E-values for anytime-valid inference with exponential families

Hao, Y.

Chapter 1

Introduction

In today's world, data surrounds us like air. For instance, devices like the Apple Watch continuously collect data about our daily routines, exercise habits, and more. However, this vast amount of data must be organized and analyzed before meaningful conclusions can be drawn. Solving practical problems with data is at the core of the science called *statistics*. Over the years, statistics has demonstrated its value across various research fields, such as bioinformatics and sociological research. In the era of big data, where new data constantly accumulates, we need statistical methods that can handle these continuous data flows and enable real-time decision-making.

This dissertation focuses on an *anytime-valid method* (called the *e-value method*), a powerful approach designed to tackle hypothesis testing problems within streaming data contexts. Throughout this work, we develop numerous mathematical results concerning the theory of the e-value method for hypothesis testing. In this introductory chapter, we introduce the key topics covered in the dissertation.

First, Section 1.1 introduces the hypothesis testing problem and discusses how it is addressed by classical methods. We then illustrate the problems that arise when these traditional methods are used sequentially as the data come in. In Section 1.2, we present the core concept of the e-value method and explain why it works "safely" for real-time decision-making. Section 1.3 covers preliminary knowledge that is frequently referenced in later chapters but not introduced in detail in those chapters. Section 1.4 provides an introduction to the exponential families, which are a central focus of this dissertation. Lastly, in Section 1.5, we offer an outline of each chapter of the dissertation.

1.1 Hypothesis testing

Hypothesis testing is a common practice in everyday life. For example, during winter, I often feel discomfort in my stomach after having lunch at the CWI canteen. Since I rarely ate cold food in China during winter, I suspect that the cold vegetable salad might be causing this discomfort. To *test* my suspicion, I eat the cold salad on some

random days over the course of a month and monitor how my stomach feels each day. If my stomach consistently feels bad after eating the cold salad but feels fine on other days, this would provide strong evidence that my suspicion is correct. Of course, one could argue that it might just be coincidence, but such coincidence would have an extremely low probability if my suspicion is wrong.

To explain hypothesis testing further, suppose we have collected n observations, denoted $X_{(1)}, X_{(2)}, \ldots$. We are interested in whether these observations are consistent with one of two hypotheses: the *null hypothesis* ($\mathcal{H}_0$) or the *alternative hypothesis* ($\mathcal{H}_1$). In the example above, $\mathcal{H}_0$ might be "Eating cold vegetable salad does NOT cause stomach discomfort", while $\mathcal{H}_1$ would be "Eating cold vegetable salad does cause stomach discomfort". In general, $\mathcal{H}_0$ represents a status quo assumption or a standard model that the data might conform to, while $\mathcal{H}_1$ represents a departure from $\mathcal{H}_0$.

$\mathcal{H}_0$ and $\mathcal{H}_1$ are usually formalized as probability distributions. For example, the data $X_{(1)}, X_{(2)}, \ldots$ are independent and identically distributed (i.i.d.). Each $X_{(i)}$ is of the form $(Y_{(i)}, G_{(i)})$ with $Y_{(i)} \in \{\text{Feel good}, \text{Feel bad}\}$ and $G_{(i)} \in \{\text{Eat salad}, \text{Not eat salad}\}$. Then $\mathcal{H}_0$ is the set of conditional distributions with

$$\begin{aligned}&\Pr(Y_{(i)} = \text{Feel good} | G_{(i)} = \text{Eat salad})\\ =&\Pr(Y_{(i)} = \text{Feel good} | G_{(i)} = \text{Not eat salad});\end{aligned}$$

and $\mathcal{H}_1$ is the set of conditional distributions, in which the first probability is smaller than the second one.

There are many approaches to hypothesis testing, roughly categorized as *frequentist* or *Bayesian* methods, as explained in detail by Royall [74]. The frequentist approach is further divided into *Fisherian* and *Neyman-Pearson tests*. Fisherian testing focuses on measuring the evidence against $\mathcal{H}_0$ using the p-value (defined later); the smaller the p-value, the stronger the evidence against $\mathcal{H}_0$. In this framework, there is no explicit $\mathcal{H}_1$, as the test focuses solely on $\mathcal{H}_0$. On the other hand, the Neyman-Pearson approach explicitly compares two hypotheses, $\mathcal{H}_0$ and $\mathcal{H}_1$, with the goal of choosing one over the other. We will omit Bayesian methods here because we do not use them so much in the thesis.

In some chapters in the thesis, we use $\mathcal{H}_0$ to denote the statement defining the null hypothesis (Chapter 2 and Chapter 3), and $\mathcal{P}$ to denote the corresponding set of distributions. Similarly, we use $\mathcal{H}_1$ to denote the statement defining the alternative, and $\mathcal{Q}$ to denote the distributions. In other chapters including this introduction, we use $\mathcal{H}_0$ and $\mathcal{H}_1$ to directly denote both the set of distributions and their defining statement.

The general goal in all chapters of this thesis is to choose between a null hypothesis $\mathcal{H}_0$ and an alternative $\mathcal{H}_1$ based on the observations $x_{(1)}, x_{(2)}, \ldots$. Both $\mathcal{H}_1$ and $\mathcal{H}_0$ can be composite, meaning they may consist of multiple possible distributions rather than a single, fixed one. For example, consider testing whether a coin is fair. The null hypothesis $\mathcal{H}_0$ asserts that the coin is fair, meaning the probability of getting "heads" in a coin toss is exactly 0.5. In this case, the null model space is simple, corresponding to a Bernoulli distribution with parameter 0.5. $\mathcal{H}_0$ is referred to as a simple hypothesis. On the other hand, the alternative hypothesis $\mathcal{H}_1$ claims the coin is not fair. This means the probability of getting "heads" could be any value except 0.5, specifically any

real number $\theta \in [0, 0.5) \cup (0.5, 1]$. Since this allows for a range of different probabilities, $\mathcal{H}_1$ is referred to as a composite hypothesis.

