UNIVERSITÀ
DEGLI STUDI
DI PADOVA

Sede Amministrativa: Università degli Studi di Padova

Dipartimento di Scienze Statistiche
SCUOLA DI DOTTORATO DI RICERCA IN SCIENZE STATISTICHE
CICLO XXVI

QUANTILE INFERENCE IN GENOMIC STUDIES

Direttore della Scuola: Prof. Monica Chiogna

Supervisore: Prof. Monica Chiogna

Dottorando: Lorenzo Maragoni

July 31, 2014

To the Maragoni family,
who rocks.

Abstract

Genomic studies aim at identifying genes' location and function across the genome of living organisms of interest. In the last decades, this field of research has been the object of lively interest, motivated by the introduction of microarray and sequencing technologies, capable of providing huge amounts of data concerning several aspects of the genome. Statistical tools have proven to be necessary in this context, to support and sometimes lead biological investigation, impractical or impossible to be conducted over the whole set of data provided by the above mentioned technologies. In this Thesis, we introduce novel statistical tools for dealing with well-known problems in the genomic field, such as identifying differential expression in microarray data, and evaluating differential binding in the context of ChIP-Seq data. Our specific interest will be inference on quantiles, motivated by their interpretability, even for irregularly shaped distributions of the data, and by the fact that they allow to compare different aspects of the whole distribution of the data. We propose Studentized and pseudo-Studentized statistics, whose structure resembles closely that of a classic t-test, and evaluate their performances via simulated studies and application to real data.

Sommario

Gli studi di genomica hanno l'obiettivo di identificare posizione e funzione dei geni all'interno del genoma di organismi oggetto di interesse. Negli ultimi vent'anni, questo campo di ricerca è stato oggetto di vivace interesse, motivato dall'introduzione di microarray e tecnologie di sequenziamento, capaci di produrre enormi quantità di dati riguardanti diversi aspetti del genoma. In questo contesto, gli strumenti statistici si sono dimostrati necessari per supportare e in alcuni casi guidare la ricerca biologica, poco pratica o impossibile da condurre sull'intero insieme di dati prodotto dalle tecnologie di cui sopra. In questa Tesi, si introdurranno nuovi strumenti statistici per affrontare problemi noti nell'ambito genomico, come l'identificazione di geni differenzialmente espressi tramite dati di microarray, e l'analisi dei siti di legame nel contesto dei dati di ChIP-Seq. L'interesse specifico sarà l'inferenza sui quantili, motivato dalla loro interpretabilità, anche per distribuzioni dei dati dalle forme irregolari, e dal fatto che permettono di confrontare differenti aspetti della distribuzione dei dati. Si proporranno statistiche Studentizzate e pseudo-Studentizzate, la cui struttura richiama da vicino quella di un t-test classico, e si valuterà il loro comportamento attraverso studi di simulazione e applicazione su dati reali.

Acknowledgements

For having been working on fascinating and challenging topics, for having had the possibility to visit some of the best research institute of the world, and for much of what my overall PhD experience has been, I owe a lot to my Supervisor, prof. Monica Chiogna. Her ability to give advice, and find always new and motivating topics of research, has been of great inspiration for me. Moreover, her capacity of understanding and support through the very different times that have characterised my three+ years of PhD have been an invaluable help. For all of this I am very thankful to her.

I also wish to thank the PhD School and the Department of Statistics of the University of Padua, that have made this great learning adventure possible, including my wonderful colleagues, Akram, Darda, Erlis, Ivan, Luca, Roberta and Shireen, from whom I received encouragement and support over the whole time together.

I would like to thank prof. Matteo Bottai from Karolinska Institutet, who has mentored me through my early attempts at quantile inference, and by whose passion in researching and communicating statistics I have been really inspired, prof. Sandrine Dudoit from University of California at Berkeley, who has introduced me to how statistical methods meet application, and shared with me her fascination for the genomic world, and prof. Chiara Romualdi and her lab for the support received in the last months, for which I am very grateful.

I wish to thank Davide and Vale, for having done a great job making me feel at home in Berkeley. I have many wonderful memories of the time spent there, and I owe them a big part of it. For the same reason I wish to thank Paolo, with whom I have shared the best parts of my experience in Stockholm.

I wish to thank the people that are and have been in my life, that have been exceptionally supportive during this whole long journey, encouraging me through the hard times, being present to share with me the good ones, and being understanding for the ones I disappeared from social life. Elena, Emeline, Vanessa, Anna, Mauro, Luca, Bea, Giulia, Giorgia, Paola, Vera, and all the guys in Amor Vacui. You have been there for me, and have contributed a lot to make this PhD possible.

Finally, I wish to thank the awesome Maragoni family, who rocks.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Overview

Comparison of two (or more) independent groups has always been a fundamental problem in statistics, and in applications to the genomic field. It often arises in the setting where quantitative information is collected from two samples, and it can be thought of as a realisation of some underlying random variables, say X_A and X_B, with density functions p_{X_A} and p_{X_B}, respectively. In this setting, one might be interested in summarising the information in the samples by use of appropriate statistics, and in using them to test hypotheses on specific aspects of the underlying distributions. When doing this, one of the main points of concern is how to take into account variability of the measures. Since the introduction of the t-test by Student (1908), a multitude of methods has been developed for this task, including comparison of means, medians, variances, and ranks of the observations. Statistical hypotheses of interest are often expressed as:

$$H_0 : \theta(p_{X_A}) = \theta(p_{X_B}),$$

where $\theta(p_{X_A})$ and $\theta(p_{X_B})$ are parameters of interest.

It is apparent how this simple problem finds application to numerous different settings. In the last twenty years, the field of genomics has been a vast and ever-changing experimental area for the application of statistical tools for testing this kind of hypotheses. Since the development of technologies capable of providing huge amounts of data, such as microarray (Schena et al., 1995) and ChIP-Seq data

(Johnson et al., 2007), the pace at which new statistical tests have been released has been very intense. For example, Pan (2002) and Jeffery et al. (2006) give an idea of the impressive number and variety of statistical methods for identifying differential expression between two groups in microarray studies, while Wilbanks & Facciotti (2010a) report how in the three years that followed the first public release of ChIP-Seq data, at least 31 dedicated open source programs were developed and released. Of course, this complexity generates difficulties in finding shared guidelines to choose the appropriate tool for every problem of interest. However, statistical tests for group comparison have been a crucial help for fundamental discoveries in biological and medical research concerning, for example, the relation between gene expression and the development of different kinds of tumours (Lambert et al., 2013; Curry et al., 2013).

This Thesis aims to introduce novel quantile-based test statistics for group comparison, with an interest driven by application to genomic studies. The specific interest in inference on quantiles, firstly presented in a systematic way by Koenker & Bassett Jr (1978), is motivated by their interpretability, even for irregularly shaped distributions of the data, and by the fact that they allow to compare different aspects of the whole distribution of the data. We will initially propose Studentized statistics inspired by the work of Chung et al. (2013), and then move to the definition of novel pseudo-Studentized statistics, whose structure resembles closely that of a classic t-test. In Chapter 2, we introduce the test statistics, derive their most important properties, and discuss their main advantages and limitations over existing methods. We also provide simulation studies to test their performances in terms of control of type I error rate and power under a variety of statistical models. In Chapter 3 and Chapter 4, we provide application to differential expression analysis for microarray data and to the analysis of peaks obtained from ChIP-Seq data, respectively. Chapter 5 contains a general discussion of the results and some possible future directions to be explored.

1.2 Main contributions of the Thesis

With respect to the current literature, the main contributions of this Thesis may be summarized as follows:

1. analysis of the relationships existing between different parameters of transformed variables on different scales and their consequences on hypotheses testing;

2. introduction of novel Studentized test statistics for group comparison and evaluation of their performances from a theoretical and numerical point of view;

3. introduction of novel pseudo-Studentized test statistics, with simple structure, that might be used in everyday lab practice for investigating a variety of statistical hypotheses on the distribution of the populations, and assessment of their properties;

4. application of the above-mentioned tools to data from arrays, with the aim to identify differentially expressed genes between different biological conditions;

5. application of the same methods to data from sequencing, with the aim to compare the distribution of the reads in binding sites identified in different samples from ChIP-Seq data.

Chapter 2

Quantile-based Inference for Two Samples Comparison

In genomic studies, data are often collected on a continuous scale for two groups, and can be thought of as realizations of some underlying absolutely continuous random variables, X_A and X_B, with density functions p_{X_A} and p_{X_B}, respectively, i.e., $X_A \sim p_{X_A}$ and $X_B \sim p_{X_B}$. Before conducting statistical inference on some parameters of interest, data often undergo some transformation, say $g_A(\cdot)$ and $g_B(\cdot)$, to minimize the effect of experimental and/or technical variations and to achieve desirable distributional properties, such as normality or symmetry, on the transformed random variables $Y_A = g_A(X_A)$ and $Y_B = g_B(X_B)$. Usually, the transformation for both samples is the same, i.e., $Y_A = g(X_A)$ and $Y_B = g(X_B)$; depending on the context and on the nature of the data, a frequently used transformation for obtaining normality is the natural logarithm, as it is commonly accepted that "relative or absolute gene expression measurements are approximately normal on the log scale" (Tai et al., 2006). When normality is not necessarily desired, but symmetry is, a modification of the Box-Cox transformation is an option (Draper & Cox, 1969).

After data have been transformed, inference is conducted, often in the form of a test of statistical hypotheses. A common hypothesis of interest is equality of the means μ_{Y_K}, $K \in \{A, B\}$ of the transformed variables Y_K, i.e.,

$$H_0 : \mu_{Y_A} = \mu_{Y_B}. \tag{2.1}$$

When testing (2.1), we are implicitly testing a hypothesis on some parameters of the distribution of the original variables X_A and X_B, too, although, in general, these parameters do not have a direct interpretation. In fact, testing (2.1) is in general different from testing

$$H_0 : \mu_{X_A} = \mu_{X_B},$$

due to Jensen's inequality. However, under certain conditions, it is possible to retrieve a meaningful interpretation of (2.1) on the X scale, too. Starting from these simple considerations, in this Chapter we will try to understand under which conditions it is possible to establish an explicit relationship between testing hypotheses on the X and on the Y scale. In fact, a solid knowledge of what is the true statistical hypothesis being tested at any time of the data analysis process seems a crucial matter from the point of view of the investigators, who should be aware of the correct interpretation of the statistical procedure applied to the data. After having investigated these relationships, we will investigate possible ways to exploit them to define new test statistics that test directly hypotheses on the original scale X, without necessarily having to transform the data, while mantaining the desired interpretation.

In Section 2.1, we investigate the relations between parameters of interest for statistical models on different scales. In Section 2.2, we derive several novel Studentized statistics for group comparison, following the approach of Chung et al. (2013). In Sections 2.3, we discuss a key case in applications, i.e., a commonly encountered situation in genomic studies, that also allows for simplifications in the expression of the test statistics. For this key case, we assess the properties of the proposed test statistics in terms of type I error rate control and power for a variety of statistical models. In Section 2.4 we introduce a novel "pseudo-Studentized" statistic, i.e., an alternative test statistic completely based on quantiles on the X scale, whose structure recalls closely that of a classic t-test. Control of type I error rate and power properties are evaluated by means of simulation studies for this test statistic, too.

2.1 Relations between Parameters of Statistical Models on Different Scales

Our first aim is to explore whether it is possible to establish a connection between the means on the transformed random variables Y_A and Y_B, μ_{Y_A} and μ_{Y_B}, and the τ-quantiles, with $\tau \in (0,1)$, of the original variables, X_A and X_B, in the following denoted as $\xi_{X_A}(\tau)$ and $\xi_{X_B}(\tau)$. Our choice is motivated by several considerations:

a) the means on the Y scale are key quantities, as they are the parameters most frequently investigated in practice, and allow easy inference based on the sample mean estimators;

b) on the other hand, quantiles are very interpretable parameters, even for irregularly shaped distributions, and allow comparison of different aspects of the whole distribution of the data;

c) taking into account the previous two points, we wish to formulate procedures which are based on the sample mean estimators, and exploit the relations between parameters, to test hypotheses concerning quantiles of the distributions either on the transformed or on the original scale.

