

Stephen F. Austin State University

SFA ScholarWorks

Electronic Theses and Dissertations

8-2024

A UNIFORMLY MOST POWERFUL TEST FOR THE MEAN OF A BETA DISTRIBUTION

Richard Ntiamoah Kyei

Stephen F Austin State University, kyeirn@jacks.sfasu.edu

Follow this and additional works at: <https://scholarworks.sfasu.edu/etds>



Part of the [Applied Statistics Commons](#), [Mathematics Commons](#), [Other Physical Sciences and Mathematics Commons](#), [Other Statistics and Probability Commons](#), [Statistical Methodology Commons](#), and the [Statistical Theory Commons](#)

[Tell us](#) how this article helped you.

Repository Citation

Kyei, Richard Ntiamoah, "A UNIFORMLY MOST POWERFUL TEST FOR THE MEAN OF A BETA DISTRIBUTION" (2024). *Electronic Theses and Dissertations*. 565.

<https://scholarworks.sfasu.edu/etds/565>

This Thesis is brought to you for free and open access by SFA ScholarWorks. It has been accepted for inclusion in Electronic Theses and Dissertations by an authorized administrator of SFA ScholarWorks. For more information, please contact cdsscholarworks@sfasu.edu.

A UNIFORMLY MOST POWERFUL TEST FOR THE MEAN OF A BETA DISTRIBUTION

Creative Commons License



This work is licensed under a [Creative Commons Attribution-Noncommercial-No Derivative Works 4.0 License](https://creativecommons.org/licenses/by-nc-nd/4.0/).

A UNIFORMLY MOST POWERFUL TEST FOR THE MEAN OF A BETA
DISTRIBUTION

by

RICHARD NTIAMOAH KYEI, B.Sc.

Presented to the Faculty of the Graduate School of

Stephen F. Austin State University

In Partial Fulfillment

of the Requirements

For the Degree of

Master of Science

STEPHEN F. AUSTIN STATE UNIVERSITY

August 2024

A UNIFORMLY MOST POWERFUL TEST FOR THE MEAN OF A BETA
DISTRIBUTION

by

RICHARD NTIAMOAH KYEI, B.Sc.

APPROVED:

Jacob Turner, Ph.D., Thesis Director

Robert Henderson, Ph.D., Committee Member

Jonathan Mitchell, Ph.D., Committee Member

Jeremy Becnel, Ph.D., Committee Member

Forest Lane Ph.D.
Dean of Research and Graduate Studies

ABSTRACT

The beta distribution is used in numerous real-world applications, including areas such as manufacturing (quality control) and analyzing patient outcomes in health care. It also plays a key role in statistical theory, including multivariate analysis of variance (MANOVA) and Bayesian statistics. It is a flexible distribution that can account for many different characteristics of real data. To our surprise, there has been very little work or discussion on performing statistical hypothesis testing for the mean when it is reasonable to assume that the population is beta distributed. Many analysts conduct traditional analyses using a t-test or nonparametric approach, try transformations, or use standard maximum likelihood-based approaches. We showed via simulations that these tools cannot appropriately control type-I error rates for various situations.

Additionally, this research has set out to construct a uniformly most powerful test using saddle point approximations. These approximations tend to have better accuracy than traditional likelihood-based methods, even when sample sizes are quite low. We provide the necessary methodology development to perform the test. Further simulation studies on power of test are conducted to compare our new method to traditional approaches and illustrate the superiority of our test in many situations. We also provided recommendations on the best way to use this new approach.

ACKNOWLEDGEMENTS

I would like to express my profound gratitude to the following individuals who have supported me on this journey:

Jacob Turner, Ph.D., my thesis supervisor, for his time, effort and commitment in helping me complete this research, and his constructive criticism and patience when reviewing my codes and writing. With this experience with my thesis supervisor, I can confidently say that I know statistics. I would also like to thank Mrs. Turner for allowing my supervisor all the time he needed to work with me during off-school hours.

My committee members, Robert Henderson, Ph.D., Jonathan Mitchell, Ph.D., and Jeremy Becnel, Ph.D., for their guidance and suggestions.

The members of the faculty of the Department of Mathematics and Statistics at Stephen F. Austin State University for their enormous support.

The Balusek Family and Department of Mathematics and Statistics at Stephen F. Austin State University for their financial support.

My classmates, this journey may not be accomplished without them. I appreciate their presence, great minds and all the discussions and arguments. I am glad to have had this experience going through graduate school with you all.

My family especially Florence Akosua Addae, Mrs. Dona Afriyie, Princess Kyei and Skeels Kwadwo Kyei for all the love, prayers and support. My friends especially James & Seth, Williams, Katina, Albert, Henry, Nana Sefa and Charles. The Nacogdoches community for their unwavering support. This is not possible without your contribution.

Lastly, I dedicate this work to my late parents, Mr. and Mrs. Kyei. I love you all.

CONTENTS

ABSTRACT	iii
ACKNOWLEDGEMENTS	iv
1 INTRODUCTION	1
1.1 Mean and Precision Parameterization	2
2 STATISTICAL THEORY	5
2.1 Exponential Family and Canonical Parameters	5
2.2 Sufficient Statistics	6
2.3 Adjustments to Notation	8
2.4 UMP Tests for Exponential Families	8
2.5 Maximum Likelihood Estimators for Exponential Families	9
2.6 Fisher's Information	11
2.7 Likelihood Ratio Tests	12
3 SADDLEPOINT APPROXIMATIONS	14
3.1 Saddlepoint Approximation for Probability Density Function	14
3.1.1 Conditional Saddlepoint PDF	16
3.1.2 Maximum Likelihood Estimators and Saddlepoint Approximations	18
3.2 Saddlepoint Approximation for Cumulative Density Functions	20
4 APPROXIMATE UMP TEST USING SADDLEPOINT APPROXIMATION	22
4.1 Summary of Brakefield 2020: Inference on μ with ϕ Known	22

4.2	Inference on μ with ϕ Unknown	25
4.2.1	The Support of $T_1 T_2$	29
4.2.1.1	Support of $T_1 T_2$ Proof	33
4.2.2	Examining the Saddlepoint Approximations of $T_1 T_2$	34
4.2.3	Decision Rule of the UMP Size α Test	39
5	SIMULATION STUDIES	40
5.1	Type I Error Rate Simulations	41
5.1.1	Right Tailed Tests	42
5.1.2	Left Tailed Tests	46
5.1.3	Type I Error Simulation Summary	49
5.2	Power Simulations	49
5.2.1	Power of Right Tailed Tests	51
5.2.2	Power of Left-Tailed Tests	55
5.3	Summary	58
6	CONCLUSION AND FUTURE RESEARCH	60
	BIBLIOGRAPHY	61
	VITA	63

LIST OF FIGURES

1.1	PDF for the Beta Distribution (α, β Parameterization)	2
1.2	PDF for the Beta Distribution (μ, ϕ Parameterization)	3
4.1	Approximating Distribution of Sufficient Statistic T of μ	24
4.2	Support and Distribution of $T_1 T_2$	30
4.3	Jensen's Inequality Defining the Boundary of the Joint Distribution between the Sufficient statistics	32
4.4	Demonstration of Jensen's Inequality in Different Scenarios	33
4.5	A Beta Population and Joint Distribution of $T_1 T_2$ $B(\mu = 0.1, \phi =$ $10), n = 10$	35
4.6	The Distribution of $T_1 T_2$ based on the Beta Population $B(\mu = 0.1, \phi =$ $10), n = 10$ for Statistical Inference of μ	36
4.7	Beta Population Histogram, Support of T_1, T_2 and Distribution of UMP Test under the Condition $B(\mu = 0.1, \phi = 5), n = 10$	37
4.8	Beta Population Histogram, Support of T_1, T_2 and Distribution of UMP Test under the Condition $B(\mu = 0.1, \phi = 20), n = 10$	38
4.9	Beta Population Histogram, Support of T_1, T_2 and Distribution of UMP Test under the Condition $B(\mu = 0.5, \phi = 10), n = 10$	39
5.1	Estimated Type I Error Rates for Right-Tailed Tests ($\phi = 5$)	42
5.2	Estimated Type I Error Rates for Right-Tailed Tests ($\phi = 10$)	43
5.3	Sampling Distribution of -2log-LRT-Statistic $B(\mu = 0.5, \phi = 10) n = 5$	45
5.4	Sampling Distribution of T-Statistic $B(\mu = 0.1, \phi = 10) n = 10$	46
5.5	Estimated Type I Error Rates for Left-Tailed Tests ($\phi = 5$)	47
5.6	Estimated Type I Error Rates for Left-Tailed Tests ($\phi = 10$)	48

5.7	Simulated Power Curves for Right-Tailed Tests ($\mu_0 = 0.1, \phi = 5, n = 5, 10, 25$)	52
5.8	Size α Tests for Right-Tailed Tests ($\mu_0 = 0.1, \phi = 5, n = 5, 10, 25$) . .	53
5.9	Simulated Power Curves for Right-Tailed Tests ($\mu_0 = 0.1, \phi = 10, n = 5, 10, 25$)	54
5.10	Size α Tests for Right-Tailed Tests ($\mu_0 = 0.1, \phi = 10, n = 5, 10, 25$) . .	55
5.11	Simulated Power Curves for Right-Tailed Tests ($\mu_0 = 0.1, \phi = 5, n = 5, 10, 25$)	56
5.12	Simulated Power Curves for Right-Tailed Tests ($\mu_o = 0.1, \phi = 10, n = 5, 10, 25$)	57
5.13	Size α Tests for Right-Tailed Tests ($\mu_o = 0.1, \phi = 10, n = 5, 10, 25$) . .	58

1 INTRODUCTION

The beta distribution is a continuous probability distribution defined on an open interval $(0, 1)$. It is a versatile distribution commonly used in Bayesian statistics to model the uncertainty of random variables that represent proportions or probabilities. The applications of the distribution are vast including its use as a prior distribution in Bayesian statistics, Multivariate Analysis of Variance, Biostatistics, and quality control [4, 3, 8]. For a large number of examples and applications see [7].

The probability density function (PDF) of the beta distribution is given by:

$$f(x; \alpha, \beta) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha, \beta)}, \quad 0 < x < 1$$

where $B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha + \beta)}$ is the beta function, a normalization factor to ensure that the probability density function integrates to 1. Note that, $\Gamma(t)$ is the gamma function evaluated at $t > 0$. The beta distribution is particularly flexible at modeling different curves within the interval, including symmetrical, left and right-skewed, U and inverted U shapes, and straight lines.

For a random variable X that follows the beta distribution, the probability density function (PDF) is

$$f(x|\alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}, \quad 0 < x < 1. \quad (1.1)$$

The distribution has shape parameters $\alpha > 0$ and $\beta > 0$. The mean of X is $E(X) = \mu = \frac{\alpha}{\alpha + \beta}$, $0 < \mu < 1$. The variance of X is $\text{Var}(X) = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}$. Under different values of the parameters α, β , the beta distribution exhibits different forms. Figure 1.1 displays the many shapes of beta distribution under different α and β which highlight its flexibility.

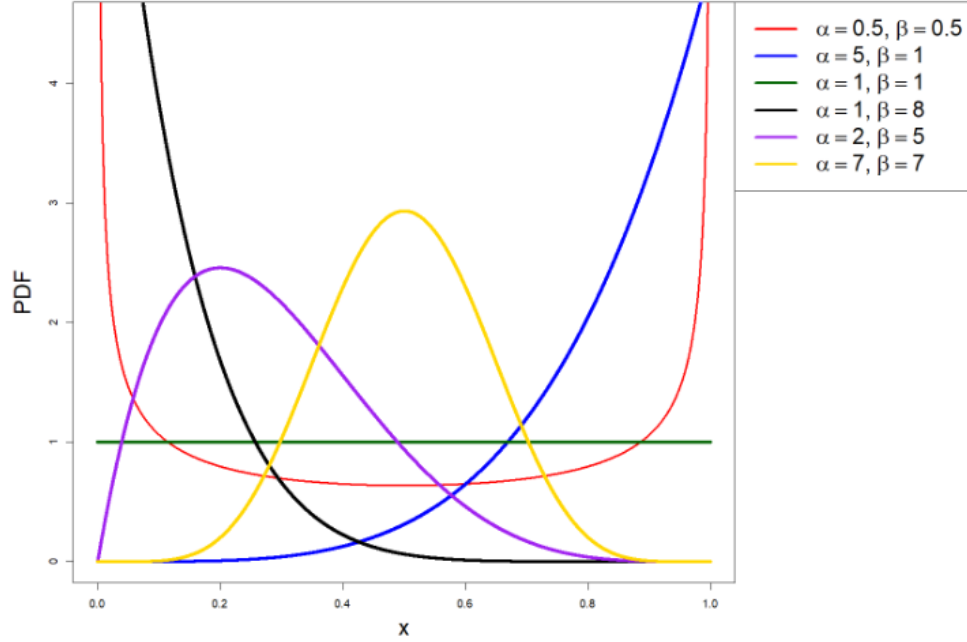


Figure 1.1: PDF for the Beta Distribution (α, β Parameterization)

The PDF can be strictly increasing ($\alpha > 1, \beta = 1$), strictly decreasing ($\alpha = 1, \beta > 1$), U-shaped ($\alpha < 1, \beta < 1$), or unimodal ($\alpha > 1, \beta > 1$). The case $\alpha = \beta$ yields a PDF symmetric about $\frac{1}{2}$ with mean $\frac{1}{2}$ and variance $\frac{1}{4(2\alpha+1)}$. The PDF becomes more concentrated as α increases, but stays symmetric. If $\alpha = \beta = 1$, the beta distribution reduces to the uniform(0,1) distribution, showing that the uniform can be considered to be a member of the beta family.

1.1 Mean and Precision Parameterization

Since the main focus of this research is to provide statistical inference for the mean of the beta distribution, it is helpful to consider the “mean/precision” parameterization of the distribution. The mean is defined as $\mu = \frac{\alpha}{\alpha+\beta}$ and the precision $\phi = \alpha + \beta$. Therefore, $\alpha = \mu\phi$ and $\beta = \phi(1 - \mu)$. As we re-parameterize for $X \sim B(\mu, \phi)$, the

probability density function is

$$f(x|\mu, \phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma(\phi(1-\mu))} x^{\mu\phi-1} (1-x)^{\phi(1-\mu)-1}, \quad 0 < x < 1 \quad (1.2)$$

where $E(X) = \mu$, $0 < \mu < 1$ and $\phi > 0$ is the precision which is in control of the variability of the distribution. However, the variance of X is $\text{Var}(X) = \frac{\mu(1-\mu)}{\phi+1}$. There is an obvious relationship between the variance of X and the parameters μ and ϕ . We can see that the variance becomes smaller as precision ϕ increases while μ is held constant. As we vary among the new parameters μ and ϕ , the shape of the beta distribution changes. To demonstrate this behaviour, Figure 1.2 provides the theoretical PDFs for particular combinations of μ and ϕ .

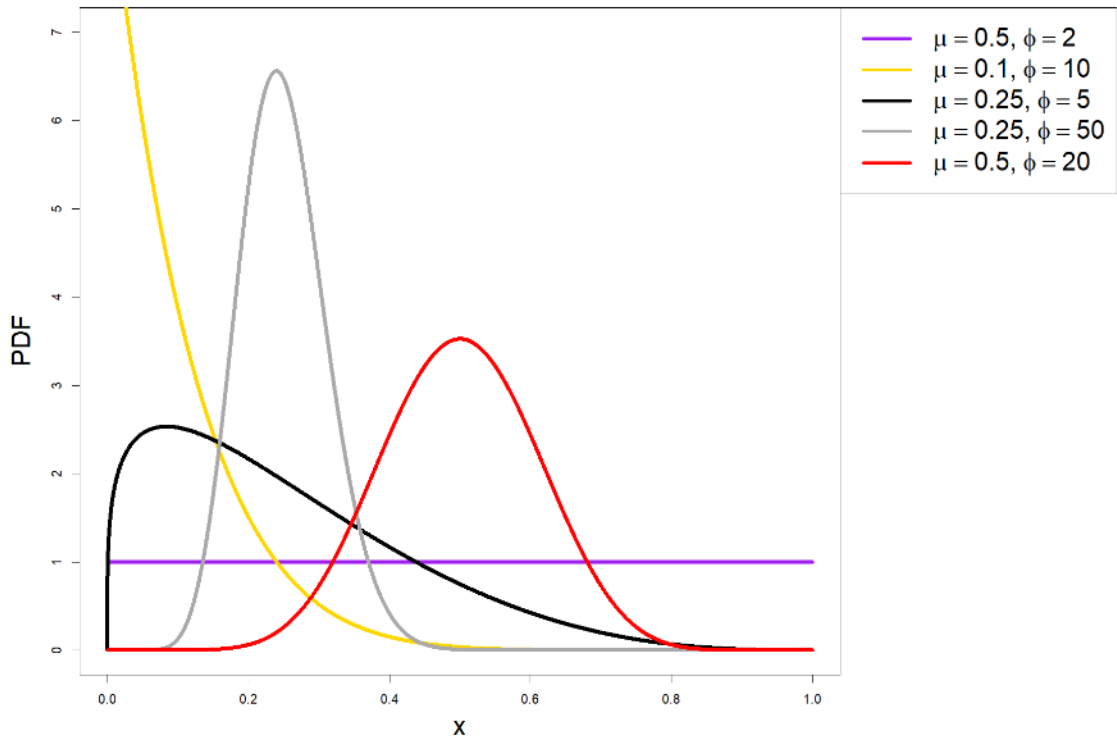


Figure 1.2: PDF for the Beta Distribution (μ, ϕ Parameterization)

We can see from Fig 1.2 (purple line) that the PDF can become the standard uniform continuous distribution ($\mu = 0.5, \phi = 2$). The PDF can become symmetric

(red curve) ($\mu = 0.5, \phi = 20$) or right skewed ($\mu = 0.25, \phi = 5$). We can verify from Figure 1.2 (black and grey curves) that as we hold μ fixed, the variability/shape of the distribution narrows as precision ϕ gets larger.

