if } \frac{f(\mathbf{x}:\theta_1)}{f(\mathbf{x}:\theta_0)} = k, \\ 0 & \text{if } \frac{f(\mathbf{x}:\theta_1)}{f(\mathbf{x}:\theta_0)} < k, \end{cases} \end{aligned} \tag{1.3}$$

where,[2] $k = g_{\theta_0,\theta_1}(c) \in (0, \infty)$. Now, by the NP-lemma, the representation in (1.3) shows that $\phi(\mathbf{x})$ is MP of size $\alpha = \mathbf{E}_{\theta_0}\phi(\mathbf{X}) \in (0, 1)$ with power $\beta(\theta_1) = \mathbf{E}_{\theta_1}\phi(\mathbf{X})$ (if, the distributions are distinct, we must have $\beta > \alpha$, with the inequality being strict). If, $\alpha = 1$, we must have $\beta(\theta_1) = \mathbf{E}_{\theta_1}\phi(\mathbf{X}) = 1$.

Now, in order to prove that ϕ is UMP, consider any other ϕ_1 satisfying, $\sup_{\theta \leqslant \theta_0} \mathbf{E}_\theta \phi_1(\mathbf{X}) \leqslant \alpha$. This implies, $\mathbf{E}_{\theta_0}\phi_1(\mathbf{X}) \leqslant \alpha$. But as ϕ is MP for $H_0 : \theta = \theta_0$ against $H_1 : \theta = \theta_1 (> \theta_0)$, one must have,

[1] These hypotheses should be given different names, but I am using the same notation.

[2] Otherwise, $k = +\infty$, will lead to the size of ϕ being zero.

$\mathbf{E}_{\theta_1}\phi(\mathbf{X}) \geqslant \mathbf{E}_{\theta_1}\phi_1(\mathbf{X})$. Since θ_1 is arbitrary, it implies, $\mathbf{E}_{\theta}\phi(\mathbf{X}) \geqslant \mathbf{E}_{\theta}\phi_1(\mathbf{X})$, for any $\theta > \theta_0$. This proves, ϕ is UMP size α (we need to use the non-decreasing property of the power function of ϕ to claim this) for testing $H_0 : \theta \leqslant \theta_0$ against $H_1 : \theta > \theta_0$.

Note that the original form of $\phi(\mathbf{x})$ in (1.2) is independent of the choice of θ_0 or θ_1. But, the representation in (1.3) is dependent on these choices. So, for every distinct pair (θ_*, θ_{**}), with $\theta_* < \theta_{**}$, the test $\phi(\mathbf{x})$ in (1.2) will be MP of size $\mathbf{E}_{\theta_*}\phi(\mathbf{X})$ with power $\mathbf{E}_{\theta_{**}}\phi(\mathbf{X})$. In all cases, we must be able to say (using the fact that size < power for MP tests) that $\mathbf{E}_{\theta_*}\phi(\mathbf{X}) \leqslant \mathbf{E}_{\theta_{**}}\phi(\mathbf{X})$. So, this shows that the power function of ϕ is non-decreasing in θ (this will become strictly increasing if distributions under each θ are distinct). This is due to the fact that for each choice of the pair (θ_*, θ_{**}), the representation in (1.3) changes. Since the power function is non-decreasing, we must have,

$$\sup_{\theta \leqslant \theta_0} E_{\theta}\phi(\mathbf{X}) \leqslant \mathbf{E}_{\theta_0}\phi(\mathbf{X}) = \alpha,$$

which shows ϕ has size α. The choice of c and γ in (1.2) (for the case of $\alpha \in (0,1)$) is given by,

$$c = c_\alpha = \inf\left\{t : P_{\theta_0}\left(T(\mathbf{X}) \leqslant t\right) \geqslant 1-\alpha\right\} \in \mathbb{R}, \quad \text{if } \alpha \in (0,1), \text{ and}$$

$$\gamma = \gamma_\alpha = \begin{cases} \left\{(1-\alpha) - P_{\theta_0}(T(\mathbf{X}) \leqslant c)\right\}/P_{\theta_0}(T(\mathbf{X}) = c) & \text{if } P_{\theta_0}(T(\mathbf{X}) = c) > 0, \\ 0 & \text{o.w.} \end{cases}$$

Note, for $\alpha \in (0,1]$, the choices of c_α and γ_α are unique. This completes the proof. □

Remark. For testing $H_0' : \theta \leqslant \theta_0'$ against $H_1' : \theta > \theta_0'$, this test $\phi(\mathbf{x})$ in (1.2) is UMP of its size $\alpha' = \mathbf{E}_{\theta_0'}\phi(\mathbf{X})$.

Remark. The test $\phi(\mathbf{x})$ in (1.2) minimizes the power for all $\theta \leqslant \theta_0$, among all tests for testing H_0 against H_1. This is because, $\phi(\mathbf{x})$ is designed in such a way so that it is least-powerful for testing $H_0 : \theta = \theta_1$ agains $H_1 : \theta = \theta_0$. You can verify by constructing the least powerful test for testing this simple hypothesis using the NP-lemma.

Example 4. Now, let us consider the case of $\alpha = 0$. We will present an example. Let $\{X_1, \ldots, X_n\}$ be i.i.d. from Uniform $(0, \theta)$, with $\theta > 0$. It is known that the family has MLR in $T(\mathbf{X}) = X_{(n)}$. We want to test $H_0 : \theta \leqslant 1$ against $H_1 : \theta > 1$. Consider the test,

$$\phi_c(\mathbf{x}) = \begin{cases} 1 & \text{if } X_{(n)} > c, \\ 0 & \text{o.w.} \end{cases}$$

This is of the form shown in (1.2) (with $\gamma = 0$). Then,

$$\mathbf{E}_{\theta=1}\phi_c(\mathbf{X}) = P_{\theta=1}(X_{(n)} > c) = 0, \quad \text{for all } c \geqslant 1.$$

However, only $c = 1$ gives the UMP test of size $\alpha = 0$. To verify this, let $c_1 > 1$. Then, at any $\theta > c_1$, the powers will be,

$$\mathbf{E}_{\theta}\phi_1(\mathbf{X}) = 1 - \left(1/\theta\right)^n > \mathbf{E}_{\theta}\phi_{c_1}(\mathbf{X}) = 1 - \left(c_1/\theta\right)^n.$$

This shows, even though $\phi_{c_1}(\mathbf{x})$ is also of the form (1.2) and has size $\alpha = 0$, it has lower power than $\phi_1(\mathbf{x})$. So, for $\alpha = 0$, it is not enough to write form in (1.2) and solve for the constants c and γ.

For $\alpha = 0$, a test of the form (1.2), need not be necessarily UMP of size $\alpha = 0$, but it can be shown that there will be some c_0 and γ_0, such that a test $\phi(\mathbf{x})$ of the form (1.2), with $c = c_0$ and $\gamma = \gamma_0$, will be UMP of size $\alpha = 0$. This is contrary to the case of $\alpha \in (0,1)$, where we have shown that the choices of c and γ (for given α) are unique.

Remark. It should be noted that the test ϕ in (1.2) is not only UMP of its size for $H_0 : \theta \leqslant \theta_0$ against $H_1 : \theta > \theta_0$, but is also MP of its size for testing $H' : \theta = \theta'$ against $H'' : \theta = \theta''$, for any $\theta' < \theta'' \in \Theta$, provided the size (at θ') is between 0 and 1. This follows from direct application of the construction of ϕ.

2 UMP and UMPU tests in one-parameter exponential family

We will provide a generalization of the usual NP-Lemma.

Theorem 2 (Generalized NP Lemma). *Let $g_1, \ldots, g_{m+1}$ denoted real valued functions on the sample space $\mathcal{X}$ and are integrable, i.e. $\int |g_i(x)| \, d\mu(x) < \infty$ for all $i = 1, \ldots, (m+1)$. Suppose $\{c_1, \ldots, c_m\}$ is a given sequence of constants and let*

$$\mathcal{C} = \{\phi \mid \phi : \mathcal{X} \mapsto [0,1] \text{ such that} \int \phi(x) g_i(x) \, d\mu(x) = c_i, \, i = 1, \ldots, m.\} \tag{2.1}$$

(a) Among all members of $\mathcal{C}$, there exists a function ϕ such that it maximizes $\int \phi(x) g_{m+1}(x) \, d\mu(x)$.

(b) A sufficient condition for ϕ to maximize $\int \phi(x) g_{m+1}(x) \, d\mu(x)$, is the existence of constants $\{k_1, \ldots, k_m\}$ such that,

$$\phi(x) = \begin{cases} 1 & \text{if } g_{m+1}(x) > \sum_{i=1}^m k_i g_i(x), \\ 0 & \text{if } g_{m+1}(x) < \sum_{i=1}^m k_i g_i(x). \end{cases} \tag{2.2}$$

(c) If a member of $\mathcal{C}$ satisfies (2.2) with all $k_i \geqslant 0$, then it maximizes $\int \phi(x) g_{m+1}(x) \, d\mu(x)$ among all ϕ satisfying $\int \phi(x) g_i(x) \, d\mu(x) \leqslant c_i$, $i = 1, \ldots, m$.

(d) Consider the set,

$$A = \left\{ \left(\int \phi g_1, \ldots, \int \phi g_m \right) : \ \phi \text{ is a test function} \right\}.$$

Then, A is closed and convex. If $(c_1, \ldots, c_m)$ is an interior point of A, then there exists constants $(k_1, \ldots, k_m)$ and a test function $\phi^* \in \mathcal{C}$ and satisfying (2.2), and a necessary condition for a member of $\mathcal{C}$ to to maximize $\int \phi(x) g_{m+1}(x) \, d\mu(x)$ is that (2.2) holds a.e. μ.

Proof of Theorem 2. Omitted. See Theorem 3.6.1 of Lehmann and Romano (2005). □

Remark. Part (a) is known as the existence part of the GNP-Lemma. Part (b) is the sufficiency part. Part (c) deals with superiority among 'level α' type test functions. Part (d) ensures the existence of such a test function and also its uniqueness. For practical purposes, part (b) provides us the test function which can maximize power (w.r.t. g_{m+1}), subject to m constraints. The choices of k_i's can be found by using these constraint conditions.

We can think of different types of hypothesis of interest when the parameter under consideration is real-valued. For example, we can think of,

$$\begin{aligned} H_0 : \theta \leqslant \theta_1 \text{ or } \theta \geqslant \theta_2 \quad &\text{against} \quad H_1 : \theta_1 < \theta < \theta_2, \\ H_0 : \theta_1 \leqslant \theta \leqslant \theta_2 \quad &\text{against} \quad H_1 : \theta < \theta_1 \text{ or } \theta > \theta_2, \\ H_0 : \theta = \theta_0 \quad &\text{against} \quad H_1 : \theta \neq \theta_0. \end{aligned}$$

Surely, there could more complicated examples, but we will focus on these three.

2.1 UMP test for $H_0 : \theta \leqslant \theta_1$ or $\theta \geqslant \theta_2$ against $H_1 : H_0$ is false

Consider the one-parameter exponential family:

$$f(\mathbf{x} : \theta) = c(\theta)h(\mathbf{x})\exp\{\eta(\theta)T(\mathbf{x})\},\ \theta \in \Theta \subseteq \mathbb{R}. \tag{2.3}$$

We assume that $\eta(\theta)$ is a strictly increasing function in θ and hence the family has MLR in $T(\mathbf{X})$. This representation is slightly different from the one used in class, but we will use this for convenience.

Theorem 3. *For testing the hypothesis $H_0 : \theta \leqslant \theta_1$ or $\theta \geqslant \theta_2$, (with $\theta_1 < \theta_2 \in \Theta$) against $H_1 : \theta_1 < \theta < \theta_2$, in the one-parameter exponential family given in (2.3), there exists a UMP test given by,*

$$\phi_\alpha(\mathbf{x}) = \begin{cases} 1 & \text{if } c_1 < T(\mathbf{x}) < c_2, \text{ with } c_1 < c_2, \\ \gamma_i & \text{if } T(\mathbf{x}) = c_i,\ i = 1, 2, \\ 0 & o.w. \end{cases}, \tag{2.4}$$

where the c_i's and γ_i's are determined by

$$\mathbf{E}_{\theta_1}\phi_\alpha(\mathbf{X}) = \mathbf{E}_{\theta_2}\phi_\alpha(\mathbf{X}) = \alpha. \tag{2.5}$$

This test ϕ_α, minimizes $\mathbf{E}_\theta\phi(\mathbf{X})$, for all $\theta < \theta_1$ and $\theta > \theta_2$, among all tests ϕ which satisfy the size conditions (2.5).

Proof. We assume $\alpha \in (0, 1)$. Consider all points of the form

$$A = \{(\mathbf{E}_{\theta_1}\psi(\mathbf{X}), \mathbf{E}_{\theta_2}\psi(\mathbf{X})) : \psi \text{ is any test function}\}.$$

Then, A consists of all possible size and power values for the testing problem $K_0 : \theta = \theta_1$ against $K_1 : \theta = \theta_2$, with $\theta_1 < \theta_2$. Since $f(\mathbf{x} : \theta_1)$ and $f(\mathbf{x} : \theta_2)$ are distinct, the MP test of size α for K_0 against K_1 must have power $> \alpha$. Hence, A must contain points[3] of the form (α, u_1) and (α, u_2) where $u_1 < \alpha < u_2$, for any $\alpha \in (0, 1)$. Hence, the point (α, α) is an interior point of the set A.