To explore the relation between μ_{Y_K} and $\xi_{X_K}(\tau)$, for $K \in \{A, B\}$, we start from the simple consideration that testing (2.1) is in general equivalent to testing

$$H_0 : h(\mu_{Y_A}) = h(\mu_{Y_B}), \tag{2.2}$$

for any strictly monotone function $h(\cdot)$ that does not contain other nuisance parameters, i.e., that is a reparameterization of μ_{Y_K}, for $K \in \{A, B\}$. If the distribution of Y_K is such that there exist a relationship between μ_{Y_K} and its τ-level quantile, denoted as $\xi_{Y_K}(\tau)$, i.e., if there exists a strictly monotone function $h(\cdot)$ such that

$$\xi_{Y_K}(\tau) = h(\mu_{Y_K}), \tag{2.3}$$

then (2.1) has a straightforward interpretation in terms of the quantiles of Y_K, for $K \in \{A, B\}$. In fact, when (2.3) holds, it is immediate to see that (2.2) is equivalent to

$$H_0 : \xi_{Y_A}(\tau) = \xi_{Y_B}(\tau). \tag{2.4}$$

If one now assumes that $Y_A = g(X_A)$ and $Y_B = g(X_B)$, and thanks to the fact that quantiles enjoy the property of invariance with respect to strictly monotone transformations[1], i.e., with our notation,

$$\xi_{Y_K}(\tau) = g(\xi_{X_K}(\tau)),$$

for $K \in \{A, B\}$, then (2.4) is equivalent to

$$H_0 : \xi_{X_A}(\tau) = \xi_{X_B}(\tau). \tag{2.5}$$

In other words, the hypothesis of equality of means on the Y scale is equivalent to the hypothesis of equality of the τ-quantiles on the X scale, and, by testing (2.1), one is implicitely testing (2.5), too.

Before moving on to the next steps, it is worth noting that the test of equality of the quantiles $\xi_{X_A}(\tau)$ and $\xi_{X_B}(\tau)$ could be addressed with the classic tools of quantile inference (Koenker & Bassett Jr, 1978). In particular, assume to have two independent simple random samples of size n_A and n_B for the two populations, respectively, and let $\hat{\xi}_U(\tau)$ be the sample quantile estimator for the generic random variable $U \sim p_U$. On recalling that the asymptotic distribution of $\hat{\xi}_U(\tau)$ satisfies

$$\sqrt{n_U}(\hat{\xi}_U(\tau) - \xi_U(\tau)) \dot{\sim} N\left(0, \sigma_U^2(\tau)\right),$$

where

$$\sigma_U^2(\tau) = \frac{\tau(1-\tau)}{p_U(\xi_U(\tau))^2} \tag{2.6}$$

and n_U is the sample size, Chung et al. (2013) propose a statistic for median comparison. This statistic is based on the quantity

$$\sqrt{n}(\hat{\xi}_{X_A}(1/2) - \hat{\xi}_{X_B}(1/2)),$$

where $n = n_A + n_B$ is the total sample size. The Authors define a Studentized statistic S as follows:

$$S = \frac{\sqrt{n}(\hat{\xi}_{X_A}(1/2) - \hat{\xi}_{X_B}(1/2))}{\sqrt{\frac{n}{n_A}\hat{\sigma}^2_{X_A}(1/2) + \frac{n}{n_B}\hat{\sigma}^2_{X_B}(1/2)}},$$

[1] Invariance of quantiles with respect to strictly monotone transformations is easily seen by considering the definition of ξ_{X_K} as the quantity that satisfies $P(X_K \leq \xi_{X_K}(\tau)) = \tau$. In fact, this equality can be rewritten in terms of the quantiles of Y_K as: $P(g^{-1}(Y_K) \leq \xi_{X_K}(\tau)) = P(Y_K \leq g(\xi_{X_K}(\tau))) = \tau$. Then, by applying the definition of $\xi_{Y_K}(\tau)$, one has: $\xi_{Y_K}(\tau) = g(\xi_{X_K}(\tau))$.

where the quantity $\hat{\sigma}^2_{X_K}(1/2)$ is a consistent estimator for the asymptotic variance of $\hat{\xi}_{X_K}(1/2)$, i.e., for the quantity $1/(4p_{X_K}(\xi_{X_K}(\tau))^2)$, for $K \in \{A, B\}$, and prove that its permutation distribution coincides asymptotically with the its true unconditional distribution under 2.5 while retaining exactness property for finite samples when $p_{X_A} = p_{X_B}$.

However, the use of quantile inference tools might be not immediate when n is low. In fact, estimating the variance of the median estimators (or, more in general, of τ-quantile estimators) can be a cumbersome task, as it includes estimating the density of the distribution of the variable to which it refers at the quantile of interest. This is usually done via kernel methods (Delaigle et al., 2011) or resampling methods such as the bootstrap (Efron, 1979) or the smoothed bootstrap (Hall et al., 1989), that for low sample sizes, or for extreme values of $\tau \in (0, 1)$, might fail to provide accurate estimates of the variances. Moreover, these methods can be computationally quite intensive, and the choice of the procedure to be applied requires some degree of subjectivity from the investigators. These considerations might make such methods not completely attractive for an everyday use in lab practice. Nevertheless, test statistics based on quantiles have also very appealing properties, most notably the fact that they do not require any (additional) hypothesis on the distribution of the variables of interest (in our case, X_A and X_B), in order to be used. In fact, the two populations might even have different probability laws, giving room for application to the wide set of practical problems where the assumption of normality or homoscedasticity of the data, for example, do not hold.

These considerations motivate us in moving further and defining statistical tools, with performances comparable to those of Chung et al. (2013), that test (2.5) without needing estimation of the density functions p_{X_K}. Relationships (2.1) through (2.5) suggest that it is possible to make use of mean estimators on the Y scale for this task. We will exploit this idea to define a consistent estimator $\tilde{\xi}_{X_K}(\tau)$ of $\xi_{X_K}(\tau)$, alternative to $\hat{\xi}_{X_K}(\tau)$, in the following way:

$$\tilde{\xi}_{X_K}(\tau) = g^{-1}(h(\overline{Y}_K)).$$

Then, the consistency of the sample mean estimator and the strict monotonicity of both $g(\cdot)$ and $h(\cdot)$ guarantee the consistency of $\tilde{\xi}_{X_K}(\tau)$ with respect to $\xi_{X_K}(\tau)$. In

fact,

$$\tilde{\xi}_{X_K}(\tau) \xrightarrow{p} g^{-1}(h(\mu_{Y_K})) = \xi_{X_K}(\tau),$$

where the $\xrightarrow{p}$ operator indicates convergence in probability. The asymptotic distribution of $\tilde{\xi}_{X_K}(\tau)$ can be easily recovered by an application of the delta method to the asymptotic distribution of $\overline{Y}_K$, therefore opening the possibility of defining novel test statistics for quantile comparison.

Remark: An alternative way to obtain a consistent estimator of the quantile of interest is available if the likelihood function of the model is known. In fact, consider a generic random variable $U \sim f_U(u;\theta)$, with distribution function $F_U(u;\theta)$ depending only on the parameter θ. The definition of its τ-quantile, for $\tau \in (0,1)$, which is

$$\xi_U(\tau) = F_U^{-1}(\tau;\theta),$$

is in fact a reparameterization of the parameter θ. The theory of likelihood assures that, under regularity conditions, the maximum likelihood estimator $\hat{\theta}$ is asymptotically normal, i.e.,

$$\hat{\theta} \dot{\sim} N(\theta, j(\hat{\theta})^{-1}),$$

where $j(\cdot)$ denotes the observed information. Then, a simple application of the delta method shows that

$$\hat{\xi}_U(\tau) = F_U^{-1}(\tau;\hat{\theta}) \dot{\sim} N\left(\xi_U(\tau), j(\hat{\xi}_U(\tau))^{-1} \frac{1}{(F'_U(F_U^{-1}(\tau;\theta)))^2}\right),$$

which can be rewritten as:

$$\hat{\xi}_U(\tau) = F_U^{-1}(\tau;\hat{\theta}) \dot{\sim} N\left(\xi_U(\tau), j(\hat{\xi}_U(\tau))^{-1} \frac{1}{f_U(\xi_U(\tau))^2}\right),$$

therefore obtaining an asymptotic result for the distribution of the sample quantile estimator, based on a likelihood result, and a consistent estimator for $\xi_U(\tau)$. However, also in this case the density of U at the quantile of interest should be estimated in order to have an estimator of the asymptotic variance.

The definition of the test statistics, derivation of their asymptotic behaviour and study of their properties will be object of the next Section. However, before moving to the construction of the test statistics, three key questions arise. A first point of

interest could be wondering how common relationships such as (2.3) are. Actually, there are many notable cases of distributions for which this kind of relationships holds, for given quantiles, as shown in the following examples.

Example 2.1.1. Assume $Y \sim \text{Exp}(\lambda)$. In this case, $\mu_Y = 1/\lambda$ and $\xi_Y(\tau) = -\log(1-\tau)/\lambda$, which implies that $\xi_Y(\tau) = -\mu_Y \log(1-\tau)$. In particular, for $\tau = 1/2$, it holds that $\xi_Y(1/2) = \mu_Y \log(2)$, i.e., that $h(t) = t\log(2)$. △

Example 2.1.2. Assume $Y \sim f_Y$, with f_Y symmetric. In this case, it is well-known that $\xi_Y(1/2) = \mu_Y$, i.e., that for $\tau = 1/2$, $h(\cdot)$ is the identity function. △

In other cases, an appropriate $h(\cdot)$ exists if some other model parameters can be assumed to be known, as shown in the following examples.

Example 2.1.3. Assume $Y \sim \text{N}(\mu, \sigma_0^2)$. In this case, $\mu_Y = \mu$, and the following relationship holds:

$$\xi_Y(\tau) = \mu_Y + \sigma_0 \Phi^{-1}(\tau),$$

where $\Phi(\cdot)$ is the distribution function of a standard Normal random variable. Then, if σ_0^2 was known, we would have that $h(t) = t + \sigma_0 \Phi^{-1}(\tau)$, for any fixed $\tau \in (0,1)$. △

Example 2.1.4. Assume $Y \sim \text{LN}(\mu, \sigma_0^2)$, i.e., Y is distributed as a Log-Normal random variable whose logarithmic transformation is a Normal random variable with mean μ and variance σ_0^2. In this case, $\mu_Y = e^{\mu}$ and $\xi_Y(1/2) = e^{\mu + \sigma_0^2/2}$. Then, the following relationship holds:

$$\xi_Y(1/2) = \mu_Y e^{\sigma_0^2/2}.$$

Then, if σ_0^2 was known, we would have that $h(t) = te^{\sigma_0^2/2}$, for $\tau = 1/2$. A strictly monotone relation between $\xi_Y(1/2)$ and σ_0^2 exists also if μ was known. △

Example 2.1.5. Assume $Y \sim \text{Logistic}(\mu, s)$. In this case, $\mu_Y = \mu$ and $\xi_Y(\tau) = \mu - s\log((1-\tau)/\tau)$. Then,the following relationship holds:

$$\xi_Y(\tau) = \xi_Y(\tau) = \mu_Y - s\log\left(\frac{1-\tau}{\tau}\right).$$

Then, if s was known, we would have that $h(t) = t - s\log((1-\tau)/\tau)$, for a fixed $\tau \in (0,1)$. Of course, for $\tau = 1/2$, the relationship reduces to the identity, being the Logistic distribution symmetric. $\triangle$

Example 2.1.6. Assume $Y \sim$ Uniform(a,b). In this case, $\mu_Y = (a+b)/2$ and $\xi_Y(\tau) = a + (b-a)\tau$. Then, the following relationship holds:

$$\xi_Y(\tau) = 2\mu_Y(1-\tau) - b(1-2\tau).$$

Then, if b was known, we would have that $h(t) = 2t(1-\tau) - b(1-2\tau)$, for $\tau \in (0,1)$. Of course, for $\tau = 1/2$, the relationship reduces to the identity, being the Uniform distribution symmetric. $\triangle$