This is one of the most common settings today. While we will not cover all classical hypothesis testing methods, we will introduce one of the most well-known: the p-value method.

p-value A *p-value* is the probability of obtaining test results at least as extreme as the result actually observed, under the assumption that the null hypothesis $\mathcal{H}_0$ is true. Mathematically, a strict p-value is a random variable p such that for any given significance level $\alpha \in [0, 1]$ and for all $P_0 \in \mathcal{H}_0$,

$$P_0(p \leq \alpha) = \alpha.$$

In practice, for a set of observations $x^{(n)} := x_{(1)}, \ldots, x_{(n)}$, we *reject* $\mathcal{H}_0$ if the p-value computed from these observations is less than or equal to α, which is commonly set at 0.05. Otherwise, we fail to reject $\mathcal{H}_0$.

To better understand this, let us look at a couple of classical examples.

Example 1. [Correct coin toss test] This test examines whether a coin is fair. The same example is discussed in Pérez-Ortiz's PhD dissertation [69]. We toss the coin n times, and from these observations $x^{(n)} = x_{(1)}, \ldots, x_{(n)}$, we record each time whether "heads" or "tails" appears. The empirical mean is defined as:

$$t(x^{(n)}) = \frac{1}{n} \#\{\text{heads in } x^{(n)}\}.$$

The number of heads follows a binomial distribution, and by the Central Limit Theorem (CLT), if $\mathcal{H}_0$ (the coin is fair) is true, $t(X^{(n)})$ approximately complies with a Gaussian distribution.

Suppose we have an observation $x^{(n)}$ and get $t(x^{(n)}) = 1/2 + t^*$. Then the event "If we would replicate the experiment, we get a result $t(X^{(n)})$ that is at least as extreme as the result actually observed" is given by $t(X^{(n)}) \in (-\infty, 1/2 - t^*] \cup [1/2 + t^*, \infty)$. Note that now $X^{(n)}$ refers to "new" data, whereas $x^{(n)}$ refers to the actually observed data—see the red intervals in Figure 1.1. Then the p-value corresponding to statistic $t(x^{(n)})$ is defined as the probability that $t(X^{(n)})$ falls within the red intervals. For each value of t^*, there is a corresponding p-value. If $t^* = 0.98/\sqrt{n}$, then the p-value is 0.05.

The p-value method would reject $\mathcal{H}_0$ if $t(X^{(n)})$ falls inside these red intervals, resulting in a *Type-I error* (i.e., false rejection of $\mathcal{H}_0$) guarantee with a probability of 0.05. That is

$$\text{Type-I error} := \Pr(\textsc{reject } \mathcal{H}_0 | \mathcal{H}_0 \textsc{ is true}),$$

which equals 0.05 in this example. This demonstrates the validity of the p-value method in this example, as it controls the Type-I error at a significance level 0.05.

Example 2. [Incorrect coin toss test] If we extend the previous example to an online streaming case, the p-value method violates the Type-I error control. Here, we keep tossing the coin, collecting new observations, and computing $t(x^{(n)})$ at each step

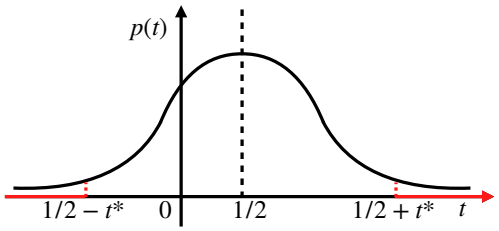

Figure 1.1: If the p-value $p = 0.05$, then $t^* = 0.98/\sqrt{n}$.

until $t(x^{(n)})$ eventually falls outside the interval $[1/2 - 0.98/\sqrt{n}, 1/2 + 0.98/\sqrt{n}]$. Due to the randomness in sampling, extreme values will occur eventually, leading us to reject $\mathcal{H}_0$ even if $\mathcal{H}_0$ is true. In this scenario, the Type-I error becomes 1, showing that the p-value method is no longer valid. This will be illustrated in Figure 1.2.

We will demonstrate that the e-value method remains valid in the scenario described in the above examples, to be discussed in Section 1.2. Furthermore, to illustrate why the Type-I error exceeds α in Example 2, we provide a clearer explanation in the next example.

Example 3. [Multi-stage tests] Consider the following multi-stage experiment. Initially, researchers collect a dataset $X^{(n)}$ and compute the p-value p_1. They reject $\mathcal{H}_0$ if $p_1 < 0.05$, and accept $\mathcal{H}_0$ if $p_1 \geq 0.1$. However, if $0.05 \leq p_1 < 0.1$, they deem the evidence inconclusive but promising. Therefore, they collect a new sample $X'^{(m)}$ and compute a new p-value p_2 based on joining datasets $X^{(n)}$ and $X'^{(m)}$. They reject $\mathcal{H}_0$ if $p_2 \leq 0.05$, otherwise they accept $\mathcal{H}_0$.

Let us represent the event where $0.05 < p_1 < 0.1$ as G. The total Type-I error exceeds 0.05 because:

$$\text{Type-I error} = 0.05 + P_0(G) \cdot P_0(p_2 < 0.05 \mid G) > 0.05.$$

This shows that the p-value method fails to control the Type-I error in the multi-stage testing, which already happens when there are just 2 stages. In practice, there are often multiple stages, which further increases the overall Type-I error.

In the next subsection, we will show that the e-value method consistently succeeds in the scenarios presented in these examples. More generally, the test process using the e-value method can be halted at any time without requiring a predefined stopping rule, offering greater flexibility compared to traditional methods.