Fix any $\theta' \in (\theta_1, \theta_2)$. Using part (d) of GNP Lemma it follows, there exists constants k_1, k_2, and a test function ϕ_0 which satisfies (2.5), maximizes $\int \phi_0 f_{\theta'}$, and can be written as,

$$\phi_0(\mathbf{x}) = \begin{cases} 1 & \text{if } f(\mathbf{x} : \theta') > k_1 f(\mathbf{x} : \theta_1) + k_2 f(\mathbf{x} : \theta_2), \\ 0 & \text{if } f(\mathbf{x} : \theta') < k_1 f(\mathbf{x} : \theta_1) + k_2 f(\mathbf{x} : \theta_2). \end{cases}$$

Using (2.3), we can simplify the rejection region of ϕ_0 as,

$$\begin{aligned} & f(\mathbf{x} : \theta') > k_1 f(\mathbf{x} : \theta_1) + k_2 f(\mathbf{x} : \theta_2) \\ \Leftrightarrow\ & c(\theta')h(\mathbf{x})\exp\{\eta(\theta')T(\mathbf{x})\} > k_1 c(\theta_1)h(\mathbf{x})\exp\{\eta(\theta_1)T(\mathbf{x})\} + k_2 c(\theta_2)h(\mathbf{x})\exp\{\eta(\theta_2)T(\mathbf{x})\} \\ \Leftrightarrow\ & c(\theta')\exp\{\eta(\theta')T(\mathbf{x})\} > k_1 c(\theta_1)\exp\{\eta(\theta_1)T(\mathbf{x})\} + k_2 c(\theta_2)\exp\{\eta(\theta_2)T(\mathbf{x})\},\ (\text{since } h(\mathbf{x}) > 0) \\ \Leftrightarrow\ & 1 > \frac{k_1 c(\theta_1)}{c(\theta')}\exp\{(\eta(\theta_1) - \eta(\theta'))T(\mathbf{x})\} + \frac{k_2 c(\theta_2)}{c(\theta')}\exp\{(\eta(\theta_2) - \eta(\theta'))T(\mathbf{x})\} \\ \Leftrightarrow\ & 1 > a_1 e^{b_1 T} + a_2 e^{b_2 T}, \end{aligned}$$

[3] This is due to an earlier result (which states that size-power diagram consists of only the line $y = x$ only if the distributions under the null and alternative are same).

where, $b_1 = \eta(\theta_1) - \eta(\theta') < 0$, $b_2 = \eta(\theta_2) - \eta(\theta') > 0$ (due to strictly increasing η), $a_1 = k_1 c(\theta_1)/c(\theta')$ and $a_2 = k_2 c(\theta_2)/c(\theta')$. Hence, in terms of the statistic T,

$$\phi_0(\mathbf{x}) = \phi_0(T(\mathbf{x})) = \phi_0(t) = \begin{cases} 1 & \text{if } a_1 e^{b_1 t} + a_2 e^{b_2 t} < 1, \\ 0 & \text{if } a_1 e^{b_1 t} + a_2 e^{b_2 t} > 1. \end{cases}$$

We will simplify the rejection region of $\phi_0(t)$. Define, $g(t) = a_1 e^{b_1 t} + a_2 e^{b_2 t}$, for all $t \in \mathbb{R}$. Note, $g'(t) = a_1 b_1 e^{b_1 t} + a_2 b_2 e^{b_2 t}$, for all $t \in \mathbb{R}$. We claim, $a_1, a_2 > 0$.

(a) If $a_1, a_2 \leqslant 0$, then $g(t) \leqslant 0$, for all t. Hence, $\phi_0(t) = 1$, for all t, and ϕ_0 will have size 1. This violates the assumption $\alpha \in (0, 1)$.

(b) If $a_1 < 0$ and $a_2 \geqslant 0$, then $a_1 b_1 > 0$ and $a_2 b_2 \geqslant 0$. This implies, $g'(t) > 0$, for all t. Hence, $g(t)$ is strictly increasing. Hence, for some t_0 (as g is continuous), $\{t : g(t) < 1\} = (-\infty, t_0)$, which implies

$$\phi_0(t) = \begin{cases} 1 & \text{if } t < t_0, \\ 0 & \text{if } t > t_0, \end{cases}$$

which resembles the test function obtained using MLR for a one-sided testing problem (the family (2.3) has MLR in $T(\mathbf{X})$). We know, power functions of such tests are strictly decreasing, hence such a ϕ_0 cannot satisfy (2.5) at two points $\theta_1 < \theta_2$. A similar argument works if the signs of a_1, a_2 are reversed. As a result, opposite signs of a_i's are not possible.

(c) So, the only possible option is, $a_1, a_2 > 0$. If we assume $a_1 = 0$ and $a_2 > 0$ (or vice-versa), then, $g(t) = a_2 e^{b_2 t}$, is strictly monotone and raises same issues as in part (b) above.

Now, $g'() = 0$, has an unique solution at $t^* = (b_2 - b_1)^{-1} \log(-a_1 b_1 / a_2 b_2)$, and $g''(t) = a_1 b_1^2 e^{b_1 t} + a_2 b_2^2 e^{b_2 t} > 0$, for all t. Hence, t^* is a global minima of g. Thus,[4] $\{t : g(t) < 1\} = \{t : c_1 < t < c_2\}$, for some choice of $c_1 < c_2$. On the equality part, we assign two values γ_1, γ_2 such that,

$$\phi_0(\mathbf{x}) = \phi_0(t) = \begin{cases} 1 & \text{if } c_1 < t < c_2, \\ \gamma_i & \text{if } t = c_i,\ i = 1, 2, \\ 0 & \text{if } t < c_1 \text{ or } t > c_2, \end{cases}$$

satisfies the size conditions (2.5). Since, $c(\theta)$ (cf. (2.3)) is non-negative, and since $a_i > 0$, $i = 1, 2$, it implies $k_1, k_2 > 0$. Hence, part (c) of GNP Lemma implies, $\phi_0(t)$ will maximize $\mathbf{E}_{\theta'}\phi$, among all level-α tests ϕ satisfying, $\mathbf{E}_{\theta_i}\phi \leqslant \alpha$, $i = 1, 2$. Applying (2.5), it is clear the choices of c_i, γ_i, $i = 1, 2$, will only depend on θ_1 and θ_2, and will not depend on θ'. However, the unknown k_1, k_2, used in the original formulation of ϕ_0, will indeed depend on θ'. This follows by noting that (c_1, c_2) depends on the solution of $g(t) = 1$ and the choice of α, which in turn depends on (a_i, b_i), $i = 1, 2$, which in turn depends on $c(\theta_i), c(\theta'), \eta(\theta_i), \eta(\theta')$, $i = 1, 2$. Apparently, this gives the impression that c_i's indeed should depend on θ'. But, a corresponding change in k_i's balances out the changes made in the values of b_i's and a_i's (at some other θ''), so that c_i's are not affected, if the maximization of power of ϕ_0 was attempted at an alternate parameter value θ''. In a nutshell, the regions $\{f(\mathbf{x} : \theta') > \sum_{i=1}^{2} k_i f(\mathbf{x} : \theta_i)\}$, change as k_1, k_2 depend on θ', and hence $\phi_0(t)$ has different representations at the triplets, $\{(\theta', k_1(\theta'), k_2(\theta')) : \theta' \in (\theta_1, \theta_2)\}$, but the overall function ϕ_0 remains same over its domain.

[4] If, $\{t : g(t) < 1\} = \varnothing$, then $\mathbf{E}_\theta \phi_0(T) = 0$, for all θ. This violates the assumption of $\alpha \in (0, 1)$.

Now, consider the case where $\theta_* < \theta_1$. We want to find a test function ϕ_*, among all test functions ϕ, such that $\mathbf{E}_{\theta_i}\phi_(\mathbf{X}) = \alpha$, for $i = 1, 2$, and $\mathbf{E}_{\theta_*}\phi(\mathbf{X})$ is minimized. Applying the GNP in reverse, the test that will minimize $\mathbf{E}_{\theta_*}\phi(\mathbf{X})$, will be of the form

$$\phi_*(\mathbf{x}) = \begin{cases} 0 & \text{if } f(\mathbf{x} : \theta_*) > d_1 f(\mathbf{x} : \theta_1) + d_2 f(\mathbf{x} : \theta_2), \\ 1 & \text{if } f(\mathbf{x} : \theta_*) < d_1 f(\mathbf{x} : \theta_1) + d_2 f(\mathbf{x} : \theta_2), \end{cases}$$

where, d_1, d_2 will be constants specified by the size conditions. We can write,

$$\begin{aligned} & f(\mathbf{x} : \theta_*) > d_1 f(\mathbf{x} : \theta_1) + d_2 f(\mathbf{x} : \theta_2) \\ \Leftrightarrow\ & c(\theta_*) \exp\{\eta(\theta_*)T(\mathbf{x})\} > d_1 c(\theta_1) \exp\{\eta(\theta_1)T(\mathbf{x})\} + d_2 c(\theta_2) \exp\{\eta(\theta_2)T(\mathbf{x})\} \\ \Leftrightarrow\ & \frac{c(\theta_*)}{d_1 c(\theta_1)} \exp\big(T(\mathbf{x})\{\eta(\theta_*) - \eta(\theta_1)\}\big) > 1 + \frac{d_2 c(\theta_2)}{d_1 c(\theta_1)} \exp\big(T(\mathbf{x})\{\eta(\theta_2) - \eta(\theta_1)\}\big) \\ \Leftrightarrow\ & \bar{a}_1 e^{\bar{b}_1 T(\mathbf{x})} + \bar{a}_2 e^{\bar{b}_2 T(\mathbf{x})} > 1, \end{aligned}$$

where, $\bar{b}_1 = \eta(\theta_*) - \eta(\theta_1) < 0$, $\bar{b}_2 = \eta(\theta_2) - \eta(\theta_1) > 0$ and $\bar{a}_i$, $i = 1, 2$, are constants. Hence, $\phi_*(t)$ is,

$$\phi_*(t) = \begin{cases} 0 & \text{if } \bar{a}_1 e^{\bar{b}_1 T(\mathbf{x})} + \bar{a}_2 e^{\bar{b}_2 T(\mathbf{x})} > 1, \\ 1 & \text{if } \bar{a}_1 e^{\bar{b}_1 T(\mathbf{x})} + \bar{a}_2 e^{\bar{b}_2 T(\mathbf{x})} < 1. \end{cases}$$

Write, $\bar{g}(t) = \bar{a}_1 e^{\bar{b}_1 t} + \bar{a}_2 e^{\bar{b}_2 t}$, for all $t \in \mathbb{R}$. If, $\bar{a}_1, \bar{a}_2 < 0$, then $\bar{g}(t) < 0$, and $\phi_*(t) = 1$ for all t. This will violate the size condition $\alpha \in (0, 1)$. If, $\bar{a}_1$ and $\bar{a}_2$ have opposite signs, then, $\bar{g}'(t)$ will be either strictly positive or strictly negative, for all t. Hence, $\{t : \phi_*(t) = 1\} = (-\infty, t_0)$ or (t_0, ∞), for some t_0. Using the argument used in part (b) earlier, the power function of ϕ_* will be strictly monotone and will violate the size conditions. Carefully argue the cases where one or both of the $\bar{a}_i$'s can be zero. So by the existence part of GNP lemma (part (iv)), there must exist such constants $\bar{a}_1$ and $\bar{a}_2$, and (by excluding all other possibilities) these constants must satisfy, $\bar{a}_1, \bar{a}_2 > 0$. Then, $\bar{g}'(t)$ will have a global minima at some point $\bar{t}_*$. Hence, $\{t : \bar{g}(t) < 1\} = \{t : \bar{c}_1 < t < \bar{c}_2\}$, for some constants $\bar{c}_1, \bar{c}_2$, which are determined by the size conditions.

The same argument works if we choose any $\theta_* > \theta_2$ and try to construct a test which satisfies the size conditions and minimizes power at θ_*. We will obtain a similar rejection region for that test. Since the size conditions used for finding c_i and $\bar{c}_i$, $i = 1, 2$, are same, the tests $\phi_0(t)$ and $\phi_*(t)$ (and the one that you can develop for the case of $\theta_* > \theta_2$, mentioned above) will be the same. This implies, the test $\phi_0(t)$ maximizes power at any $\theta \in (\theta_1, \theta_2)$ and minimizes power at any $\theta \in (-\infty, \theta_1) \cup (\theta_2, \infty)$, among all tests which satisfy the size conditions $\mathbf{E}_{\theta_i}\phi(\mathbf{X}) = \alpha$, $i = 1, 2$. Now, comparing $\phi_0(t)$ with the trivial test $\phi_\alpha(t) = \alpha$, for all t, we have,

$$\mathbf{E}_\theta \phi_0(T) \leqslant \alpha = \mathbf{E}_\theta \phi_\alpha(T), \quad \text{for all } \theta \notin [\theta_1, \theta_2].$$

This implies, ϕ_0 is size α, *i.e.*, $\sup_{\theta \notin (\theta_1, \theta_2)} \mathbf{E}_\theta \phi_0(T) = \alpha$. We still need to show that ϕ_0 is the best test among all level α tests.