Example 2.1.7. Assume $Y \sim$ Laplace(μ, b), with $b > 0$. In this case, $\mu_Y = \mu$ and $\xi_Y(\tau) = \mu + b\log(2\tau)$ if $\xi_Y(\tau) \geq \mu$, and $\xi_Y(\tau) = \mu + b\log(2(1-\tau))$ if $\xi_Y(\tau) \leq \mu$. Then, the following relationships hold:

$$\begin{aligned} \xi_Y(\tau) &= \mu_Y + b\log(2\tau) && \text{if } \xi_Y(\tau) \geq \mu_Y \\ \xi_Y(\tau) &= \mu_Y + b\log(2(1-\tau)) && \text{if } \xi_Y(\tau) \leq \mu_Y. \end{aligned}$$

Then, if b was known, we would have that:

$$\begin{aligned} h(t) &= t + b\log(2\tau) && \text{if } \xi_Y(\tau) \geq t \\ h(t) &= t + b\log(2(1-\tau)) && \text{if } \xi_Y(\tau) \leq t, \end{aligned}$$

for a fixed $\tau \in (0,1)$. Of course, for $\tau = 1/2$, the relationship reduces to the identity, being the Laplace distribution symmetric. $\triangle$

As a second point of interest, it is worth noting that relationship (2.3) could actually be different in the two populations, i.e.,

$$\begin{aligned} \xi_{Y_A}(\tau) &= h_A(\mu_{Y_A}) \\ \xi_{Y_B}(\tau) &= h_B(\mu_{Y_B}), \end{aligned}$$

with h_K strictly monotone for $K \in \{A, B\}$, but $h_A(\cdot) \neq h_B(\cdot)$. In this case, testing (2.5) starting from the quantities $\overline{Y}_A$ and $\overline{Y}_B$ might require additional steps. We will illustrate this case in the following example.

Example 2.1.8. Assume $X_A \sim \text{LGa}(1,1)$, $X_B \sim \text{LLaplace}(\log(2),1)$, where $\text{LGa}(\alpha,\lambda)$ indicates a random variable whose logarithmic transformation has a Gamma distribution with shape α and rate λ and $\text{LLaplace}(\mu,b)$ indicates a random variable whose logarithmic transformation has a Laplace distribution with mean μ and scale b. Let $\tau = 1/2$, and assume $Y_A = \log(X_A)$ and $Y_B = \log(X_B)$. In this case, it holds that $\xi_{Y_A}(1/2) = \xi_{Y_B}(1/2) = \log(2)$ (see examples 2.1.1 and 2.1.7), while $1 = \mu_{Y_A} \neq \mu_{Y_B} = \log(2)$. Therefore,

$$h_A(t) = t\log(2)$$
$$h_B(t) = t,$$

and testing

$$H_0 : \mu_{Y_A} = \mu_{Y_B}$$

is equivalent to testing

$$H_0 : h_A^{-1}(\xi_{Y_A}(1/2)) = h_B^{-1}(\xi_{Y_B}(1/2)),$$

i.e.,

$$H_0 : \frac{\xi_{Y_A}(1/2)}{\log(2)} = \xi_{Y_B}(1/2).$$

The relation linking (2.1) to (2.4) cannot be retrieved in this case. However, let us consider the transformed random variables:

$$Y_A^* = h_A(Y_A) = Y_A\log(2)$$
$$Y_B^* = h_B(Y_B) = Y_B.$$

It is easy to see that the sample mean estimator $\overline{Y}_K^*$, based on the transformed random variable Y_K^*, is consistent for $\xi_{Y_K}(1/2)$, for $K \in \{A, B\}$. In fact:

$$\overline{Y}_A^* = \overline{Y}_A\log(2) \xrightarrow{p} \mu_{Y_A}\log(2) = h_A^{-1}(\xi_{Y_A}(1/2))\log(2) = \xi_{Y_A}(1/2)$$
$$\overline{Y}_B^* = \overline{Y}_B \xrightarrow{p} \mu_{Y_B} = h_B^{-1}(\xi_{Y_B}(1/2)) = \xi_{Y_B}(1/2).$$

This suggests that $\overline{Y}_A$ and $\overline{Y}_B$ can still be used to test

$$H_0 : \xi_{Y_A}(1/2) = \xi_{Y_B}(1/2),$$

and implicitly

$$H_0 : \xi_{X_A}(1/2) = \xi_{X_B}(1/2),$$

provided that appropriate transformations are applied. Consider to this aim the transformed estimators $g^{-1}(\overline{Y}^*_A)$ and $g^{-1}(\overline{Y}^*_B)$, where $g(t) = \log(t)$. Then, it holds that:

$$g^{-1}(\overline{Y}^*_A) = e^{\overline{Y}^*_A} = e^{\overline{Y}_A \log(2)} \xrightarrow{p} e^{\mu_{Y_A}\log(2)} = e^{\xi_{Y_A}(1/2)} = \xi_{X_A}(1/2)$$
$$g^{-1}(\overline{Y}^*_B) = e^{\overline{Y}^*_B} = e^{\overline{Y}_B} \xrightarrow{p} e^{\mu_{Y_B}} = e^{\xi_{Y_B}(1/2)} = \xi_{X_B}(1/2),$$

and therefore $g^{-1}(\overline{Y}^*_K)$ is a consistent estimator for $\xi_{X_K}(1/2)$, and we might write $\tilde{\xi}_{X_K}(1/2) = g^{-1}(\overline{Y}^*_K)$, for $K \in \{A, B\}$. △

Of course, example 2.1.8 can be generalised to any situation where $h_A(\cdot) \neq h_B(\cdot)$, provided they are both strictly monotone, and to a generic quantile level $\tau \in (0, 1)$. Formally, if there exist $h_A(\cdot)$ and $h_B(\cdot)$ such that $\xi_{Y_A}(\tau) = h_A(\mu_{Y_A})$ and $\xi_{Y_B}(\tau) = h_B(\mu_{Y_B})$, then one can define the auxiliary random variables $Y^*_A = h_A(Y_A)$ and $Y^*_B = h_B(Y_B)$, and use the sample mean estimators based on the transformed variables, $\overline{Y}^*_A$ and $\overline{Y}^*_B$, to define the consistent estimators $\tilde{\xi}_{X_A}(\tau) = g^{-1}(\overline{Y}^*_A)$ and $\tilde{\xi}_{X_B}(\tau) = g^{-1}(\overline{Y}^*_B)$, which can be used for testing hypothesis (2.5).

The third observation refers to the transformation function $g(\cdot)$, and leads to a further generalisation. In fact, it is possible to retrieve estimators of the class $\tilde{\xi}_{X_K}(\tau)$ even when the initial transformations applied to X_A and X_B are different between the two groups, i.e., $g_A(\cdot)$ and $g_B(\cdot)$, provided that they are strictly monotone. With the help of the following example, we will see how to generalise the possibility of using $\overline{Y}_A$ and $\overline{Y}_B$ to test $H_0 : \xi_{X_A}(\tau) = \xi_{X_B}(\tau)$ under the more general setting:

$$Y_A = g_A(X_A)$$
$$Y_B = g_B(X_B),$$

where $g_A(\cdot) \neq g_B(\cdot)$, provided that both functions are still strictly monotone.

Example 2.1.9. Assume that $X_A \sim$ IG$(1, 1)$ and $X_B \sim$ LLaplace$(\log(2), 1)$, where $IG(\alpha, \gamma)$ indicates an Inverse Gamma distribution with shape α and rate λ. Let $\tau = 1/2$, and assume that $Y_A = 1/X_A$ and $Y_B = \log(X_B)$. In this case,

consider again the transformed variables (see example 2.1.8):

$$Y_A^* = h_A(Y_A)$$
$$Y_B^* = h_B(Y_B).$$

Then, the following results hold:

$$g_A^{-1}(\overline{Y}_A^*) = \frac{1}{\overline{Y}_A \log(2)} \xrightarrow{p} \frac{1}{\mu_{Y_A}\log(2)} = \frac{1}{\xi_{Y_A}(1/2)} = \xi_{X_A}(1/2)$$
$$g_B^{-1}(\overline{Y}_B^*) = e^{\overline{Y}_B} \xrightarrow{p} e^{\mu_{Y_B}} = e^{\xi_{Y_B}(1/2)} = \xi_{X_B}(1/2),$$

and we can define $\tilde{\xi}_{X_K}(\tau) = g_K^{-1}(\overline{Y}_K^*)$, for $K \in \{A, B\}$. △

Of course, example 2.1.9 can be generalised to any situation where $h_A(\cdot) \neq h_B(\cdot)$ and $g_A(\cdot) \neq g_B(\cdot)$, provided they are all strictly monotone functions, and to a generic quantile level $\tau \in (0, 1)$. In these cases, it is possible to define the consistent estimators $\tilde{\xi}_{X_A}(\tau) = g_A^{-1}(\overline{Y}_A^*)$ and $\tilde{\xi}_{X_B}(\tau) = g_B^{-1}(\overline{Y}_B^*)$, which can be used for testing hypothesis (2.5).

It seemed necessary to make these three considerations before defining the actual test statistics, to give an idea of the spectrum of situations where they could be applied. Even if some of these settings might occur quite rarely in practice, we feel that it was important to emphasise the possibility to generalise them to a wider framework of application. To summarise, it seems possible to conclude that, if g_K and h_K are strictly monotone functions, for $K \in \{A, B\}$, then it is possible to find suitable transformations

$$f_A = g_A^{-1} \circ h_A$$
$$f_B = g_B^{-1} \circ h_B,$$

where the $\circ$ operator denotes composition of functions, that allow to exploit $\overline{Y}_A$ and $\overline{Y}_B$ to test (2.5). The key result concerning consistency of the estimators can be expressed as:

$$\tilde{\xi}_{X_K}(\tau) = f_K(\overline{Y}_K) \xrightarrow{p} g_K^{-1}(h_K(\mu_{Y_K})) = g_K^{-1}(\xi_{Y_K}(\tau))) = \xi_{X_K}(\tau). \qquad (2.7)$$

The next Section will formalize the results obtained so far to develop novel test statistics for inference on $\xi_{X_K}(\tau)$, for $K \in \{A, B\}$, and give the main results concerning their properties and asymptotic behaviour.

2.2 Studentized Statistics

Our aim in this Section is to test $H_0 : \xi_{X_A}(\tau) = \xi_{X_B}(\tau)$, starting from the sample means $\overline{Y}_A$ and $\overline{Y}_B$. Let $X_{K1}, \ldots, X_{Kn_K}$ be two simple random samples of size n_K, independent from each other, from the density function p_{X_K}, having quantiles $\xi_{X_K}(\tau)$, $\tau \in (0,1)$, for $K \in \{A, B\}$. Moreover, let $n = n_A + n_B$ be the total sample size. In the following, we will assume that all the transformations applied to the quantities involved are such that the resulting random variables have finite variances. Let $Y_{K1}, \ldots, Y_{Kn_K}$ be the transformed samples on the Y scale, i.e., $Y_{Ki} = g_K(X_{Ki})$, with density function p_{Y_K}, mean μ_{Y_K} and variance $\sigma^2_{Y_K} < \infty$, for $i = 1, \ldots, n_K$ and $K \in \{A, B\}$. Let $\xi_{Y_K}(\tau) = h_K(\mu_{Y_K})$, where $h_K(\cdot)$ is a reparameterization of μ_{Y_K}. Let $\overline{Y}_K = \frac{1}{n_K}\sum_{i=1}^{n_K} Y_{Ki}$ be the sample mean estimator of μ_{Y_K} and $S^2_{Y_K} = \frac{1}{n_K - 1}\sum_{i=1}^{n_K}(Y_{Ki} - \overline{Y}_K)^2$ be the unbiased sample variance estimator of $\sigma^2_{Y_K}$, for $K \in \{A, B\}$.