1.2 Anytime-valid tests: e-value, e-process

In Section 1.1, we discussed a flaw of the p-value in sequential testing. The issue of sampling until a significant result is obtained has actually been debated by statisticians since at least the 1940s. Feller in 1940 [34] observed this issue in studies of extra-sensory

perception, and Anscombe in 1954 [4] famously called it "sampling to a foregone conclusion." Robbins in 1952 [73] also pointed out this problem. However, it was not until 2019 that a fully general framework emerged to address it. In that year, four papers from different research groups were published on arXiv, laying the foundation for what would soon become known as the concept of the *e-value* [42, 95, 91, 76]. We will now introduce this general framework.

e-variable, e-value Consider a batch of data (a random vector) X, which can be collected sequentially or all at once. We define a nonnegative statistic $S(X)$, that is a function of the observed data. Let $\mathcal{H}_0 = \{P_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$ be the set of distributions for X defined within their parameter space $\boldsymbol{\Theta}$. If the statistic $S(X)$ satisfies the condition:

$$\text{for all } P \in \mathcal{H}_0 : \qquad \mathbb{E}_P\left[S(X)\right] \leq 1, \tag{1.2.1}$$

then we refer to $S(X)$ as an *e-variable relative to* $\mathcal{H}_0$. The value of $S(X)$ computed from the observed data is called the *e-value*.

Similar to how we test with p-values, we first set a significance level $\alpha \in [0, 1]$ before conducting an e-value-based test. Given a sample X, we construct an e-variable $S(X)$ using the entire sample at once, and we reject $\mathcal{H}_0$ if and only if $S(X) \geq 1/\alpha$. This is analogous to the traditional p-value approach. This approach ensures a Type-I error guarantee, as can be explained by Markov's inequality:

$$\text{Type-I error} = \Pr(S(X) \geq 1/\alpha | \mathcal{H}_0 \text{ is true}) \leq \frac{\sup_{P \in \mathcal{H}_0} \mathbb{E}_P[S(X)]}{1/\alpha} \leq \alpha.$$

However, data is sometimes collected sequentially, requiring us to make decisions at any time using the data collected so far. A specific example of this is provided in Example 2 and Example 3. In contrast to p-value methods, the e-value method can also be applied for testing with a data stream while maintaining a Type-I error guarantee. We divide this situation into two types: *optional continuation* and *optional stopping*.

Optional continuation Consider a stream of data batches $X_{(1)}, X_{(2)}, \ldots$. We assume the $X_{(i)}$ are independent. Let $S(X_{(i)})$ be an e-variable based on $X_{(i)}$. Then, for any positive integer $N \in \mathbb{N}^+$, we define the product $S^{(N)} := \prod_{i=1}^{N} S(X_{(i)})$, which remains an e-variable. In other words,

$$\text{for all } P \in \mathcal{H}_0 : \qquad \mathbb{E}_P\left[S^{(N)}\right] \leq 1, \tag{1.2.2}$$

which can be shown as follows: since $S(X_{(1)}), \ldots, S(X_{(N)})$ are independent, it follows that for all $P \in \mathcal{H}_0$,

$$\mathbb{E}_P\left[S^{(N)}\right] = \prod_{i=1}^{N} \mathbb{E}_P\left[S(X_{(i)})\right] \leq 1.$$

Then, if we reject $\mathcal{H}_0$ when $S^{(N)} \geq 1/\alpha$, the Type I error is bounded by level α. This is because, by Markov's inequality,

$$\text{Type-I error} = \Pr\left(S^{(N)} \geq 1/\alpha \middle| \mathcal{H}_0 \text{ IS TRUE}\right) \leq \frac{\sup_{P \in \mathcal{H}_0} \mathbb{E}_P\left[S^{(N)}\right]}{1/\alpha} \leq \alpha. \tag{1.2.3}$$

This inequality holds for every fixed N. However, as shown by Grünwald et al [42], (1.2.3) still holds if N is any data dependent stopping time. A stopping time is a time determined by a rule that, at each step N, decides—based on the data observed so far, $X_1, X_2, \ldots, X_N$—whether to stop or continue. For example: "stop as soon as $S^{(N)} \geq 1/\alpha$", or "$N = 10$", or "stop if $X_{(N)}$ contains 0". For a formal definition of stopping time, we refer to Ramdas et al. [71].

In more detail, Grünwald et al [42] show that $S^{(1)}, S^{(2)}, \ldots$ is a *test supermartingale* [77], which is a nonnegative supermartingale with $\mathbb{E}_P[S^{(1)}] \leq 1$. This implies (1.2.3) and also implies that Ville's inequality [90] holds, then further implies, for all $P \in \mathcal{H}_0$, we have

$$\text{Type-I error} \leq \Pr\left(\sup_{N:N>0} S^{(N)} \geq 1/\alpha \middle| \mathcal{H}_0 \text{ IS TRUE}\right) \overset{(*)}{\leq} \frac{\mathbb{E}_P[S^{(1)}]}{1/\alpha} \leq \alpha, \tag{1.2.4}$$

where $(*)$ follows from Ville's inequality. We say that the e-value method preserves Type-I error guarantees under *optional continuation*.

Optional stopping Now suppose each $X_{(i)}$ is a single (just one) data point and we have an e-variable $S(X_{(i)})$ for each point $X_{(i)}$. But now by (1.2.4), we can do optional stopping - stop at any point we like, and still preserve Type-I error guarantees.

e-process So far, we have assumed that the data stream $X_{(1)}, X_{(2)}, \ldots$ is composed of independent observations, which also led to the independence of $S(X_{(1)}), S(X_{(2)}), \ldots$ by construction. However, in practice, they can be dependent. In this thesis, the $X_{(i)}$ in the data streams considered will always be independent, but we may use different e-variables $S^{(i)}$ that could be dependent. We directly define a stochastic process $S^{(1)}, S^{(2)}, \ldots$, with $S^{(i)}$ a nonnegative statistic of $X_{(1)}, \ldots, X_{(i)}$, satisfying

$$\text{for all } P \in \mathcal{H}_0 : \qquad \sup_{N \in \mathcal{T}} \mathbb{E}_P\left[S^{(N)}\right] \leq 1, \tag{1.2.5}$$

where $\mathcal{T}$ is the set of all valid stopping times. We call such a process an *e-process*.