If $\psi(\mathbf{x})$ denotes a level α test, then $\mathbf{E}_{\theta_i}\psi(\mathbf{X}) \leqslant \alpha$, for $i = 1, 2$. Since $k_1, k_2 > 0$ (for any $\theta' \in (\theta_1, \theta_2)$), using part (c) of GNP Lemma, we can claim ϕ_0 will provide the highest power at θ', among all tests ψ, which satisfy level α condition at θ_1 and θ_2. This is true for all $\theta' \in (\theta_1, \theta_2)$ (using various choices of k_1, k_2 depending on θ'). Thus, $\mathbf{E}_{\theta'}\phi_0(\mathbf{X}) \geqslant \mathbf{E}_{\theta'}\psi(\mathbf{X})$, for all $\theta' \in (\theta_1, \theta_2)$, among all ψ satisfying $\sup_{\theta \notin (\theta_1, \theta_2)} \mathbf{E}_\theta \psi(\mathbf{X}) \leqslant \alpha$. Hence, $\phi_0(T) = \phi_0(\mathbf{X})$, is the required UMP level α test. □

Remark. It should be noted the power function for $\phi_0(\mathbf{x})$ is,

$$\beta_{\phi_0}(\theta) = \int \phi_0(\mathbf{x}) \cdot c(\theta) \exp\{\eta(\theta)T(\mathbf{x})\}h(\mathbf{x})\ d\mathbf{x}.$$

Using a result about exponential families (cf. Theorem 2.7.1 of Lehmann and Romano (2005)), it follows that $\theta \mapsto \beta_{\phi_0}(\theta)$ is continuous. Infact, the power function $\beta_\phi(\theta)$ has a maxima at some point $\theta_* \in (\theta_1, \theta_2)$, and keeps decreasing as $|\theta - \theta_*|$ increases, except if there exists two values t_1, t_2, such that $P_\theta(T = t_1) + P_\theta(T = t_2) = 1$, for all θ. The proof of this statement is given in Lehmann and Romano (2005) (cf. Theorem 3.7.1). Another useful text which discusses these results is Schervish (1995).

Remark. It is to be noted, the whole proof hinges crucially on equality of power ($= \alpha$) at both θ_1 and θ_2. Typically, one can always decide to choose different values at θ_i. The arguments will not hold in that case. For the case of $\alpha = 0$, we need to develop arguments separately, but typically in all texts the proofs are based (but never stated explicitly) on the assumption that, $\alpha \in (0, 1)$.

Remark. The solutions of c_i, γ_i, $i = 1, 2$, are found by trial and error. If you consider the second and third hypothesis in the end of page 1, there will be no UMP tests. But, there will be UMPU tests. However, even though Theorem 3 provides an useful test which satisfies nice optimality properties, it has very narrow applicability. Notice, for simple against simple testing, NP lemma was usable for any pairs of hypothesis, provided the densities had either pdf or pmf. In case of one-sided hypothesis of the form $H_0 : \theta \leqslant \theta_0$, firstly we narrowed down the parameter space to a subset of the real line and we also required the MLR property to hold. More complicated hypothesis, like the one discussed in Theorem 3, requires even more stringent assumption, the family of distributions forms an one-parameter exponential family.

Example 5. Let $\{X_1, \ldots, X_n\}$ be an i.i.d. sample from $N(\mu, 1)$. We want to test, $H_0 : \mu \leqslant \mu_1$ or $\mu \geqslant \mu_2$, against $H_1 : \mu \in (\mu_1, \mu_2)$. It can be easily checked that the family of distributions of $\{X_1, \ldots, X_n\}$ have MLR in $T(\mathbf{X}) = \bar{X}_n$. Following the theorem, the UMP size α test is,

$$\phi(\mathbf{x}) = \begin{cases} 1 & \text{if } c_1 < \bar{x}_n < c_2, \\ 0 & \text{o.w.} \end{cases}$$

The constants c_1, c_2 are found by using, $P_{\mu=\mu_i}\big(c_1 < \bar{X}_n < c_2\big) = \alpha$, for $i = 1, 2$. This leads to two equations,

$$\Phi\big(\sqrt{n}(c_2 - \mu_1)\big) - \Phi\big(\sqrt{n}(c_1 - \mu_1)\big) = \alpha = \Phi\big(\sqrt{n}(c_2 - \mu_2)\big) - \Phi\big(\sqrt{n}(c_1 - \mu_2)\big) = \alpha.$$

Let us use, $\alpha = 0.1$, $\mu_1 = -1$, $\mu_2 = 0.5$ and $n = 16$. Then we have,

$$\Phi\big(4c_2 + 4\big) - \Phi(4c_1 + 4) = 0.1 = \Phi(4c_2 - 2) - \Phi(4c_1 - 2).$$

After a fair amount of effort with the computer, we obtain, $c_1 = -0.679615$ and $c_2 = 0.1797$. The test function is, $\phi(\mathbf{x}) = \mathbf{1}(-0.679615 < \bar{x}_n < 0.1797)$. The power is, $\beta_\phi(\mu) = \Phi\left(4\big(0.1797 - \mu\big)\right) - \Phi\left(4\big(-0.679615 - \mu\big)\right)$. Using MLR theory, the size α ($= 0.1$) UMP tests for $K_0 : \mu \leqslant -1$ against $K_1 : \mu > -1$ and $K_0 : \mu \geqslant 1/2$ against $K_1 : \mu < -1/2$ are, $\phi_L(\mathbf{x}) = \mathbf{1}(\bar{x}_n > -0.6796121)$ and $\phi_R(\mathbf{x}) = \mathbf{1}(\bar{x}_n < 0.1796121)$, respectively. The power functions of ϕ (blue), ϕ_R (black) and ϕ_L (green) are plotted in Figure 1.

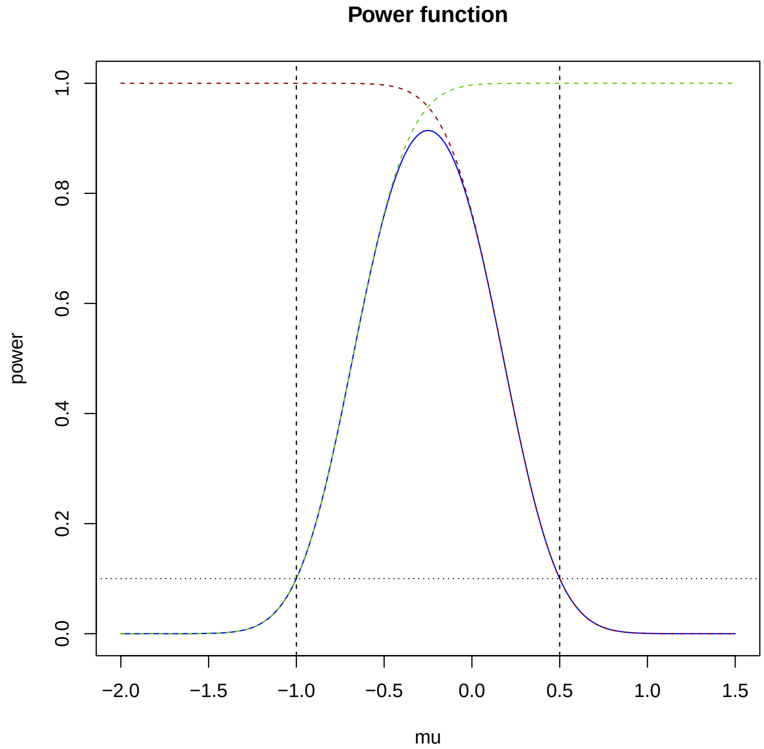

Figure 1: Power function of the UMP test found in Example 1 and UMP tests for one-sided hypothesis.

Example 6. $\{X_1, \ldots, X_n\}$ are i.i.d. $f(x : \theta) = \theta e^{-\theta x} \cdot \mathbf{1}(x > 0)$, and $\theta \in (0, \infty)$. We want a size α UMP test for $H_0 : \theta \leqslant 1$ or $\theta \geqslant 2$, against $H_1 : H_0$ is false.

Example 7. $\{X_1, \ldots, X_n\}$ are i.i.d. binomial $(1, \theta)$, and $\theta \in (0, 1)$. We want a size α UMP test for $H_0 : \theta \leqslant 1/4$ or $\theta \geqslant 3/4$, against $H_1 : H_0$ is false. In this case, depending on the choice of α and n, we might need to randomize. For the particular choice of $n = 10$ and $\alpha = 0.1$, find the required test and plot the power function.

2.2 UMPU tests for two-sided hypothesis

Firstly, we define unbiased and UMPU tests.

Definition 2 (Unbiased test and UMPU test)**.** A test ϕ for testing $H_0 : \theta \in \Theta_0$ against $H_1 : \theta \in \Theta_1$, where $\Theta_0 \cap \Theta_1 = \varnothing$, is a size α **unbiased test** if,

$$\sup_{\theta \in \Theta_0} \mathbf{E}_\theta \phi(\mathbf{X}) = \alpha \quad \text{and} \quad \mathbf{E}_{\theta_1} \phi(\mathbf{X}) \geqslant \alpha, \quad \text{for all } \theta_1 \in \Theta_1.$$

It is a level α unbiased test, if $\sup_{\theta \in \Theta_0} \mathbf{E}_\theta \phi(\mathbf{X}) \leqslant \alpha$. A test ϕ is called **Uniformly Most Powerful Unbiased (UMPU)** size α test for H_0 against H_1, if it is UMP among all unbiased tests of level[5] α.

[5] Let ψ be a particular unbiased level-α test, with size $= \sup_{\theta \in \Theta_0} \mathbf{E}_\theta \psi = \alpha/2$ $(< \alpha)$, but $\mathbf{E}_\theta \psi \geqslant \alpha$, for all $\theta \in \Theta_1$. So, 'unbiased at level-α' condition ensures the power never drops below α at alternatives (it does not mean power never drops (at

Now consider the two hypothesis,

$$H_0 : \theta \in [\theta_1, \theta_2] \quad \text{v/s} \quad H_1 : H_0 \text{ is false} \quad \text{and} \quad H_0 : \theta = \theta_0 \quad \text{v/s} \quad H_1 : \theta \neq \theta_0. \tag{2.6}$$

where, $\theta_1 < \theta_2$ and θ_0 are points in the parameter space $\Theta \subset \mathbb{R}$. It is not possible to construct[6] UMP size-α tests, for *all* one-parameter exponential family distributions (cf. (2.3)), for testing the two types of null hypothesis in (2.6). Instead, UMPU tests are available.

2.2.1 Why UMP test does not exist for $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$

We begin with a simple result on uniqueness of MP tests (the proof is available in Lemma 4.38 of Schervish (1995)), which is similar in nature to the last assertion of NP-Lemma. Consider testing $H_0 : P = P_0$ against $H_1 : P = P_1$, where both are simple hypothesis. Let f_i denote the density of P_i, $i = 0, 1$, and ϕ_0 is the MP size α test based on NP-Lemma. We assume,

$$P_i(f_1(\mathbf{X}) = kf_0(\mathbf{X})) = 0, \quad \text{for } i = 0, 1, \text{ and any } k \in [0, \infty].$$

Then, for any other size α test ϕ_*, either ϕ_* has lower power than ϕ_0 or ϕ_* is same (with probability 1 under both P_i) under ϕ_0. Note, since the *equality* part has probability zero under both P_i, hence the conclusion is more stronger than the uniqueness assertion of NP-Lemma (which does not say anything about the *equality* region). The idea is intuitively clear, any other test with same power (and same size), must match the MP test ϕ_0, if equality regions have zero probability.

Now, consider the case of testing $H_0 : \theta = \theta_0$ against $H_1 : \theta > \theta_0$, and assume that the underlying family $\{f(\mathbf{x} : \theta) : \theta \in \Theta\}$ with $\Theta \subset \mathbb{R}$, has MLR in $T(\mathbf{X})$. We know, the UMP size α test will be of the form, $\phi_0(\mathbf{x}) = \mathbf{1}(T(\mathbf{x}) > c) + \gamma \cdot \mathbf{1}(T(\mathbf{x}) = c)$. Assume,

$$P_\theta\left(f(\mathbf{X} : \theta) = kf(\mathbf{X} : \theta_0)\right) = 0, \quad \text{for all } \theta > \theta_0 \text{ and all } k \in [0, \infty]. \tag{2.7}$$

This is an extension of the earlier assumption, for all alternatives $\theta > \theta_0$. This will be true for one-parameter exponential families, if $T(\mathbf{x})$ has continuous distribution under all θ. As earlier, any other size α test ϕ_*, must have either lower power than ϕ_0 at all alternatives, or ϕ_* must match ϕ_0, *i.e.*, $P_{\theta_1}(\phi_* = \phi_0) = 1$, for each $\theta_1 > \theta_0$. This follows because, ϕ_0 itself is MP size α for $H_0 : \theta = \theta_0$ against $K_1 : \theta = \theta_1$, for each $\theta_1 > \theta_0$. So, any MLR based UMP test for a one-sided alternative hypothesis of the form given in H_1 is essentially unique. Same argument works if we consider $H_1 : \theta < \theta_0$.