To test (2.1), we could consider the natural estimator of μ_{Y_K}, i.e., the sample mean estimator $\overline{Y}_K$. A well-known result due to the central limit theorem, i.e.,

$$\sqrt{n_K}(\overline{Y}_K - \mu_{Y_K}) \xrightarrow{d} N(0, \sigma^2_{Y_K}),$$

gives rise to the following approximation for the asymptotic distribution of $\overline{Y}_K$:

$$\overline{Y}_K \dot{\sim} N\left(\mu_{Y_K}, \frac{\sigma^2_{Y_K}}{n_K}\right). \tag{2.8}$$

However, we are interested in how we could use this result to make inference on the quantity $\xi_{X_K}(\tau)$, which is related to μ_{Y_K} through the following relationship:

$$\xi_{X_K}(\tau) = g_K^{-1}(h_K(\mu_{Y_K})). \tag{2.9}$$

Therefore, we aim to obtain a consistent estimator of $g_K^{-1}(h_K(\mu_{Y_K}))$.

For this task, let $f_K = g_K^{-1} \circ h_K$, which is strictly monotone as it is a composition of strictly monotone functions, and let $\tilde{\xi}_{X_K}(\tau) = f_K(\overline{Y}_K)$. On noting that

$$\tilde{\xi}_{X_K}(\tau) \xrightarrow{p} \xi_{X_K}(\tau),$$

as seen in (2.7), let

$$W = \tilde{\xi}_{X_A}(\tau) - \tilde{\xi}_{X_B}(\tau),$$

for which

$$W \xrightarrow{p} \xi_{X_A}(\tau) - \xi_{X_B}(\tau)$$

holds, thanks to the consistency of sample quantile estimators, $\tilde{\xi}_{X_K}(\tau)$. The asymptotic distribution of W descends immediately from (2.8) through the delta method. In fact,

$$f(\overline{Y}_K) \dot{\sim} N\left(\xi_{X_K}(\tau), \frac{\sigma^2_{Y_K}}{n_K} f'(\mu_{Y_K})^2\right),$$

and therefore,

$$W \dot{\sim} N(\xi_{X_A}(\tau) - \xi_{X_B}(\tau), V(W)), \tag{2.10}$$

with

$$V(W) = \frac{\sigma^2_{Y_A}}{n_A} f'_A(\mu_{Y_A})^2 + \frac{\sigma^2_{Y_B}}{n_B} f'_B(\mu_{Y_B})^2. \tag{2.11}$$

Result (2.10) gives rise to the approximately standard Normal Z random variable, defined as:

$$Z = \frac{W - (\xi_{X_A}(\tau) - \xi_{X_B}(\tau))}{\sqrt{V(W)}} \dot{\sim} N(0,1).$$

In order to be able to use Z for inference, we need an estimator for the variance of W, which is not known. For this purpose, we introduce the estimator $\hat{V}(W)$, defined as:

$$\hat{V}(W) = \frac{S^2_{Y_A}}{n_A} f'_A(\overline{Y}_A)^2 + \frac{S^2_{Y_B}}{n_B} f'_B(\overline{Y}_B)^2.$$

Since Y_K has mean and finite variance, we can recall that $\overline{Y}_K \xrightarrow{p} \mu_{Y_K}$ and $S^2_{Y_K} \xrightarrow{p} \sigma^2_{Y_K}$. We also note that, since $f_K(\cdot)$ is strictly monotone because it is a composition of strictly monotone functions, then it is continuous almost everywhere in its domain. Then, by means of the continuous mapping theorem (Mann & Wald, 1943), it holds that

$$f'(\overline{Y}_K)^2 \xrightarrow{p} f'(\mu_{Y_K})^2.$$

Therefore, since convergence in probability of the generic random variables U_1 and U_2 implies convergence in probability of the random vector (U_1, U_2), and the continuous mapping theorem applies to vectors, too, then we can write:

$$\frac{S^2_{Y_K}}{n_K} f'(\overline{Y}_K)^2 \xrightarrow{p} \frac{\sigma^2_{Y_K}}{n_K} f'(\mu_{Y_K})^2. \tag{2.12}$$

Application of the continuous mapping theorem to the sum of the quantities defined in (2.12) for $K \in \{A, B\}$, proves that $\hat{V}(W)$ is a consistent estimator for the true

variance of W. In fact, it holds that

$$\frac{S^2_{Y_A}}{n_A} f'_A(\overline{Y}_A)^2 + \frac{S^2_{Y_B}}{n_B} f'_B(\overline{Y}_B)^2 \xrightarrow{p} \frac{\sigma^2_{Y_A}}{n_A} f'_A(\mu_{Y_A})^2 + \frac{\sigma^2_{Y_B}}{n_B} f'_B(\mu_{Y_B})^2,$$

i.e.,

$$\hat{V}(W) \xrightarrow{p} V(W). \tag{2.13}$$

By applying the Slutsky's theorem, it is possible to show that the asymptotic distribution of the statistic T, with

$$T = \frac{W - (\xi_{X_A}(\tau) - \xi_{X_B}(\tau))}{\sqrt{\hat{V}(W)}} \tag{2.14}$$

is standard Normal, so that it is possible to use it to compute an approximate level of significance for testing (2.5).

Note that the expression (2.11) hides quantiles either on the Y or on the X scale. In fact, let us compute $f'_K(\mu_{Y_K})$:

$$f'_K(\mu_{Y_K}) = \frac{\partial}{\partial y} g_K^{-1}(h_K(y))\bigg|_{y=\mu_{Y_K}} = \frac{h'_K(\mu_{Y_K})}{g'_K(g_K^{-1}(h_K(\mu_{Y_K})))}. \tag{2.15}$$

This expression can be written in terms of $\xi_{Y_K}(\tau)$ as:

$$f'_K(\mu_{Y_K}) = \frac{h'_K(h_K^{-1}(\xi_{Y_K}(\tau)))}{g'_K(g_K^{-1}(\xi_{Y_K}(\tau)))}, \tag{2.16}$$

or in terms of $\xi_{X_K}(\tau)$ as:

$$f'_K(\mu_{Y_K}) = \frac{h'_K(h_K^{-1}(g(\xi_{X_K}(\tau))))}{g'_K(\xi_{X_K}(\tau))}, \tag{2.17}$$

where the invertibility and differentiability of $h_K(\cdot)$ and $g_K(\cdot)$ is guaranteed by their strict monotonicity. The above equalities introduce the possibility of defining different estimators for the variance of W, such as, for example,

$$\hat{V}(W) = \sum_{K\in\{A,B\}} \frac{S^2_{Y_K}}{n_K} \left(\frac{h'_K(h_K^{-1}(\overline{Y}_K))}{g'_K(g_K^{-1}(\overline{Y}_K))} \right)^2,$$

which is based on (2.15), or

$$\hat{V}(W) = \sum_{K\in\{A,B\}} \frac{S^2_{Y_K}}{n_K} \left(\frac{h'_K(h_K^{-1}(\hat{\xi}_{Y_K}(\tau)))}{g'_K(g_K^{-1}(\hat{\xi}_{Y_K}(\tau)))} \right)^2,$$

which is based on (2.16), or

$$\hat{V}(W) = \sum_{K\in\{A,B\}} \frac{S^2_{Y_K}}{n_K} \left(\frac{h'_K(h_K^{-1}(g(\hat{\xi}_{X_K}(\tau))))}{g'_K(\hat{\xi}_{X_K}(\tau))} \right)^2,$$

which is based on (2.17). Plugging either of these expressions in (2.14) does not affect the asymptotic distribution of T, which remains standard Normal, and can be used to define a pivot for testing (2.5).

It is worth noting that the test statistic T introduced above satisfies the conditions of Theorem 2.2. in Chung et al. (2013), with $\hat{V}(W)$ replaced by any of the consistent estimators for the variance of W that we have defined, and therefore its permutation distribution coincides with its asymptotic counterpart for sufficiently large sample sizes under (2.5), while retaining exactness property for finite samples when $p_{X_A} = p_{X_B}$. For the reader's convenience, Theorem 2.2. in Chung et al. (2013) is reported in the Appendix of the present Chapter, along with a proof of its applicability to T. In the next Section, we will introduce a special case, which is frequently encountered in practice, which also allow for simplification of the expressions introduced above.

2.3 A Key Case

Consider the case $g_A(\cdot) = g_B(\cdot) = g(\cdot)$ and $\xi_{Y_K}(\tau) = \mu_{Y_K}$ for $K \in \{A, B\}$, i.e., $h_A(\cdot) = h_B(\cdot) = h(\cdot)$ is the identity function. In other words, consider the case where it is possible to find a common transformation $g(\cdot)$ for which $\xi_{Y_K}(\tau) = \mu_{Y_K}$ for $K \in \{A, B\}$. This situation is common in applications, often with $\tau = 1/2$, and allows to construct test statistics with an easy structure, based on results of the previous Section. Some examples of such cases are listed below, chosen for comparison with Chung et al. (2013). The models are chosen in such a way that, although the two population do not share the same distribution, in all cases the relationship $\mu_{Y_K} = \xi_{Y_K}(1/2) = e^{\xi_{X_K}(1/2)}$ holds for $K \in \{A, B\}$.

Example 2.3.1. Let $X_A \sim LN(\mu_A, \sigma^2_A)$, $X_B \sim LN(\mu_B, \sigma^2_B)$; $Y_A = \log(X_A)$ and $Y_B = \log(X_B)$ (see example 2.1.4). Then $\mu_{Y_K} = e^{\xi_{X_K}(1/2)}$, for $K \in \{A, B\}$.

△

Example 2.3.2. Let $X_A \sim LN(\mu_A, \sigma_A^2)$, $X_B \sim Lt(\mu_B, d)$; $Y_A = \log(X_A)$ and $Y_B = \log(X_B)$, where $Lt(\mu_B, d)$ indicates a random variable whose logarithmic transformation has a Student's t distribution with non-centrality parameter μ_B and d degrees of freedom. Then $\mu_{Y_K} = e^{\xi_{X_K}(1/2)}$, for $K \in \{A, B\}$. △

Example 2.3.3. Let $X_A \sim LLog(\mu, s)$, $X_B \sim LU(a, b)$; $Y_A = \log(X_A)$ and $Y_B = \log(X_B)$, where $LLog(\mu, s)$ indicates a random variable whose logarithmic transformation has a Logistic distribution with location μ and scale s, and $LU(a, b)$ indicates a random variable whose logarithmic transformation has a Uniform distribution on the interval (a, b). Then $\mu_{Y_K} = e^{\xi_{X_K}(1/2)}$, for $K \in \{A, B\}$.△

Example 2.3.4. Let $X_A \sim LGa(\alpha, \lambda)$, $X_B \sim LLap(\mu, b)$; $Y_A = \log(X_A)$ and $Y_B = \log(X_B)$ (see example 2.1.8). Then $\mu_{Y_K} = e^{\xi_{X_K}(1/2)}$, for $K \in \{A, B\}$. △

In these cases, it is easy to see that $f_A(\cdot) = f_B(\cdot) = f(\cdot) = g^{-1}(\cdot)$; $f'(\cdot) = 1/(g' \circ g^{-1})(\cdot)$, and some useful simplifications of the results seen so far occur. In fact, if $\xi_{Y_K}(\tau) = \mu_{Y_K}$, the variance of W becomes:

$$V(W) = \sum_{K \in \{A,B\}} \frac{\sigma_{Y_K}^2}{n_K} \frac{1}{g'(g^{-1}(\mu_{Y_K}))^2}. \tag{2.18}$$

or, equivalently,

$$V(W) = \sum_{K \in \{A,B\}} \frac{\sigma_{Y_K}^2}{n_K} \frac{1}{g'(g^{-1}(\xi_{Y_K}(\tau)))^2}, \tag{2.19}$$

or, equivalently,

$$V(W) = \sum_{K \in \{A,B\}} \frac{\sigma_{Y_K}^2}{n_K} \frac{1}{g'(\xi_{X_K}(\tau))^2}. \tag{2.20}$$

Therefore, examples of consistent estimators for the variance of W are:

$$\hat{V}_1(W) = \sum_{K \in \{A,B\}} \frac{S_{Y_K}^2}{n_K} \frac{1}{g'(g^{-1}(\overline{Y}_K))^2},$$

or

$$\hat{V}_2(W) = \sum_{K\in\{A,B\}} \frac{S^2_{Y_K}}{n_K} \frac{1}{g'(g^{-1}(\hat{\xi}_{Y_K}(\tau)))^2},$$

or

$$\hat{V}_3(W) = \sum_{K\in\{A,B\}} \frac{S^2_{Y_K}}{n_K} \frac{1}{g'(\hat{\xi}_{X_K}(\tau))^2}.$$

Although the condition $\xi_{Y_K}(\tau) = \mu_{Y_K}$, implying that the mean is equal to the τ-level quantile on the Y scale, might be difficult to find in practice, it assumes a more meaningful interpretation when $\tau = 1/2$. In fact, as we have recalled in example 2.1.2, this is the case of Y_K having a symmetric density function (although not necessarily Normal) for $K \in \{A, B\}$, which might be of special interest in applications.