Ramdas et al. [72] show that if a stochastic process $S^{(1)}, S^{(2)}, \ldots$ is an e-process, then for all $P \in \mathcal{H}_0$, there is a test martingale $M_P^{(1)}, M_P^{(2)}, \ldots$ such that $S^{(i)} \leq M_P^{(i)}$ holds for every $i > 0$. Therefore, an e-process ensures that the Type-I error remains bounded, which is derived by Ville's inequality again, as in (1.2.4).

Let us run a simple simulation comparing an e-process with the traditional p-value method, as illustrated in Example 2. The results of this simulation are shown in Figure 1.2. In this example, the p-value method fails to maintain the Type-I error

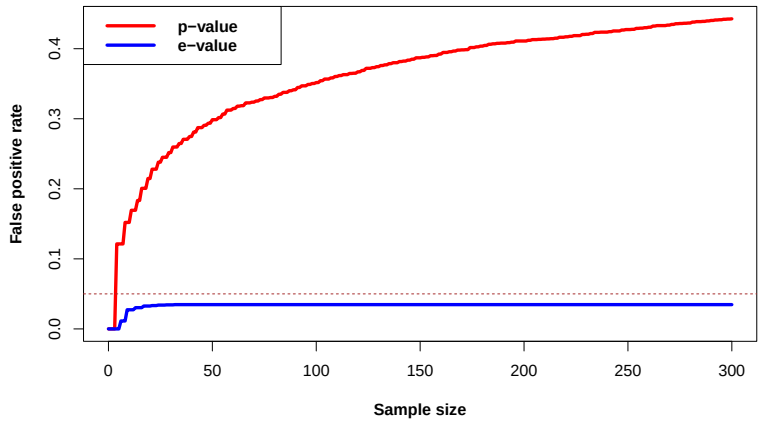

Figure 1.2: A toy simulation for Example 2 is conducted. Similar simulations are also presented by Ly et al. [65] and Turner et al. [88]. In this setup, samples are drawn from a Bernoulli(1/2) distribution. We define $S_{(i)} := \frac{p_{0.9}(X_{(i)})}{p_{0.5}(X_{(i)})}$, where $p_{0.9}(X_{(i)})$ denotes the probability mass function (pmf) of a Bernoulli(0.9) distribution. $S_{(i)}$ is an e-variable because it can easily be checked that it complies with (1.2.1). We obtain a new observation at each time step, then calculate the p-value $p(t(X^{(t)}))$ and e-value $S^{(t)} := \prod_{i=1}^{t} S_{(i)}$ at time t using the first t samples observed. This process continues until all samples are examined. If, at any time $t \in \{1, 2, \ldots, 300\}$, $p(t(X^{(t)})) \leq 0.05$, we reject $\mathcal{H}_0$; similarly, we reject $\mathcal{H}_0$ based on the e-value if there exists a time $t' \in \{1, 2, \ldots, 300\}$ such that $S^{(t')} \geq 20$. This simulation is repeated 3000 times, and we compute the rejection rates for both the p-value and e-value. This simulation shows that the Type-I error of sequentially testing $S^{(t)} \geq \frac{1}{0.05}$ remains forever below 0.05, whereas the Type-I error of sequentially using the $p(t(X^{(t)})) \leq 0.05$ violates the level 0.05.

guarantee, while the e-process method successfully controls the Type-I error rate as expected. This demonstrates the robustness of the e-process approach in contrast to the p-value method, which can be prone to inflation of Type-I error in certain scenarios.

In this thesis, we rely on various pieces of preliminary knowledge, though they may not be explicitly mentioned in each corresponding chapter. Therefore, we have included essential preliminary knowledge in the following subsections.

1.3 Preliminary knowledge: RIPr and e-power

RIPr Suppose we have a distribution Q and a set of distributions $\mathcal{P}$. The goal is to find the distribution in $\mathcal{P}$ that is 'closest' to Q. One common way to measure the divergence between two distributions is by using Kullback-Leibler (KL) Divergence, denoted as $D(Q\|P)$, which is defined as:

$$D(Q\|P) := \mathbb{E}_{X\sim Q}\left[\log \frac{q(X)}{p(X)}\right],$$

where $q(X)$ and $p(X)$ are the probability densities of Q and P, respectively.

We define $\mathcal{W}(\mathcal{P})$ to be the Choquet convex hull of $\mathcal{P}$. This means that $\mathcal{W}(\mathcal{P})$ is a convex set, and distribution $P_W \in \mathcal{W}(\mathcal{P})$, if and only of there is a proper prior W on $\mathcal{P}$ such that:

$$p_W(X) = \int p(X) dW(p).$$

It is clear that $\mathcal{P} \subseteq \mathcal{W}(\mathcal{P})$ because, for any $P \in \mathcal{P}$, we can place all the prior mass on P. We define P^*, the *Reverse Information Projection* (RIPr) of Q onto $\mathcal{P}$ [60], [27], as the distribution in $\mathcal{W}(\mathcal{P})$ that minimizes the KL-divergence from Q to $\mathcal{W}(\mathcal{P})$. This means that, if the minimum in $\mathcal{W}(\mathcal{P})$ can be attained, P^* is the distribution that is 'closest' to Q in the KL-divergence sense:

$$P^* = \underset{P\in\mathcal{W}(\mathcal{P})}{\arg\min}\, D(Q\|P). \tag{1.3.1}$$

In our simplified introductory statement here, the RIPr is undefined if there is no distribution in $\mathcal{W}(\mathcal{P})$ that minimizes the KL divergence. However, the RIPr can be defined for such cases as well - see [42], [43] and [58].

e-power In traditional hypothesis testing with a p-value, we use the term 'power' to describe the probability that a test rejects the null hypothesis $\mathcal{H}_0$ when the alternative hypothesis $\mathcal{H}_1$ is true. Similarly, in the context of e-values, we aim to measure a test's effectiveness through a concept called 'e-power'. The *e-power* of an e-variable $S = S(X)$ based on a data X and alternative $\mathcal{H}_1 = \{Q\}$ is defined as the expected logarithm of

S under the alternative distribution Q:

$$\mathbb{E}_{X\sim Q}[\log S(X)] := \int \log S(x)\, dQ(x),$$

as introduced by [42] and [92].