Now, consider the case of one-parameter exponential family (cf. (2.3)) with MLR in $T(\mathbf{X})$, $H_0 : \theta = \theta_0$, and the alternative hypotheses, $K_1 : \theta > \theta_0$ and $K_2 : \theta < \theta_0$. Assume, (2.7) holds for all $\theta \neq \theta_0$. Suppose, ϕ_0 is the UMP size α test for H_0 against $H_1 : \theta \neq \theta_0$. Using MLR theory, there are UMP size α tests ϕ_1 and ϕ_2 for testing H_0 against K_1 and K_2, respectively. Note, ϕ_1 and ϕ_2 are also size α (hence level α) tests for H_0 against H_1. But, ϕ_0 is UMP for H_0 against H_1, so

$$\beta_{\phi_0}(\theta) \geqslant \beta_{\phi_1}(\theta) \quad \text{for } \theta > \theta_0, \quad \text{and} \quad \beta_{\phi_0}(\theta) \geqslant \beta_{\phi_2}(\theta) \quad \text{for } \theta < \theta_0. \tag{2.8}$$

alternatives) below the actual size of ψ, which can be $\alpha/2$). If the power function of ψ is continuous and there is a common boundary θ_* between Θ_0 and Θ_1, unbiasedness implies $\mathbf{E}_{\theta_*}\psi \geqslant \alpha$, but ψ is size $\alpha/2$, hence there must be a jump in the power function at θ_*. This is impossible if $\beta_\psi(\theta)$ is continuous. Hence in this particular situation, all such tests are effectively ruled out (from the class of level-α unbiased tests). But, if the power function is not necessarily continuous, tests like ψ will remain members of the class of level-α unbiased tests.

[6] This non-existence also holds for slightly general families.

Conversely, ϕ_0 will also[7] be a level α test for H_0 against K_1. So, by the previous discussion, we must have either

$$P_\theta\big(\phi_0 = \phi_1\big) = 1, \quad \text{for all } \theta > \theta_0,$$

or, there is a $\theta_1 > \theta_0$, such that ϕ_0 has strictly lower power than ϕ_1 at θ_1. The second possibility is contradicted by our earlier statement in (2.8). Similarly, we can conclude,

$$P_\theta\big(\phi_0 = \phi_2\big) = 1, \quad \text{for all } \theta < \theta_0.$$

Assume, $P_\theta(A) = 0$ for all $\theta < \theta_0$ (or $> \theta_0$), implies, $P_\theta(A) = 0$, for all θ. This is again true if $\mathbf{X}$ has same support under all $\theta \in \Theta$ (for example: Normal, exponential distributions). As a result,

$$P_\theta(\phi_1 = \phi_2) = 1, \quad \text{for all } \theta.$$

But, $\phi_1(\mathbf{x}) = \mathbf{1}(T(\mathbf{x}) > c_1)$ and $\phi_2(\mathbf{x}) = \mathbf{1}(T(\mathbf{x}) < c_2)$, using MLR theory. This is impossible for any finite c_1, c_2. Hence, there is no UMP size α test for $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$.

A similar, but possibly more complicated argument could be carried out to show that[8] in general no UMP size α test exists for $H_0 : \theta \in [\theta_1, \theta_2]$ against $H_1 : \theta \notin [\theta_1, \theta_2]$.

Remark. The real cause of not finding an UMP test for the two sided null $H_0 : \theta = \theta_0$ is there are two different best size α tests ϕ_1 and ϕ_2 on the regions (alternatives) $\theta > \theta_0$ and $\theta < \theta_0$, respectively. But, both of them are the worst in terms of power, on the regions $\theta < \theta_0$ and $\theta > \theta_0$, respectively. Our requirement is to have best power on any $\theta \neq \theta_0$. Thus, we can constraint our search on tests which atleast have power $\geqslant \alpha$, for all θ. For such tests, the power function will be higher on the alternative, than on the null.

Remark. Possibly, there could be UMP test for $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$, in some other families (and indeed there are). The above discussion shows an example about a case where such UMP tests will not exist. It uses quite a few assumptions (zero probability in equality region, continuous distribution of $\mathbf{X}$, and possibly same support of $\mathbf{X}$ under all parameters), but it suffices to provide a counterexample against existence of such tests. Hence, we could not hope to find a general theory that works for all one-parameter exponential families. Instead, we can attempt to develop a theory for UMPU tests in such cases.

2.2.2 UMPU tests and α-similar tests

In case of exponential families given in (2.3), power functions of any test are differentiable (and hence continuous) everywhere in the interior of the parameter space for θ (cf. Theorem 2.7.1 of Lehmann and Romano (2005)). This will be used to construct UMPU tests.

We define the notion of α-**similar** tests. Consider any testing problem, $H_0 : \theta \in \Theta_0$ against $H_1 : \theta \in \Theta_1$. Write,

$$\bar{\Theta}_i = \text{closure of } \Theta_i = \Theta_i \cup \big(\text{limit points of } \Theta_i\big), \quad i = 0, 1.$$

Also write, $\bar{\Theta}_{0,1} = \bar{\Theta}_0 \cap \bar{\Theta}_1$.

(i) If, $\Theta_0 = \{\theta_0\}$ and $\Theta_1 = \mathbb{R}\backslash\{\theta_0\}$, then $\bar{\Theta}_0 = \{\theta_0\}$, $\bar{\Theta}_1 = \mathbb{R}$ and $\bar{\Theta}_{0,1} = \{\theta_0\}$.

[7] Since ϕ_1 is UMP for this alternative, $\beta_{\phi_0}(\theta) \leqslant \beta_{\phi_1}(\theta)$, for all $\theta > \theta_0$. Hence, $\beta_{\phi_0}(\theta) = \beta_{\phi_1}(\theta)$, for all $\theta > \theta_0$.

[8] I have not worked it out, but one can think of one-sided tests for the nulls $K_0 : \theta \geqslant \theta_1$ and $M_0 : \theta \leqslant \theta_2$, and hopefully use similar arguments. Note, in case $\theta_1 = \theta_2 = \theta_0$, this reduces to the point null hypothesis $H_0 : \theta = \theta_0$.

(ii) If, $\Theta_0 = [\theta_1, \theta_2]$, $\Theta_1 = (-\infty, \theta_1) \cup (\theta_2, \infty)$, then $\bar{\Theta}_0 = [\theta_1, \theta_2]$, $\bar{\Theta}_1 = (-\infty, \theta_1] \cup [\theta_2, \infty)$ and $\bar{\Theta}_{0,1} = \{\theta_1, \theta_2\}$.

Definition 3 (Similar tests)**.** A test ϕ is called α-similar on $\bar{\Theta}_{0,1}$ (assuming this set is non-empty) if,

$$\mathbf{E}_\theta \phi(\mathbf{X}) = \alpha, \quad \text{for all } \theta \in \bar{\Theta}_{0,1}. \tag{2.9}$$

The next lemma establishes a connection between α-similar tests and UMPU tests.

Lemma 4. *Consider testing $H_0 : \theta \in \Theta_0$ against $H_1 : \theta \in \Theta_1$. Suppose, for every test function ϕ, the power function $\beta_\phi(\theta)$ is a continuous in θ. Then, if ϕ_0 is the UMP test among all tests satisfying (2.9) and having level α, then ϕ_0 will be the UMPU size α test.*

Proof of Lemma 4. Consider the class of level α and α-similar tests,

$$\mathcal{C}_1 = \left\{ \phi : \mathbf{E}_\theta \phi(\mathbf{X}) = \alpha, \text{ for all } \theta \in \bar{\Theta}_{0,1}, \sup_{\theta \in \Theta_0} \mathbf{E}_\theta \phi(\mathbf{X}) \leqslant \alpha \right\}. \tag{2.10}$$

Also define the class of level α unbiased tests,

$$\mathcal{C}_2 = \left\{ \phi : \sup_{\theta \in \Theta_0} \mathbf{E}_\theta \phi(\mathbf{X}) \leqslant \alpha, \ \mathbf{E}_\theta \phi(\mathbf{X}) \geqslant \alpha, \text{ for all } \theta \in \Theta_1 \right\}.$$

Claim: $\mathcal{C}_2 \subseteq \mathcal{C}_1$. Suppose, $\theta_* \in \bar{\Theta}_{0,1}$. Then, there exists sequences $\{\theta_{j,m} : m \geqslant 1\} \in \Theta_j$, $j = 0, 1$, such that, $\theta_{j,m} \to \theta_*$, as $m \to \infty$. By continuity assumption, for any $\phi \in \mathcal{C}_2$, level α assumption implies, $\alpha \geqslant \beta_\phi(\theta_{0,m}) \to \beta_\phi(\theta_*)$. Hence, $\beta_\phi(\theta_*) \leqslant \alpha$. Similarly, the unbiasedness condition implies, $\alpha \leqslant \beta_\phi(\theta_{1,m}) \to \beta_\phi(\theta_*)$. Hence, $\beta_\phi(\theta_*) \geqslant \alpha$. This implies, $\beta_\phi(\theta_*) = \alpha$. Hence, ϕ is α-similar and $\phi \in \mathcal{C}_1$.

Now, ϕ_0 is the best test in $\mathcal{C}_1$. But, the trivial test $\phi_1(\mathbf{x}) = \alpha$, is also a member of $\mathcal{C}_1$. But, as ϕ_0 is UMP in $\mathcal{C}_1$, $\mathbf{E}_\theta \phi_0(\mathbf{X}) \geqslant \alpha$, and hence ϕ_0 is unbiased. So, $\phi_0 \in \mathcal{C}_2$. So, the best level α and α-similar test is actually an unbiased test. Hence, ϕ_0 is the UMPU test among all level α tests. □

2.2.3 UMPU test for $H_0 : \theta \in [\theta_1, \theta_2]$ against $H_1 : H_0$ is false

Consider finding an UMPU size α test for $H_0 : \theta \in [\theta_1, \theta_2]$ against $H_1 : \theta \notin [\theta_1, \theta_2]$, when the underlying family is given by (2.3). For any test ϕ, $\beta_\phi(\theta) = \mathbf{E}_\theta \phi(\mathbf{X})$, will be continuous in θ. Hence, Lemma 4 will be applicable.

Lemma 5. *Consider the family of distributions in (2.3). For testing $H_0 : \theta \in [\theta_1, \theta_2]$ against $H_1 : H_0$ is false, the UMPU size α test is given by,*

$$\psi_\alpha(\mathbf{x}) = \begin{cases} 1 & \text{if } T(\mathbf{x}) < c_1 \text{ or } T(\mathbf{x}) > c_2 \\ \gamma_i & \text{if } T(\mathbf{x}) = c_i,\ i = 1, 2, \\ 0 & \text{if } c_1 < T(\mathbf{x}) < c_2, \end{cases} \tag{2.11}$$

where, c_i, γ_i, $i = 1, 2$, are found by using the size conditions,

$$\mathbf{E}_{\theta_i} \psi_\alpha(\mathbf{X}) = \alpha, \quad \text{for } i = 1, 2. \tag{2.12}$$

Proof of Lemma 5. Note, in this case $\bar{\Theta}_{0,1} = \{\theta_1, \theta_2\}$. Let $\phi_{1-\alpha}$ denote the UMP test found in Theorem 3 (cf. (2.4)) for $K_0 : \theta \leqslant \theta_1$ or $\theta \geqslant \theta_2$ against $H_1 : \theta \in (\theta_1, \theta_2)$, with size $= (1 - \alpha)$ (cf. (2.5)). Let,

$\psi_\alpha = 1 - \phi_{1-\alpha}$. Then, $\mathbf{E}_{\theta_i}\psi_\alpha = 1 - \mathbf{E}_{\theta_i}\phi_{1-\alpha} = 1 - (1-\alpha) = \alpha$, for $i = 1, 2$. Hence, ψ_α is α-similar. Also, $\phi_{1-\alpha}$ is unbiased (as it is UMP of its size), hence

$$\sup_{\theta \in [\theta_1, \theta_2]} \mathbf{E}_\theta \psi_\alpha = 1 - \inf_{\theta \in [\theta_1, \theta_2]} \mathbf{E}_\theta \phi_{1-\alpha} = 1 - (1-\alpha) = \alpha.$$

Thus, ψ_α is size-α. Let, $\psi^\dagger$ be another test for H_0 against H_1 which satisfies, $\mathbf{E}_{\theta_i}\psi^\dagger = \alpha$, for $i = 1, 2$, and[9] $\sup_{\theta \in [\theta_1, \theta_2]} \mathbf{E}_\theta \psi^\dagger \leqslant \alpha$. We want to show, $\mathbf{E}_\theta\psi_\alpha \geqslant \mathbf{E}_\theta\psi^\dagger$, for all $\theta < \theta_1$ and $\theta > \theta_2$.