If the null hypothesis (2.5) holds, it is possible to exploit it to further simplify the expression of the variance of W under H_0. Note that the simplification can be achieved without introducing any new hypotheses on the distribution for the two populations, such as, for example, homogeneity of the variances. Let μ_{Y_P} be the common mean of the two populations on the Y scale, i.e., $\mu_{Y_P} = \mu_{Y_A} = \mu_{Y_B}$, let $\xi_{Y_P}(\tau)$ be the common τ-level quantile of the two populations on the Y scale, i.e., $\xi_{Y_P}(\tau) = \xi_{Y_A}(\tau) = \xi_{Y_B}(\tau) = \mu_{Y_P}$, and let $\xi_{X_P}(\tau)$ be the common τ-level quantile of the two populations on the X scale, i.e., $\xi_{X_P}(\tau) = \xi_{X_A}(\tau) = \xi_{X_B}(\tau) = g^{-1}(\mu_{Y_P})$. Then, under (2.5), expression (2.18) becomes:

$$V(W) = \frac{1}{g'(g^{-1}(\mu_{Y_P}))^2} \left(\frac{\sigma^2_{Y_A}}{n_A} + \frac{\sigma^2_{Y_B}}{n_B} \right);$$

expression (2.19) becomes:

$$V(W) = \frac{1}{g'(g^{-1}(\xi_{Y_P}(\tau)))^2} \left(\frac{\sigma^2_{Y_A}}{n_A} + \frac{\sigma^2_{Y_B}}{n_B} \right);$$

and expression (2.20) becomes:

$$V(W) = \frac{1}{g'(\xi_{X_P}(\tau))^2} \left(\frac{\sigma^2_{Y_A}}{n_A} + \frac{\sigma^2_{Y_B}}{n_B} \right).$$

It is possible therefore to construct simpler estimators for the variance of W, valid under (2.5), from which to derive pivots for hypothesis testing. We stress again the fact that, also in this case, we do not require any additional assumption, such as, for example, homoscedasticity, on the distribution of the two populations. An example of such estimators is:

$$\hat{V}_4(W) = \frac{1}{g'(g^{-1}(\overline{Y}_P))^2}\left(\frac{S^2_{Y_A}}{n_A} + \frac{S^2_{Y_B}}{n_B}\right), \tag{2.21}$$

where

$$\overline{Y}_P = \frac{1}{n}\sum_{K\in\{A,B\}}\sum_{i=1}^{n_K} Y_{Ki},$$

i.e., $\overline{Y}_P$ is the mean estimator of the sample $Y = (Y_{A1}, \ldots, Y_{An_A}, Y_{B1}, \ldots, Y_{Bn_B})$, and a consistent estimator for the common mean μ_{Y_P}. When (2.5) is true, the weak law of large numbers guarantees that $\overline{Y}_P \xrightarrow{p} \mu_{Y_P}$, and the continuous mapping theorem proves convergence in probability, under (2.5), of $\hat{V}_4(W)$ to $V(W)$, similarly to what we have seen in (2.13). Other possible estimators for $V(W)$ under (2.5) are:

$$\hat{V}_5(W) = \frac{1}{g'(g^{-1}(\hat{\xi}_{Y_P}(\tau)))^2}\left(\frac{S^2_{Y_A}}{n_A} + \frac{S^2_{Y_B}}{n_B}\right), \tag{2.22}$$

where $\hat{\xi}_{Y_P}(\tau)$ is the sample τ-quantile of the pooled sample Y, as well as

$$\hat{V}_6(W) = \frac{1}{g'(\hat{\xi}_{X_P}(\tau))^2}\left(\frac{S^2_{Y_A}}{n_A} + \frac{S^2_{Y_B}}{n_B}\right), \tag{2.23}$$

where $\hat{\xi}_{X_P}(\tau)$ is the sample τ-quantile of the pooled sample $X = (X_{A1}, \ldots, X_{An_A}, X_{B1}, \ldots, X_{Bn_B})$. Consistency under (2.5) of the estimators $\hat{\xi}_{Y_P}(\tau)$ and $\hat{\xi}_{X_P}(\tau)$ for $\xi_{Y_P}(\tau)$ and $\xi_{X_P}(\tau)$, respectively, can be proven via the asymptotic representation of quantiles of pooled samples proposed by Liu & Yin (1994), which is based on the Bahadur representation of quantiles (Bahadur, 1966).

These alternative estimators of the variance allow to construct several test statistics, which we define under (2.5) as:

$$T_j = \frac{W}{\sqrt{\hat{V}_j(W)}}, \qquad j = 1, \ldots, 6,$$

and whose asymptotic normality under (2.5) is guaranteed by the Slutsky's theorem in analogy to what we have seen as a general result for T. Of course, different test statistics might behave differently in terms of type I error rate control and power relative to the sample sizes considered. Before evaluating these aspects via simulation studies in the next Subsection, we show that the statistics assume a quite simple form when the transformation function is $g(\cdot) = \log(\cdot)$, which is also a case frequently encountered in practice, possibly making their application more appealing for everyday lab practice.

Example 2.3.5. One of the most commonly applied transformations in practical applications is the natural logarithm. When $h(\cdot)$ is the identity function and $g(\cdot) = \log(\cdot)$, the equivalence (2.9) becomes:

$$\xi_{X_K}(\tau) = e^{\xi_{Y_K}(\tau)} = e^{\mu_{Y_K}},$$

and the quantity W is:

$$W = e^{\overline{Y}_A} - e^{\overline{Y}_B}.$$

Under (2.5), the variance estimators that we have seen throughout Section 2.3 become, respectively:

$$\begin{aligned}
\hat{V}_1(W) &= \sum_{K\in\{A,B\}} \frac{S^2_{Y_K}}{n_K} e^{2\overline{Y}_K}; \\
\hat{V}_2(W) &= \sum_{K\in\{A,B\}} \frac{S^2_{Y_K}}{n_K} e^{2\hat{\xi}_{Y_K}(\tau)}; \\
\hat{V}_3(W) &= \sum_{K\in\{A,B\}} \frac{S^2_{Y_K}}{n_K} \hat{\xi}^2_{X_K}(\tau); \\
\hat{V}_4(W) &= e^{2\overline{Y}_P} \left(\frac{S^2_{Y_A}}{n_A} + \frac{S^2_{Y_B}}{n_B} \right); \\
\hat{V}_5(W) &= e^{2\hat{\xi}_{Y_P}(\tau)} \left(\frac{S^2_{Y_A}}{n_A} + \frac{S^2_{Y_B}}{n_B} \right); \\
\hat{V}_6(W) &= \hat{\xi}^2_{X_P}(\tau) \left(\frac{S^2_{Y_A}}{n_A} + \frac{S^2_{Y_B}}{n_B} \right).
\end{aligned}$$

The corresponding test statistics from T_1 to T_6 all follow asymptotically a standard Normal distribution under (2.5). △

Example 2.3.6. Useful simplifications can be obtained with the logarithmic transformation even if some of the assumptions in this Section are not met. For example, assume $X_A \sim LLap(\log(2), 1)$, $X_B \sim LGa(1, 1)$ (see example 2.3.4). In this case, $h_A(t) = t\log(2)$, while $h_B(t) = t$. The quantity W becomes:

$$W = e^{\overline{Y}_A \log(2)} - e^{\overline{Y}_B},$$

and its asymptotic variance is

$$V(W) = \frac{\sigma^2_{Y_A}}{n_A} e^{2\mu_{Y_A}\log(2)} \log(2)^2 + \frac{\sigma^2_{Y_B}}{n_B} e^{2\mu_{Y_B}},$$

which can be rewritten as:

$$V(W) = \frac{\sigma^2_{Y_A}}{n_A} e^{2\xi_{Y_A}(1/2)} \log(2)^2 + \frac{\sigma^2_{Y_B}}{n_B} e^{2\xi_{Y_B}(1/2)}$$

or as:

$$V(W) = \frac{\sigma^2_{Y_A}}{n_A} \xi^2_{X_A}(1/2) \log(2)^2 + \frac{\sigma^2_{Y_B}}{n_B} \xi^2_{X_B}(1/2).$$

Possible estimators of $V(W)$, in analogy to the ones seen in example 2.3.5, are then:

$$\begin{aligned}
\hat{V}_1(W) &= \frac{S^2_{Y_A}}{n_A} e^{2\overline{Y}_A \log(2)} \log(2)^2 + \frac{S^2_{Y_B}}{n_B} e^{2\overline{Y}_B}; \\
\hat{V}_2(W) &= \frac{S^2_{Y_A}}{n_A} e^{2\hat{\xi}_{Y_A}(1/2)} \log(2)^2 + \frac{S^2_{Y_B}}{n_B} e^{2\hat{\xi}_{Y_B}(1/2)}; \\
\hat{V}_3(W) &= \frac{S^2_{Y_A}}{n_A} \hat{\xi}^2_{X_A}(1/2) \log(2)^2 + \frac{S^2_{Y_B}}{n_B} \hat{\xi}^2_{X_B}(1/2);
\end{aligned}$$

and, upon recalling that $Y^*_A = Y_A \log(2)$ and $Y^*_B = Y_B$, also:

$$\begin{aligned}
\hat{V}_4(W) &= e^{2\overline{Y}^*_P} \left(\frac{S^2_{Y_A}}{n_A} + \frac{S^2_{Y_B}}{n_B} \right); \\
\hat{V}_5(W) &= e^{2\hat{\xi}_{Y^*_P}(\tau)} \left(\frac{S^2_{Y_A}}{n_A} + \frac{S^2_{Y_B}}{n_B} \right); \\
\hat{V}_6(W) &= \hat{\xi}^2_{X^*_P}(\tau) \left(\frac{S^2_{Y_A}}{n_A} + \frac{S^2_{Y_B}}{n_B} \right),
\end{aligned}$$

where $\overline{Y}^*_P$ and $\hat{\xi}_{Y^*_P}(\tau)$ are computed over the pooled sample Y^* and $\hat{\xi}^2_{X^*_P}(\tau)$ is computed over the pooled sample X^*. $\triangle$

In the next Subsection, we will evaluate control of the type I error rate provided by test statistics from T_1 to T_6 via simulation studies, comparing the asymptotic type I error rates based on the standard Normal approximation, α_{asymp}, and the permutation type I error rates α_{perm}, to the nominal type I error rate, α, for a variety of statistical models of interest.

2.3.1 Simulation Studies: Type I Error Rate Control

In this Subsection, we assume the conditions of the key case and in particular of example 2.3.5, i.e., $h_A(\cdot) = h_B(\cdot) = h(\cdot)$ is the identity function and $g_A(\cdot) = g_B(\cdot) = g(\cdot) = \log(\cdot)$. Under these assumptions, we test (2.5) on simulated data, assuming $\tau = 1/2$. For comparison purposes, we choose simulation settings similar to those in Chung et al. (2013), to which we add a mixture model that introduces a small contamination for studying robustness, and a model that does not satisfy one of the conditions of the key case, namely $h_A(\cdot) = h_B(\cdot)$. The models, which are chosen in such a way that $Y_K = \log(X_K)$ have a well-known distribution, for $K \in \{A, B\}$, are the following ones:

1. $X_A \sim LN(0, 1)$, $X_B \sim LN(0, 5)$ (see example 2.3.1);
2. $X_A \sim LN(0, 1)$, $X_B \sim Lt(0, 5)$ (see example 2.3.2);
3. $X_A \sim LLog(0, 1)$, $X_B \sim LU(-10, 10)$ (see example 2.3.3).