This thesis is focused on providing statistical inference for the mean of the beta distribution by extending the original work in [2], which used a highly accurate saddlepoint approximation for the uniformly most powerful (UMP) test for the mean of the beta distribution μ assuming the precision ϕ is known. The outline of the remainder of the thesis is as follows. Chapter 2 summarizes key statistical results derived from exponential families of which the beta distribution belongs. Chapter 3 summarizes key chapters within [3] involving saddlepoint approximation theory and additional definitions and notations. Chapter 4 summarizes the previous thesis work in [2]. We then provide the main contributions for our work, which constructs a UMP test for the mean of the beta distribution when the precision parameter is unknown. Chapter 5 provides summaries of type I error and statistical power simulations investigating the performance of the approximation in comparison to standard likelihood ratio test (LRT) and the t-test. Chapter 6 provides a general summary of the thesis and discusses future avenues of research.

2 STATISTICAL THEORY

2.1 Exponential Family and Canonical Parameters

A family of probability density or mass function(s) is called an exponential family if it can be expressed as

$$f(x|\theta) = h(x)e^{\sum_{j=1}^k w_j(\theta)t_j(x) - c(\theta)} \quad (2.1)$$

where $h(x) \geq 0$, $c(\theta)$, $w_j(\theta)$, and $t_j(x)$ are real-valued functions that depend on either x or θ exclusively. In general, x is a random vector and θ is a k -dimensional vector of parameters. When $k = 1$, we have a one parameter exponential family. It is often notationally helpful to let $\eta_j = w_j(\theta)$ and $\eta = (\eta_1, \dots, \eta_k)$. This is commonly called the natural or canonical parameterization of the distribution. Substituting η into Equation (2.1) we have

$$f(x|\eta) = h(x)e^{\sum_{j=1}^k \eta_j t_j(x) - A(\eta)} \quad (2.2)$$

where $A(\eta) = c(w^{-1}(\eta))$ for $\theta = w^{-1}(\eta)$. While not obvious in this report, canonical parameterizations provide a notationally “clean” process in which to provide saddle-point approximations for various statistics and is utilized in this thesis [3]. We now provide several distributional examples which are members of the exponential family.

Example 2.1. The PDF of the exponential distribution, given as

$$f(x|\lambda) = \lambda e^{-\lambda x}, \quad x > 0, \quad \lambda > 0 \quad (2.3)$$

is of exponential family form with $h(x) = I_{\{x>0\}}$ where I_E is an indicator function such that

$$I_E = \begin{cases} 1 & \text{if } x \in E \\ 0 & \text{if } x \notin E \end{cases}$$

such that E is the domain of the random variable X , $t(x) = x$, and $c(\lambda) = -\ln \lambda$, $w(\lambda) = -\lambda$. Therefore, $f(x|\lambda)$ can be expressed as

$$f(x|\lambda) = I_{\{x>0\}} e^{w(\lambda)x - c(\lambda)}, \quad x > 0 \quad (2.4)$$

and is thus a member of the one parameter exponential family. Upon reparameterization, the canonical parameter is $\eta = w(\lambda) = -\lambda$, and $A(\eta) = -\ln(-\eta)$.

Example 2.2. The PDF of the beta distribution with parameters α, β is also of exponential family form, given as

$$f(x|\alpha, \beta) = h(x) e^{\alpha \ln x + \beta \ln(1-x) - c(\alpha, \beta)}, \quad 0 < x < 1 \quad (2.5)$$

where $h(x) = \frac{1}{x(1-x)}$, $0 < x < 1$, $c(\alpha, \beta) = -\ln \left(\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \right)$, $t_1(x) = \ln x$, $t_2(x) = \ln(1-x)$, $w_1(\alpha, \beta) = \alpha$, $w_2(\alpha, \beta) = \beta$. In this case, we can see that the beta distribution is a two parameters exponential family. It should be noted that mean/precision parameterization expressed in Equation (1.2) is not of canonical form.

2.2 Sufficient Statistics

In this section, we defined and briefly discuss the role of sufficient statistics within exponential families, their joint distributions, and their role in estimation and inference. We acknowledge that for many of the results described in this section, various regularity conditions must be met in order for the results to hold. When working with exponential families, as is the case with the beta distribution, these regularity conditions are met and thus are not a concern for this thesis. For an extensive review on the necessary regularity conditions see [4].

Suppose X_1, X_2, \dots, X_n are independent and identically distributed with probability density $f(x|\theta)$ and represent a future random sample taken from a population. A statistic $T = t(x_1, x_2, \dots, x_n)$ is a function of observed values of the sample. Statistics condense the original sample space down to typically a few values. This data reduction allows one to summarize the key “information” that the sample contains. The

principle idea of a sufficient statistic is that the data reduction made is computed in such a way that no information about the parameter θ is lost. Formally, the statistic T is a sufficient statistic if and only if the conditional distribution of X_1, X_2, \dots, X_n given T is free of the population's parameter.

When working with exponential families under a canonical parameterization, the sufficient statistics for each parameter are readily obtained by writing the joint distribution of the sample of observed values in exponential family form,

$$f(x_1, x_2, \dots, x_n | \eta) = \prod_{i=1}^n h(x_i) e^{\sum_{j=1}^k \eta_j \sum_{i=1}^n t_j(x_i) - A(\eta)}. \quad (2.6)$$

It is easily shown that the statistics $T_j = \sum_{i=1}^n t_j(x_i)$ within Equation (2.6) are the sufficient statistics for the respective η_j 's [4]. Additionally, the joint distribution of the sufficient statistics can be expressed in a general form, but in many settings, obtaining a closed form expression is difficult. However, the cumulant generating function (CGF) for sufficient statistics is readily available and will be discussed next.

Suppose continuous random vector $X = (X_1, X_2, \dots, X_n)$ has joint density f with domain $D \subset \mathbb{R}^n$. The moment generating function (MGF) associated with the joint density f is defined as

$$M(s) = E(e^{\sum_{i=1}^n s_i X_i}) = \int \dots \int_D e^{\sum_{i=1}^n s_i X_i} f(x_1, \dots, x_n) dx_1 \dots dx_n \quad (2.7)$$

where the vector $s = (s_1, s_2, \dots, s_n)$ is an element in \mathbb{R}^n such that the integral converges. Note that the convergence is always assured at the origin due to f being a proper probability density. The CGF of f is defined as

$$K(s) = \ln M(s) \quad (2.8)$$

The CGF for sufficient statistics under canonical exponential families, denoted $K_T(s)$, is discussed in [3]. It turns out, $K_T(s)$ can be expressed completely in terms of the function $A(\eta)$ obtained by examining Equation (2.6) and is

$$K_T(s) = nA(\eta + s) - nA(\eta) \quad (2.9)$$

The utility of this simple and compact result should not be taken lightly. Equation (2.7) highlights the fact that the MGF is closely related to the Laplace transform of f . In future chapters, we will utilize saddlepoint methods, which provide a mechanism to approximate an inverse Laplace transform. In doing so, the distribution of sufficient statistics can be approximated with relatively high accuracy using the CGF alone which is readily available when working with exponential families.

2.3 Adjustments to Notation

To effectively disseminate key results throughout the remainder of this thesis, it is helpful to make a few notation adjustments to remain consistent with [3]. While the results provided are more generalizable to a broader class of distributions, moving forward, we are strictly working with exponential families with either one or two parameters only. Second, the parameters within the exponential family are assumed to be canonical and is denoted as ψ (first parameter) and χ (second parameter). Occasionally, we make note when the results provided hold true even when the parameters need not be canonical.

To illustrate these changes, consider the following example. For the two parameter exponential family, the canonical parameter vector $\eta = (\eta_1, \eta_2)$ is now denoted as $\eta = (\psi, \chi)$. Using Equation (2.9), the joint CGF of the sufficient statistics, $T = (T_1, T_2)$, for canonical parameters $\eta = (\psi, \chi)$ is expressed as

$$K_T(s_1, s_2) = nA(\psi + s_1, \chi + s_2) - nA(\psi, \chi) \quad (2.10)$$

2.4 UMP Tests for Exponential Families

A uniformly most powerful (UMP) test is a hypothesis test which controls the type I error rate at a pre-specified level, α , and is the most powerful compared to

all tests that have type I error control at the same prescribed level. Based on the definition of UMP, only one sided tests can be UMP. Additional restrictions on the testing procedure must be placed for two sided tests and is not the focus of this thesis. For canonical exponential families, a general theory and framework exists to construct UMP tests. We briefly described the two most pertinent cases for this thesis. The first situation is for exponential families with one canonical parameter. The second situation considers two canonical parameter exponential families in which the hypothesis of interest includes one of the parameters and the remaining parameter is a nuisance parameter. Similar to dealing with an unknown population standard deviation when conducting statistical inference for the mean of a Normal distribution.

The following results can be found in [3, 9]. For a one parameter exponential family, the UMP test for ψ should be based on the sufficient statistic T . For the two parameter case, the parameter of interest is still ψ but χ is an unknown nuisance parameter. Let T_1 and T_2 be the sufficient statistics for ψ and χ respectively. A UMP test involving ψ should be based on the conditional distribution $T_1|T_2$ [9].

2.5 Maximum Likelihood Estimators for Exponential Families

Sufficient statistics are directly linked to parameter estimation via maximum likelihood. For the one parameter case involving ψ , a sample of n identically distributed and independent observations has joint PDF

$$f(x_1, \dots, x_n|\psi) = \prod_{i=1}^n h(x_i) e^{\{\psi \sum_{i=1}^n t(x_i) - nA(\psi)\}}.$$

Upon observing the data, the joint distribution becomes only a function of the parameter ψ and is referred to as the likelihood function, $\mathcal{L}(\psi)$ and is typically provided on the log scale,

$$l(\psi) = \ln \mathcal{L}(\psi) = \sum_{i=1}^n \ln(h(x_i)) + \psi \sum_{i=1}^n t(x_i) - nA(\psi). \quad (2.11)$$

Notationally, the log-likelihood is also expressed as $l(\psi|\mathbf{x})$ where $\mathbf{x} = (x_1, x_2, \dots, x_n)$ to emphasize the fact that it is a function of ψ conditional on the fact that the sample \mathbf{x} was observed. This notation is also used to define the random variable $l(\psi|\mathbf{X})$ where \mathbf{X} represents the random variables associated with the unobserved sample. One could also create random variables by first taking derivatives of l with respect to ψ and then replacing \mathbf{x} with \mathbf{X} (ie. $\frac{\partial}{\partial \psi} l(\psi|\mathbf{X})$). These types of random variables play a key role in multiple computations covered in the remaining sections of this chapter.

The maximum likelihood estimator (MLE) $\hat{\psi}$ is the value that maximizes Equation (2.11) obtained by solving the equation $\frac{d}{d\psi} l(\psi) = 0$ which is expressed as

$$nA'(\hat{\psi}) = \sum_{i=1}^n t(x_i). \quad (2.12)$$

The resulting differential equation presented in Equation (2.12) highlights the fact that the MLE $\hat{\psi}$ must be a function of the sufficient statistic $T = \sum_{i=1}^n t(x_i)$.

For the two parameter case, $\eta = (\psi, \chi)$, using Equation (2.2), the log-likelihood is

$$l(\psi, \chi) = \ln \mathcal{L}(\psi, \chi) = \sum_{i=1}^n \ln(h(x_i)) + \psi T_1 + \chi T_2 - nA(\psi, \chi) \quad (2.13)$$

where $T_1 = \sum_{i=1}^n t_1(x_i)$ and $T_2 = \sum_{i=1}^n t_2(x_i)$. Maximizing the log-likelihood resorts to solving a system of two equations with two unknowns by differentiating with respect to ψ and χ ,

$$T_1 - nA_\psi(\hat{\psi}, \hat{\chi}) = 0 \quad (2.14)$$

$$T_2 - nA_\chi(\hat{\psi}, \hat{\chi}) = 0 \quad (2.15)$$

where $A_\psi(\psi, \chi) = \frac{\partial}{\partial \psi} A(\psi, \chi)$ and $A_\chi(\psi, \chi) = \frac{\partial}{\partial \chi} A(\psi, \chi)$. The solution to this system of equations corresponds to the MLEs $\hat{\psi}$ and $\hat{\chi}$. It is important to note that the solution to the linear systems depicted in both the one and two parameter cases may yield closed-form solutions, while others must be solved numerically. This depends on the functional form of both $A(\psi)$ and $A(\psi, \chi)$ and therefore depends on

the specific distribution that is a member of the exponential family such as Normal, Gamma, Beta, etc.

2.6 Fisher's Information

When constructing confidence intervals using maximum likelihood estimators, Wald intervals can be constructed by using the asymptotic properties of MLEs and require the computation of Fisher's Information. While confidence intervals are not the focus of this thesis, a definition of Fisher's information is warranted for the developments in Chapter 3. The general definition for Fisher's information for a one parameter case, which does not require a canonical parameterization, is expressed as

$$j(\psi) = E \left[\left(\frac{\partial}{\partial \psi} l(\psi) \right)^2 \right]. \quad (2.16)$$

Recall previously that the log-likelihood, $l(\psi)$, is also expressed as $l(\psi|\mathbf{x})$ and as a random variable, $l(\psi|\mathbf{X})$. The expectation defined within Fishers information is with respect to the random variable $\frac{\partial}{\partial \psi} l(\psi|\mathbf{X})$. If the likelihood satisfies additional regularity conditions as stated in [4], Fisher's information can be expressed in an easier computational form as

$$j(\psi) = -E \left(\frac{\partial^2 l(\psi)}{\partial \psi^2} \right). \quad (2.17)$$

For the two parameter case, Fisher's information is a 2×2 matrix of partial derivatives of the log-likelihood. We provide the definition using the two parameters ψ and χ , but again they need not be canonical

$$j(\psi, \chi) = \begin{bmatrix} j_{\psi\psi} & j_{\psi\chi} \\ j_{\chi\psi} & j_{\chi\chi} \end{bmatrix} = \begin{bmatrix} -E \left(\frac{\partial^2 l(\psi, \chi)}{\partial \psi^2} \right) & -E \left(\frac{\partial^2 l(\psi, \chi)}{\partial \chi \partial \psi} \right) \\ -E \left(\frac{\partial^2 l(\psi, \chi)}{\partial \psi \partial \chi} \right) & -E \left(\frac{\partial^2 l(\psi, \chi)}{\partial \chi^2} \right) \end{bmatrix} \quad (2.18)$$

When working with exponential families under canonical parameters, Fisher's information computations are drastically simplified. After differentiation of either Equation (2.11) or (2.13), the resulting expressions do not have the random variable X in

it any longer. This fact completely removes the need for the expectation computation since the expected value of a constant is just the constant. To see this, consider computing Fisher's information for the one parameter case. Computing $\frac{\partial^2 l(\psi)}{\partial \psi^2}$ using Equation (2.11) yields $nA''(\psi)$. By the definition of exponential families, $A(\psi)$ is a function of ψ alone and is thus a constant with respect to \mathbf{X} . Similarly, the Fisher's information matrix for the two parameter case reduces to the matrix of partial derivatives of $nA(\psi, \chi)$.

2.7 Likelihood Ratio Tests

Likelihood ratio tests (LRTs) provide a simple mechanism for constructing hypothesis test while dealing with or without nuisance parameters. For the one parameter case, suppose we wish to test $H_0 : \psi = \psi_0$ versus $H_1 : \psi \neq \psi_0$. The LRT statistic is defined as

$$\lambda(\mathbf{x}) = \frac{\mathcal{L}(\psi_0)}{\mathcal{L}(\hat{\psi})} \quad (2.19)$$

where $\hat{\psi}$ is the MLE for ψ . Assuming the null hypothesis is true, $\lambda(\mathbf{x})$ should be close to one as the estimate from the data $\hat{\psi}$ should be close to ψ_0 . If the alternative is true, $\lambda(\mathbf{x})$ should be lower than one relative to the degree of the true alternative value for ψ . This leads one to consider rejecting H_0 by selecting a value c such that $P(\lambda(\mathbf{x}) < c | H_0) = \alpha$.

A similar approach is taken for the two parameter case. When testing $H_0 : \psi = \psi_0$ versus $H_1 : \psi \neq \psi_0$, with nuisance parameter χ , the LRT statistics is defined as

$$\lambda(\mathbf{x}) = \frac{\mathcal{L}(\psi_0, \hat{\chi}_{\psi_0})}{\mathcal{L}(\hat{\psi}, \hat{\chi})} \quad (2.20)$$

where $\hat{\psi}, \hat{\chi}$ are the “unrestricted” MLEs of ψ and χ as discussed previously. The additional quantity, $\hat{\chi}_{\psi_0}$, is the “restricted” MLE of χ while holding ψ fixed at ψ_0 and is obtained by solving Equation (2.15) where ψ_0 is plugged in for ψ .

The challenge with performing the LRT test typically comes from determining the value of c to ensure the type-I error rate is correctly specified as α . This requires us to know the distribution of $\lambda(\mathbf{x})$ directly or transform the quantity to a known distribution. Success in doing such a task largely depends on the distribution f assumed that the data was sampled from. In practice and for the two cases considered here, it is common to use an asymptotic result which states that the test statistic $-2 \ln \lambda(\mathbf{x})$ follows a Chi-square distribution with 1 degree of freedom [4].

3 SADDLEPOINT APPROXIMATIONS

3.1 Saddlepoint Approximation for Probability Density Function

The saddlepoint approximation method was initially proposed by [5]. It is a formula that approximates a probability density or mass function when the moment generating function (MGF) or cumulant generating function (CGF) of the random variable is known [3]. It can be applied in both univariate and multivariate cases as well as approximating conditional distributions. The approximation is popular due to its general high accuracy. Traditional approximation approaches used in techniques such as the LRT described in Chapter 2 rely on approximations of order $O(n^{-\frac{1}{2}})$ while saddlepoint approximations generally are order $O(n^{-1})$ but can have up to order $O(n^{-\frac{3}{2}})$ in some cases [3, 6].