Consider the test, $\phi^\dagger = 1 - \psi^\dagger$. Then, $\mathbf{E}_{\theta_i}\phi^\dagger = 1 - \alpha$, for $i = 1, 2$, hence $\phi^\dagger$ satisfies the same size condition (2.5), as the test $\phi_{1-\alpha}$. Since both $\phi^\dagger$ and $\phi_{1-\alpha}$ satisfy (2.5) (with $(1-\alpha)$), using the second assertion of Theorem 3, it follows, $\mathbf{E}_\theta\phi^\dagger \geqslant \mathbf{E}_\theta\phi_{1-\alpha}$, for each $\theta < \theta_1$ and $\theta > \theta_2$. Thus, $\mathbf{E}_\theta\psi_\alpha \geqslant \mathbf{E}_\theta\psi^\dagger$, for each $\theta < \theta_1$ and $\theta > \theta_2$. This proves ϕ_α is the required UMP test for H_0 against H_1. □

2.2.4 UMPU test for $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$

Now, we discuss the finding[10] an UMPU size α test for $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$. Here, the power function of every test will be differentiable (again using Theorem 2.7.1 of Lehmann and Romano (2005)) and $\bar{\Theta}_{0,1} = \{\theta_0\}$. We aim to find the UMP size-α test among the class of all unbiased level-α tests,

$$\mathcal{C}_3 = \{\phi : \mathbf{E}_{\theta_0}\phi \leqslant \alpha,\ \mathbf{E}_\theta\phi \geqslant \alpha, \text{ for all } \theta \neq \theta_0\}. \tag{2.13}$$

Let ϕ_α be the UMP size-α test in $\mathcal{C}_3$. As $\Theta_0 = \{\theta_0\}$, $\beta_{\phi_\alpha}(\theta)$ has a minima[11] at $\theta = \theta_0$. Since all power functions are differentiable, ϕ_α satisfies $\beta'_{\phi_\alpha}(\theta_0) = 0$. Using (2.3), for any test function ϕ,

$$\frac{\partial}{\partial\theta}\beta_\phi(\theta) = \beta'_\phi(\theta) = \eta'(\theta) \cdot \mathbf{E}_\theta\big(\phi(\mathbf{X})T(\mathbf{X})\big) + \frac{c'(\theta)}{c(\theta)} \cdot \beta_\phi(\theta). \tag{2.14}$$

If we use, $\phi \equiv \alpha$, then $\beta'_\phi(\theta) = 0$ at all θ. Hence, at $\theta = \theta_0$,

$$0 = \eta'(\theta_0)\mathbf{E}_{\theta_0}(\alpha \cdot T(\mathbf{X})) + \frac{c'(\theta_0)}{c(\theta_0)} \cdot \alpha \quad \Rightarrow \quad \frac{c'(\theta_0)}{c(\theta_0)} = -\eta'(\theta_0)\mathbf{E}_{\theta_0}\big(T(\mathbf{X})\big).$$

Since $\beta'_{\phi_\alpha}(\theta_0) = 0$, the desired test ϕ_α also satisfies another condition (other than the size-α condition),

$$\eta'(\theta_0)\mathbf{E}_{\theta_0}\big(\phi_\alpha(\mathbf{X})T(\mathbf{X})\big) - \eta'(\theta_0) \cdot \alpha\mathbf{E}_{\theta_0}\big(T(\mathbf{X})\big) = 0 \quad \Rightarrow \quad \mathbf{E}_{\theta_0}\big(\phi_\alpha(\mathbf{X})T(\mathbf{X})\big) = \alpha\mathbf{E}_{\theta_0}\big(T(\mathbf{X})\big). \tag{2.15}$$

Fix a $\theta' \neq \theta_0$. So, in terms of the GNP-Lemma, we are considering the following problem, find a test function ϕ, such that ϕ maximizes

$$\left.\begin{aligned} &\phi \text{ maximizes } \int \phi(\mathbf{x})\ f(\mathbf{x} : \theta')\ d\mu(\mathbf{x}), \\ &\text{subject to, } \int \phi(\mathbf{x})\ f(\mathbf{x} : \theta_0)\ d\mu(\mathbf{x}) = \alpha \quad \text{and} \quad \int \phi(\mathbf{x})\ T(\mathbf{x}) f(\mathbf{x} : \theta_0)\ d\mu(\mathbf{x}) = \alpha\mathbf{E}_{\theta_0}(T). \end{aligned}\right\} \tag{2.16}$$

The GNP Lemma based choices are, $m = 2$, $g_1(\mathbf{x}) = f(\mathbf{x} : \theta_0)$, $g_2(\mathbf{x}) = T(\mathbf{x})f(\mathbf{x} : \theta_0)$, $c_1 = \alpha$, $c_2 = \alpha\mathbf{E}_{\theta_0}(T)$ and $g_3(\mathbf{x}) = f(\mathbf{x} : \theta')$. Note, $\int |g_2(\mathbf{x})|\ d\mu(\mathbf{x}) < \infty$, if T has finite expectation (which is usually the case).

[9] Since $\theta_i \in \Theta_0$, for $i = 1, 2$, the α-similar condition automatically ensures any level-α test has size $= \alpha$.

[10] We must assume $\theta_0 \in \mathrm{int}(\Theta)$, where Θ is the parameter space in the family (2.3). Usually, we should use the natural parametrization in an exponential family. We can write, $\eta(\theta) = \eta$, and reparametrize the density in (2.3) in terms of η, and the parameter space becomes the natural parameter space. If we consider the null $H_0 : \theta = \theta_0$, the corresponding null in terms of η is, $H_0 : \eta = \eta_0$, where $\eta_0 = \eta(\theta_0)$. These will be equivalent if $\theta \mapsto \eta(\theta)$ is strictly monotone and preferably continuous. Most texts directly use $\eta(\theta) = \theta$, to avoid this sort of questions about the behavior of η. We will use η as any link, and we will impose suitable conditions on its smoothness (differentiability etc.).

[11] If not, then ϕ_α becomes worse than the trivial test $\phi \equiv \alpha$, and $\phi \in \mathcal{C}_1$.

If we use the trivial test with power function having zero derivative everywhere, and apply it to (2.14), we obtain, $c'(\theta)/c(\theta) = -\eta'(\theta)\mathbf{E}_\theta(T)$. As a result, for any test function ϕ,

$$\beta'_\phi(\theta) = \eta'(\theta)\left[\mathbf{E}_\theta(\phi \cdot T) - \mathbf{E}_\theta(\phi) \cdot \mathbf{E}_\theta(T)\right] = \eta'(\theta) \cdot \mathbf{cov}_\theta\big(\phi(\mathbf{X}), T(\mathbf{X})\big). \tag{2.17}$$

Lemma 6. *For testing $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$, in the family (2.3), the UMPU size-α test will be $\psi_\alpha(\mathbf{x})$ (cf. (2.11)), where c_i, γ_i, $i = 1, 2$, are determined using, $\mathbf{E}_{\theta_0}\psi_\alpha(\mathbf{X}) = \alpha$ and the condition given in (2.15).*

Proof of Lemma 6. Consider the region,

$$A = \left\{\left(\int \phi g_1 \, d\mu, \int \phi g_2 \, d\mu\right) : \phi \text{ is a test function}\right\} = \left\{(\mathbf{E}_{\theta_0}(\phi), \mathbf{E}_{\theta_0}(\phi \cdot T)) : \phi \text{ is a test function}\right\}.$$

where, g_1, g_2 are defined earlier. Then, A is convex and contains all points $(u, u\mathbf{E}_{\theta_0}(T))$, for all $0 < u < 1$ (this follows by considering all trivial tests of size-u). It also contains a point (α, u_2), with $u_2 > \alpha\mathbf{E}_{\theta_0}(T)$. This will follow if we show a test ϕ exists which satisfies, $\alpha = \mathbf{E}_{\theta_0}(\phi)$ and $(u_2 - \alpha) = \mathbf{E}_{\theta_0}(\phi T) - \alpha\mathbf{E}_{\theta_0}(T) = \mathbf{cov}_{\theta_0}(\phi, T) > 0$. As, $\eta'(\theta) > 0$, for all θ, this is equivalent to showing the existence of size-α test ϕ satisfying,

$$\beta'_{\theta_0}(\phi) = \eta'(\theta_0) \cdot \mathbf{cov}_\theta\big(\phi, T\big) > 0. \tag{2.18}$$

We are assuming differentiability[12] of $\eta(\theta)$ on the interior of the parameter space. Consider the UMP size-α MLR based test ξ_α for $K_0 : \theta = \theta_0$ against $K_1 : \theta > \theta_0$. This test is of the form (cf. Theorem 1),

$$\xi_\alpha(\mathbf{x}) = \begin{cases} 1 & \text{if } T(\mathbf{x}) > d_\alpha, \\ \rho_\alpha & \text{if } T(\mathbf{x}) = d_\alpha, \\ 0 & \text{o.w.} \end{cases} \tag{2.19}$$

So, $\int \xi_\alpha g_1 = \alpha$. Further, all power functions are differentiable everywhere, hence for any $\epsilon > 0$, and any other size-α test ξ (for K_0 against K_1),

$$\begin{aligned} &\beta_{\xi_\alpha}(\theta_0 + \epsilon) \geqslant \beta_\xi(\theta_0 + \epsilon) \\ \Rightarrow \; &\beta_{\xi_\alpha}(\theta_0 + \epsilon) - \beta_{\xi_\alpha}(\theta_0) \geqslant \beta_\xi(\theta_0 + \epsilon) - \beta_\xi(\theta_0) \\ \Rightarrow \; \beta'_{\xi_\alpha}(\theta_0) = \lim_{\epsilon \downarrow 0} &\frac{\beta_{\xi_\alpha}(\theta_0 + \epsilon) - \beta_{\xi_\alpha}(\theta_0)}{\epsilon} \geqslant \lim_{\epsilon \downarrow 0} \frac{\beta_\xi(\theta_0 + \epsilon) - \beta_\xi(\theta_0)}{\epsilon} = \beta'_\xi(\theta_0) \end{aligned}$$

Hence, the UMP test ξ_α has the highest derivative at θ_0, among all size-α tests for K_0 against K_1. Although this is an useful fact, it still does not prove $\beta'_{\xi_\alpha}(\theta_0) > 0$. But, we have proved this fact in Lemma 8. Thus, the existence of such a point $(\alpha, u_2) \in A$, with $u_2 > \alpha\mathbf{E}_{\theta_0}(T)$, is established. A similar argument can be extended to show there are points $(\alpha, u_1) \in A$, with $u_1 < \alpha\mathbf{E}_{\theta_0}(T)$. Hence, the point $\big(\alpha, \alpha\mathbf{E}_{\theta_0}(T)\big)$ is an interior point of A.

Using GNP-Lemma part (d), the required test (subject to the constraints in (2.16)) that maximizes power at some $\theta' \neq \theta_0$, will be of the form

$$\psi_\alpha(\mathbf{x}) = \begin{cases} 1 & \text{if } f(\mathbf{x} : \theta') > k_1 f(\mathbf{x} : \theta_0) + k_2 T(\mathbf{x}) f(\mathbf{x} : \theta_0), \\ 0 & \text{if } f(\mathbf{x} : \theta') < k_1 f(\mathbf{x} : \theta_0) + k_2 T(\mathbf{x}) f(\mathbf{x} : \theta_0), \end{cases} \tag{2.20}$$

[12] This is needed if we work the original parametrization in (2.3), but is not needed if we use the natural parameterization, or use $\eta = \eta(\theta) = \theta$.

where, k_1, k_2 are found by the size conditions. Using (2.3), we obtain

$$\begin{aligned} f(\mathbf{x}:\theta') &> k_1 f(\mathbf{x}:\theta_0) + k_2 T(\mathbf{x}) f(\mathbf{x}:\theta_0) \\ \Rightarrow\ c(\theta')\exp\{\eta(\theta')T(\mathbf{x})\} &> k_1 c(\theta_0)\exp\{\eta(\theta_0)T(\mathbf{x})\} + k_2 T(\mathbf{x}) c(\theta_0)\exp\{\eta(\theta_0)T(\mathbf{x})\} \\ \Rightarrow\ \exp\{b\cdot T(\mathbf{x})\} &> a_1 + a_2\cdot T(\mathbf{x}), \end{aligned}$$

where, $b = \eta(\theta') - \eta(\theta_0)$, $a_1 = k_1 c(\theta_0)/c(\theta')$ and $a_2 = k_2 c(\theta_0)/c(\theta')$. Depending on the location of θ', either $b < 0$ or $b > 0$. Consider the case of $b > 0$. Write,

$$C = \{t : e^{bt} > a_1 + a_2 t\}.$$

Let $a_2 > 0$. Then, either $C = \varnothing$ or $C = (c_1, c_2)$, for some $c_1 < c_2$, irrespective of the sign of a_1. If, $a_2 = 0$, then either $C = \varnothing$ or $C = (c_0, \infty)$, for some $c_0 \in \mathbb{R}$. If, $a_2 < 0$, then $C = (-\infty, c_0)$, for some $c_0 \in \mathbb{R}$. In case $C = \varnothing$, the test ψ_α always rejects and has size 0, which is not allowed in consideration. Moreover, it will fail to satisfy the size condition. Similarly, if $C = (c_0, \infty)$ or $(-\infty, c_0)$, the test resembles a MLR based UMP test for a one-sided null $K_0 : \theta \leqslant \theta_0$ against $K_1 : \theta > \theta_0$ (or vice-versa). Such tests have strictly monotone power functions, and hence the slope condition at θ_0 will fail. The only possible alternative is to have $C = (c_1, c_2)$. A similar argument for $b < 0$ shows the only possible region is, $C = (c_1, c_2)$, for some $c_1 < c_2$. Note, we can not claim anything about the signs[13] of a_1, a_2 (and k_1, k_2) in this case. As a result, the test in (2.20) reduces to,

$$\psi_\alpha(\mathbf{x}) = \psi_\alpha(t) = \begin{cases} 1 & \text{if } t < c_1 \text{ or } t > c_2, \\ \gamma_i & \text{if } t = c_i,\ i = 1, 2, \\ 0 & \text{if } c_1 < t < c_2. \end{cases}$$

A similar argument works[14] if we use $b < 0$. The choices of c_i, γ_i, will only depend on θ_0 (but as earlier, the choices of k_i's depend on θ'). Definitely, ψ_α is unbiased, by comparison with the trivial test $\psi = \alpha$ (which satisfies both size constraints, and ψ_α must be better than ψ at all alternatives).