The model with a small contamination is:

4. $X_A \sim LN(-0.003, 1)$, $X_B \sim MD(0, 1, -3, 1, 0.999)$, where $MD(\mu_1, \sigma_1^2, \mu_2, \sigma_2^2, \epsilon)$ indicates an exponential transformation of a mixture of two Normal random variables with means μ_1 and μ_2, variances σ_1^2 and σ_2^2, and weights $\epsilon \in (0, 1)$ and $1 - \epsilon$, respectively.

The model that does not satisfy the assumptions of the key case is:

5. $X_A \sim LLap(\log(2), 1)$, $X_B \sim LGa(1, 1)$ (see example 2.3.6).

Each simulation experiment is ran for different sample sizes, n_A and n_B. In each experiment, $m = 999$ permutations are used, except when $n_A = n_B = 5$; in that case, all available $m = 252$ permutations are used. The number of replicates in each experiment is 10000. Samples are generated under the null hypothesis (2.5).

We expect that the asymptotic type I error rate approaches the nominal one, which is chosen to be $\alpha = 0.05$, for sufficiently large sample sizes. We expect a similar behaviour from the permutation type I error rate, thanks to the result of Chung et al. (2013), although we do not expect the permutation result to be exact for finite sample, as $p_{X_A} \neq p_{X_B}$.

In the following, we display the results obtained from simulation studies. Each Table refers to one of the five models of interest above mentioned, and displays permutation and asymptotic type I error rates. Different test statistics are displayed by row, while every column refers to a different sample size.

Table 2.1 shows Monte Carlo results for the $X_A \sim LN(0,1)$ vs. $X_B \sim LN(0,5)$ case. It seems that all the test statistics achieve the nominal type I error rate for a sufficiently large sample size. Moreover, the asymptotic p-value performs in a comparable way to the permutation one for several statistics, even for moderate to low sample sizes. The best performances, in terms of smaller sample size needed to approximate α, are obtained by the test statistics that exploit the null hypothesis in estimating the variance of W, namely T_4, T_5 and T_6. It is worth noting that these tests provide exactly the same permutation p-value. An explanation of this phenomenon is provided in the final remarks of the Section.

		$n_A = n_B$				
		5	11	101	1001	5001
T_1	α_{perm}	0.0847	0.1063	0.0729	0.0576	0.0501
	α_{asymp}	0.0926	0.0929	0.0712	0.0588	0.0513
T_2	α_{perm}	0.0825	0.1114	0.0840	0.0622	0.0520
	α_{asymp}	0.1005	0.0904	0.0714	0.0596	0.0508
T_3	α_{perm}	0.0865	0.1122	0.0843	0.0627	0.0527
	α_{asymp}	0.1005	0.0904	0.0714	0.0596	0.0508
T_4	α_{perm}	0.0536	0.0542	0.0487	0.0504	0.0454
	α_{asymp}	0.0990	0.0716	0.0520	0.0525	0.0479
T_5	α_{perm}	0.0536	0.0542	0.0487	0.0504	0.0454
	α_{asymp}	0.0647	0.0452	0.0392	0.0475	0.0451
T_6	α_{perm}	0.0536	0.0542	0.0487	0.0504	0.0454
	α_{asymp}	0.0466	0.0425	0.0486	0.0521	0.0478

Table 2.1: Monte Carlo permutation and asymptotic type I error rates for the $LN(0,1)$ vs. $LN(0,5)$ case under $\xi_{X_A}(1/2) = \xi_{X_B}(1/2)$. Nominal type I error rate $\alpha = 0.05$.

Table 2.2 shows Monte Carlo results for the $X_A \sim LN(0,1)$ vs. $X_B \sim Lt(0,5)$ case. Results seem better than in the previous scenario for all the statistics, in the sense that both the permutation and the asymptotic p-value are closer to the nominal one at lower sample sizes than in the previous setting. This might be due to the fact that the distributions of X_A and X_B are actually more similar in this case, being a $Lt(0,5)$ distribution closer to a $LN(0,1)$ than a $LN(0,5)$ is. This fact might enhance both the performance of the asymptotic type I error rate (which approaches α for lower sample sizes) and those of the permutation type I error rate (which is more likely to be conditionally "almost" exact for finite samples if the two distributions are very similar). Again, T_4, T_5 and T_6 perform the best, but the difference to the other test statistics is greatly reduced.

		$n_A = n_B$				
		5	11	101	1001	5001
T_1	α_{perm}	0.0546	0.0445	0.0544	0.0501	0.0459
	α_{asymp}	0.0543	0.0526	0.0544	0.0510	0.0472
T_2	α_{perm}	0.0427	0.0544	0.0537	0.0503	0.0458
	α_{asymp}	0.0567	0.0538	0.0521	0.0504	0.0461
T_3	α_{perm}	0.0550	0.0460	0.0539	0.0509	0.0459
	α_{asymp}	0.0538	0.0567	0.0521	0.0504	0.0461
T_4	α_{perm}	0.0427	0.0495	0.0503	0.0482	0.0453
	α_{asymp}	0.0739	0.0636	0.0528	0.0496	0.0465
T_5	α_{perm}	0.0427	0.0495	0.0503	0.0482	0.0453
	α_{asymp}	0.0689	0.0611	0.0489	0.0480	0.0458
T_6	α_{perm}	0.0427	0.0495	0.0503	0.0482	0.0453
	α_{asymp}	0.0492	0.0505	0.0514	0.0495	0.0465

Table 2.2: Monte Carlo permutation and asymptotic type I error rates for the $LN(0,1)$ vs. $Lt(0,5)$ case under $\xi_{X_A}(1/2) = \xi_{X_B}(1/2)$. Nominal type I error rate $\alpha = 0.05$.

Table 2.3 shows Monte Carlo results for the $X_A \sim LLog(0,1)$ vs. $X_B \sim LU(-10,10)$ case. The distributions are quite different, and, as a consequence, more marked differences between the test statistics seem to emerge. Test statistics T_4, T_5 and T_6 perform much better than the others, that possibly require larger sample sizes than those considered to approach the nominal level, although the type I error rates are closer to the nominal one as the sample sizes increase. Test statistic T_6 seems to have a conservative behaviour (low type I error rate) for low sample sizes, which might be due to the increased differences between the distributions.

		$n_A = n_B$				
		5	11	101	1001	5001
T_1	α_{perm}	0.0847	0.1063	0.0729	0.0691	0.0611
	α_{asymp}	0.0926	0.0929	0.0712	0.0614	0.0614
T_2	α_{perm}	0.1236	0.2038	0.1753	0.0991	0.0747
	α_{asymp}	0.1255	0.1312	0.1496	0.0891	0.0713
T_3	α_{perm}	0.0865	0.1122	0.0843	0.1002	0.0755
	α_{asymp}	0.1005	0.0904	0.0714	0.0891	0.0713
T_4	α_{perm}	0.0573	0.0511	0.0535	0.0480	0.0521
	α_{asymp}	0.1500	0.0999	0.0625	0.0488	0.0529
T_5	α_{perm}	0.0573	0.0511	0.0535	0.0480	0.0521
	α_{asymp}	0.0716	0.0360	0.0265	0.0348	0.0475
T_6	α_{perm}	0.0573	0.0511	0.0535	0.0480	0.0521
	α_{asymp}	0.0054	0.0012	0.0419	0.0465	0.0522

Table 2.3: Monte Carlo permutation and asymptotic type I error rates for the $LLog(0,1)$ vs. $LU(-10,10)$ case under $\xi_{X_A}(1/2) = \xi_{X_B}(1/2)$. Nominal type I error rate $\alpha = 0.05$.

Table 2.4 shows Monte Carlo results for the $X_A \sim LN(-0.003, 1)$ vs. $X_B \sim MD(0, 1, -3, 1, 0.999)$ case, i.e., for the small contamination scenario. As expected, since the contamination is very small and samples of size $n_K < 500$ are likely to have been unaffected by it, all of the considered statistics, with some minor differences, seem to perform well, proving to be robust to this specific kind of small deviation from the null hypothesis.

		$n_A = n_B$				
		5	11	101	1001	5001
T_1	α_{perm}	0.0467	0.0462	0.0506	0.0490	0.0487
	α_{asymp}	0.0494	0.0407	0.0498	0.0498	0.0497
T_2	α_{perm}	0.0467	0.0490	0.0501	0.0488	0.0485
	α_{asymp}	0.0597	0.0478	0.0497	0.0495	0.0493
T_3	α_{perm}	0.0506	0.0495	0.0505	0.0491	0.0493
	α_{asymp}	0.0597	0.0478	0.0497	0.0495	0.0493
T_4	α_{perm}	0.0460	0.0457	0.0501	0.0490	0.0489
	α_{asymp}	0.0797	0.0596	0.0519	0.0501	0.0497
T_5	α_{perm}	0.0460	0.0457	0.0501	0.0490	0.0489
	α_{asymp}	0.0826	0.0652	0.0510	0.0500	0.0495
T_6	α_{perm}	0.0460	0.0457	0.0501	0.0490	0.0489
	α_{asymp}	0.0589	0.0489	0.0510	0.0499	0.0497

Table 2.4: Monte Carlo permutation and asymptotic type I error rates for the $LN(-0.003, 1)$ vs. $MD(0, 1, -3, 1, 0.999)$ case under $\xi_{X_A}(1/2) = \xi_{X_B}(1/2)$. Nominal type I error rate $\alpha = 0.05$.

Table 2.5 shows Monte Carlo results for the $X_A \sim LLap(\log(2), 1)$ vs. $X_B \sim LGa(1, 1)$ case. The majority of the statistics seem to provide good control of the type I error rate, even for moderate sample sizes, being the only exceptions test statistics T_2 and T_3, which show a conservative behaviour for large sample sizes.

		$n_A = n_B$				
		5	11	101	1001	5001
T_1	α_{perm}	0.0720	0.0574	0.0462	0.0481	0.0448
	α_{asymp}	0.0380	0.0287	0.0426	0.0491	0.0450
T_2	α_{perm}	0.0670	0.0485	0.0337	0.0362	0.0345
	α_{asymp}	0.0805	0.0704	0.0649	0.0664	0.0625
T_3	α_{perm}	0.0700	0.0493	0.0345	0.0367	0.0360
	α_{asymp}	0.0805	0.0704	0.0649	0.0664	0.0625
T_4	α_{perm}	0.0699	0.0642	0.0545	0.0509	0.0458
	α_{asymp}	0.1008	0.0785	0.0564	0.0524	0.0468
T_5	α_{perm}	0.0699	0.0642	0.0545	0.0509	0.0458
	α_{asymp}	0.1489	0.1272	0.0924	0.0828	0.0780
T_6	α_{perm}	0.0699	0.0642	0.0545	0.0509	0.0458
	α_{asymp}	0.0748	0.0651	0.0545	0.0521	0.0468

Table 2.5: Monte Carlo permutation and asymptotic type I error rates for the $LLap(\log(2), 1)$ vs. $LGa(1, 1)$ case under $\xi_{X_A}(1/2) = \xi_{X_B}(1/2)$. Nominal type I error rate $\alpha = 0.05$.

Overall, it seems that for sufficiently large sample sizes both the permutation and the asymptotic type I error rates for most test statistics approach the nominal one for the proposed models. In more than one case, we remark that both the asymptotic and the permutation type I error rates approach the nominal level even for moderate to low sample sizes. The test statistics that perform most consistently well across models are T_4, T_5 and T_6, i.e., those that exploit (2.5) to estimate the variance of W. The improvement over test statistics T_1, T_2 and T_3 is particularly evident for the cases where the two distributions are more different from each other.