The saddlepoint approximation can be used for both discrete and continuous cases. Without loss of generality, suppose X is a continuous random variable with CGF, $K_X(s)$, defined on $s \in (a, b)$. The saddlepoint density approximation to the density function of X , denoted $\hat{f}_X(x)$, is given as

$$\hat{f}_X(x) = \frac{1}{\sqrt{2\pi K_X''(\hat{s})}} e^{K_X(\hat{s}) - \hat{s}x}. \quad (3.1)$$

The symbol $\hat{s} = \hat{s}(x)$ denotes the unique solution to the equation

$$K_X'(\hat{s}) = x, \quad \hat{s} \in (a, b) \quad (3.2)$$

and it is an implicitly defined function of x . The previous equation is referred to as the saddlepoint equation and \hat{s} the saddlepoint associated with the value x . The symbol $K_X''(\hat{s})$ is the second derivative of the CGF evaluated at \hat{s} .

Since, \hat{f}_X is an approximation of the density function, the approximation does not typically integrate to one. However one can “force a density” by integration and then

scaling. Letting $c = \int_{-\infty}^{\infty} \hat{f}_X(x)dx$, a normalized saddlepoint approximation is defined as

$$\bar{f}_X(x) = \frac{\hat{f}_X}{c} \quad (3.3)$$

and is a proper density function since $\int_{-\infty}^{\infty} \bar{f}_X(x)dx = 1$. In practice, the process of finding the normalizing constant c is typically accomplished using numerical integration of $\hat{f}_X(x)$ over the support of X .

To illustrate the utility of the saddlepoint approximation, we considered two examples.

Example 3.1.1 Consider the normal distribution whose CGF is

$$K_X(s) = \mu s + \frac{\sigma^2 s^2}{2}. \quad (3.4)$$

Taking the first two derivatives yields

$$K'_X(s) = \mu + \sigma^2 s$$

$$K''_X(s) = \sigma^2.$$

The solution to the saddlepoint equation, $K'_X(\hat{s}) = \mu + \sigma^2 \hat{s} = x$, can be solved analytically and is

$$\hat{s}(x) = \hat{s} = \frac{x - \mu}{\sigma^2} \quad (3.5)$$

Substituting $K_X(\hat{s})$ and $K''_X(\hat{s})$ in Equation (3.1), the saddlepoint approximation for the normal density is expressed as

$$\hat{f}_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\mu + \sigma^2 \hat{s}^2 - (\frac{x-\mu}{\sigma^2})x}$$

and reduces to

$$\hat{f}_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}, \quad x \in \mathbb{R} \quad (3.6)$$

In this particular case, the saddlepoint approximation is exact. This is not typically the case as demonstrated with the next example.

Example 3.1.2 The saddlepoint approximation could be used to estimate the PDF of discrete random variables. To demonstrate this, consider the Poisson distribution $X \sim \text{Poisson}(\lambda)$ whose CGF is

$$K_X(s) = \lambda(e^s - 1). \quad (3.7)$$

Its first and second derivatives are

$$K'_X(s) = K''_X(s) = \lambda e^s.$$

The saddlepoint equation, $K'_X(\hat{s}) = \lambda e^{\hat{s}} = x$, has the solution

$$\hat{s}(x) = \hat{s} = \ln\left(\frac{x}{\lambda}\right). \quad (3.8)$$

The saddlepoint approximation for the Poisson PDF is expressed as

$$\hat{f}_X(x) = \frac{1}{\sqrt{2\pi x}} e^{[x - \lambda - x \ln \frac{x}{\lambda}]}$$

when substituting $K_X(\hat{s})$ and $K''_X(\hat{s})$ in Equation (3.1). The PDF approximation can be simplified further to

$$\hat{f}_X(x) = \lambda^x e^{-\lambda} \left[\frac{1}{e^{-x} (\sqrt{2\pi}) (x^{x+0.5})} \right].$$

Note that, $e^{-x} (\sqrt{2\pi}) (x^{x+0.5})$ is Sterling's approximation for $x!$ and thus the saddlepoint approximation is accurate up to the degree of accuracy of Sterling's approximation.

3.1.1 Conditional Saddlepoint PDF

The saddlepoint approximation method can also be used to approximate the probability density function of a conditional random variable. Suppose we have two continuous random variables X, Y . The conditional PDF of $Y|X$ is

$$f_{Y|X}(y|x) = \frac{f_{X,Y}(x, y)}{f_X(x)},$$

where $f_{X,Y}(x, y)$ is the joint PDF and $f_X(x)$ is the marginal PDF. The main idea here is to provide saddlepoint approximations for the joint distribution, $\hat{f}_{X,Y}(x, y)$, and an additional saddlepoint approximation for the marginal distribution, $\hat{f}_X(x)$. After obtaining the approximations, the saddlepoint density of the conditional random variable $Y|X$ is obtained by taking the ratio,

$$\hat{f}_{Y|X}(y|x) = \frac{\hat{f}_{X,Y}(x, y)}{\hat{f}_X(x)}. \quad (3.9)$$

The two approximations in Equation (3.9) can be accomplished if the joint CGF $K_{X,Y}(s_1, s_2)$ is known. Details of the derivation can be found in [3]. We briefly summarized the results here. Using $K_{X,Y}(s_1, s_2)$ directly, the saddlepoint approximation of the joint PDF is

$$\hat{f}_{X,Y}(x, y) = (2\pi)^{-1} |K''_{X,Y}(\hat{s}_1, \hat{s}_2)|^{-1/2} e^{\{K_{X,Y}(\hat{s}_1, \hat{s}_2) - \hat{s}_1 x - \hat{s}_2 y\}}. \quad (3.10)$$

Here the 2-dimensional saddlepoint (\hat{s}_1, \hat{s}_2) solves the set of two equations from the CGF $K_{X,Y}(s_1, s_2)$ obtained from Equation (2.10)

$$nA_{s_1}(\psi + \hat{s}_1, \chi + \hat{s}_2) = X \quad (3.11)$$

$$nA_{s_2}(\psi + \hat{s}_1, \chi + \hat{s}_2) = Y \quad (3.12)$$

and the second derivative of the CGF yields

$$K''_{X,Y}(s_1, s_2) = \begin{bmatrix} nA_{s_1 s_1}(\psi + s_1, \chi + s_2) & nA_{s_2 s_1}(\psi + s_1, \chi + s_2) \\ nA_{s_1 s_2}(\psi + s_1, \chi + s_2) & nA_{s_2 s_2}(\psi + s_1, \chi + s_2) \end{bmatrix} \quad (3.13)$$

Note that $|\cdot|$ is the absolute value of the determinant if \cdot is a matrix. However, if \cdot is just a scalar, then $|\cdot|$ is the magnitude of the value \cdot or simply the absolute value of \cdot . The marginal CGF for X is obtained from the joint CGF, $K_X(s_0) = K_{X,Y}(s_1 = s_0, s_2 = 0)$ and the saddlepoint approximation provided in Equation (3.1) applies and can be expressed as

$$\hat{f}_X(x) = (2\pi)^{-1/2} |K''_X(\hat{s}_0, 0)|^{-1/2} e^{\{K_X(\hat{s}_0, 0) - \hat{s}_0 x\}} \quad (3.14)$$

where \hat{s}_0 is the saddlepoint that solves the single equation from $K_X(s_0, 0)$

$$nA_{s_0}(\psi + \hat{s}_0) = X$$

and $K_X''(\hat{s}_0, 0) = nA_{s_0 s_0}(\psi + s_0)$ is the second derivative with respect to s_0 of the marginal CGF. Finally, the conditional saddlepoint PDF is obtained by the ratio of the joint and marginal approximations, $\frac{\hat{f}_{X,Y}(x,y)}{\hat{f}_X(x)}$. Upon algebraic manipulations, we have

$$\hat{f}_{Y|X}(y|x) = (2\pi)^{-0.5} \left\{ \frac{|K_{X,Y}''(\hat{s}_1, \hat{s}_2)|}{|K_X''(\hat{s}_0, 0)|} \right\}^{-0.5} e^{[\{K_{X,Y}(\hat{s}_1, \hat{s}_2) - \hat{s}_1 x - \hat{s}_2 y\} - \{K_X(\hat{s}_0, 0) - \hat{s}_0 x\}]}. \quad (3.15)$$

3.1.2 Maximum Likelihood Estimators and Saddlepoint Approximations

When working with exponential families, there is an interesting relationship between MLEs and saddlepoint approximations which enables the saddlepoint density approximations in the previous sections to be expressed completely in terms of MLEs, Fisher's Information, and Likelihood functions. A comprehensive discussion can be found in [3]. For our purposes, we provided the details for the univariate case to gain insight. We simply provide the key result for the conditional distribution case.

For the one parameter exponential family case, the CGF for the sufficient statistic T is

$$K_T(s) = nA(\psi + s) - nA(\psi). \quad (3.16)$$

When using this form to compute a saddlepoint approximation for the PDF of T , f_T , the saddlepoint equation is

$$K_T'(\hat{s}) = nA_s(\psi + \hat{s}) = T. \quad (3.17)$$

Therefore, by using Equation (3.1), the saddlepoint approximation can be rewritten as

$$\hat{f}_T(t) = \frac{1}{\sqrt{2\pi n A_{ss}(\psi + \hat{s})}} e^{nA(\psi + \hat{s}) - nA(\psi) - \hat{s}t}. \quad (3.18)$$

When discussing MLEs for the one parameter case, recall in Chapter 2 that maximizing the log-likelihood with respect to ψ yielded the differential equation presented in Equation (2.12) and is presented below for the reader as:

$$nA_\psi(\hat{\psi}) = T$$

where $T = \sum_{i=1}^n t(x_i)$. Upon comparison of the two differential equations in Equation (2.12) and (3.17), it is clear that the MLE $\hat{\psi}$ must be equal to $\psi + \hat{s}$. This implies that, for a given sufficient statistic value $T = t$, the saddlepoint solution can be expressed completely in terms of its corresponding MLE as $\hat{s} = \hat{\psi} - \psi$. Using this expression, in addition to the likelihood form of Equation (2.11) and the fact that $j(\hat{\psi}) = nA_{ss}(\psi + \hat{s})$, Equation (3.18) can be rewritten as

$$\hat{f}_T(t|\psi) = \frac{1}{\sqrt{2\pi j(\hat{\psi})}} \frac{\mathcal{L}(\psi)}{\mathcal{L}(\hat{\psi})} \quad (3.19)$$

This particular form of the saddlepoint approximation of the PDF of T is solely a function of t , through the MLE $\hat{\psi}$ and only depends on the parameter ψ . This particular form is extremely helpful when implementing the saddlepoint approximation in software such as R software, as packages to compute MLEs for a given sufficient statistic are readily available.

When working with two parameter exponential families with canonical interest parameter ψ and nuisance parameter χ , inference should be conducted on the sufficient statistics for ψ , T_1 , given the sufficient statistic for χ , T_2 . Thus a saddlepoint approximation for the density of $T_1|T_2$ is needed to compute p-values. Like the univariate setting, the approximation can be completely written in terms of likelihood

quantities as

$$\hat{f}_{T_1|T_2}(t_1|t_2; \psi) = (2\pi)^{-\frac{1}{2}} \left\{ \frac{|j(\hat{\chi}, \hat{\psi})|}{|j_{\chi\chi}(\hat{\chi}_\psi, \psi)|} \right\}^{-\frac{1}{2}} \frac{\mathcal{L}(\hat{\chi}_\psi, \psi)}{\mathcal{L}(\hat{\chi}, \hat{\psi})}. \quad (3.20)$$

While not apparently obvious, the conditional approximation is a function of t_1 for a fixed value of t_2 and ψ . The distribution doesn't depend on χ at all, which is a result from conditioning on the sufficient statistic T_2 .

3.2 Saddlepoint Approximation for Cumulative Density Functions

The cumulative density function (CDF) of T and $T_1|T_2$ can be approximated using one of two approaches. The first is via numerical integration, and the second is to use the approximation results from Luganni-Rice and Skovgaard [3]. These approximations can be used to directly compute p-values in which the test statistic involves sufficient statistics.

The saddlepoint approximation for the CDF of the sufficient statistic T in a one parameter exponential family developed by Lugannani and Rice [3] can be expressed as

$$\hat{F}_T(t) = \begin{cases} \Phi(\hat{w}) + \otimes(\hat{w})(\frac{1}{\hat{w}} - \frac{1}{\hat{u}}) & \text{if } t \neq K'_T(0) \\ \frac{1}{2} + \frac{K''_T(0)}{6\sqrt{2\pi K''_T(0)^{3/2}}} & \text{if } t = K'_T(0) \end{cases} \quad (3.21)$$

where

$$\hat{w} = \text{sgn}(\hat{\psi} - \psi) \sqrt{-2 \ln \frac{\mathcal{L}(\psi)}{\mathcal{L}(\hat{\psi})}} \quad (3.22)$$

and

$$\hat{u} = (\hat{\psi} - \psi) \sqrt{j(\hat{\psi})} \quad (3.23)$$

are functions of t . The symbols \otimes and Φ represent the standard normal density and CDF respectively and $\text{sgn}(\hat{\psi} - \psi)$ captures the sign \pm of $\hat{\psi} - \psi$. The bottom expression in (3.21) defines the approximation at the mean of T or when $\hat{s} = \hat{\psi} - \psi = 0$. In this case, $\hat{w} = 0 = \hat{u}$, and the last factor in the top expression of (3.21) is undefined.

Fortunately the discontinuity is removable. As $t \rightarrow K'_T(0)$ in either direction, the limiting value is $\frac{1}{2} + \frac{K_T'''(0)}{6\sqrt{2\pi K_T''(0)^{3/2}}}$. See [3] for additional details on the derivations. The CDF expression is continuous and, more generally, continuously differentiable or smooth. Apart from the theoretical smoothness, any practical computation that uses software is vulnerable to numerical instability when making the computation of Equation (3.21) for t quite close to $K'_T(0)$. When making computations within a neighborhood of $K'_T(0)$, $\hat{F}_T(t)$ can be approximated with several interpolation methods [1].

The saddle point approximation of the CDF of the conditional distribution $T_1|T_2$ for the two parameter exponential family was developed by Skovgaard [3] and can be expressed as

$$\hat{F}_{T_1|T_2}(t_1|t_2) = \Phi(\hat{w}) + \bigotimes(\hat{w})\left(\frac{1}{\hat{w}} - \frac{1}{\hat{u}}\right) \quad (3.24)$$

where

$$\hat{w} = \text{sgn}(\hat{\psi} - \psi) \sqrt{-2 \ln \frac{\mathcal{L}(\psi, \hat{\chi}_\psi)}{\mathcal{L}(\hat{\psi}, \hat{\chi})}} \quad (3.25)$$

and

$$\hat{u} = (\hat{\psi} - \psi) \sqrt{\frac{|j(\hat{\psi}, \hat{\chi})|}{|j_{\chi\chi}(\psi, \hat{\chi}_\psi)|}}. \quad (3.26)$$

The CDF is undefined when $\hat{\psi} = \psi$.

4 APPROXIMATE UMP TEST USING SADDLEPOINT APPROXIMATION

In this chapter, a summary of the results in [2] for conducting inference on the mean of the beta distribution, μ , where the precision parameter ϕ is known is shown. We then extended the testing procedure by developing an additional approach to handle the situation when ϕ is unknown. Note that for this chapter, the interest parameter ψ is now μ for a one parameter case.

4.1 Summary of Brakefield 2020: Inference on μ with ϕ Known

Bryn Brakefield's thesis work [2] provides a detailed explanation of statistical inference for the mean μ of the beta distribution while the precision parameter ϕ is known using traditional likelihood-based methods and implementing a saddlepoint approximation. Under this setting, it turns out the parameter μ is a canonical parameter and the one parameter methods covered in Chapter 3 can be applied. In this section, a general summary of the saddlepoint derivations of [2] are provided along with an illustrative example.

The hypothesis of interest is defined as

$$H_0 : \mu = c_0, \quad H_1 : \mu \neq c_0. \quad (4.1)$$

One-sided tests of (4.1) were also considered. In the work by [2], there were comparison of the standard large sample likelihood-based tests, a standard t-test, and a saddlepoint approximation for the UMP test for the one-sided alternative scenarios. Through simulation work, she provided empirical evidence for various aspects of the hypothesis tests and her corresponding confidence intervals for μ under a wide variety

of values for c_0 as well as small to moderate sample sizes. She offered some examples of direct applications of her methods to real-world problems, including a flow cytometry (FCM) data set for testing and sample size determination.

The mean/precision parameterization of the beta density function in Equation (1.2) while treating ϕ as a known quantity (constant), can be written in exponential family form,

$$f_X(x|\mu) = \frac{1}{x(1-x)^{1-\phi}} e^{\mu\phi \ln(\frac{x}{1-x}) - A(\mu)}, \quad 0 < x < 1. \quad (4.2)$$

where $A(\mu) = -\ln\left(\frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma(\phi(1-\mu))}\right)$. The parameter μ can be considered a canonical parameter with its corresponding sufficient statistic $T = \phi \sum_{i=1}^n \ln(\frac{x_i}{1-x_i})$.

Using Equation (2.9), the CGF of T is

$$K_T(s) = nA(\mu + s) - nA(\mu) = n \ln \left(\frac{\Gamma(\phi\mu + \phi s)\Gamma(\phi - \phi\mu - \phi s)}{\Gamma(\mu\phi)\Gamma(\phi - \phi\mu)} \right) \quad (4.3)$$

respectively.

The first two derivatives of $K_T(s)$ are

$$K'_T(s) = nA'_s(\mu + s) = n\phi(\Psi(\phi(\mu + s)) - \Psi(\phi(1 - \mu - s))) \quad (4.4)$$

$$K''_T(s) = nA''_s(\mu + s) = n\phi^2(\Psi'(\phi(\mu + s)) + \Psi'(\phi(1 - \mu - s))). \quad (4.5)$$

where Ψ and Ψ' are the digamma and trigamma functions. They are the first and second derivatives of $\ln(\Gamma(y))$ respectively, where $\Gamma(y)$ is the gamma function evaluated at $y > 0$. for every positive integer of y . Note that the solution to the saddlepoint equation $K'_T(\hat{s}) = t$ does not have a closed form and must be obtained numerically using a root finder. The solution can also be found by computing the MLE through $nA'(\hat{\mu}) = t$ numerically and solving for the saddlepoint using $\hat{s} = \hat{\mu} - \mu$ as discussed in Chapter 3. The PDF formula for the saddlepoint approximation for T , $\hat{f}_T(t|\mu)$, can be obtained by substituting Equations (4.3) and (4.5) into (3.1).