Let $\mathcal{H}_1 = \{Q\}$ and $\mathcal{H}_0 = \mathcal{P}$. Grünwald et al. [42] prove that $S_{\text{RIP}} = \frac{q(X)}{p^*(X)}$ is an e-variable, where P^* is specified as in (1.3.1). They also demonstrate that among all e-variables for $\mathcal{H}_0$, the RIPr e-variable S_{RIP} yields the highest e-power relative to the alternative $\mathcal{H}_1$ and the null $\mathcal{H}_0$.

In p-value testing, rejecting $\mathcal{H}_0$ requires a small p-value, but in the e-value framework, $\mathcal{H}_0$ is rejected when $S > 1/\alpha$. Therefore, if $\mathcal{H}_1$ is true, we want S to be as large as possible. The e-power captures this by measuring the strength of S in providing evidence against $\mathcal{H}_0$.

1.4 Preliminary knowledge: Exponential family

The probability density function (pdf) of the exponential distribution is well-known and can be expressed as:

$$p_\lambda(x) = \lambda e^{\lambda x}, x \in [0, \infty), \lambda \in (0, \infty), \tag{1.4.1}$$

where x represents the data, and λ is the rate parameter. Interestingly, many other families, such as the Gaussian, Poisson and Beta distributions, can be written in a similar form to the exponential distribution. These types of distributions are part of a broader class known as *exponential families*. We will now explain this concept in more detail.

Definition of exponential family Consider a set of probability distributions $\mathcal{P} \in \{P_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$ for data U, where $\boldsymbol{\Theta}$ is the parameter space of $\mathcal{P}$. If there exists a re-parametrization $\mathcal{P} = \{P_{\boldsymbol{\beta}} : \boldsymbol{\beta} \in \mathtt{B}\}$ with $\mathtt{B} \subset \mathbb{R}^d$ for some $d \in \mathbb{N}^+$, and a random vector X that is a function of U and the probability density (or mass) functions can be written as:

$$p_{\boldsymbol{\beta}}(U) = \frac{1}{Z(\boldsymbol{\beta})} \exp\left(\boldsymbol{\beta}^\top X\right), \tag{1.4.2}$$

where $Z(\boldsymbol{\beta})$ is a normalizing factor, then $\mathcal{P}$ is called an exponential family. We call X the *sufficient statistic* for $\boldsymbol{\beta}$, which can be verified easily using the *Fisher–Neyman factorization theorem*. $\boldsymbol{\beta}$ is the *natural* or *canonical parameter* of the distribution. (Note that in Chapter 4, we use λ for the canonical parameter, whereas in Chapter 5, we use θ.) When the functions X and $\boldsymbol{\beta} := \boldsymbol{\beta}(\boldsymbol{\theta})$ are fixed, they define a specific exponential family. Some people prefer to write the exponential family form as:

$$p_{\boldsymbol{\beta}}(U) = \frac{1}{Z(\boldsymbol{\beta})} \exp\left(\boldsymbol{\beta}^\top X\right) h(U),$$

where $h(U)$ is called the *carrier function*. However, this is essentially the same as the previous form since it can be rewritten as:

$$p_{\boldsymbol{\beta}}(U) = \frac{1}{Z(\boldsymbol{\beta})} \exp\left(\boldsymbol{\beta}'^{\top} X'\right),$$

where $\boldsymbol{\beta}' = (\boldsymbol{\beta}^\top, 1)^\top$ and $X' = (X^\top, \log h(U))^\top$.

This re-parametrization highlights how various familiar distributions can be unified under the exponential family framework. We explain it using the Gaussian example.

Gaussians are an exponential family For simplicity, we only show the one-dimensional case here. In this example, the standard parameterization would be $\boldsymbol{\theta} = (\mu, \sigma^2)$ and $\boldsymbol{\Theta} = \mathbb{R} \times \mathbb{R}^+$. The pdf of a Gaussian can be written as

$$\begin{aligned} p_{\mu,\sigma^2}(u) =& \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(u-\mu)^2\right) \\ =& \exp\left(-\frac{1}{2\sigma^2}u^2 + \frac{\mu}{\sigma^2}u - \frac{1}{2\sigma^2}\mu^2 - \log\sigma - \frac{1}{2}\log(2\pi)\right) \\ =& \exp\left(\boldsymbol{\beta}^\top x - \log Z(\boldsymbol{\beta})\right), \end{aligned}$$

where the last equality holds if we let $\boldsymbol{\beta} = (\frac{\mu}{\sigma^2}, -\frac{1}{2\sigma^2})^\top$, $x = (u, u^2)^\top$ and $\log Z(\boldsymbol{\beta}) = \log \int \exp\left(\boldsymbol{\beta}^\top x\right) dx = \frac{1}{2\sigma^2}\mu^2 + \log\sigma + \frac{1}{2}\log(2\pi)$.

In this section, we introduced the basic concept of the exponential family. However, throughout this dissertation, more advanced knowledge about the exponential family is frequently used, though not explicitly explained in the corresponding chapters. We have placed these more technical details in Appendix 1.A.

1.5 Outline

We provide a brief overview of the content of each chapter in the following subsections. Each chapter reports our research on various e-variables for exponential family null hypothesis and/or alternative hypothesis, addressing different problems. Additionally, Chapter 5 introduces a novel concentration inequality for multivariate exponential families.

Chapter 2: Conditions for the existence of simple e-variable

In many academic fields, such as social sciences, biology, and physics, researchers often aim to infer the underlying distribution of data. To do this, they may assume a specific model structure. For example, they might hypothesize that the data follow a binomial distribution, then estimate the parameters that best fit the data. However, the data may come from a different set of distributions, such as a set of negative binomial distribution. In mathematics, a correct assumption is referred to as a well-specified

model assumption. Therefore, determining whether a model is well-specified is crucial for accurately learning the structure of the data. This involves testing whether the observed data are actually distributed according to an element of the given set of distributions.