Remark: As mentioned above, the permutation behaviour of T_4, T_5 and T_6 is exactly the same for the three test statistics. We recall that, under (2.5), variance estimators for W are expressed in (2.21), (2.22) and (2.23), respectively. Both W and the term

$$\frac{S^2_{Y_A}}{n_A} + \frac{S^2_{Y_B}}{n_B}$$

take in general different values for each permutation, but their values are the same for the three test statistics considered. The other term of T_j, which is $\overline{Y}_P$, $\hat{\xi}_{Y_P}(\tau)$ and $\hat{\xi}_{X_P}(\tau)$ for $j = 4, 5, 6$, respectively, is, instead, constant across permutations, as it is computed on the pooled sample Y for T_4 and T_5 and on the pooled sample X for T_6. Therefore, the permutation p-value is exactly the same for the three statistics considered.

In the following Subsection, the power of the proposed test statistics will be evaluated via simulation for the same models.

2.3.2 Simulation Studies: Power

In order to assess the power of the proposed statistics, we choose the same generating models proposed in the previous Subsection (with the exception of the contaminated model), but we assume the two populations to have different medians on the X scale, i.e., we simulate data under the alternative hypothesis. The shift in the medians is chosen in such a way that the behaviour of the estimators can be compared in terms of power for moderate to low sample sizes. In the model description, recall that the parameters are often referred to the distribution after the logarithmic transformation; this should be kept into account in order to have a correct idea of the scale of the shift. For example, if $X_A \sim LN(0, 1)$ and $X_B \sim LN(-1, 5)$, then $\mu_{Y_A} - \mu_{Y_B} = \xi_{Y_A}(1/2) - \xi_{Y_B}(1/2) = 1$, but $\xi_{X_A}(1/2) - \xi_{X_B}(1/2) = 1 - e^{-1}$. The models are chosen as follows:

1. $X_A \sim LN(0, 1)$, $X_B \sim LN(-1, 5)$;
2. $X_A \sim LN(0, 1)$, $X_B \sim Lt(-1, 5)$;
3. $X_A \sim LN(0, 1)$, $X_B \sim Lt(-0.5, 5)$;
4. $X_A \sim LLog(0, 1)$, $X_B \sim LU(-10, 5)$;

5. $X_A \sim LLap(2,1)$, $X_B \sim LGa(1,1)$.

Each simulation experiment is ran for different sample sizes, n_A and n_B. In each experiment, $m = 999$ permutations are used, except when $n_A = n_B = 5$; in that case, all available $m = 252$ permutations are used. The number of replicates in each experiment is 10000. We expect the power to increase with increasing sample sizes, and to approach 1 if the test is consistent.

In the following, we display the results obtained from simulation studies. Each Table refers to one of the five models of interest above mentioned, and displays permutation and asymptotic power. Different test statistics are displayed by row, while every column refers to a different sample size.

Table 2.6 shows Monte Carlo results for the $X_A \sim LN(0,1)$ vs. $X_B \sim LN(-1,5)$ case. While at large sample sizes all of the test statistics seem to approach 1, some differences can be seen at smaller sample sizes. It seems that the test statistics T_4, T_5 and T_6, that controlled better than the other ones the type I error rate, are less powerful for the settings with sample sizes lower than 51. This result could be due to the fact that these test statistics exploit (2.5), which, in these setting, is not true, to estimate the variance of W.

		$n_A = n_B$				
		5	11	21	51	101
T_1	α_{perm}	0.2868	0.5320	0.7194	0.9365	0.9953
	α_{asymp}	0.2973	0.4979	0.6924	0.9296	0.9954
T_2	α_{perm}	0.2560	0.5315	0.7277	0.9416	0.9960
	α_{asymp}	0.3003	0.4662	0.6688	0.9214	0.9480
T_3	α_{perm}	0.2644	0.5326	0.7286	0.9417	0.9961
	α_{asymp}	0.3003	0.4662	0.6688	0.9214	0.9948
T_4	α_{perm}	0.2122	0.3713	0.5847	0.8908	0.9916
	α_{asymp}	0.3290	0.4380	0.6236	0.9009	0.9923
T_5	α_{perm}	0.2122	0.3713	0.5847	0.8908	0.9916
	α_{asymp}	0.2171	0.3033	0.4856	0.8516	0.9884
T_6	α_{perm}	0.2122	0.3713	0.5847	0.8908	0.9916
	α_{asymp}	0.1606	0.3167	0.5544	0.8888	0.9918

Table 2.6: Monte Carlo permutation and asymptotic power for the $LN(0,1)$ vs. $LN(-1,5)$ case.

Table 2.7 shows Monte Carlo results for the $X_A \sim LN(0,1)$ vs. $X_B \sim Lt(-1,5)$ case. The power is uniformly higher for all of the tests than in the previous case, and is above 0.90 even for $n_A = n_B = 21$. This could be due to the lesser variability of a $Lt(-1,5)$ distribution with respect to a $LN(-1,5)$ distribution, which makes it more clearly separated from a $LN(0,1)$ distribution.

		$n_A = n_B$				
		5	11	21	51	101
T_1	α_{perm}	0.3975	0.7621	0.9537	0.9996	1.0000
	α_{asymp}	0.3891	0.7462	0.9501	0.9996	1.0000
T_2	α_{perm}	0.3393	0.7078	0.9414	0.9996	1.0000
	α_{asymp}	0.3840	0.6858	0.9313	0.9996	1.0000
T_3	α_{perm}	0.3494	0.7088	0.9417	0.9996	1.0000
	α_{asymp}	0.3840	0.6858	0.9313	0.9993	1.0000
T_4	α_{perm}	0.4124	0.7555	0.9472	0.9950	1.0000
	α_{asymp}	0.5492	0.7974	0.9543	0.9950	1.0000
T_5	α_{perm}	0.4124	0.7555	0.9472	0.9950	1.0000
	α_{asymp}	0.4860	0.7418	0.9378	0.9995	1.0000
T_6	α_{perm}	0.4124	0.7555	0.9472	0.9995	1.0000
	α_{asymp}	0.3991	0.7379	0.9447	0.9995	1.0000

Table 2.7: Monte Carlo permutation and asymptotic power for the $LN(0,1)$ vs. $Lt(-1,5)$ case.

Table 2.8 shows Monte Carlo results for the $X_A \sim LN(0,1)$ vs. $X_B \sim Lt(-0.5,5)$ case. This setting is reported to compare how power changes with respect to the previous case when a smaller shift in the median difference occurs. As expected, the power is consistently lower across all of the test statistics, but it is above 0.8 for all of the test statistics for sample sizes equal to or larger than 51.

		$n_A = n_B$				
		5	11	21	51	101
T_1	α_{perm}	0.1653	0.3319	0.5281	0.8347	0.9794
	α_{asymp}	0.1835	0.3225	0.5176	0.8319	0.9795
T_2	α_{perm}	0.1513	0.3117	0.5163	0.8311	0.9787
	α_{asymp}	0.1864	0.2933	0.4840	0.8191	0.9784
T_3	α_{perm}	0.1582	0.3129	0.5186	0.8316	0.9788
	α_{asymp}	0.1864	0.2933	0.4840	0.8191	0.9784
T_4	α_{perm}	0.1686	0.3152	0.5050	0.8190	0.9769
	α_{asymp}	0.2652	0.3663	0.5313	0.8308	0.9783
T_5	α_{perm}	0.1686	0.3152	0.5050	0.8190	0.9769
	α_{asymp}	0.2359	0.3299	0.4965	0.8140	0.9771
T_6	α_{perm}	0.1686	0.3152	0.5050	0.8190	0.9769
	α_{asymp}	0.1794	0.3079	0.5007	0.8227	0.9778

Table 2.8: Monte Carlo permutation and asymptotic power for the $LN(0,1)$ vs. $Lt(-0.5,5)$ case.

Table 2.9 shows Monte Carlo results for the $X_A \sim LLog(0,1)$ vs. $X_B \sim LU(-10,5)$ case. In this case, quite different results occur for different test statistics. In fact, as in the first setting, it seems that T_1, T_2 and T_3 perform better than T_4, T_5 and T_6, with T_1 achieving the highest power for all sample sizes. However, all of the test statistics seem to achieve a reasonably good results (permutation power over 0.7) for sample sizes equal to or larger than 21.

		$n_A = n_B$				
		5	11	21	51	101
T_1	α_{perm}	0.4410	0.7609	0.9116	0.9959	0.9999
	α_{asymp}	0.1994	0.4539	0.8257	0.9940	0.9999
T_2	α_{perm}	0.3316	0.6909	0.8803	0.9922	0.9999
	α_{asymp}	0.2729	0.4255	0.6912	0.9668	0.9989
T_3	α_{perm}	0.3316	0.6909	0.8803	0.9922	0.9999
	α_{asymp}	0.2729	0.4255	0.6912	0.9668	0.9989
T_4	α_{perm}	0.2431	0.5032	0.7713	0.9847	0.9999
	α_{asymp}	0.4774	0.6422	0.8281	0.9889	0.9999
T_5	α_{perm}	0.2431	0.5032	0.7713	0.9847	0.9999
	α_{asymp}	0.2343	0.2842	0.4419	0.8445	0.9959
T_6	α_{perm}	0.2431	0.5032	0.7713	0.9847	0.9999
	α_{asymp}	0.0240	0.0523	0.4753	0.9758	0.9999

Table 2.9: Monte Carlo permutation and asymptotic power for the $LLog(0,1)$ vs. $LU(-10,5)$ case.

Table 2.10 shows Monte Carlo results for the $X_A \sim LLap(2,1)$ vs. $X_B \sim LGa(1,1)$ case. In this setting, it seems that exploiting the null hypothesis (2.5) when estimating the variance of W does not have a negative influence on the power of the test statistics. For moderate to low sample sizes it even seems to have a positive effect, which is apparently in opposition to results obtained for previous settings. Possibly, this might be due to the fact that in this setting variance estimation is more stable even when computed under the null hypothesis.

		$n_A = n_B$				
		5	11	21	51	101
T_1	α_{perm}	0.3939	0.5861	0.7854	0.9806	0.9998
	α_{asymp}	0.2340	0.4170	0.7080	0.9780	0.9998
T_2	α_{perm}	0.3333	0.4821	0.6541	0.9442	0.9992
	α_{asymp}	0.3382	0.4836	0.6924	0.9626	0.9996
T_3	α_{perm}	0.3333	0.4821	0.6541	0.9442	0.9992
	α_{asymp}	0.3382	0.4836	0.6924	0.9626	0.9996
T_4	α_{perm}	0.3840	0.6100	0.8113	0.9846	0.9999
	α_{asymp}	0.4851	0.6566	0.8302	0.9867	0.9998
T_5	α_{perm}	0.3840	0.6100	0.8113	0.9846	0.9999
	α_{asymp}	0.5430	0.7072	0.8653	0.9906	0.9999
T_6	α_{perm}	0.3840	0.6100	0.8113	0.9846	0.9999
	α_{asymp}	0.3839	0.5960	0.8071	0.9851	0.9998

Table 2.10: Monte Carlo permutation and asymptotic power for the $LLap(2,1)$ vs. $LGa(1,1)$ case.

Overall, it seems that all proposed test statistic show a reasonable power for all the considered settings, with the best results obtained by T_1, T_2 and T_3.