To demonstrate the accuracy of the approximations, we simulated 10000 observations of $T = \phi \sum_{i=1}^n \ln(\frac{x_i}{1-x_i})$. Each of these observations were generated by taking

simple random samples from a beta distribution, $X_1, X_2, \dots, X_n \sim B(\mu = 0.2, \phi = 5)$, considering sample sizes of $n = 2$ and $n = 1000$. Histograms are presented in Figure 4.1 to visualize the distributional shapes of T . Additionally, we overlaid the unnormalized and normalized saddlepoint probability density curves as well as a normal approximation curve on the T distribution. The normal approximation is derived by using the first two derivatives of the CGF of T , $K'_T(0)$ and $K''_T(0)$, which correspond to the mean and variance of T respectively.

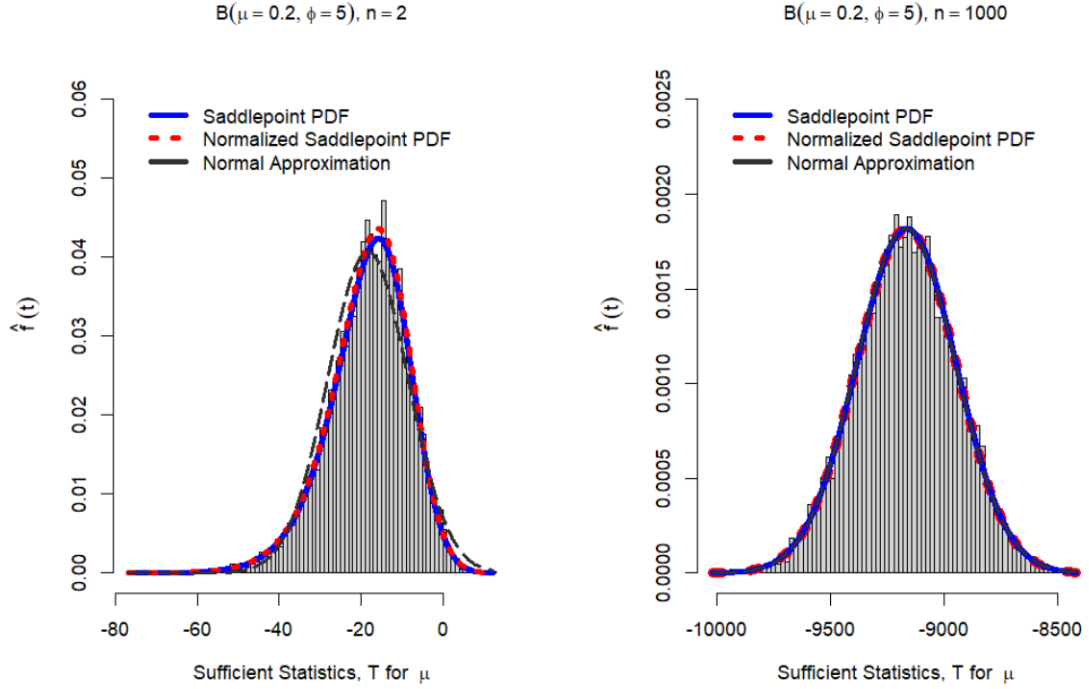


Figure 4.1: Approximating Distribution of Sufficient Statistic T of μ

We can see from Figure 4.1 that, for smaller sample sizes such as $n = 2$, the normal approximation PDF (black curve) fails to approximate the true distribution of T , especially at the tails. However, the saddlepoint approximations (normalized and unnormalized) do a better job capturing the key distributional characteristics of T . As expected, for $n = 1000$, the distribution of T closely follows a Normal

distribution due to the Central Limit Theorem (CLT). There is little to no difference between the saddlepoint and normal approximation in this case.

Figure 4.1 highlights one of the real strengths of saddlepoint approximations when it comes to dealing with smaller sample sizes. In practical settings where data collection is difficult, the concern for using results involving CLT is potential lack of type-I error control. As figure 4.1 depicts, using a normal approximation to compute a tail probability (p-value) when working with smaller sample sizes could lead one to draw an incorrect conclusion at the specified α level.

To test for $H_0 : \mu \leq c_0, H_1 : \mu > c_0$ for a given observed sample x_1, x_2, \dots, x_n , the UMP size α test rejects H_0 for $T > t^*$ (where t^* is a critical value) such that $P(T > t^* | \mu = c_0) = \alpha$. Switching the inequalities yields the left-tailed version. P-values are obtained for an observed data set by first computing the sufficient statistic $t = \phi \sum_{i=1}^n \log(\frac{x_i}{1-x_i})$ and then computing $P(T > t | \mu = c_0)$. This can be done by using the Lugannani-Rice \hat{F}_T provided in Equation (3.21). Without loss of generality, similar approach can be used for testing $H_1 : \mu < c_0$.

4.2 Inference on μ with ϕ Unknown

We now focus our attention to testing for the mean μ of the beta distribution with ϕ unknown. The hypothesis being tested is $H_0 : \mu \leq c_0$ vs $H_1 : \mu > c_0$, where $0 < c_0 < 1$ or its left tailed version. The ultimate goal here is to find canonical parameters ψ and χ such that interest parameter ψ corresponds to μ and the nuisance parameter χ corresponds to ϕ . Upon doing so, we can then determine the sufficient statistics T_1 and T_2 and use the saddlepoint approximations presented in Chapter 3 to approximate the distribution of $T_1|T_2$ and use it to compute p-values.

Equation (4.6) provides the beta distribution with parameters μ and ϕ in expo-

nential family form

$$f_X(x|\mu, \phi) = \frac{1}{x(1-x)} e^{\mu\phi \ln(\frac{x}{1-x}) + \phi \ln(1-x) + \ln\left(\frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma(\phi(1-\mu))}\right)}. \quad (4.6)$$

We can see that μ and ϕ are not canonical parameters under this parameterization of the distribution, but it is close. The first canonical parameter is $\mu\phi$ which is equal to α under the α, β parameterization. The second canonical parameter is ϕ and corresponds to our nuisance parameter. If we were to blindly work under this parameterization, the interest parameter is $\psi = \mu\phi = \alpha$ and the nuisance parameter is $\chi = \phi$. The null hypothesis would be of the form $H_0 : \alpha \leq c_0$ or $H_0 : \alpha \geq c_0$ and thus, not be a direct equivalent test for μ .

In the search of canonical parameters that directly corresponded to μ , consider rewriting Equation (4.6), using canonical parameters α and ϕ

$$f_X(x|\alpha, \phi) = \frac{1}{x(1-x)} e^{\alpha \ln(\frac{x}{1-x}) + \phi \ln(1-x) + \ln\left(\frac{\Gamma(\phi)}{\Gamma(\alpha)\Gamma(\phi-\alpha)}\right)}. \quad (4.7)$$

Now consider adding the following “clever zero”, $-c_0\phi \ln(\frac{x}{1-x}) + c_0\phi \ln(\frac{x}{1-x})$, into the exponent of Equation (4.7). Here c_0 is a real valued constant taken from the unit interval $(0, 1)$. Upon some algebraic rearranging, we can examine a new exponential family form

$$f_X(x|\alpha, \phi) = \left[\frac{1}{x(1-x)} \right] e^{[(\alpha - c_0\phi) \ln(\frac{x}{1-x}) + \phi(c_0 \ln(\frac{x}{1-x}) + \ln(1-x))] + \ln\left[\frac{\Gamma(\phi)}{\Gamma(\alpha)\Gamma(\phi-\alpha)}\right]}. \quad (4.8)$$

In Equation (4.8), let $\psi = \alpha - c_0\phi$ and $\chi = \phi$ denote the canonical parameters. Fortunately in this case, testing the hypothesis $H_0 : \psi \leq 0$ vs $H_1 : \psi > 0$ is equivalent to testing the hypothesis $H_0 : \mu \leq c_0$ vs $H_1 : \mu > c_0$. Now that we have an exponential family with canonical parameters that directly correspond to our hypothesis of interest, we can use the saddlepoint approximation results within Chapter 3 to provide a formal statistical test.

The density function of the beta random variable, reparameterized in terms of $\psi = \alpha - c_0\phi$ and $\chi = \phi$, is

$$f_X(x|\psi, \chi) = \left[\frac{1}{x(1-x)} \right] e^{\psi t_1(x) + \chi t_2(x) - A(\psi, \chi)} \quad (4.9)$$

where

$$t_1(x) = \ln \left(\frac{x}{1-x} \right),$$

$$t_2 = c_0 \ln \left(\frac{x}{1-x} \right) + \ln(1-x) = \ln \left(\frac{x^{c_0}}{[1-x]^{c_0-1}} \right),$$

and

$$A(\psi, \chi) = -\ln \left(\frac{\Gamma(\chi)}{\Gamma(\psi + c_0\chi)\Gamma(\chi - c_0\chi - \psi)} \right).$$

For a given sample of data, the sufficient statistics for ψ and χ are

$$T_1 = \sum_{i=1}^n \ln \left(\frac{x_i}{1-x_i} \right) \quad (4.10)$$

and

$$T_2 = \sum_{i=1}^n \ln \left(\frac{x_i^{c_0}}{[1-x_i]^{c_0-1}} \right) \quad (4.11)$$

respectively.

According to [9] and summarized in Chapter 3, a UMP size α test should be based on the probability distribution of the conditional random variable $T_1|T_2$. While we do not know the conditional PDF of $T_1|T_2$ directly, saddlepoint approximations are readily available due to the fact we know the joint cumulant generating function of T_1 and T_2 which is obtained by using $A(\psi, \chi)$. Additionally, we summarized the very useful connection between the CGF and saddlepoint equations with likelihood quantities and MLEs. These results allow for one to streamline the approximation process. We now derive the likelihood, Fisher's information, and the system of equations to compute MLEs needed for the saddlepoint PDF and CDF formulas provided in Equations (3.20) and (3.24).

The likelihood function for a simple random sample of beta distributed random variables under the (ψ, χ) canonical parameterization is

$$\mathcal{L}(\psi, \chi) = \left[\frac{\Gamma(\chi)}{\Gamma(\psi + c_o\chi)\Gamma((1 - c_o)\chi - \psi)} \right]^n \left[\prod_{i=1}^n \left(\frac{1}{x_i(1 - x_i)} \right) \right] e^{\{\psi T_1 + \chi T_2\}}. \quad (4.12)$$

Fisher's information $j(\psi, \chi)$, obtained by computing the matrix of partial derivatives of $A(\psi, \chi)$, is

$$j(\psi, \chi) = \begin{bmatrix} j_{\psi\psi} & j_{\psi\chi} \\ j_{\chi\psi} & j_{\chi\chi} \end{bmatrix} = \begin{bmatrix} nA_{\psi\psi}(\psi, \chi) & nA_{\chi\psi}(\psi, \chi) \\ nA_{\psi\chi}(\psi, \chi) & nA_{\chi\chi}(\psi, \chi) \end{bmatrix} \quad (4.13)$$

where

$$\begin{aligned} nA_{\psi\psi}(\psi, \chi) &= n [\Psi''(\psi + c_o\chi) + \Psi''((1 - c_o)\chi - \psi)] \\ nA_{\chi\psi}(\psi, \chi) &= n [c_o\Psi''(\psi + c_o\chi) - (1 - c_o)\Psi''((1 - c_o)\chi - \psi)] \\ nA_{\psi\chi}(\psi, \chi) &= n [c_o\Psi''(\psi + c_o\chi) - (1 - c_o)\Psi''((1 - c_o)\chi - \psi)] \\ nA_{\chi\chi}(\psi, \chi) &= n [-\Psi''(\chi) + c_o^2\Psi''(\psi + c_o\chi) + (1 - c_o)^2\Psi''((1 - c_o)\chi - \psi)]. \end{aligned}$$

Note again that,

$$\begin{aligned} \Psi(y) &= \frac{d}{dy} \ln \Gamma(y) \text{ is digamma,} \\ \Psi'(y) &= \frac{d^2}{dy^2} \ln \Gamma(y) \text{ is trigamma and} \\ \Psi''(y) &= \frac{d^3}{dy^3} \ln \Gamma(y) \text{ is polygamma function of order 2.} \end{aligned}$$

Taking the log of Equation (4.12), the log-likelihood is

$$l(\psi, \chi) = n \ln \left[\frac{\Gamma(x)}{\Gamma(\psi + c_o\chi)\Gamma((1 - c_o)\chi - \psi)} \right] + \left[\sum_{i=1}^n \ln \left(\frac{1}{x_i(1 - x_i)} \right) \right] + \psi T_1 + \chi T_2 \quad (4.14)$$

The derivatives of the log-likelihood with respect to ψ and χ respectively are

$$\frac{\partial}{\partial \psi} l(\psi, \chi) = n [\Psi'((1 - c_o)\chi - \psi) - \Psi'(\psi + c_o\chi)] + T_1 \quad (4.15)$$

$$\frac{\partial}{\partial \chi} l(\psi, \chi) = n [\Psi'(\chi) - c_o\Psi'(\psi + c_o\chi) - (1 - c_o)\Psi'((1 - c_o)\chi - \psi)] + T_2. \quad (4.16)$$

The unrestricted MLEs of the canonical parameters, $\hat{\psi}$ and $\hat{\chi}$ are obtained by setting Equations (4.15) and (4.16) to 0, and solving them numerically (equations are not of closed form). The restricted MLE $\hat{\chi}_{\psi=0}$ is obtained by substituting the null hypothesized value $H_0 : \psi = 0$ for ψ in Equation (4.16), equating and then solving for χ numerically. Both the unrestricted and restricted MLE solutions can be obtained using the STATS4 package within R.

The previous derivations are plugged into Equation (3.20) to produce the saddlepoint conditional density of $T_1|T_2$,

$$\hat{f}_{T_1|T_2}(t_1|t_2; \psi = 0) = (2\pi)^{-\frac{1}{2}} \left\{ \frac{|j(\hat{\psi}, \hat{\chi})|}{|j_{\chi\chi}(\psi, \hat{\chi}_{\psi=0})|} \right\}^{-\frac{1}{2}} \frac{\mathcal{L}(\psi, \hat{\chi}_{\psi=0})}{\mathcal{L}(\hat{\psi}, \hat{\chi})}. \quad (4.17)$$

The normalized saddlepoint PDF technique in Equation (3.3) was implemented to ensure that the saddlepoint approximation of the conditional PDF integrated to 1. To accomplish the normalization of the PDF, we integrated the (4.17) on the appropriate domain to obtain the normalization constant, which was then used for the adjustment such that the normalized conditional PDF integrated to 1. This is only critical for graphing the approximate conditional densities as we will use the Skovgard approximation to the CDF when computing tail probabilities. An important question to ask is “what is an appropriate domain?” for integration. We address this very important question next.

4.2.1 The Support of $T_1|T_2$

Note that in both the one parameter and two parameter cases, the sufficient statistic for the interest parameter is $T = T_1 = \sum_{i=1}^n \ln(\frac{x_i}{1-x_i})$ with the one parameter case having constant ϕ as part of its sufficient statistic. While the support of T marginally is $(-\infty, \infty)$ for the one parameter case, this is not true for $T_1|T_2$. It is critical that we have the correct support obtained to only evaluate the approximations via Equation

(4.17) over that interval. For any other values, the conditional density should be 0. If this is not carefully considered, numerical errors or NaN are produced.

To demonstrate the unique joint support of T_1 and T_2 , we simulated 100,000 sufficient statistics each obtained by sampling from a beta population with the parameters $(\mu = 0.9, \phi = 10)$ with a sample size of $n = 10$. We set $c_0 = 0.9$ within the calculation of T_2 such that $\psi = 0$ and thus represents a null distribution case. The joint distribution of T_1, T_2 were plotted on the left side of Figure 4.2. We can clearly see that the scatter plot between the sufficient statistics is not the whole of the \mathbb{R}^2 space but rather a subset of it.

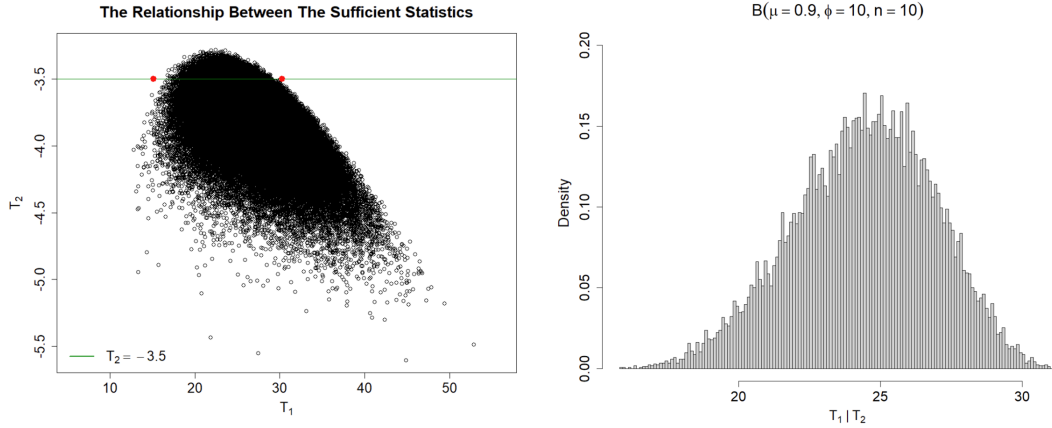


Figure 4.2: Support and Distribution of $T_1|T_2$

This shows that T_1 and T_2 are dependent variables. However, since the variable of interest is $T_1|T_2$, we are not interested in the whole domain of T_2 but just at a specific value of T_2 . If we fix the value at $T_2 = -3.5$ as depicted by the horizontal green line in Figure 4.2, we can see that the plausible values that T_1 can take is bounded above and below as indicated by the red dots.