Chapter 2 addresses this task using e-variables for exponential families. Specifically, we may want to test if a certain parameter in an exponential family is zero, or not—this includes linear regression testing as a particular case. In this chapter, we study the GRO e-variable for this task, for a single outcome U. As was announced in Section 1.3, the GRO e-variable is closely connected with the RIPr. The simplest example of GRO e-variables is the likelihood ratio between simple alternative and simple null hypotheses. However, for composite hypotheses, the situation becomes more complex. Nevertheless, GRO e-variables in the form of a likelihood ratio involving a single, specific element of the composite null hypothesis can sometimes still be found. We refer to such GRO e-variables as 'simple' e-variables. As we will demonstrate, their existence is closely linked to properties of the aforementioned RIPr.

When simple e-variables exist, they can be easily computed and are known to be optimal in terms of e-power [53, 42]. In the context of repeated experiments with a fixed stopping rule for data collection, and a simple alternative, using a simple e-variable will, asymptotically, provide the strongest evidence against the null hypothesis compared to other e-variables. Therefore, it is important to determine when simple e-variables exist in specific contexts. Chapter 2 offers a set of equivalent conditions under which simple e-variables exist for exponential family null hypotheses.

Chapter 3: General exponential family test

Chapter 3 continues from Chapter 2. In this chapter, we explore the scenario where a condition 'opposite' to the previous conditions applies, which we refer to as the *anti-simple case*, meaning that simple e-variables do not exist. For both cases—whether simple e-variables exist or not—we analyze common types of e-variables and e-processes related to composite exponential family nulls, but now for sequences of outcomes rather than a single one: we examine and compare their *e-power* [94] for i.i.d. data $U_{(1)}, U_{(2)}, \ldots$. Recall that e-power plays a pivotal role, as it is maximized by the optimal e-variable (i.e. GRO e-variable) across all e-variables defined on $U^{(n)}$. As we announced in Section 1.3, it can be determined using the *reverse information projection (RIPr)*. We denote this optimal e-variable as S_{RIP}. Additionally, we consider a sequentialized version of the RIPr e-variable, $S_{\text{SEQ-RIP}}$, which is optimal at the individual outcome level but not necessarily over the entire sample. We also investigate a *conditional* e-variable, S_{COND}, based on conditioning on the sufficient statistic, along with a well-known version of the *universal inference* e-variable, S_{UI} [95].

Instantiating such e-variables requires specifying an alternative hypothesis. We begin by considering a simple alternative $\mathcal{Q} = \{Q\}$. Our results demonstrate that the RIPr prior W that achieves the minimum in (3.1.2) is approximately Gaussian with variance $O(1/n)$ in an asymptotic sense, and exactly if $\mathcal{H}_0$ is a Gaussian location family and Q is also Gaussian. To our knowledge, this is the first time that insights into a nondegenerate RIPr prior have been obtained for the case of a parametric, non-convex

null.

This result is made possible by our key theoretical insight: the conditional e-variable S_{COND} can be analyzed using a local central limit theorem with explicit bounds on the error terms [16]. Consequently, we derive not only explicit $o(1)$ bounds on its e-power but also establish that S_{COND} is closely related to S_{RIP} (in the Gaussian anti-simple case, they even coincide). We extend these results to other types of e-variables, not only under the 'true' alternative Q but also in the *misspecified case*, where the data are sampled i.i.d. from a distribution $R \neq Q$.

We employ two standard methods to design e-variables for composite alternatives $\mathcal{Q}$: the sequential plug-in method [38] and the method of mixtures [71]. We observe that, when using the method of mixtures and equipping the alternative with a prior W_1, under regularity conditions, the RIPr prior W in (3.1.2) is, approximately, the *same* prior W_1, regardless of whether we are in the simple case or not. Summarizing some of our main findings for the composite case, we derive the following relationships. Under appropriate (though mild) regularity conditions on $\mathcal{H}_0$ and $\mathcal{H}_1$, for all $Q \in \mathcal{H}_1$, (in the chapter we use $\mathcal{P}$ and $\mathcal{Q}$ because it fits better with other notations.) we have:

$$\begin{aligned}
\mathbb{E}_Q[\log S^{(n)}_{\text{RIP}}/S^{(n)}_{\text{COND}}] &= o(1).\\
\mathbb{E}_Q[\log S^{(n)}_{\text{COND}}/S^{(n)}_{\text{UI}}] &= \frac{d}{2}\log n + O(1).\\
\mathbb{E}_Q[\log S^{(n)}_{\text{SEQ-RIP}}/S^{(n)}_{\text{UI}}] &= \frac{d_{qp}}{2}\log n + O(1)\\
&\quad \text{with } 0 < d_{qp} < d, \text{ in the strict simple case.}\\
\mathbb{E}_Q[\log S^{(n)}_{\text{SEQ-RIP}}/S^{(n)}_{\text{COND}}] &\leq -n\epsilon\\
&\quad \text{for some } \epsilon > 0, \text{ all large } n, \text{ in the strict anti-simple case.}
\end{aligned}$$

Here, d represents the dimensionality of the exponential family, and d_{qp} is a measure of 'effective dimension', whose exact value depends on Q. Of course, we provide precise definitions of "strict simple, anti-simple" in Chapter 3.

Chapter 4: k-sample tests with exponential families

A k-sample test is the general version of a two-sample test. It involves analyzing data from k independent random samples, each drawn from a possibly different population. For example when $k = 2$, in studying the effectiveness of a new treatment (such as a new blood pressure medication), patients are divided into two groups: a treatment group and a placebo group. Researchers track the number of recoveries in each group after a set treatment period. If the treatment is effective, a higher recovery rate is expected in the treatment group compared to the placebo group. The statistical test is used to determine if the observed difference in recovery rates between the two groups is significant, or if it could have occurred by chance. Two-sample tests are designed to model such situations. Mathematically, the objective is to determine whether the observed difference between the two populations is statistically significant, meaning whether the difference is likely due to chance or represents a true difference between

the populations.