Note, in this case the continuity of power functions simplifies[15] the class $\mathcal{C}_3$ (cf. (2.13)). Hence,

$$\mathcal{C}_3 = \{\psi : \mathbf{E}_{\theta_0}\psi = \alpha,\ \mathbf{E}_\theta\psi \geqslant \alpha, \quad \text{for all } \theta \neq \theta_0\} = \text{all size-}\alpha \text{ unbiased tests}.$$

Let $\mathcal{C}_4 = \{\psi : \mathbf{E}_{\theta_0}\psi = \alpha,\ \psi \text{ satisfies } (2.15)\}$. If $\psi \in \mathcal{C}_3$, then it implies $\psi \in \mathcal{C}_4$. Hence, $\mathcal{C}_3 \subseteq \mathcal{C}_4$. Also, ψ_α is the best test in $\mathcal{C}_4$, but $\psi_\alpha \in \mathcal{C}_3$. Hence, it is the required UMPU size-α test. □

Remark. A simplification of this test can be done if the distribution of T is symmetric about a point. Further, the UMPU tests for both these hypothesis are strictly unbiased, with power $> \alpha$, at alternatives. The reasons are given in Lehmann and Romano (2005) (pp. 112-113). It should be noted, α-similarity is not used in deriving the UMPU test for $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$.

Lemma 7 (Lemma 5 of Khatri (1967)**).** *Consider a r.v. X and two functions $g(x)$, $h(x)$ (with same domains). Assume, g and h satisfy the following condition,*

$$g(x_1) \geqslant g(x_2) \quad \textit{and} \quad h(x_1) \geqslant h(x_2), \quad \textit{for any } x_1, x_2.$$

Then, $\mathbf{cov}(g(X), h(X)) \geqslant 0$.

[13] We do not claim that ψ_α has to be best in the class of all tests with level α and satisfying $\int \phi g_2 \leqslant c_2$.

[14] It should be noted, the explicit use of $\eta(\theta)$, instead of reparametrizing it as $\eta(\theta) = \theta$, is not an issue. All that is needed is the strictly increasing property of η and its differentiability.

[15] If any test $\psi \in \mathcal{C}_3$ satisfies, $\mathbf{E}_{\theta_0}\psi < \alpha$, then, it can not satisfy the unbiasedness criteria, $\mathbf{E}_\theta\psi \geqslant \alpha$, for all $\theta \neq \theta_0$, without becoming discontinuous at θ_0.

Proof of Lemma 7. Let X_1, X_2 be i.i.d. copies of X. Then, $\{g(X_1) - g(X_2)\}\{h(X_1) - h(X_2)\} \geqslant 0$, w.p. 1. Hence,

$$0 \leqslant \mathbf{E}\{g(X_1) - g(X_2)\}\{h(X_1) - h(X_2)\} = 2\left[\mathbf{E}g(X)h(X) - \mathbf{E}g(X) \cdot \mathbf{E}h(X)\right] = 2 \cdot \mathbf{cov}\left(g(X), h(X)\right),$$

and the proof follows. This result also holds for random vectors $\mathbf{X}$. □

Lemma 8. *Consider the UMP size-α $(0 < \alpha < 1)$ test ξ_α (cf. (2.19)), for $K_0 : \theta = \theta_0$ against $K_1 : \theta > \theta_0$, when the underlying family is (2.3), $\theta_0 \in int(\Theta)$ and $\theta \mapsto \eta(\theta)$ is strictly increasing and differentiable everywhere on the interior of Θ. Then, $\beta'_{\xi_\alpha}(\theta_0) > 0$.*

Proof of Lemma 8. Using (2.17) and the assumptions on η, it is enough to show $\mathbf{cov}_{\theta_0}(\xi_\alpha, T) > 0$. Following the approach in Lemma 7, and since ξ_α is a function of $T(\mathbf{x})$, for all $t \in \mathbb{R}$, we write, $g(t) = t$, $h(t) = \xi_\alpha(t)$. Let, T_1, T_2, be i.i.d. copies of T. To simplify notation, we write $d_\alpha = d$, $\rho_\alpha = \rho$, $\xi_\alpha(\cdot) = \xi(\cdot)$, $P_\theta = P$ and $\mathbf{cov}_{\theta_0} = \mathbf{cov}$ (cf. (2.19)), since α and θ_0 are fixed. Then, if $\rho \in (0, 1)$, for any $t_1, t_2 \in \mathbb{R}$,

$$(t_2 - t_1)\big(\xi(t_2) - \xi(t_1)\big) = \begin{cases} 0 & \text{if } t_1 < d,\ t_2 < d, \\ (d - t_1) \cdot \rho\ (> 0) & \text{if } t_1 < d,\ t_2 = d, \\ (t_2 - t_1) \cdot 1\ (> 0) & \text{if } t_1 < d,\ t_2 > d, \\ -(t_2 - d) \cdot \rho\ (> 0) & \text{if } t_1 = d,\ t_2 < d, \\ 0 & \text{if } t_1 = d,\ t_2 = d, \\ (t_2 - d) \cdot (1 - \rho)\ (> 0) & \text{if } t_1 = d,\ t_2 > d, \\ -(t_2 - t_1) \cdot 1\ (> 0) & \text{if } t_1 > d,\ t_2 < d, \\ (d - t_1) \cdot (\rho - 1)\ (> 0) & \text{if } t_1 > d,\ t_2 = d, \\ 0 & \text{if } t_1 > d,\ t_2 > d. \end{cases} \tag{2.21}$$

If $\rho = 0$, then the above term reduces to,

$$(t_2 - t_1)\big(\xi(t_2) - \xi(t_1)\big) = \begin{cases} 0 & \text{if } t_1 < d,\ t_2 \leqslant d, \\ (t_2 - t_1) \cdot 1\ (> 0) & \text{if } t_1 < d,\ t_2 > d, \\ 0 & \text{if } t_1 = d,\ t_2 \leqslant d, \\ (t_2 - d) \cdot 1\ (> 0) & \text{if } t_1 = d,\ t_2 > d, \\ -(t_2 - t_1) \cdot 1\ (> 0) & \text{if } t_1 > d,\ t_2 \leqslant d, \\ 0 & \text{if } t_1 > d,\ t_2 > d. \end{cases}$$

A similar reduction can be achieved in case of $\gamma = 1$. If we consider the decomposition in (2.21), immediately it follows,

$$2 \cdot \mathbf{cov}\left(\xi, T\right) > 0,$$

unless,

$$P(T_1 < d, T_2 < d) + P(T_1 = d, T_2 = d) + P(T_1 > d, T_2 > d) = 1.$$

However, as $T_i \overset{d}{=} T$, and $P(T < d) + P(T = d) + P(T > d) = 1$, for any $d \in \mathbb{R}$, this is equivalent to knowing if there exists constants $b_1, b_2, b_3 \in [0, 1]$, which satisfy

$$b_1 + b_2 + b_3 = 1 \quad \text{and} \quad b_1^2 + b_2^2 + b_3^2 = 1.$$

Note, for any sequence of non-negative b_i's, the following inequality is true, $\sum_{i=1}^m b_i^2 \leqslant \left(\sum_{i=1}^m b_i\right)^2$, and equality only holds if there exists some i_0, such that $b_{i_0} > 0$ and $b_j = 0$, for $j \neq i_0$. The proof follows by noting that if we expand the rhs above, the crossproduct terms $b_i b_j$ are non-negative, and can $\sum_{i \neq j} b_i b_j = 0$, only when the above condition holds.As a result, if atleast one of $b_1 = P(T < d)$, $b_2 = P(T = d)$ or $b_3 = P(T > d)$, is < 1, the proof will be complete.

If on contrary, say $b_1 = 1$. Then $\xi(t) = 1$, for all t (w.p. 1). So, the size becomes 1, which is not allowed. Similar argument works in case of $b_3 = 1$. If $b_2 = 1$, $\xi(t) = \rho \in (0, 1)$, for all t (w.p. 1). In this case, ξ is same as the trivial test of size ρ, and the power function of ξ is not strictly increasing (using MLR theory, cf. Theorem 1), which will violate identifiability in the family (2.3) (identifiability is already ensured by strictly increasing property of η). Hence, none of the b_i's can be equal to 1. So, we must have $\mathbf{cov}(\xi, T) > 0$.

In case of $\rho = 0$, similar arguments will show that strict inequality in the Lemma will only be false, if there are constants $b_1, b_2 \in [0, 1]$, which satisfy, $b_1 + b_2 = 1$ and $b_1^2 + b_1^2 = 1$ (in this case, $b_1 = P(T \leqslant d)$ and $b_2 = P(T > d)$). Again these can be negated by similar arguments as above. The case of $\rho = 1$ can be dealt similarly. This completes the proof. □

3 UMPU tests in multiparameter exponential families

In some testing problems, the family of distributions depend on several parameters. However the hypothesis of interest concerns a single scalar parameter. The remaining parameters are then called *nuisance* parameters. For example, with the $N(\mu, \sigma^2)$ distribution, if the hypothesis is concerning μ, then the unknown σ becomes a nuisance parameter. In this situation, for multiparameter exponential families, UMPU tests can be found for some specific testing problems. The approach is to find a statistic $\mathbf{T}$, such that conditional distribution of the data given $\mathbf{T}$ is dependent on a one-dimensional parameter.

Consider a family of distributions $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ and a testing problem $H_0 : \theta \in \Theta_0$ against $H_1 : \theta \in \Theta_1$. Suppose, $\mathbf{T}$ is a sufficient statistic for the family $\mathcal{P}$. Consider tests ϕ satisfying

$$\mathbf{E}\left[\phi(\mathbf{X}) \mid \mathbf{T}\right] = \alpha, \quad \text{w.p. 1, for all } \theta \in \bar{\Theta}_{0,1}. \tag{3.1}$$

A test satisfying (3.1) is said to have *Neyman structure* with respect to $\mathbf{T}$, with respect to the parameters space $\bar{\Theta}_{0,1}$. This implies,

$$\{\phi : \phi \text{ satisfies (3.1)}\} \subseteq \{\phi : \phi \text{ is } \alpha\text{-similar on } \bar{\Theta}_{0,1}\}. \qquad (\star)$$

The reverse inclusion holds in $(\star)$ under a simple condition. Assume $\mathbf{X} \sim P_\theta$ and $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$. We say $\mathcal{P}$ is *boundedly complete*, if for all bounded functions f,

$$\mathbf{E}_\theta f(\mathbf{X}) = 0 \quad \text{for all } \theta, \text{ implies} \quad P_\theta(f(\mathbf{X}) = 0) = 1 \quad \text{for all } \theta.$$

Surely, if $\mathcal{P}$ is complete (usual definition of completeness), then it is boundedly complete, but the converse can be false.

Lemma 9. *Suppose,* $\mathbf{T}(\mathbf{X})$ *is sufficient for the family* $\mathcal{P}_{0,1} = \{P_\theta : \theta \in \bar{\Theta}_{0,1}\}$. *Then a necessary and sufficient condition for all* α*-similar tests on* $\bar{\Theta}_{0,1}$*, to have Neyman structure with respect to* $\mathbf{T}$ *is that* $\mathbf{T}$ *is boundedly complete.*

Proof of Lemma 9. Let $\mathbf{T}$ be boundedly complete for $\mathcal{P}_{0,1}$ and ϕ be α-similar on $\bar{\Theta}_{0,1}$. Then,

$$\begin{aligned} \mathbf{E}_\theta(\phi(\mathbf{X}) - \alpha) = 0, &\quad \text{for all } \theta \in \bar{\Theta}_{0,1}, \\ \Rightarrow\ \mathbf{E}_\theta\left(\mathbf{E}\left[\phi \mid \mathbf{T}\right] - \alpha\right) = 0, & \\ \Rightarrow\ \mathbf{E}_\theta g(\mathbf{T}) = 0, &\quad \text{for all } \theta \in \bar{\Theta}_{0,1}, \text{ where } g(\mathbf{T}) = \mathbf{E}(\phi \mid \mathbf{T}) - \alpha, \\ \Rightarrow\ P_\theta(g(\mathbf{T}) = 0) = 1, &\quad \text{for all } \theta \in \bar{\Theta}_{0,1}, \text{ since } g \text{ is bounded and } \mathbf{T} \text{ is complete for } \mathcal{P}_{0,1}. \end{aligned}$$

Thus, ϕ satisfies (3.1) and has Neyman structure.