To get a sense of what the conditional distribution of $T_1|T_2 = -3.5$ looks like, we subsetting out simulations to only include observed values of T_1 with corresponding T_2 values that were relatively close to -3.5 within a small margin of 0.05 below and

above $T_2 = -3.5$. The T_1 data points that were observed on the green line while holding $T_2 = -3.5$ fixed were of interest. A histogram of these values are shown on the right side of Figure 4.2. Again we can see that values for T_1 given $T_2 = -3.5$ range from about 14 to 33. The range will be much wider for example if we conditioned at $T_2 = -4.25$.

We can see from Figure 4.2 that the relationship between the sufficient statistics have a clear support. Knowing the boundary of the support can be helpful because the upper and lower limits of $T_1|T_2$ can be obtained. As shown below, the support can be defined by the following inequality.

$$T_2 \leq g(T_1), \quad \text{where} \quad g(x) = c_o x - n \ln(e^{\frac{x}{n}} + 1) \quad (4.18)$$

Figure 4.3 provides the previous support example, but with the support boundary generated from Equation (4.18) (blue curve) included. We can see that no simulated data falls outside of the support region. We verified this over numerous situations and by increasing the number of simulations into the millions.

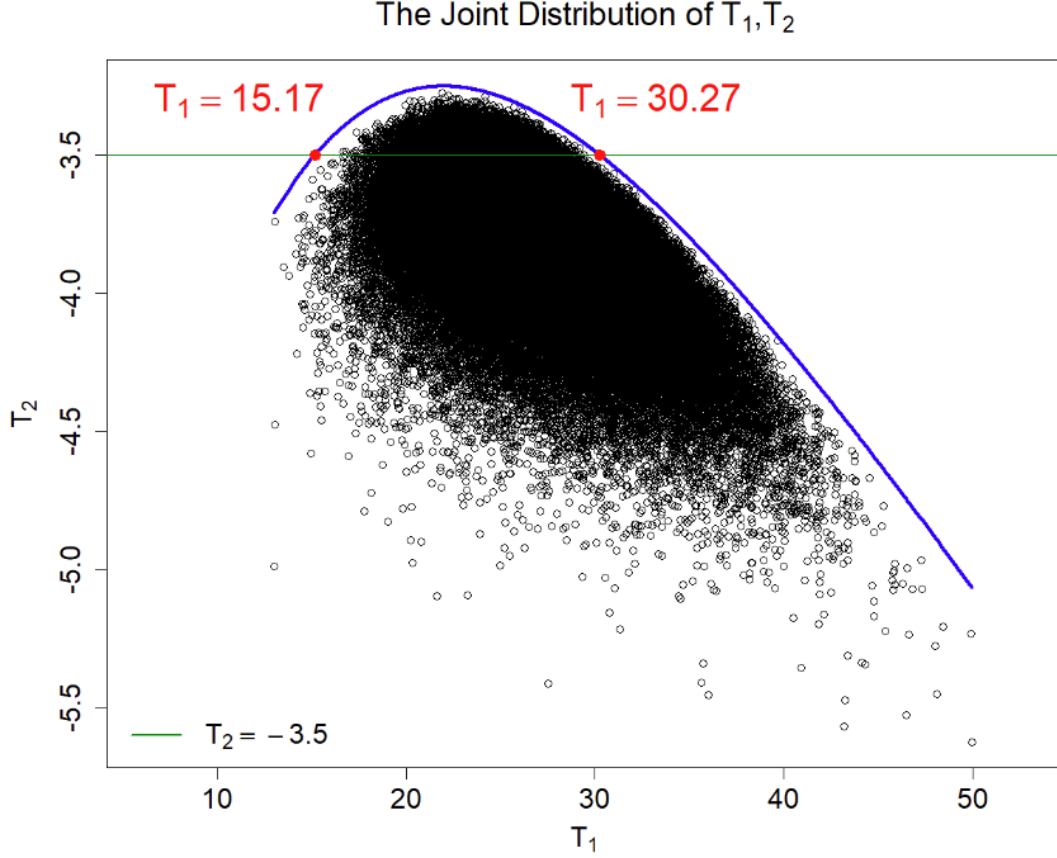


Figure 4.3: Jensen's Inequality Defining the Boundary of the Joint Distribution between the Sufficient statistics

As we condition on $T_2 = -3.5$ in Figure 4.3, we can see that the boundary and the horizontal line at $T_2 = -3.5$ intersects at two points (red dots). These correspond to the upper and lower bounds of the conditional distribution $T_1|T_2 = -3.5$. To obtain these two values, it is easily shown that $g''(x) < 0$ for all x and thus a value for T_1 that maximizes g can be obtained. We denoted this value as T_1^* . To determine the two bounds we simply apply a root finder for the function $h(x) = g(x) - (-3.5)$ over the two disjoint domains of h , $(-\infty, T_1^*)$ and (T_1^*, ∞) . For the specific example depicted in Figure 4.3, the lower and upper limits for the support given $T_2 = -3.5$

are $T_1 = 15.17$ and $T_1 = 30.27$ respectively. In general, the root finder is applied to the function $h(x) = g(x) - t_2$ where t_2 is the observed conditional value.

To illustrate the boundary further, Figure 4.4 shows different scenarios of Beta distributions, with varying sample sizes, of which the joint distribution of T_1 and T_2 were simulated. As illustrated in the figure, the joint distributions and thus the conditional distributions can be quite different depending on the values of μ , ϕ , and n .

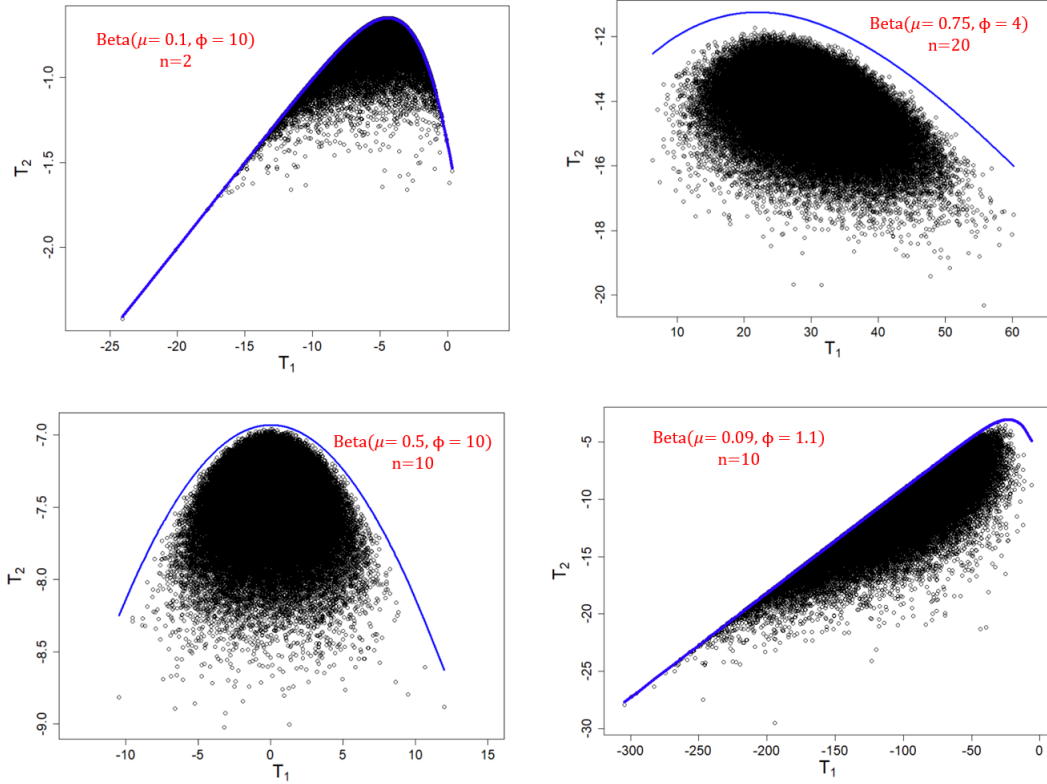


Figure 4.4: Demonstration of Jensen's Inequality in Different Scenarios

4.2.1.1 Support of $T_1|T_2$ Proof

The proof of the support depicted in Equation (4.18) is a simple exercise using Jensen's Inequality. We want to show that for a given concave function g , the statement

$g(T_1) \geq T_2$ is true.

The finite form of Jensen's inequality states that for a given concave function φ , with values x_1, x_2, \dots, x_n in its domain, we have the following inequality

$$\varphi\left(\frac{\sum_{i=1}^n x_i}{n}\right) \geq \frac{\sum_{i=1}^n \varphi(x_i)}{n}. \quad (4.19)$$

Since g in (4.18) is concave, applying Jensen's inequality and by Equation (4.10) we have

$$\begin{aligned} g\left(\frac{T_1}{n}\right) &\geq \frac{\sum_{i=1}^n g(\ln(\frac{x_i}{1-x_i}))}{n} \\ nc_0 \frac{T_1}{n} - n \ln(e^{\frac{T_1}{n}} + 1) &\geq \frac{\sum_{i=1}^n n[c_0 \ln(\frac{x_i}{1-x_i}) - \ln(e^{\ln(\frac{x_i}{1-x_i})} + 1)]}{n} \\ c_0 T_1 - n \ln(e^{\frac{T_1}{n}} + 1) &\geq \sum_{i=1}^n \ln\left(\frac{x_i^{c_0}}{(1-x_i)^{c_0-1}}\right). \end{aligned}$$

Therefore, by Equation (4.11)

$$T_2 \leq c_0 T_1 - n \ln(e^{\frac{T_1}{n}} + 1).$$

4.2.2 Examining the Saddlepoint Approximations of $T_1|T_2$

The purpose of this section is to visually demonstrate the performance of the approximation and the utility of the boundary function. We considered the beta population $B(\mu = 0.1, \phi = 10)$ and respective sample of size 10 first. The left graphic within Figure 4.5 shows the histogram of the actual population (simulation size was 1 million). The right graphic within Figure 4.5 shows the joint distribution of T_1 and T_2 along with their support boundary. We also consider conditioning on two scenarios: $T_2 = -3.7$ (green line) and $T_2 = -4.5$ (yellow line).

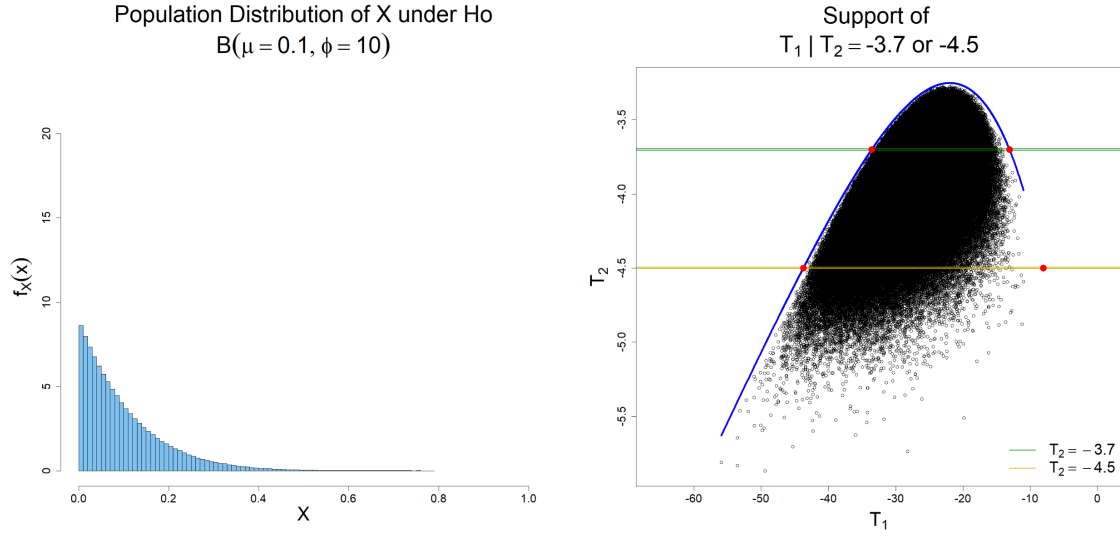


Figure 4.5: A Beta Population and Joint Distribution of $T_1|T_2$ $B(\mu = 0.1, \phi = 10), n = 10$

Figure 4.6 provides subsets of the data depicted in Figure 4.5 (right sided plot) just looking at values of T_1 when T_2 is close to -3.7 (green line) and -4.5 (yellow line). These graphs represent simulation of the conditionals and the saddlepoint approximations are overlayed on top. Upon examination, we can see that there is a great agreement between the simulated conditional data and the saddlepoint approximation. A key advantage of the saddlepoint approximations is their accuracy in the tails of the distributions where reliable p-value computations are needed most.

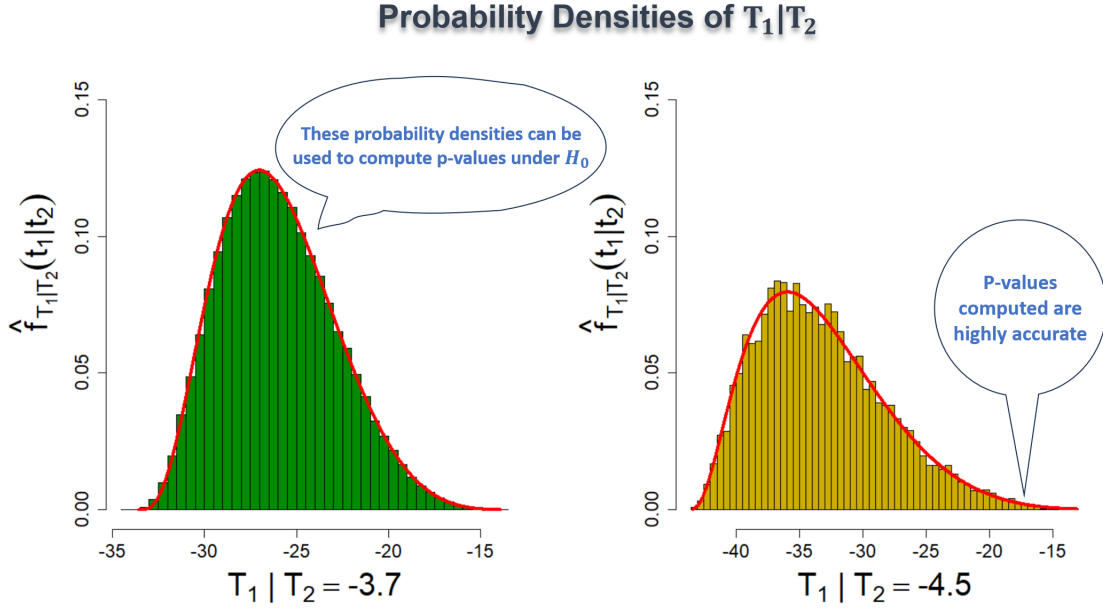


Figure 4.6: The Distribution of $T_1|T_2$ based on the Beta Population $B(\mu = 0.1, \phi = 10), n = 10$ for Statistical Inference of μ

Also, we considered situations where $\mu = 0.1$ with varied values of ϕ and considered $\phi = 5, 10$, and 20 . Figure 4.7, from left to right, provides the beta population, joint distribution of T_1 & T_2 and boundary plot, and a conditional density approximation for $T_2 = -5$ while setting $\phi = 5$ and keeping the sample size at 10 .

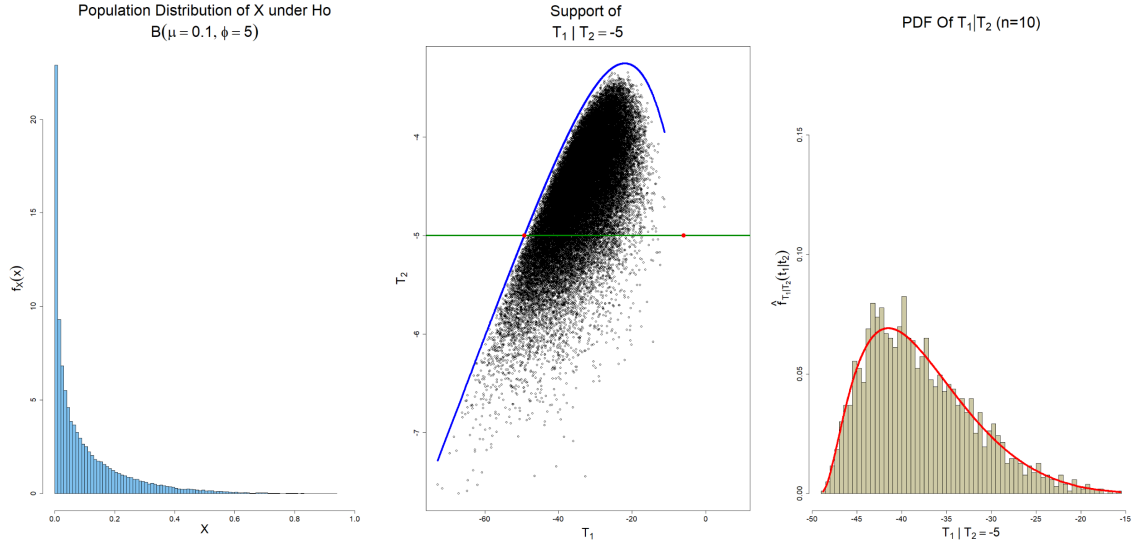


Figure 4.7: Beta Population Histogram, Support of T_1, T_2 and Distribution of UMP Test under the Condition $B(\mu = 0.1, \phi = 5), n = 10$

Similarly, Figure 4.8, repeats the process of the previous graphic, while setting $\phi = 20$ and conditioning at $T_2 = -3.6$. Note that the sample size is still fixed at 10 and the mean parameter is fixed at $\mu = 0.1$. We can see from both Figures 4.7 and 4.8 that the saddlepoint conditional PDF (red curve) trends accurately on the simulated distributions of $T_1 | T_2$ under different scenarios.