Chapter 4 centers on k-sample tests. Some of the results presented in the chapter are special cases of those from Chapter 3. However, it remains valuable to include them in this chapter because we provide full details that were not covered in Chapter 3. We develop four (pseudo-) e-variables for k-sample tests in exponential families: $S_{\text{RIP}}, S_{\text{COND}}, S_{\text{MIX}}$, and S_{PSEUDO}. $S_{\text{RIP}}, S_{\text{COND}}, S_{\text{MIX}}$ are real e-variables, while S_{PSEUDO} is an e-variable only when it coincides with S_{RIP}, which happens whenever the latter is computationally straightforward; in other instances, it is not a true e-variable but remains useful for our theoretical analysis. Suppose the (shortest) ℓ_2-distance between the k-dimensional parameter of the alternative and the null parameter space is denoted by δ. Our results show that, for any two of the aforementioned e-variables S and S', the difference in e-power is given by $\mathbb{E}[\log S - \log S'] = O(\delta^4)$.

Chapter 5: GROW e-variables and a novel concentration inequality

The link between optimal rejection regions for anytime-valid tests at a fixed level α and optimal anytime-valid concentration inequalities is well-documented [47]. Chapter 5 explores a variation of this connection, focusing on a simple multivariate null hypothesis and a range of composite alternatives. We examine both absolute and relative *GROW* ('growth-rate optimal in the worst-case') e-variables as introduced by Grünwald et al. [42]. Further, we illustrate how these e-variables connect to a concentration inequality, which we refer to as the *Csiszár-Sanov-Chernoff* (CSC) inequality.

To start, we analyze the GROW e-variable S_{GROW} within this framework, considering cases where $\mathcal{H}_1$ is either the set $\mathcal{P}_1$ of all distributions with means in a specified *convex* set $\mathtt{M}_1$, the set $\mathcal{E}_1$ of all distributions in the exponential family generated by P_0 with means in $\mathtt{M}_1$, or any $\mathcal{H}_1$ for which $\mathcal{E}_1 \subset \mathcal{H}_1 \subset \mathcal{P}_1$. Remarkably, the GROW e-variables coincide across all such $\mathcal{H}_1$. We derive this result by applying the well-known Csiszár-Topsøe Pythagorean theorem for relative entropy, which leads us to the fundamental CSC concentration inequality. This section's focus is primarily on rephrasing established findings, familiar to the information-theoretic community but perhaps less so to those working with e-values.

Chapter 5 then introduces a novel approach, examining cases where the *complement* of $\mathtt{M}_1$ forms a connected, bounded set containing P_0 — a scenario more commonly encountered in practical applications and more aligned with the multivariate central limit theorem (CLT). This configuration, which we call the *surrounding* $\mathcal{H}_1$ case because P_0 is "surrounded" by $\mathcal{H}_1$, has rarely been considered in the derivation of CSC bounds, with an exception being the variation studied by Kaufmann and Koolen [51].

We extend the previous S_{GROW} e-variable to this surrounding $\mathcal{H}_1$ case in two ways. The first approach is a straightforward *absolute* extension of the GROW e-variable to the multivariate case, still denoted as S_{GROW}. Or we can determine a *relatively* optimal GROW e-variable S_{REL} that is as close as possible to the largest S_{GROW} among all e-variables S_{GROW} that can be defined on convex subsets of $\mathcal{H}_1$, where we define relative optimality in a minimax-regret sense. We characterize S_{GROW} for the univariate case ($d = 1$) while leaving the multidimensional case ($d > 1$) as an open problem, and we fully characterize S_{REL} for general dimensions. We then show that S_{REL} leads again

to a CSC bound — and this CSC bound is new.

Appendix 1.A More exponential family preliminaries

Mean parameterization Since X is the sufficient statistic for parameter, it is often more useful to directly study X rather than U, as we often do in this dissertation. From equation (1.4.2), we know that X has the same dimension as the canonical parameter $\boldsymbol{\beta}$. We already know that when the functions $X(U)$ and $\boldsymbol{\beta}(\boldsymbol{\theta})$ are fixed, we may define a specific class of exponential family distributions, denoted by $\mathcal{P}$ (e.g., the Gaussian family). A natural approach is to represent $P_{\boldsymbol{\theta}}$ using the expectation of X, defined as $\boldsymbol{\mu}(\boldsymbol{\theta}) := \mathbb{E}_{U\sim P_{\boldsymbol{\theta}}}[X]$, which is called the *mean parameterization*, and $\boldsymbol{\mu}(\boldsymbol{\theta})$ is the *mean-value parameter*. It can be shown that for $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \boldsymbol{\Theta}$, ($\boldsymbol{\Theta}$ is the standard parameter space of the distribution family $\mathcal{P}$), if $P_{\boldsymbol{\theta}} \neq P'_{\boldsymbol{\theta}}$, then $\boldsymbol{\mu}(\boldsymbol{\theta}) \neq \boldsymbol{\mu}(\boldsymbol{\theta}')$, ensuring a distinct probability model for each parameter.

Canonical parameterization For simplicity, we use the notation $\boldsymbol{\beta} := \boldsymbol{\beta}(\boldsymbol{\theta}), \boldsymbol{\mu} := \boldsymbol{\mu}(\boldsymbol{\theta})$ going forward. Since $Z(\boldsymbol{\beta})$ is the normalizing factor, we have:

$$Z(\boldsymbol{\beta}) = \int \exp\left(\boldsymbol{\beta}^\top x\right) dx,$$

where the integral becomes a sum in the discrete case. Then taking the first derivative of $\log Z(\boldsymbol{\beta})$ w.r.t. $\boldsymbol{\beta}$ gives us the mean of X under $P_{\boldsymbol{\beta}}$:

$$\frac{d\log Z(\boldsymbol{\beta})}{d\boldsymbol{\beta}} = \frac{\int x \exp\left(\boldsymbol{\beta}^\top x\right) dx}{\int \exp\left(\boldsymbol{\beta}^\top x\right) dx} = \int x \cdot \frac{1}{Z(\boldsymbol{\beta})} \exp\left(\boldsymbol{\beta}^\top x\right) dx = \mathbb{E}_{P_{\boldsymbol{\beta}}}[X] := \boldsymbol{\mu}(\boldsymbol{\beta}). \tag{1.A.1}$$