Now, assume $\mathbf{T}$ is not boundedly complete for $\mathcal{P}_{0,1}$, but all α-similar tests satisfy (3.1). As $\mathbf{T}$ is not boundedly complete, there is a function $h(\mathbf{t})$ such that $|h(\mathbf{t})| \leqslant c_0$, for some finite c_0, $\mathbf{E}_\theta h(\mathbf{T}) = 0$, for all $\theta \in \bar{\Theta}_{0,1}$, but $P_\theta(h(\mathbf{T}) = 0) < 1$, for some $\theta \in \bar{\Theta}_{0,1}$. Define the test,

$$\phi(\mathbf{x}) = \alpha + k \cdot h(\mathbf{T}(\mathbf{x})) = \alpha + k \cdot h(\mathbf{t}), \quad \text{where, } k = \min\{\alpha, 1-\alpha\}/c_0.$$

Then, $|\phi(\mathbf{x})| \leqslant \alpha + \min\{\alpha, 1-\alpha\} \leqslant 1$, for all $\mathbf{x}$. Also, ϕ is α-similar, because

$$\mathbf{E}_\theta \phi(\mathbf{X}) = \alpha + k \cdot \mathbf{E}_\theta h(\mathbf{T}) = 0, \quad \text{for all } \theta \in \bar{\Theta}_{0,1}.$$

But, this implies (using the claim about h),

$$\mathbf{E}\left(\phi(\mathbf{X}) \mid \mathbf{T} = \mathbf{t}\right) = \alpha + k \cdot h(\mathbf{t}) \neq \alpha, \quad \text{w.p. } > 0, \text{ for some } \theta \in \bar{\Theta}_{0,1}.$$

As a result, ϕ no longer satisfies the Neyman structure condition, which is a contradiction to our original assumption. Hence, $\mathbf{T}$ must be boundedly complete. □

We now consider multiparameter exponential families of the form,

$$f(\mathbf{x} : \theta, \boldsymbol{\eta}) = C(\theta, \boldsymbol{\eta}) h(\mathbf{x}) \cdot \exp\left\{\theta \cdot U(\mathbf{x}) + \sum_{i=1}^{k} \eta_i \cdot T_i(\mathbf{x})\right\}, \quad \text{where } (\theta, \boldsymbol{\eta}) \in \mathcal{T}, \tag{3.2}$$

and $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_k)^T$ is a k-dimensional parameter. We assume the density in (3.2) is in the natural exponential family form, with natural parameter space $\mathcal{T}$. Also write, $\mathbf{T} = (T_1, \ldots, T_k)^T$.

Theorem 10. *Suppose* $\mathbf{X}$ *has the density* (3.2)*. Assume the following:*

(a) For each θ_j*,* $j = 0, 1, 2$*, there exists points* $\theta \in \mathcal{T}$*, such that* θ *is* $<$ *and* $>$*,* θ_j*.*

(b) $\mathbf{T}$ *is sufficient for each of the families,* $\{P^{\mathbf{X}}_{(\theta,\boldsymbol{\eta})} : (\theta, \boldsymbol{\eta}) \in \mathcal{T},\ \theta = \theta_j\}$*,* $j = 0, 1, 2$*.*

(c) The families of distributions of $\mathbf{T}$*,* $\{P^{\mathbf{T}}_{(\theta,\boldsymbol{\eta})} : (\theta, \boldsymbol{\eta}) \in \mathcal{T},\ \theta = \theta_j\}$*,* $j = 0, 1, 2$*, are each boundedly complete.*

Then, the following results hold.

(i) For testing $H_1 : \theta \leqslant \theta_0$ *against* $K_1 : \theta > \theta_0$*, an UMPU test of size-*α *is given by,*

$$\phi_1(u, \mathbf{t}) = \begin{cases} 1 & \textit{if } u > c(\mathbf{t}), \\ \gamma(\mathbf{t}) & \textit{if } u = c(\mathbf{t}), \\ 0 & \textit{o.w.}, \end{cases} \tag{3.3}$$

where, $c(\mathbf{t})$ *and* $\gamma(\mathbf{t})$ *are determined by,* $\quad \mathbf{E}_{\theta_0}\left[\phi_1(U, \mathbf{T}) \mid \mathbf{T} = \mathbf{t}\right] = \alpha.$

(ii) *For testing* $H_0 : \theta \leqslant \theta_1$ *or* $\theta \geqslant \theta_2$ *against* $H_1 : \theta_1 < \theta < \theta_2$*, an UMPU test of size-*α *is,*

$$\phi_2(u, \mathbf{t}) = \begin{cases} 1 & \text{if } c_1(\mathbf{t}) < u < c_2(\mathbf{t}), \\ \gamma_i(\mathbf{t}) & \text{if } u = c_i(\mathbf{t}), \text{ for } i = 1, 2, \\ 0 & o.w., \end{cases} \tag{3.4}$$

where, $c_i(\mathbf{t}), \gamma_i(\mathbf{t})$*,* $i = 1, 2$*, are determined by,* $\mathbf{E}_{\theta_i}\left[\phi_2(U, \mathbf{T}) \mid \mathbf{T} = \mathbf{t}\right] = \alpha, \quad \text{for } i = 1, 2.$

(iii) *For testing* $H_0 : \theta \leqslant \theta \leqslant \theta_2$ *against* $H_1 : \theta \notin [\theta_1, \theta_2]$*, an UMPU test of size-*α *is given by,*

$$\phi_3(u, \mathbf{t}) = \begin{cases} 1 & \text{if } u < c_1(\mathbf{t}) \text{ or } u > c_2(\mathbf{t}), \\ \gamma_i(\mathbf{t}) & \text{if } u = c_i(\mathbf{t}), \text{ for } i = 1, 2, \\ 0 & o.w., \end{cases} \tag{3.5}$$

where, $c_i(\mathbf{t}), \gamma_i(\mathbf{t})$*,* $i = 1, 2$*, are determined by,* $\mathbf{E}_{\theta_i}\left[\phi_3(U, \mathbf{T}) \mid \mathbf{T} = \mathbf{t}\right] = \alpha, \quad \text{for } i = 1, 2.$

(iv) *For testing* $H_0 : \theta = \theta_0$ *against* $H_1 : \theta \neq \theta_0$*, an UMPU test of size-*α *is given by* (3.5)*, where* $c_i(\mathbf{t}), \gamma_i(\mathbf{t})$*,* $i = 1, 2$*, are determined by,*

$$\mathbf{E}_{\theta_0}\left[\phi(U, \mathbf{T}) \mid \mathbf{T} = \mathbf{t}\right] = \alpha \quad \text{and} \quad \mathbf{E}_{\theta_0}\left[U \cdot \phi(U, \mathbf{T}) \mid \mathbf{T} = \mathbf{t}\right] = \alpha \cdot \mathbf{E}_{\theta_0}\left[U \mid \mathbf{T} = \mathbf{t}\right]. \tag{3.6}$$

Before prooceeding with the proof of this result, we make some important remarks.

1. In the family (3.2), $(U, \mathbf{T})$ are sufficient statistics. For any test function $\phi(\mathbf{x})$, define, $\psi(u, \mathbf{t}) = \mathbf{E}[\phi(\mathbf{X}) \mid U = u, \mathbf{T} = \mathbf{t}]$. Then, sufficiency implies ψ is free of parameters and is itself a test, with same power as ϕ. Hence, it is enough to consider test functions based on $(U, \mathbf{T})$.

2. The joint distribution of $(U, \mathbf{T})$ has density,

$$f(u, \mathbf{t} : \theta, \boldsymbol{\eta}) = C(\theta, \boldsymbol{\eta}) h(u, \mathbf{t}) \exp\left\{\theta u + \boldsymbol{\eta}^T \mathbf{t}\right\}, \quad (\theta, \boldsymbol{\eta}) \in \mathcal{T},$$

with respect to some σ-finite measure ν. Also, $\mathbf{T}$ has marginal density

$$f(\mathbf{t} : \theta, \boldsymbol{\eta}) = C_\theta(\boldsymbol{\eta}) h(\mathbf{t}) \exp\{\boldsymbol{\eta}^T \mathbf{t}\}, \quad \text{for all } (\theta, \boldsymbol{\eta}) \in \mathcal{T},$$

with respect to some σ-finite measure ν_θ. Similarly, $[U \mid \mathbf{T} = \mathbf{t}]$ has density,

$$f(u | \mathbf{t} : \theta) = C_{\mathbf{t}}(\theta) h(u) \cdot e^{\theta u}, \quad \text{for all } \theta, \text{ with respect to some } \sigma\text{-finite measure } \nu_{\mathbf{t}}.$$

This follows from results in Section 2.7 of Lehmann and Romano (2005).

3. Define, the sub-parameter spaces,

$$\mathcal{T}_j = \{(\theta, \boldsymbol{\eta}) \in \mathcal{T} : \theta = \theta_j\}, \quad \text{for } j = 0, 1, 2.$$

The theorem requires that the family of distributions, $\{P^{\mathbf{T}}_{(\theta, \boldsymbol{\eta})} : (\theta, \boldsymbol{\eta}) \in \mathcal{T}_j\}$ is boundedly complete. Surely[16], this will be true if a k-dimensional rectangle can be inscribed inside $\mathcal{T}_j$. Also, $\mathbf{T}$ will be sufficient for the family, $\{f(u, \mathbf{t} : \theta, \boldsymbol{\eta}) : (\theta, \boldsymbol{\eta}) \in \mathcal{T}_j\}$ (because $\theta = \theta_j$ is fixed at a known value).

[16] At present, I do not know about any sufficient conditions on $\mathcal{T}$, which will guarantee this. The original paper Lehmann and Scheffé (1955) simply assumes that $\mathcal{T}_j$ contain k-dimensional open rectangles.

4. Consider a general testing problem, $H_0 : \theta \in \Theta_0$ v/s $H_1 : \theta \in \Theta_1$. Let, $\bar{\mathcal{P}} = \{P_\theta : \theta \in \bar{\Theta}_{0,1}\}$ denote the underlying family at $\bar{\Theta}_{0,1}$. $\mathbf{T}$ is assumed to be sufficient for $\bar{\mathcal{P}}$ and the family $\{P_\theta^{\mathbf{T}} : \theta \in \bar{\Theta}_{0,1}\}$ is boundedly complete. We also assume all power functions are continuous everywhere. Our goal is to find the best level-α unbiased test. Write,

$$\begin{aligned}
\mathcal{D}_1 &= \{\phi : \phi \text{ is unbiased and level-}\alpha\},\\
\mathcal{D}_2 &= \{\phi : \phi \text{ is } \alpha\text{-similar and level-}\alpha\},\\
\mathcal{D}_3 &= \{\phi : \phi \text{ has Neyman-structure and level-}\alpha\},\\
\mathcal{D}_4 &= \{\phi : \phi \text{ has Neyman-structure}\} = \{\phi : \phi \text{ has conditional power} = \alpha \text{ at } \theta \in \bar{\Theta}_{0,1}\}.
\end{aligned}$$

Lemmas 4 and 9, imply

$$\mathcal{D}_1 \subseteq \mathcal{D}_2 = \mathcal{D}_3 \subset \mathcal{D}_4.$$

Lemma 4 says, finding the best test in $\mathcal{D}_2$ will ensure we have found the best test in $\mathcal{D}_1$. But, if we find the best test in $\mathcal{D}_4$ and then show it is a member of $\mathcal{D}_1$, our job is done. This will be the desired UMPU test. Essentially, $\mathcal{D}_3$ has no role.

Finding the best test in $\mathcal{D}_4$ is easier in multiparameter exponential families, because conditioning on $\mathbf{T}$ removes the role of nuisance parameters $\boldsymbol{\eta}$. The problem reduces to finding UMP or UMPU tests in an one-parameter exponential family setup, using the conditional distribution of $[U \mid \mathbf{T} = \mathbf{t}]$. Such tests will maximize the conditional power $\mathbf{E}_\theta[\phi(U, \mathbf{T}) \mid \mathbf{T} = \mathbf{t}]$, and hence will maximize the unconditional power, $\mathbf{E}_{(\theta,\boldsymbol{\eta})}\phi(U, \mathbf{T})$, at all $\theta \in \Theta_1$ and all $\boldsymbol{\eta} \in \mathcal{T}$.

Proof of Theorem 10. We present proof of each part separately. The requirement of having points, $\theta' < \theta_j < \theta''$, $j = 0, 1, 2$, is needed to ensure that θ_j is not a boundary point of $\mathcal{T}$, so that theory for one-parameter exponential families can be applied on the conditional distribution of $[U \mid \mathbf{T}]$.

(i) In this case, $\bar{\Theta}_{0,1} = \mathcal{T}_0$ and conditions of Lemmas 4 and 9 hold. By construction, $\phi_1 \in \mathcal{D}_4$. Any other test $\psi \in \mathcal{D}_4$ satisfies, $\mathbf{E}_{\theta_0}[\psi(U, \mathbf{T}) \mid \mathbf{T} = \mathbf{t}] = \alpha$, w.p. 1, for all $\boldsymbol{\eta} \in \mathcal{T}$, *i.e.*, ψ has conditional power $= \alpha$ at θ_0.

The construction of the UMP test ϕ (cf. (1.2)) in Theorem 1 shows, our test $\phi_1(u, \mathbf{t}) = \phi_{1,\mathbf{t}}(u)$, has highest (conditional) power at any alternative $\theta > \theta_0$, among tests[17] which have conditional power $= \alpha$ at θ_0. Thus, for all $\psi \in \mathcal{D}_4$,

$$\begin{aligned}
&\mathbf{E}_\theta[\phi_1(U, \mathbf{T}) \mid \mathbf{T} = \mathbf{t}] \geqslant \mathbf{E}_\theta[\psi(U, \mathbf{T}) \mid \mathbf{T} = \mathbf{t}], \quad \text{for all } \theta > \theta_0, \text{ w.p. 1, for all } \boldsymbol{\eta} \in \mathcal{T},\\
&\Rightarrow \ \mathbf{E}_{(\theta,\boldsymbol{\eta})}\phi_1(U, \mathbf{T}) \geqslant \mathbf{E}_{(\theta,\boldsymbol{\eta})}\psi(U, \mathbf{T}), \quad \text{for all } \theta > \theta_0 \text{ and } \boldsymbol{\eta} \in \mathcal{T}.
\end{aligned}$$

Thus ϕ_1 is UMP in $\mathcal{D}_4$. Also, ϕ_1 is unbiased (comparing with the trivial test) and $\theta \mapsto \mathbf{E}_\theta[\phi_1(U, \mathbf{T}) \mid \mathbf{T} = \mathbf{t}]$ is nondecreasing (using MLR theory). Hence, ϕ_1 has unconditional level-α. Thus, $\phi_1 \in \mathcal{D}_1$. This proves ϕ_1 is UMP in $\mathcal{D}_1$.