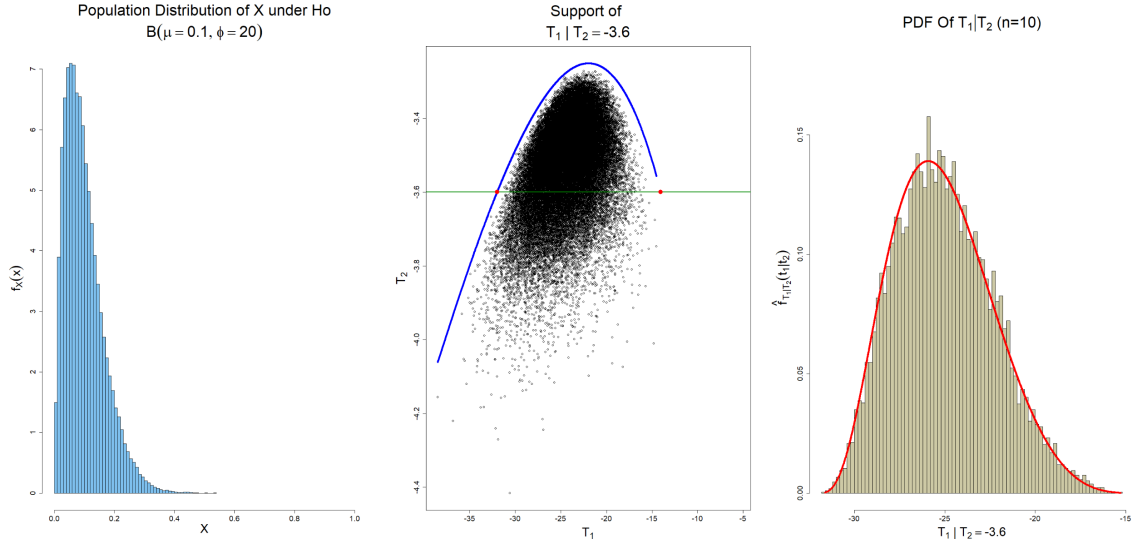


Figure 4.8: Beta Population Histogram, Support of T_1, T_2 and Distribution of UMP Test under the Condition $B(\mu = 0.1, \phi = 20), n = 10$

Of the scenarios considered so far, we have only considered $\mu = 0.1$ which produces right skewed populations. Now, we consider the case of the beta distribution with parameters $\mu = 0.5, \phi = 10$, and $n = 10$. Under this scenario the beta density is a symmetric distribution. Upon examining Figure 4.9, we see that the joint distribution of T_1 and T_2 has some symmetry to the distribution about $T_1 = 0$. This makes sense given the formulation of T_1 . With values equally likely to occur on either side of $\mu = 0.5$, $\sum_{i=1}^n \ln(\frac{x_i}{1-x_i})$ is centered at 0. As in the previous cases, the saddlepoint method performs well under this scenario.

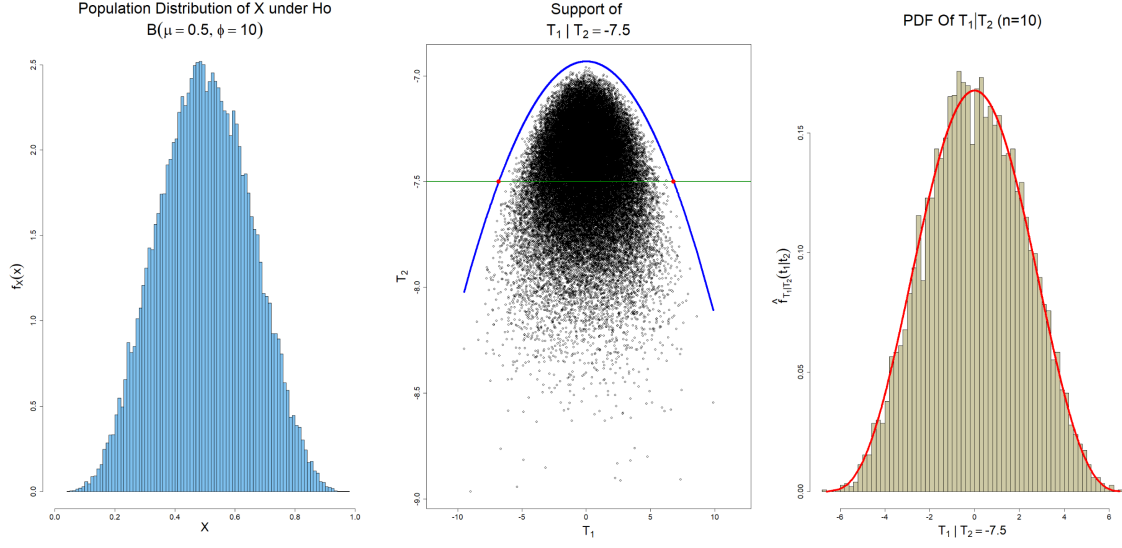


Figure 4.9: Beta Population Histogram, Support of T_1, T_2 and Distribution of UMP Test under the Condition $B(\mu = 0.5, \phi = 10), n = 10$

4.2.3 Decision Rule of the UMP Size α Test

With confidence that the conditional saddlepoint approximations are working as intended, we now describe its application for our inferencing problem. To test for $H_0 : \psi \leq 0$ vs $H_1 : \psi > 0$ which is equivalent to testing $H_0 : \mu \leq c_0$ vs $H_1 : \mu > c_0$, let t_1 and t_2 be the observed sufficient statistics obtained from the sample. The corresponding p-value, utilizing the conditional saddlepoint CDF of $T_1 | T_2$ provided in Equation (3.24), is computed by evaluating $1 - \hat{F}_{T_1 | T_2}(t_1 | t_2, \psi = 0)$. This corresponds to the right tail probability assessing how likely it is to observe values of $T_1 | T_2 = t_2$ that are as extreme or more extreme as the observed statistic $t_1 | t_2$ under $H_0 : \mu \leq c_0$. When rejecting H_0 for p-values less than a specified significance level α , the test is a highly accurate approximation to the UMP size α test for the right tailed test considered. Without loss of generality, similar computations are made for the left tailed test.

5 SIMULATION STUDIES

In this chapter, we will explore three different approaches to performing a hypothesis test for the mean of the beta distribution. The three tests include the classic Student's t-test (t-test), the Likelihood Ratio Test (LRT) described in Section 2.7 and our approach via the saddlepoint approximation method (Beta test). A comparison of each tests performance is conducted both in terms of type I error rates and statistical power.

The t-test and LRT were considered because they are standard methods that are commonly used in practice and can be easily applied to the beta distribution problem considered here. Additionally, it is well understood that the t-test, while technically assumes normally distributed data, is quite robust to this assumption but t-test may likely not be robust when the beta population is J-shaped or L-shaped (i.e. asymptotically goes to infinity on either endpoint). Assessing its robustness in the specific case of the beta distribution is an interesting study on its own. The null distribution of the LRT is often difficult to determine and the common approach is to consider a large sample approximation. We expect this test to work just fine for large sample sizes, but for smaller sample sizes, we expect the accuracy of the approximation to diminish. When assuming the precision parameter is known in advance, [2] showed via simulations that the saddlepoint approximation performed better than these standard tests. We hope that the story does not change in this case upon relaxing this assumption.

5.1 Type I Error Rate Simulations

One of the most important components of a statistical test is its ability to control type I error rates. The type I error rate is the probability of rejecting the null hypothesis given that the null hypothesis is true. A common approach to assess a test's type I error control is to utilize a simulation study. For our simulation study we compared the type I error control of our proposed method, the LRT, and t-test. For a user specified set of parameters μ_0, ϕ , and n , the simulation algorithm can be summarized as follows:

For each of 10,000 iterations:

1. Simulate x_1, x_2, \dots, x_n under the $B(\mu_0, \phi)$ model.
2. Perform the three testing procedures, setting $c_0 = \mu_0$ for a certain α .
3. Keep a running count of the number of rejected tests for each procedure.

The estimated type I error rate for each procedure is the proportion of rejected tests out of the 10,000 simulations. This estimate should be relatively close to .05 if the testing procedure is controlling the error rate as specified. Since the type I error rate result from the simulation is just an estimate, it is helpful to account for simulation error. Since the estimated rate is just a sample proportion, if the procedure is truly controlled at α , the estimates should fall within the limits $\alpha \pm 2\sqrt{\frac{\alpha(1-\alpha)}{10000}}$ roughly 95% of the time. For any type I error rate estimate outside of these limits, we consider the procedure for that given scenario to not have control of the type I error rate. For our simulations, we set the significance level to $\alpha = 0.05$.

For the simulation procedure described above, we considered the null cases, ($\mu_0 = 0.05, 0.1, 0.15, 0.2, 0.25, 0.3, 0.35, 0.4, 0.45, 0.5$). We do not consider values above 0.5 due to the symmetry of the beta distribution. If type I error is controlled at $\mu_0 = 0.25$, then it is controlled at $\mu_0 = 0.75$. Also, we consider two different values for the

precision parameter ($\phi = 5, 10$) and three different sample sizes ($n = 5, 10, 25$). For all of the scenarios listed, we considered both left and right-tailed versions of the testing procedures.

5.1.1 Right Tailed Tests

In this section, we compare the type I error rate estimations of all three test procedures under a right-tailed test : $H_0 : \mu \leq \mu_0$ vs $H_1 : \mu > \mu_0$, in order to identify which test(s) control type I error rate better. Figure 5.1 contains three plots where each plot represents one of the three tests (from left to right: beta test, LRT and t-test). Each tests ability to control type I error rates across different sample sizes for $\phi = 5$ are shown. Each graph plots the estimated type I error rate against the values of μ and is color coded and labeled by sample size. Sample sizes of 5 are coded with triangles and in red, 10 with squares in blue, and 25 with circles in green. The specified significance level, is denoted as a horizontal black line at 0.05 and accompanied by its upper and lower simulation limits in black dashed lines.

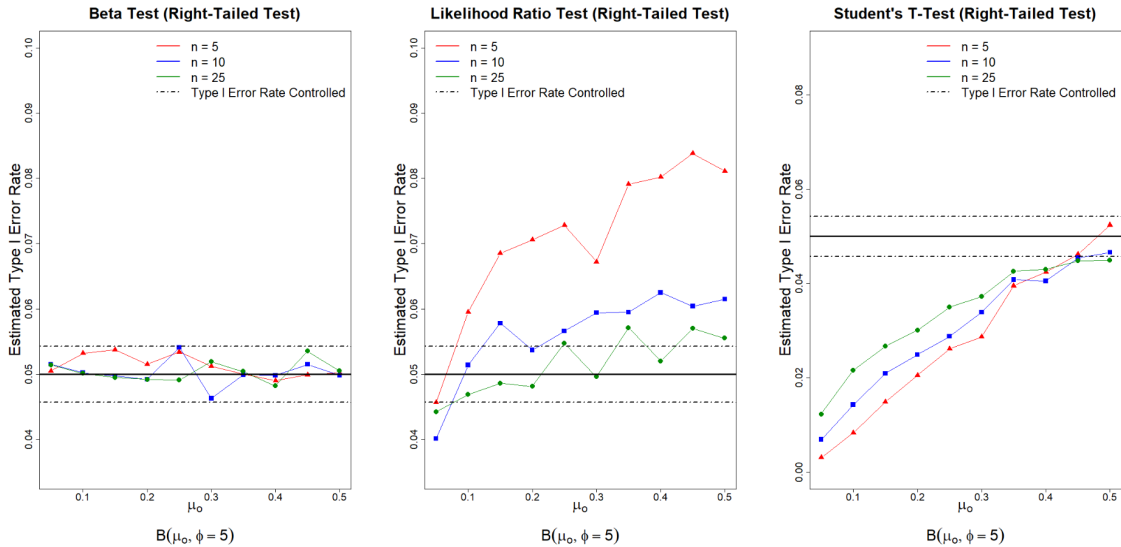


Figure 5.1: Estimated Type I Error Rates for Right-Tailed Tests ($\phi = 5$)

We can see from the right tailed beta test of Figure 5.1 that, type I error rate is controlled for all the sample sizes and null hypothesized values μ_0 under consideration. However, the LRT of Figure 5.1 is inflating type I error rate as the null hypothesized value μ_0 is getting closer to 0.5 for all sample sizes considered except for large sample ($n = 25$, green line) where type I error rate is been fairly controlled.

The student's t-test of Figure 5.1 only controls type I error rates when the null population is symmetric. Other than that, the test cannot control type I error rate. Actually, we can observe that the type I error rate is deflated as μ_0 gets closer to 0. This is due to the fact that, the t-test is not robust for extremely skewed situations and when μ_0 gets closer to 0, the density is more and more right skewed.

Also, we considered the same simulation setting as depicted in Figure 5.1 but increased the precision parameter to $\phi = 10$ so that the null beta population has less variation and in some cases much less skewed. Figure 5.2 provides the type I error rate summaries for this case.

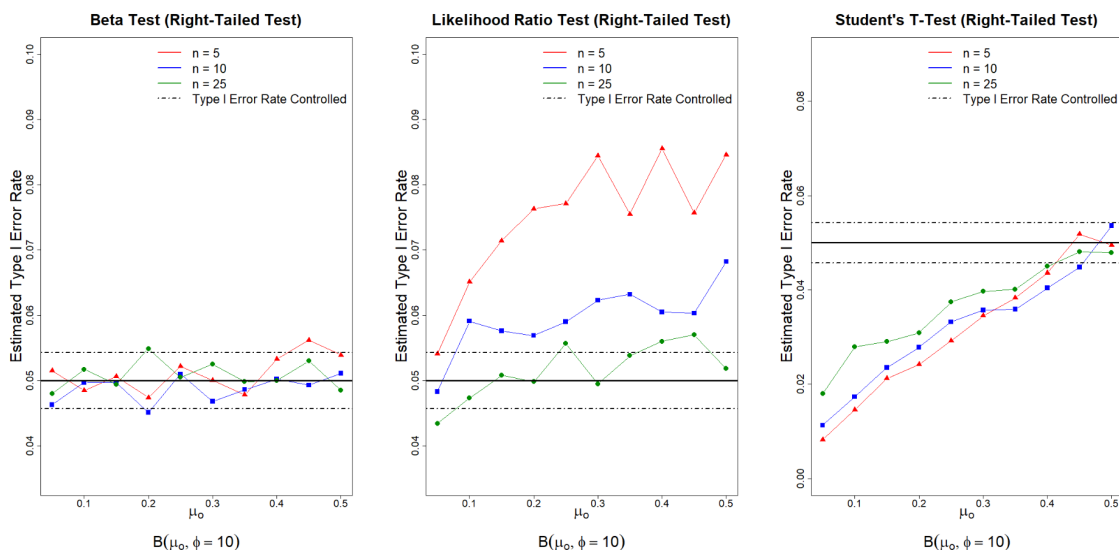


Figure 5.2: Estimated Type I Error Rates for Right-Tailed Tests ($\phi = 10$)

The results did not change for any of the tests under the parameter settings considered. The trend of behaviour for all the three tests appear to be the same. However, the t-test had an improved type I error rates although the type I error rates are still not controlled (i.e. for each sample the type I error increase by 50% from the rates on Figure 5.1).

Upon examination of both Figures 5.1 and 5.2, for a right-tailed test, the beta test under small samples for all null hypothesized values μ_0 considered, the type I error rate is well controlled regardless of the magnitude of the variability of the population.

With respect to the LRT for large sample cases and skewed populations, type I error rate is controlled regardless of the size of the population variance. Also, the type I error rate of LRT tends to inflate as the null population distribution approaches a symmetric shape (i.e. μ approaches 0.5). We can see from Figure 5.2 that for small samples ($n = 5, 10$), the LRT inflated type I error rates as high as 8.5% compared to the other sample sizes.

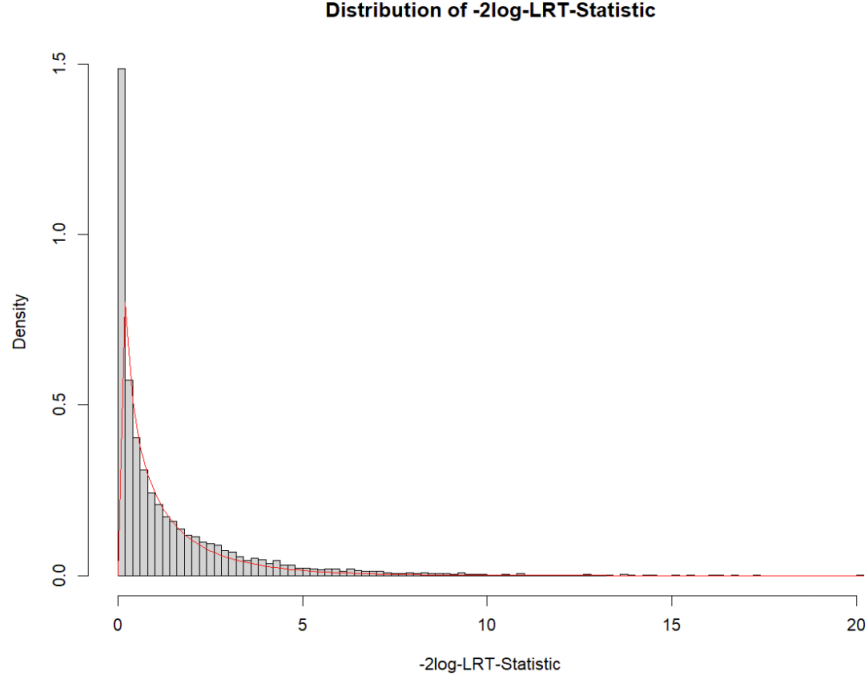


Figure 5.3: Sampling Distribution of $-2\log\text{-LRT-Statistic}$ $B(\mu = 0.5, \phi = 10)$ $n = 5$

Figure 5.3 shows that the inflated type I error rate is due to the fact that there is an overflow of $-2\log\text{-LRT-Statistics}$ at the right tail of the sampling distribution, therefore, there is more than 5% of the test statistic at the rejection region of the theoretical curve, χ_1^2 (red curve).

Lastly, for t-test, type I error rate is controlled for symmetric populations regardless of sample sizes, however, the type I error rates deflate as μ_0 approaches 0 with some estimates reaching below 0.01.