We call $\boldsymbol{\mu}(\boldsymbol{\beta})$ the mean-value parameter corresponding to $\boldsymbol{\beta}$. We continue to take the second derivative of $\log Z(\boldsymbol{\beta})$ w.r.t. $\boldsymbol{\beta}$. This gives us the covariance matrix of X under $P_{\boldsymbol{\beta}}$:

$$\Sigma_P = \frac{d^2 \log Z(\boldsymbol{\beta})}{d\boldsymbol{\beta}^2} = \frac{d\boldsymbol{\mu}(\boldsymbol{\beta})}{d\boldsymbol{\beta}}.$$

Since Σ_P is positive definite, the transformation from $\boldsymbol{\beta}$ to $\boldsymbol{\mu}$ is one-to-one, ensuring that $\mathcal{P}$ can be uniquely represented using the canonical parameter $\boldsymbol{\beta}$.

Empirical mean as MLE In the mean parameter space, the *maximum likelihood estimator (MLE)* for the sufficient statistic X generated from a data set $\mathcal{U}$ has an important property, frequently used in the following chapters. Consider a set of i.i.d. data points $x_{(1)}, x_{(2)}, \ldots, x_{(n)} := x^{(n)}$. The likelihood function is given by:

$$\mathcal{L}(\boldsymbol{\beta} \mid x^{(n)}) = \frac{1}{Z(\boldsymbol{\beta})^n} \exp\left(\boldsymbol{\beta}^\top \sum_{i=1}^{n} x_{(i)}\right).$$

To find the MLE, we take the derivative of the log-likelihood and set it to zero:

$$\frac{d \log \mathcal{L}(\boldsymbol{\beta} \mid x^{(n)})}{d\boldsymbol{\beta}} = -n \frac{d \log Z(\boldsymbol{\beta})}{d\boldsymbol{\beta}} + \sum_{i=1}^{n} x_{(i)} \overset{(a)}{=} -n\boldsymbol{\mu} + \sum_{i=1}^{n} x_{(i)} = 0. \tag{1.A.2}$$

where (a) follows from (1.A.1). If $\frac{1}{n}\sum_{i=1}^{n} x_{(i)}$ is in the interior of the the mean-value parameter space, then (1.A.2) has solution. This shows that in such cases, the MLE for the mean parameter is $\hat{\boldsymbol{\mu}} = \frac{1}{n}\sum_{i=1}^{n} x_{(i)}$, i.e., the empirical mean.

Robustness properties These properties are shown in detail in [35, Chapter 19]. We briefly introduce some of them that are used frequently in the following chapters. Let $\mathcal{P}$ be a regular exponential family, take $X = U$ and let $\mathtt{M}$ be its mean parameter space (See [13] or [19] for the definition of regular). Consider $P_{\boldsymbol{\mu}} \in \mathcal{P}$ such that $\mathbb{E}_{P_{\boldsymbol{\mu}}}[X] = \boldsymbol{\mu} \in \mathtt{M}$. Let Q be an arbitrary distribution with $\mathbb{E}_Q[X] = \boldsymbol{\mu}^* \in \mathtt{M}$. For all $P_{\boldsymbol{\mu}} \in \mathcal{P}$, we have:

$$\mathbb{E}_Q\left[\log \frac{p_{\boldsymbol{\mu}^*}(X)}{p_{\boldsymbol{\mu}}(X)}\right] = \mathbb{E}_{P_{\boldsymbol{\mu}^*}}\left[\log \frac{p_{\boldsymbol{\mu}^*}(X)}{p_{\boldsymbol{\mu}}(X)}\right] := D(P_{\boldsymbol{\mu}^*} || P_{\boldsymbol{\mu}}),$$

where $D(P_{\boldsymbol{\mu}^*} || P_{\boldsymbol{\mu}})$ is the KL-divergence between $P_{\boldsymbol{\mu}^*}$ and $P_{\boldsymbol{\mu}}$. This statement holds in the canonical parameter space ($\mathtt{B}$) of $\mathcal{P}$ as well. For all $P_{\boldsymbol{\beta}} \in \mathcal{P}$ with $\boldsymbol{\beta} \in \mathtt{B}$, we have:

$$\mathbb{E}_Q\left[\log \frac{p_{\boldsymbol{\beta}^*}(X)}{p_{\boldsymbol{\beta}}(X)}\right] = \mathbb{E}_{P_{\boldsymbol{\beta}^*}}\left[\log \frac{p_{\boldsymbol{\beta}^*}(X)}{p_{\boldsymbol{\beta}}(X)}\right] := D(P_{\boldsymbol{\beta}^*} || P_{\boldsymbol{\beta}}),$$

which is equivalent to the statement in $\mathtt{M}$ because $P_{\boldsymbol{\beta}}$ and $P_{\boldsymbol{\mu}}$ with $\boldsymbol{\mu} = \boldsymbol{\mu}(\boldsymbol{\beta})$ represent the same distribution.

Moreover, $\mathbb{E}_{P_{\boldsymbol{\beta}^*}}[-\log p_{\boldsymbol{\beta}}(X)]$ is a strictly convex function with respect to $\boldsymbol{\beta}$, achieving its unique minimum at $\boldsymbol{\beta}^*$. Since $P_{\boldsymbol{\beta}}$ and $P_{\boldsymbol{\mu}}$ represent the same distribution, $\mathbb{E}_{P_{\boldsymbol{\mu}^*}}[-\log p_{\boldsymbol{\mu}}(X)]$ is also a function of $\boldsymbol{\mu}$, achieving its unique minimum at $\boldsymbol{\mu}^*$.