(ii) In this case, $\bar{\Theta}_{0,1} = \mathcal{T}_1 \cup \mathcal{T}_2$. The Neyman-structure condition is same as requiring conditional power $= \alpha$ at θ_i, $i = 1, 2$ (cf. (3.4)). Theorem 3 states the test function ϕ_α (cf. (2.4)), which is of the same form as $\phi_2(u, \mathbf{t})$, maximizes (conditional) power at each alternative $\theta \in (\theta_1, \theta_2)$, subject to the condition in (2.5), which is same as the Neyman structure condition. So, ϕ_2 is also maximizes the unconditional power at alternatives among all tests with Neyman structure, and ϕ_2 is UMP in $\mathcal{D}_4$. But, by comparison

[17] We do not need the fact that, ϕ_1 (which is same as the test in (1.2)) is UMP among all level-α tests.

with the trivial test, ϕ_2 is unbiased (as the trivial test is in $\mathcal{D}_4$) and also minimizes conditional power at all $\theta \in \Theta_0$, among tests satisfying Neyman structure condition. Hence the unconditional power of ϕ_2 is $\leqslant \alpha$, at all $\theta \in \Theta_0$. Hence ϕ_2 is also level-α, and $\phi_2 \in \mathcal{D}_1$. This completes the proof.

(iii) The proof is similar to parts (i) and (ii).

(iv) Suppose ϕ is unbiased level-α for $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$. Then regularity conditions imply, ϕ is α-similar on $\bar{\Theta}_{0,1} = \mathcal{T}_0$. But, it also implies the power function of ϕ has a minima at θ_0, for each $\eta \in \mathcal{T}$. Now, since integration and differentiation are interchangeable in this case,

$$\begin{aligned}\frac{\partial}{\partial\theta}\,\mathbf{E}_{(\theta,\boldsymbol{\eta})}\phi(U,\mathbf{T}) &= \frac{\partial}{\partial\theta}\,\mathbf{E}_{(\theta,\boldsymbol{\eta})}\left[\mathbf{E}_\theta\left\{\phi(U,\mathbf{T}) \mid \mathbf{T}\right\}\right]\\ &= \mathbf{E}^{\mathbf{T}}_{(\theta,\boldsymbol{\eta})}\left[\frac{\partial}{\partial\theta}\int \phi(u,\mathbf{T})\cdot C_{\mathbf{T}}(\theta)h(u)e^{\theta u}\,du\right]\\ &= \mathbf{E}^{\mathbf{T}}_{(\theta,\boldsymbol{\eta})}\left[\frac{C'_{\mathbf{T}}(\theta)}{C_{\mathbf{T}}(\theta)}\mathbf{E}_\theta\left\{\phi(U,\mathbf{T}) \mid \mathbf{T}\right\} + \mathbf{E}_\theta\left\{U\cdot\phi(U,\mathbf{T}) \mid \mathbf{T}\right\}\right].\end{aligned}$$

Now, using ϕ as the trivial test, for each fixed $\mathbf{t}$, and at $\theta = \theta_0$, we have (because the inner integral is also a constant function of θ for each fixed $\mathbf{t}$, in this case),

$$\frac{C'_{\mathbf{t}}(\theta_0)}{C_{\mathbf{t}}(\theta_0)} = -\,\mathbf{E}_{\theta_0}(U \mid \mathbf{T} = \mathbf{t}).$$

Hence, at $\theta = \theta_0$, and since ϕ satisfies, $\mathbf{E}_{\theta_0}(\phi(U,\mathbf{T}) \mid \mathbf{T} = \mathbf{t}) = \alpha$, w.p. 1 for each η, we have

$$\frac{\partial}{\partial\theta}\,\mathbf{E}_{(\theta,\boldsymbol{\eta})}\phi(U,\mathbf{T})\bigg|_{\theta=\theta_0} = \mathbf{E}_{(\theta_0,\boldsymbol{\eta})}\left[U\cdot\phi(U,\mathbf{T}) - \alpha\cdot U\right] = 0, \quad \text{for all } \boldsymbol{\eta} \in \mathcal{T}. \tag{3.7}$$

The r.h.s. of (3.7) can be written as,

$$\mathbf{E}_{(\theta_0,\boldsymbol{\eta})}\left[U\cdot\phi(U,\mathbf{T}) - \alpha\cdot U\right] = \mathbf{E}^{\mathbf{T}}_{\boldsymbol{\eta}}\left[\mathbf{E}_{\theta_0}\left\{U\cdot\phi(U,\mathbf{T}) - \alpha\cdot U \mid \mathbf{T}\right\}\right] = \mathbf{E}^{\mathbf{T}}_{\boldsymbol{\eta}}g(\mathbf{T}) = 0, \quad \text{for all } \boldsymbol{\eta} \in \mathcal{T}_0.$$

We can ignore the dependence of g on θ_0, since it is fixed. The family of distributions $\{P^{\mathbf{T}}_{(\theta,\boldsymbol{\eta})} : (\theta,\boldsymbol{\eta}) \in \mathcal{T}_0\}$ is boundedly complete[18]. Hence, $g(\mathbf{T}) = 0$, w.p. 1, for all $\boldsymbol{\eta} \in \mathcal{T}_0$. Thus, any unbiased level-α test satisfies both conditions in (3.6). Let, $\mathcal{D}_5 = \{\phi : \phi \text{ satisfies constraints in } (3.6)\}$. This shows, $\mathcal{D}_1 \subset \mathcal{D}_5$. Also, $\phi_3 \in \mathcal{D}_5$ (cf. (3.5)) as per the statement of part (iv). Besides, ϕ_3 in (3.5) maximizes the conditional power at all $\theta \neq \theta_0$, among all tests satisfying (3.6), hence it maximizes the unconditional power at $\theta \neq \theta_0$, for all $\boldsymbol{\eta}$, it is also unbiased and hence $\in \mathcal{D}_1$. This completes the proof.

There is an additional technical requirement of showing that the test functions are jointly measurable in $(u, \mathbf{t})$, but we leave this part. The proof of this is given in Lehmann and Romano (2005) (p. 122). □

3.1 UMPU tests for multiparameter case in Normal families

Direct application of the results in Theorem 10 are inconvinient in case of data from multiparameter Normal distributions, since it involves calculation of the conditional distribution $[U \mid \mathbf{T}]$, which is difficult in most such cases. Henc, we aim to develop an equivalent formulation where we can work unconditional distributions in order to compute the cutoff points for the UMPU tests. The following result gives a general formulation.

[18] I have not been able to show $g(\mathbf{t})$ is bounded. If we make the stronger assumption about completeness of $\mathbf{T}$, then it will suffice. For practical purposes, we actually show completeness, rather than boundedly completeness in any family.

Theorem 11. *Suppose, $\mathbf{X}$ has density (3.2). Let, $V = V(U, \mathbf{T})$ be a statistic which is independent of $\mathbf{T}$, when $\theta = \theta_j$, $j = 0, 1, 2$.*

(i) If $V(u, \mathbf{t})$ is increasing in u for each fixed $\mathbf{t}$, then the test,

$$\psi_1(v) = \begin{cases} 1 & \textit{if } v > c, \\ \gamma & \textit{if } v = c, \\ 0 & \textit{o.w.}, \end{cases}$$

where, c, γ, are chosen to satisfy $\mathbf{E}_{\theta_0}\psi_1(V) = \alpha$, is equivalent to the UMPU test $\phi_1(u, \mathbf{t})$ given in (3.3). Hence, $\psi_1(v)$ is an UMPU size-α test for $H_0 : \theta \leqslant \theta_0$ against $H_1 : \theta > \theta_0$.

Similarly, for testing $H_0 : \theta \in (-\infty, \theta_1] \cup [\theta, \infty)$ and $H_0 : \theta \in [\theta_1, \theta_2]$ (against their corresponding alternatives), we can define tests $\psi_2(v)$ and $\psi_3(v)$, similarly as $\phi_2(u, \mathbf{t})$ and $\phi_3(u, \mathbf{t})$ (cf. (3.4) and (3.5)), by replacing $(u, \mathbf{t})$ with v, $c_i(\mathbf{t}), \gamma_i(\mathbf{t})$ with c_i, γ_i, $\mathbf{E}_{\theta_i}[\phi_j(U, \mathbf{T}) \cdot \mid \mathbf{T} = \mathbf{t}]$ with $\mathbf{E}_{\theta_i}[\psi_j(V)]$, for $i = 1, 2$ and $j = 1, 2$. The form of the test functions remains same (as ψ_1 is similar to ϕ_1).

(ii) If there are functions $a(\mathbf{t}) > 0$ and $b(\mathbf{t})$, such that, the statistic $V(u, \mathbf{t})$ satisfies

$$V(u, \mathbf{t}) = a(\mathbf{t}) \cdot u + b(\mathbf{t}),$$

then, the test $\psi_4(v)$ defined by replacing $(u, \mathbf{t})$ with v, $c_i(\mathbf{t}), \gamma_i(\mathbf{t})$ with c_i, γ_i and expectation w.r.t the distribution of $[U \mid \mathbf{T}]$ by the unconditional distribution of V, will be equivalent to the test ϕ_4 (cf. (3.6)) and will be UMPU size-α.

Proof of Theorem 11. At each fixed $\mathbf{t}$, as $v \uparrow u$, hence $u > c(\mathbf{t})$ is equivalent to $v > d(\mathbf{t})$, for some $d(\mathbf{t})$. Hence the size condition reduces to,

$$P_{\theta_0}(V > d(\mathbf{t}) \mid \mathbf{t}) + \gamma(\mathbf{t}) \cdot P_{\theta_0}(V = d(\mathbf{t}) \mid \mathbf{t}) = \alpha.$$

But, V is independent of $\mathbf{T}$ at θ_0, d and γ do not depend on $\mathbf{t}$. This shows that ψ_1 is equivalent to ϕ_1 and is also UMPU. Similar arguments works for ψ_2 and ψ_3.

In case of $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$, the test ϕ_4 in (3.6) can be written as,

$$\phi_4(u, \mathbf{t}) = \begin{cases} 1 & \text{if } u < c_1(\mathbf{t}) \text{ or } u > c_2(\mathbf{t}), \\ \gamma_i(\mathbf{t}) & \text{if } u = c_i(\mathbf{t}), \text{ for } i = 1, 2, \\ 0 & \text{o.w.} \end{cases}$$

$$\Leftrightarrow \quad \phi_4(u, \mathbf{t}) \equiv \psi_4(v) = \begin{cases} 1 & \text{if } v < d_1(\mathbf{t}) \text{ or } v > d_2(\mathbf{t}), \\ \gamma_i(\mathbf{t}) & \text{if } v = d_i(\mathbf{t}), \text{ for } i = 1, 2, \\ 0 & \text{o.w.} \end{cases} \quad \text{as } a(\mathbf{t}) > 0, \text{ for some } d_i(\mathbf{t}).$$

The linearity condition and the independence of V and $\mathbf{T}$ implies the size-condition in (3.6) reduces to, $\mathbf{E}_{\theta_0}[\psi_4(V) \mid \mathbf{T} = \mathbf{t}] = \alpha$, and

$$\mathbf{E}_{\theta_0}\left[\psi_4(V) \cdot \frac{V - b(\mathbf{T})}{a(\mathbf{T})} \mid \mathbf{T} = \mathbf{t}\right] = \alpha \cdot \mathbf{E}_{\theta_0}\left[\frac{V - b(\mathbf{T})}{a(\mathbf{T})} \mid \mathbf{T} = \mathbf{t}\right]$$
$$\Leftrightarrow \quad \mathbf{E}_{\theta_0}\left[\psi_4(V) \cdot V \mid \mathbf{T} = \mathbf{t}\right] = \alpha \cdot \mathbf{E}_{\theta_0}(V \mid \mathbf{T} = \mathbf{t}).$$

Since, V and $\mathbf{T}$ are independent under θ_0, we can discard the conditioning and the result follows. □

Remark. For Normal families, the constant γ_i will be usually zero. It should be noted that for $H_0 : \theta \in [\theta_1, \theta_2]$ or $H_0 : \theta \notin (\theta_1, \theta_2)$, it is not easy to find such statistics V, which satisfy the provided conditions in part (i) of Theorem 11.

References

Khatri, C. G. (1967). On certain inequalities for normal distributions and their applications to simultaneous confidence bounds. *Ann. Math. Statist.*, 38(6):1853–1867.

Lehmann, E. L. and Romano, J. P. (2005). *Testing Statistical Hypotheses*. Springer Texts in Statistics. Springer, New York, Third edition.

Lehmann, E. L. and Scheffé, H. (1955). Completeness, similar regions, and unbiased estimation. II. *Sankhyā*, 15:219–236.

Schervish, M. J. (1995). *Theory of statistics*. Springer Series in Statistics. Springer-Verlag, New York.