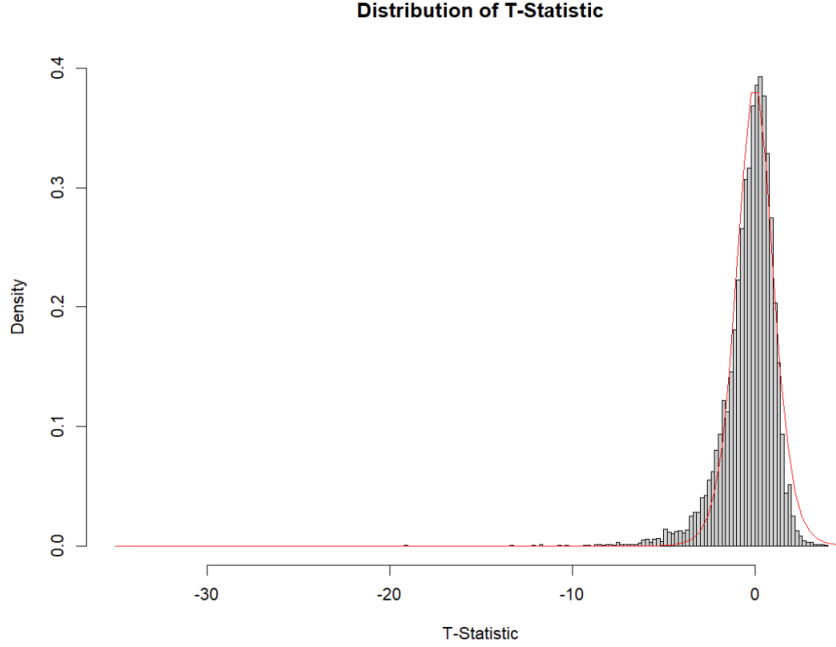


Figure 5.4: Sampling Distribution of T-Statistic $B(\mu = 0.1, \phi = 10)$ $n = 10$

This is due to the fact that the beta population becomes more heavily right-skewed as μ approaches 0. Figure 5.4 shows that the population's skewness in turn makes the sampling distribution of t-statistics left skewed and no longer symmetric, leaving very few t-statistics in the right tail. With this, the area in the right tail of the theoretical t -distribution with $df = n - 1$ (red curve) is no longer accurate for p-value calculations.

5.1.2 Left Tailed Tests

In the previous section, we studied the variation in type I error rate control for the three tests under the right-tailed test. In this section, we considered the left-tailed test: $H_0 : \mu \geq \mu_0$ vs $H_1 : \mu < \mu_0$, to assess the type I error rate control for the same testing procedures. Again, we have three plots presented in Figure 5.5 as it was in the right-tailed versions. The left plot is the beta test, the middle plot is the LRT,

and the right plot is the t-test. In this case, the precision parameter considered was $\phi = 5$.

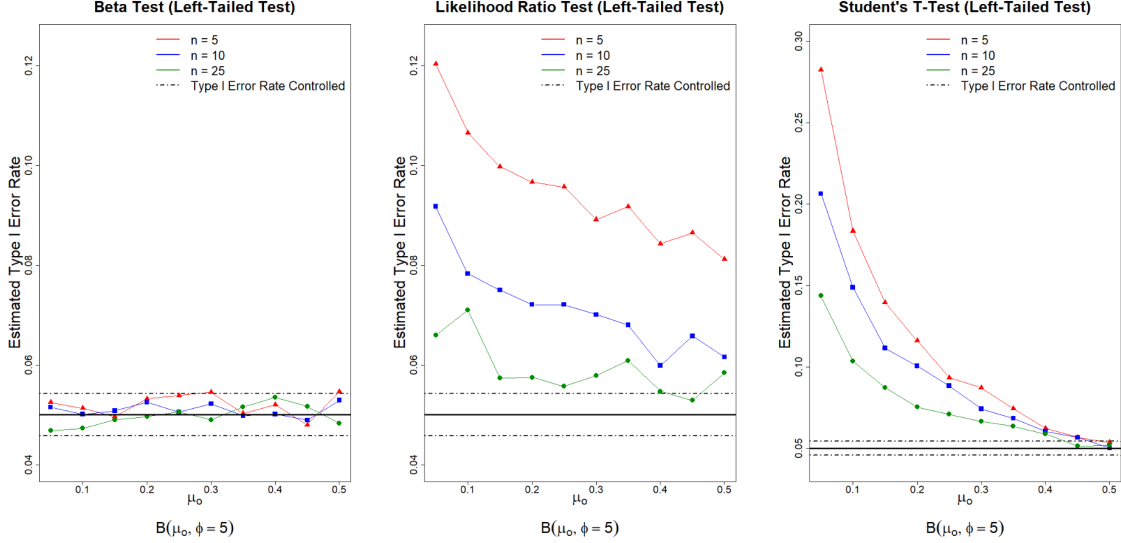


Figure 5.5: Estimated Type I Error Rates for Left-Tailed Tests ($\phi = 5$)

We can see from Figure 5.5 that for a left-tailed test, the beta test (left plot) once again controls type I error rates across all values of μ_0 and different sample sizes. With regards to the LRT, none of the tests controlled the type I error rate. In particular, the LRT applied to small samples ($n = 5$, red line) are particularly problematic where the type I error rates are above 8% but peaks at 20% as μ_0 approaches 0. In general, we can see that inflation of type I error rate for the LRT tends to increase as μ_0 approaches 0 (skewed distribution). This is very interesting because, for the right-tailed test, the LRT controls type I error rates relatively well for $n = 25$; however, the left-tailed LRT test has type I error inflation for the same sample size scenario.

The student's t-test fails to control the type I error rate for essentially all cases except for when μ_0 is close to 0.5. This has been consistent with both left and right-tailed tests. However, the t-test inflates the type I error rates as the population distribution changes from symmetric to skewed (as μ_0 approaches 0). For small

samples, the inflation of the t-test in terms of type I error rates can be extremely high in some cases. For example, when $\mu_0 = 0.05$ and 0.1 , we can see from Figure 5.5 that the t-test have type I error rates that are at least 10% but at most 28%.

Also, we considered the same left-tailed simulation setting as the previous above but increased the precision parameter to $\phi = 10$ so that the null beta population has less variation.

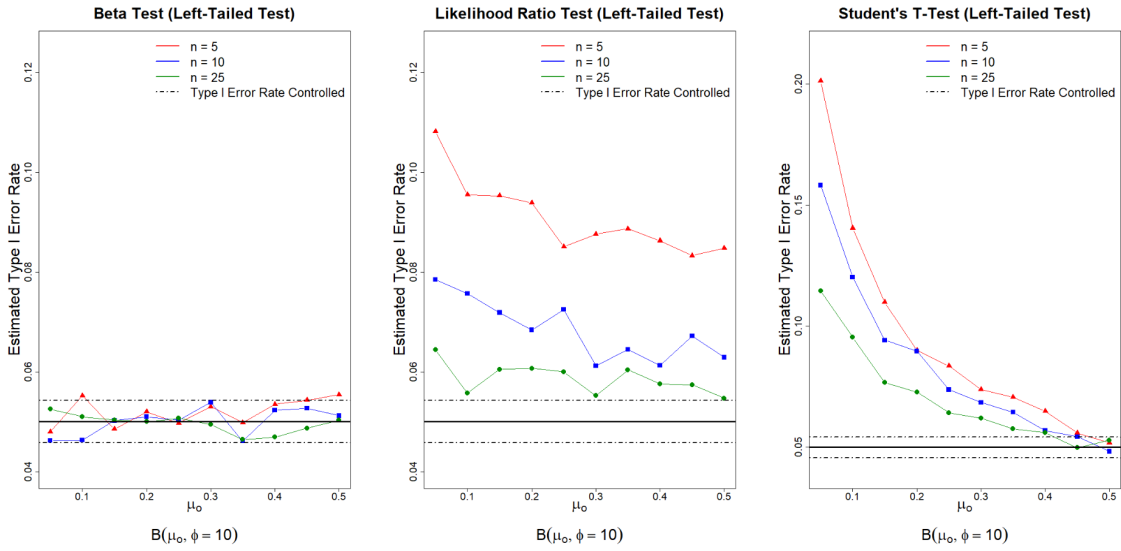


Figure 5.6: Estimated Type I Error Rates for Left-Tailed Tests ($\phi = 10$)

We can see from Figure 5.6 that, while the trend is relatively consistent as before when $\phi = 5$, there are some characteristics worth discussing. In general, the type I error rate estimates of the LRT and t-test on Figure 5.6 is low compared to the estimates when $\phi = 5$ in Figure 5.5. This suggests that when the beta distribution has a larger ϕ value, the LRT and t-test procedures can be trusted more. A deeper simulation study to determine appropriate sample sizes for ϕ and n for the LRT and t-test is warranted but we leave this for future work.

5.1.3 Type I Error Simulation Summary

In general, the t-test performed badly with regards to type I error rate control. This is due to the fact that the beta distribution does not exhibit symmetry for values of μ_0 close to 0. Therefore, the coupling of asymmetric and high variance beta distributions result in severe departures from normality and as the t-test is not robust to these types of departures, the estimated type I error rate is indeed not controlled.

The LRT uses the asymptotic properties of the MLE for ψ, χ . The apparent inflation in the estimated type I error rate for values of μ_0 relatively close to 0 is attributed to the fact that the asymptotic properties of the MLEs of ψ, χ are not necessarily maintained under skewed population and small sample conditions due to the asymmetry of the sampling distribution of the MLEs. However, the type I error rate does become slightly better controlled for relatively large sample sizes, as large sample theory would suggest.

The beta test (saddlepoint approximation method) performs best in all cases of μ, ϕ, n we considered. This provides empirical evidence substantiating the theory provided by [9, 3] regarding the use of saddlepoint approximations to perform approximate UMP tests based on the conditional random variable of the sufficient statistics $T_1|T_2$.

5.2 Power Simulations

In addition to type I error control, the statistical power of a test is also of interest to researchers. Statistical power is the probability of a test to detect a false null hypothesis. For testing procedures that all control type I error rates effectively, the test that has the higher power is preferred. Power estimate via simulation is obtained similarly to type I error rate estimates. However, in this setting, the proportion of rejections are counted while simulating under an alternative hypothesis parameter,

μ_1 , rather than the null hypothesized value μ_0 . For a right tailed test, $\mu_0 < \mu_1 < 0.5$ but for left tailed test $0 < \mu_1 < \mu_0$. For a user-specified set of parameters μ_0, μ_1, ϕ , and n , the simulation algorithm for estimating power can be summarized as follows.

For each of 10,000 iterations:

1. Simulate x_1, x_2, \dots, x_n under the $B(\mu_1, \phi)$ alternative model.
2. Perform the three testing procedures, setting $c_0 = \mu_0$ for a significance level, α .
3. Keep a running count of the number of rejected tests for each procedure.

For discussion and summary of results, we denoted the power of a test as $\beta(\mu_1)$ to denote that we are considering the power of a test as a function of the alternative mean parameter alone and thus treating the sample size and precision parameter as fixed quantities. Since there is no known theoretical power value to compare against, we provide simulation error limits for a simulated power estimate using the formula, $\hat{\beta}(\mu_1) \pm 2\sqrt{\frac{\hat{\beta}(\mu_1)(1-\hat{\beta}(\mu_1))}{10000}}$, where $\hat{\beta}(\mu_1)$ is the simulated estimate of power at μ_1 for a given simulation scenario involving μ_0, ϕ , and n .

We assessed the power of the three testing procedures considering both left and right-tailed tests when setting the null value $\mu_0 = 0.1$ and for varying ranges of ϕ, n , and alternative μ_1 . We considered the specific null value $\mu_0 = 0.1$ to assess the power of the tests on a skewed distribution under the null hypothesis and this is one of the cases where the procedures performed significantly different in terms of controlling type I error rates. Although we do not show them in this thesis, all three tests' power estimates are very similar when μ_0 is 0.5.

The farther away the selected null value $\mu_0 = 0.1$ is from the alternative value, μ_1 , the higher the power estimate should be observed in the simulation since there is a clear distinction between the test statistics observed versus what is expected under H_0 . In these cases, the test statistics observed is rare to appear when compared to the

null distribution of the test statistic being used. Therefore, higher statistical power means that there is a higher chance that the test detects real difference when it exists.

5.2.1 Power of Right Tailed Tests

In this section, we compare the power estimate of all three tests under the right-tailed setting. In this case, the precision parameter was $\phi = 5$ and 10, and in each of the three plots, a different sample size was considered as well.

For a fixed value of ϕ , we present the simulated power curves for each of the three tests in a three-panel graph, where each panel considers a different sample size that spans across $n = 5, 10, 25$. The horizontal axis is the parameter space of $\mu \in (0, 0.5)$, but we limited the space from 0 to the point where the power is 1. For a one-tailed test power curve, once the power is 1, it stays at 1 till the end, which is not interesting. So, the range of μ on the horizontal axis for each plot varies.

For a right tailed test, the power of a test is only obtained across the alternative parameter space which is $\mu_0 < \mu_1 < 0.5$. We considered 26 random values without replacement for μ_1 . The vertical axis is the power estimates $\beta(\mu)$, which can range from 0 to 1. There are three power curves on each plot. The red power curve is the beta test (saddlepoint approximation method), the blue power curve is the LRT, and the green power curve is the student's t-test. Under each condition of the assessment, there are two sets of panel plots based on a ϕ size.

The first-panel plot, Figure 5.7, views the power curve of the three tests in totality for $\phi = 5$ and $\mu_0 = 0.1$. We can see from Figure 5.7 that, for a sample size of $n = 5$, the LRT and beta tests are seen as powerful, but LRT has better power than the other two procedures. While the LRT and beta tests power estimates are relatively close, the t-test lags behind by a substantial amount. With regards to the other two plots where sample sizes are 10 and 25, we can see that LRT and beta tests become more consistent and the t-test still suffers in comparison.

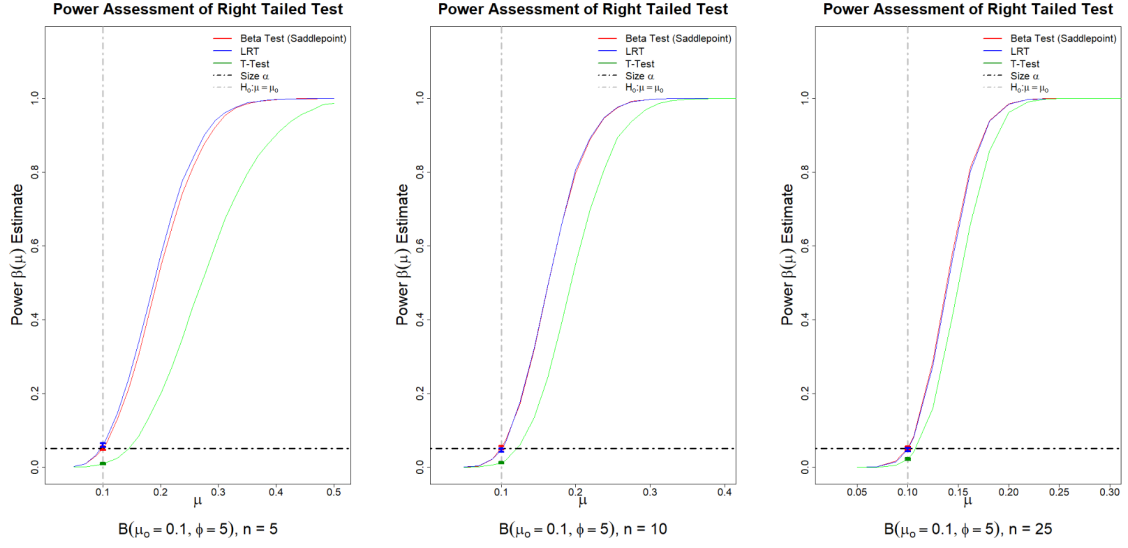


Figure 5.7: Simulated Power Curves for Right-Tailed Tests ($\mu_0 = 0.1, \phi = 5, n = 5, 10, 25$)

Figure 5.8 provides a zoomed-in plot of Figure 5.7 to again emphasize that the power of the test alone should not be the sole comparison. For these plots, simulation error limits were calculated at the power estimates of each test. While the LRT slightly edges out the beta test in terms of power as observed previously, for values of μ that are less than or equal to 0.1, the power estimates are technically type I error estimates and should be less than or equal to the 0.05 to qualify as a UMP $\alpha = 0.05$ test. For the LRT test, when $n = 5$, type I error is not controlled at $\mu = 0.1$. So, while the test has higher power, it is not a valid UMP test for the case. While the beta and t-test both maintain type I error rate estimates below 0.05, the beta testing procedures is vastly more powerful as depicted in Figure 5.8. For $n = 10$ and $n = 25$, all three tests keep type I error rates below 0.05, but in these cases, there is essentially no difference in power estimates between the LRT and beta testing procedures.

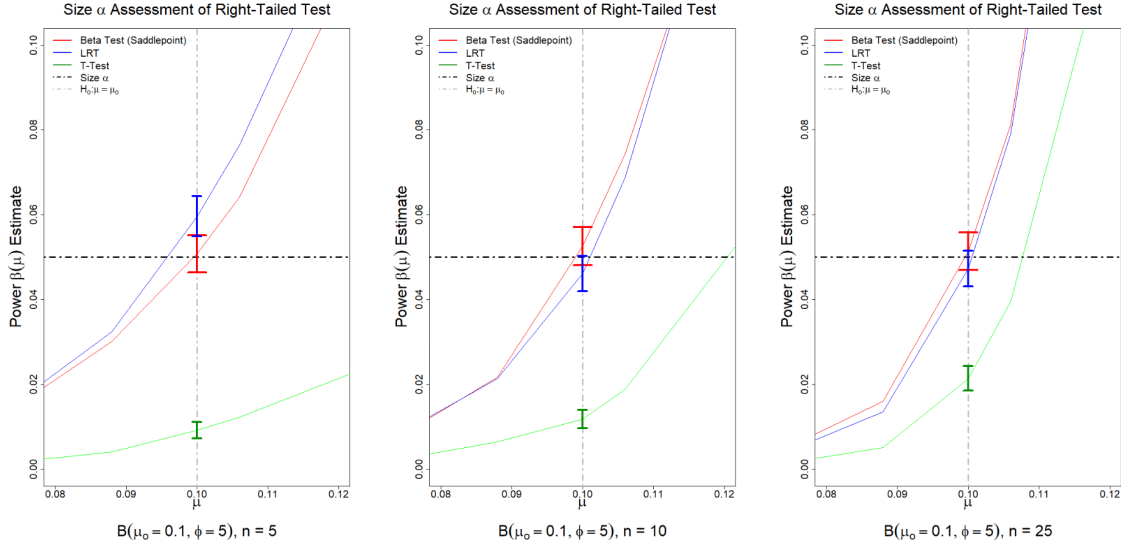


Figure 5.8: Size α Tests for Right-Tailed Tests ($\mu_0 = 0.1, \phi = 5, n = 5, 10, 25$)

We can see from Figure 5.9 that the power curves under beta populations with high precision $\phi = 10$ have a similar trend in terms of power estimates as when $\phi = 5$. For the smaller samples (left plot) we can see that LRT test has slightly higher power than the beta followed by the much less powerful t-test. For sample sizes 10 and 25 the beta test and LRT power curves look visually indistinguishable while the t-test continues to improve and obtain power much closer to the beta and LRT tests compared to when $\phi = 5$.

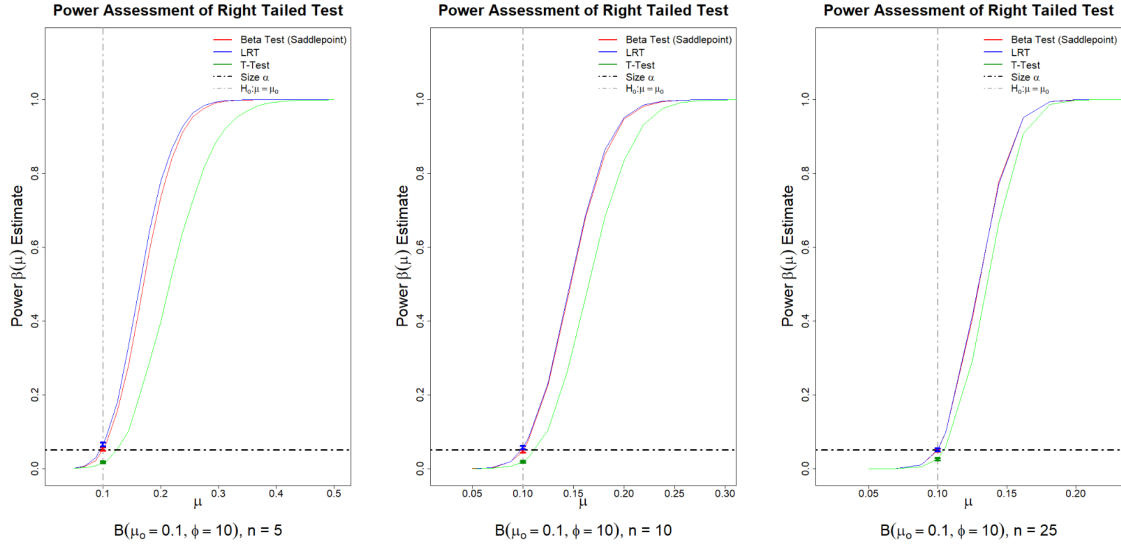


Figure 5.9: Simulated Power Curves for Right-Tailed Tests ($\mu_0 = 0.1, \phi = 10, n = 5, 10, 25$)

Figure 5.10 provides a closer look at the type I error estimates for $\phi = 10$. Similarly to the first case when $\phi = 5$, type I error is not controlled for the LRT when $n = 5$ and $n = 10$. Thus when choosing between the beta and t-test, which do control the error rate, the beta test is clearly more powerful across all scenarios.

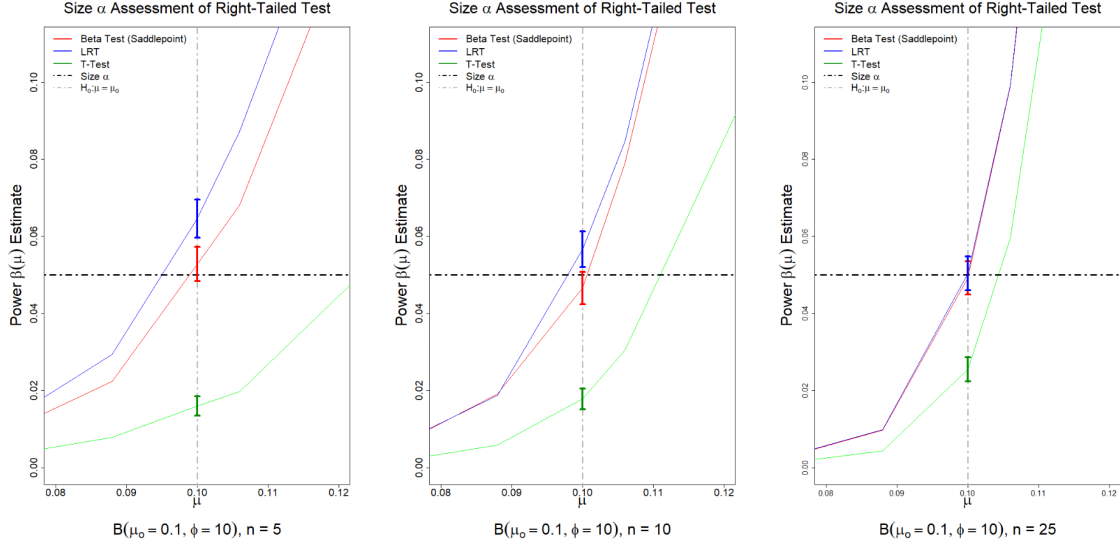


Figure 5.10: Size α Tests for Right-Tailed Tests ($\mu_0 = 0.1, \phi = 10, n = 5, 10, 25$)

In summary, the beta test is able to control type I error rate and is most powerful in all cases we considered for a right tailed test. However, for larger samples we can see that the LRT and beta tests are both reasonable approaches to consider.

5.2.2 Power of Left-Tailed Tests

In this section we compare the power estimates of all three tests under the left-tailed setting again examining values of $\phi = 5, 10$ and $n = 5, 10, 25$. For a left tailed test, the power of a test is only obtained across the alternative parameter space which is $0 < \mu_1 < \mu_0$. We considered 35 random values without replacement for μ_1 .

Figure 5.11 provides power estimates for $\phi = 5$ for all 3 tests and sample sizes. As expected, we can see that as sample size increases the power estimates of the tests increase.

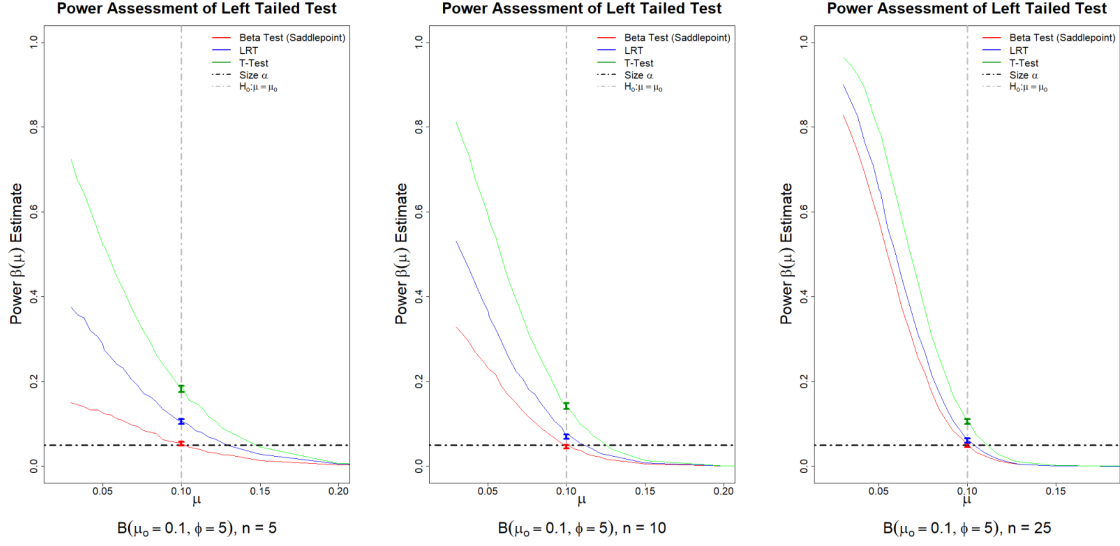


Figure 5.11: Simulated Power Curves for Right-Tailed Tests ($\mu_0 = 0.1, \phi = 5, n = 5, 10, 25$)

When examining the power estimates (when $\mu < 0.1$), the t-test has substantially higher power, followed by the LRT, and then the beta test across all settings. Unlike the right-tailed setting, it is much easier to observe the lack of type I error control for both the t-test and LRT settings and is consistent with our previous, more exhaustive type I error simulation study. While the t-test could be used as a more sensitive testing procedure, high type I error rates must be an important consideration before using it.

Figure 5.12 provides power estimates for $\phi = 10$. In terms of power, we can see that the beta test has the lowest power estimates followed by the LRT and t-test across all sample size situations. Again, we see the lack of type I error control for the LRT and t-test across all sample sizes.

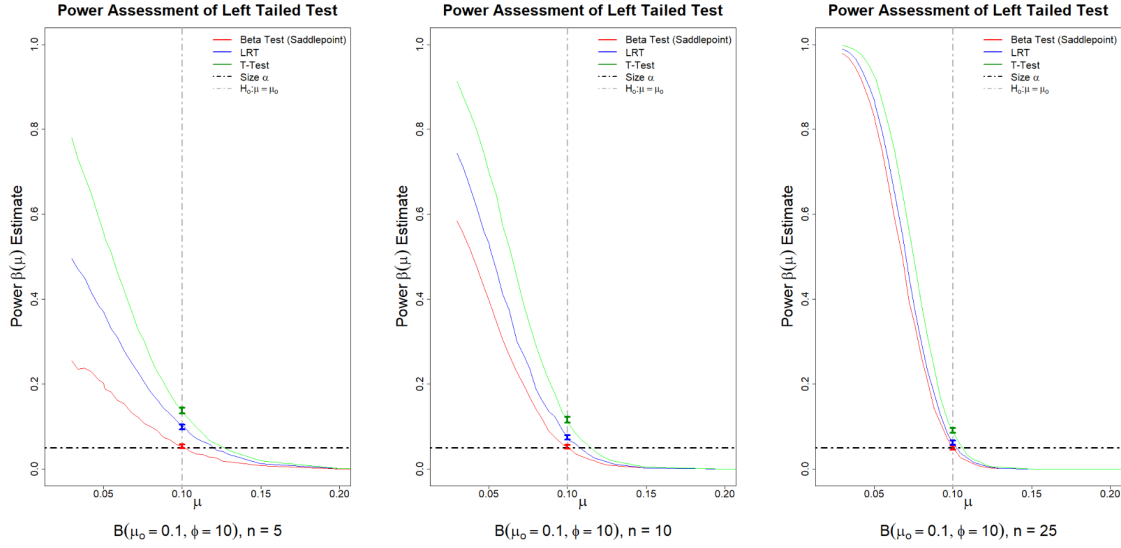


Figure 5.12: Simulated Power Curves for Right-Tailed Tests ($\mu_o = 0.1, \phi = 10, n = 5, 10, 25$)

As we increased precision to $\phi = 10$, we can see from Figure 5.13 that there was no significant change of the tests controlling type I error rate when compared to the left-tailed test with less precision ($\phi = 5$) results. The beta test is the only procedure that consistently controls the type I error rate.

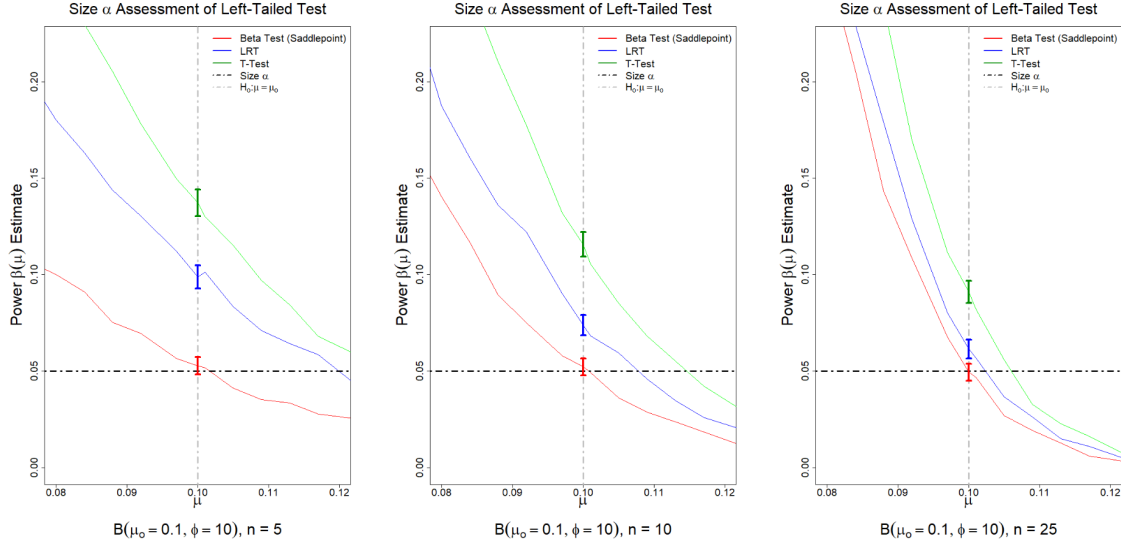


Figure 5.13: Size α Tests for Right-Tailed Tests ($\mu_0 = 0.1, \phi = 10, n = 5, 10, 25$)

5.3 Summary

It has been interesting to explore the dynamics governing the type I error rate and power estimation of the three tests based on a skewed null beta population distribution. We found that the beta test (saddlepoint approximation method) based on the conditional distribution $T_1|T_2$ controls the type I error rate in all cases of μ , ϕ , and n . For a left-tailed test, the type I error rate of the LRT and t-test is inflated as μ_0 approaches 0. For the right-tailed test, the type I error rate is inflated for LRT, whereas it is deflated for the t-test. We also found that for right-tailed tests, the LRT controls type I error rate for large samples but struggles to do so for the left-tailed setting.

In terms of power, while some tests do have better power than the saddlepoint method, they do not control the type I error rate well in all cases. The saddlepoint method is such an accurate approximation to the theoretical UMP test, we recommend its use in a general setting. As expected, when n increases, the power of all the

tests increases. This is also true for increasing values of ϕ , which makes sense given that tighter variance yields less skewed beta distributions to sample from, which helps the approximation of the CLT and the robustness of the t-test can be utilized.

6 CONCLUSION AND FUTURE RESEARCH

The purpose of this research was to develop and study a statistical testing procedure for the mean of a beta random variable while treating the precision parameter as unknown. Our approach was to apply a saddlepoint approximation for the UMP test while also considering classic likelihood-based methods and the T-test. We limited the statistical procedure to just a one tailed test (left and right tailed test).

Through simulation work, we assessed the performance of the saddlepoint approximation method, LRT and T-test. Upon our investigations, we recommend that the saddlepoint approximation method be used. While the LRT can be recommended for large samples, its ability to control type I error rates varies across sample size scenarios. The saddlepoint procedure does not have this issue. We do not recommend the use of the T-test due to its inability to control type I error rates in almost all cases we considered with the exception of symmetric beta cases. Also, we recommend that, in cases where the sample size is large under a right-tailed test, the LRT method could be implemented and may be preferred if there are issues with any of the calculations involved in the saddlepoint approximation method.

In terms of future research, a natural extension would be to derive saddlepoint approximations in the two-tailed testing scenario. While a UMP two-sided test does not exist, it is possible to construct a uniformly most powerful unbiased (UMPU) test [9, 3] for our problem. Additionally, the development of a confidence interval routine by inverting our tests would be helpful for direct interpretation and insight to what plausible values the mean can take. Also, we look forward to developing a more general two-sample comparison test in ANOVA-type settings.

BIBLIOGRAPHY

- [1] Martin Biehler, Heinz Holling, and Philipp Doeblner, *Saddlepoint approximations of the distribution of the person parameter in the two parameter logistic model*, Psychometrika **80** (2015), 665–688.
- [2] Bryn Brakefield, *Using saddlepoint approximations and likelihood-based methods to conduct statistical inference for the mean of the beta distribution*, Stephen F. Austin ScholarWorks (2020).
- [3] Ronald W. Butler, *Saddlepoint approximations with applications*, Cambridge University Press, 2007.
- [4] George Casella and Roger L. Berger, *Statistical inference*, vol. 2, Duxbury Pacific Grove, CA, 2002.
- [5] Daniels H. E., *Saddlepoint Approximations In Statistics*, The Annals Of Mathematical Statistics **25** (1954), no. 4, 631 – 650.
- [6] Constantino Goutis and George Casella, *Explaining the saddlepoint approximation*, The American Statistician **53** (1999), no. 3, 216–224.
- [7] Arjun K. Gupta and Saralees Nadarajah, *Handbook of beta distribution and its applications*, CRC press, 2004.
- [8] Teodor Tiplica, Stéphane Dufreneix, and Christophe Legrand, *A bayesian control chart based on the beta distribution for monitoring the two-dimensional gamma index pass rate in the context of patient-specific quality assurance*, Medical Physics **47** (2020), no. 11, 5408–5418.

- [9] G. A. Young and R. L. Smith, *Essentials of statistical inference*, Cambridge Series In Statistical And Probabilistic Mathematics, Cambridge University Press, 2005.

VITA

Richard Ntiamoah Kyei received a Bachelor's degree in Statistics from Kwame Nkrumah University of Science and Technology (KNUST) in 2018. He is a mathematics enthusiast. He loves to help his peers and young ones gain confidence in mathematics. His interest in public health earned him the chance to work with the best medical research center in Ghana, University of Ghana Noguchi Memorial Institute for Medical Research (UGNMIMR). He aims to earn his doctorate degree in Biostatistics at one of the best universities in United States.

Permanent Address: 15C TUC Nhyiaeso
Kumasi, Ghana 00233

This thesis was prepared by Richard Ntiamoah Kyei using L^AT_EX.