2.4 Pseudo-Studentized Statistics

In the Section 2.2, we have introduced Studentized statistics for testing hypotheses on the difference of quantiles of two populations. The main feature of these statistics is to be able to exploit the distributional properties of the sample mean estimator and relationships between parameters to test hypotheses on quantiles, therefore avoiding variance estimation of sample quantile estimators, which might produce unstable results for small sample sizes. However, with the help of a few approximations, the statistics above might assume a simpler form. In particular, our aim in this Section is to produce test statistics which are approximately equivalent to those of Section 2.2, but are entirely defined on the X scale, therefore making data transformation unnecessary. In this Section, we will introduce such statistics, which will be called "pseudo-Studentized" statistics, for testing (2.5), and will show how their structure recalls closely that of a classic t-test. The novel test statistics are defined under the conditions of the key case illustrated in Section 2.3. As a starting point, consider $\tau = 1/2$ and let us recall that, under the assumptions of Section 2.3, the variance of W can be written as

$$V(W) = \frac{\sigma^2_{Y_A}}{n_A} \frac{1}{g'(\xi_{X_A}(1/2))^2} + \frac{\sigma^2_{Y_B}}{n_B} \frac{1}{g'(\xi_{X_B}(1/2))^2}.$$

Our first step will be to try to express the quantity σ_{Y_K}, for $K \in \{A, B\}$ as a function of parameters on the X scale. We will show that, indeed, a simple relationship exists between σ_{Y_K} and the interquantile range of level $\psi_K \in (0, 1/2)$ on the X scale, i.e., the quantity $\xi_{X_K}(1-\psi_K) - \xi_{X_K}(\psi_K)$, for $K \in \{A, B\}$. Let U_K be the standardized version of Y_K, i.e.,

$$U_K = \frac{Y_K - \mu_{Y_K}}{\sigma_{Y_K}} \sim p_{U_K}(u),$$

so that $E(U_K) = 0$ and $V(U_K) = 1$. It is possible to show that linear relationship exists between the quantiles of Y_K and those of U_K. In fact, by definition of quantiles, it holds that:

$$\tau = P(U_K \leq \xi_{U_K}(\tau)) = P\left(\frac{Y_K - \mu_{Y_K}}{\sigma_{Y_K}} \leq \xi_{U_K}(\tau)\right).$$

This equivalence yields:

$$P\left(Y_K \leq \mu_{Y_K} + \sigma_{Y_K}\xi_{U_K}(\tau)\right) = \tau,$$

i.e.,

$$\xi_{Y_K}(\tau) = \mu_{Y_K} + \sigma_{Y_K}\xi_{U_K}(\tau).$$

Consider a quantile difference on the Y_K scale, i.e., a quantity such as $\xi_{Y_K}(1-\psi_K) - \xi_{Y_K}(\psi_K)$, with $\psi_K \in (0, 1/2)$. This can be written in terms of a quantile difference on the U_K scale as:

$$\begin{aligned}\xi_{Y_K}(1-\psi_K) - \xi_{Y_K}(\psi_K) &= \mu_{Y_K} + \sigma_{Y_K}\xi_{U_K}(1-\psi_K) - \mu_{Y_K} - \sigma_{Y_K}\xi_{U_K}(\psi_K) \\ &= \sigma_{Y_K}(\xi_{U_K}(1-\psi_K) - \xi_{U_K}(\psi_K)).\end{aligned}$$

On solving the above equation for σ_{Y_K}, we obtain the following result for the standard deviation:

$$\sigma_{Y_K} = \frac{\xi_{Y_K}(1-\psi_K) - \xi_{Y_K}(\psi_K)}{\xi_{U_K}(1-\psi_K) - \xi_{U_K}(\psi_K)},$$

i.e., the standard deviation of Y_K can be expressed as the ratio between an interquantile range of level $\psi_K \in (0, 1/2)$ of Y_K and the same quantity for its standardized version U_K. Note that this holds without any assumptions on the distribution of Y_K, for $K \in \{A, B\}$.

A further simplification occurs if we take into account the symmetry of Y_K, and therefore of U_K, thanks to which we can rewrite the above expression as:

$$\sigma_{Y_K} = \frac{\xi_{Y_K}(1-\psi_K) - \xi_{Y_K}(1/2)}{\xi_{U_K}(1-\psi_K)}.$$

This expression might be simplified on choosing an appropriate quantile level ψ_K^* such that $\xi_{U_K}(1-\psi_K) = 1$, i.e., $\psi_K^* = 1 - P(U_K \leq 1)$. In this case:

$$\sigma_{Y_K} = \xi_{Y_K}(1-\psi_K^*) - \xi_{Y_K}(1/2),$$

which, thanks to quantile invariance properties, can be rewritten as:

$$\sigma_{Y_K} = g_K(\xi_{X_K}(1-\psi_K^*)) - g_K(\xi_{X_K}(1/2)).$$

Therefore, we can express $V(W)$ as follows:

$$V(W) = \sum_{K\in\{A,B\}} \frac{(g_K(\xi_{X_K}(1-\psi_K^*)) - g_K(\xi_{X_K}(1/2)))^2}{n_K g'(\xi_{X_K}(1/2))^2}. \tag{2.24}$$

It is possible to provide an approximation of (2.24) by considering a Taylor expansion of the function $g(x)$ in a neighborhood of $\xi_{X_K}(1/2)$ in order to have a linear approximation for the expression of σ_{Y_K}:

$$g(x) = g(\xi_{X_K}(1/2)) + (x - \xi_{X_K}(1/2))g'(\xi_{X_K}(1/2)) + \dots$$

and therefore

$$g(x) - g(\xi_{X_K}(1/2)) \doteq (x - \xi_{X_K}(1/2))g'(\xi_{X_K}(1/2)).$$

Computation in $\xi_{X_K}(1 - \psi_K^*)$ yields:

$$g(\xi_{X_K}(1 - \psi_K^*)) - g(\xi_{X_K}(1/2)) \doteq (\xi_{X_K}(1 - \psi_K^*) - \xi_{X_K}(1/2))g'(\xi_{X_K}(1/2)),$$

where the linearization gives a good approximation if $\xi_{X_K}(1 - \psi^*)$ is close to $\xi_{X_K}(1/2)$, i.e., if $\psi^* \approx 1/2$. Note that this condition might be in conflict with the definition of ψ^* given above. In fact, $\psi^* \approx 1/2$ if $1 - P(U_K \leq 1) \approx 1/2$, i.e., if $P(U_K \leq 1) \approx 1/2$. This means that the approximation is good when the standardized random variable U_K has the median close to 1, for $K \in \{A, B\}$. The expression for the variance of W can then be approximated as:

$$V(W) \doteq \frac{(\xi_{X_A}(1 - \psi_A^*) - \xi_{X_A}(1/2))^2}{n_A} + \frac{(\xi_{X_B}(1 - \psi_B^*) - \xi_{X_B}(1/2))^2}{n_B}. \quad (2.25)$$

Approximation (2.25) has a simple structure, that resembles that of the denominator of a t-test. However, the dependence from the distribution of Y_K via the parameter ψ_K^*, for $K \in \{A, B\}$, still stands. We wish to find an estimator for this parameter, but the most natural solution, i.e., deriving it from the empirical distribution function of U_A and U_B, would be clearly impractical. A different possibility exploits the reformulation of the standard deviation in terms of quantile level, i.e., of a parameter bounded in $(0, 1/2)$. Therefore, it seems a reasonable choice to fix a working quantile level $q_K \in (0, 1/2)$, for $K \in \{A, B\}$, to approximate the variability. A good guess could be to choose $q_K \approx 1/2$, for the reasons stated above. Assuming that such a value could be found, in order to use the approximation for the variance of W stated in (2.25), which is an unknown quantity, we need an estimator for it. Thanks to continuous mapping theorem, a suitable quantity for this

task could be:

$$\hat{V}_A(W) \doteq \sum_{K \in \{A,B\}} \frac{(\hat{\xi}_{X_K}(1-q_K) - \hat{\xi}_{X_K}(1/2))^2}{n_K},$$

which we might use to build an approximated "pseudo-Studentized" test statistic T_A, which under (2.5) could be written as:

$$T_A = \frac{W}{\sqrt{\hat{V}_A(W)}}.$$

If the value of q_K provides an accurate enough linear approximation through (2.22), then goodness of the approximation relies only on goodness of the linearization of $g(\cdot)$.

However, there is further room for improvement from the interpretability point of view. In fact, the quantity W is still defined on the Y scale, and does not have an immediate "natural" interpretation. Therefore, let us move a step further, and try to replace W with an asymptotically equivalent estimator, based completely on the X scale and with an immediate interpretation. For this aim, we will keep the structure and the rationale behind it, but abandon the formal definition of a Studentized statistic provided by Chung et al. (2013), introducing a new class of tests, based on test T_A obtained in the previous Section. The basic idea is to replace W with the asymptotically equivalent quantity

$$QW(1/2) = \hat{\xi}_{X_A}(1/2) - \hat{\xi}_{X_B}(1/2).$$

As we have seen in Section 2.2, this quantity is asymptotically equivalent to W, since both are consistent for the quantile difference $\xi_{X_A}(1/2) - \xi_{X_B}(1/2)$. In the notation, we stress the dependence on the quantile level, as we will define the more general statistic $QW(\tau)$, for $\tau \in (0,1)$, further in this Section.

We introduce the pseudo-Studentized test statistic $QT(1/2, \psi_A^*, \psi_B^*)$, defined as follows:

$$QT(1/2, \psi_A^*, \psi_B^*) = \frac{\hat{\xi}_{X_A}(1/2) - \hat{\xi}_{X_B}(1/2) - (\xi_{X_A}(1/2) - \xi_{X_B}(1/2))}{\sqrt{\frac{(\xi_{X_A}(1-\psi_A^*) - \xi_{X_A}(1/2))^2}{n_A} + \frac{(\xi_{X_B}(1-\psi_B^*) - \xi_{X_B}(\psi_B))^2}{n_B}}},$$

with $\psi_K^* \in (0, 1/2)$, for $K \in \{A, B\}$. Under the null hypothesis (2.5), on fixing the working level q_K, for $K \in \{A, B\}$, and on consistently estimating the denominator of $QT(1/2, \psi_A^*, \psi_B^*)$, the pseudo-Studentized test statistic is:

$$QT(1/2, q_A, q_B) = \frac{\hat{\xi}_{X_A}(1/2) - \hat{\xi}_{X_B}(1/2)}{\sqrt{\frac{(\hat{\xi}_{X_A}(1-q_A) - \hat{\xi}_{X_A}(1/2))^2}{n_A} + \frac{(\hat{\xi}_{X_B}(1-q_B) - \hat{\xi}_{X_B}(1/2))^2}{n_B}}}. \tag{2.26}$$

We notice that the structure of the test statistic (2.26) resembles closely that of a Welch test, i.e., a generalisation of the t-test for the comparison of the means of two heteroscedastic populations, which is usually written as:

$$T_{\text{Welch}} = \frac{\overline{X}_A - \overline{X}_B}{\sqrt{\frac{1}{n_A} S_{X_A}^2 + \frac{1}{n_B} S_{X_B}^2}}.$$

Therefore the $QT(1/2, q_A, q_B)$ statistic might be interpreted as a sort of quantile version of the Welch t statistic, where the mean difference is replaced by a median difference, and the standard deviation for each group is replaced by the semi-interquantile range of level q_K, for $K \in \{A, B\}$. The relationship between different measures of variability, and in particular between standard deviation and interquartile range has been explored in the literature, for example by DasGupta & Haff (2006), who have provided the correlation between the two measures for several distributions. The analogy of structure with a well-known and widely used test statistic seems to improve the ease of interpretability, which in the genomic setting can be viewed as a crucial point: since the biological interpretation of the quantities in play is often complicated, we think it is a desirable property of the statistical methods to leave the interpretability as intact as possible.

In order to apply $QT(1/2, q_A, q_B)$ to hypothesis testing, its asymptotic distribution can be derived. This will be the object of the the next Subsection; however, we will do it for the generalised version $QT(\tau, q_A, q_B)$, for $\tau \in (0, 1)$, which we define in the following. As a preliminary step, we define

$$QW(\tau) = \hat{\xi}_{X_A}(\tau) - \hat{\xi}_{X_B}(\tau),$$

Then, under the null hypothesis (2.5), the generalised pseudo-Studentized test statistic is defined as

$$QT(\tau, q_A, q_B) = \frac{\hat{\xi}_{X_A}(\tau) - \hat{\xi}_{X_B}(\tau)}{\sqrt{\frac{(\hat{\xi}_{X_A}(1-q_A) - \hat{\xi}_{X_A}(1/2))^2}{n_A} + \frac{(\hat{\xi}_{X_B}(1-q_B) - \hat{\xi}_{X_B}(1/2))^2}{n_B}}}. \tag{2.27}$$

We will derive its asymptotic distribution and study its properties in the next Subsection.

Remark: If the sample size is low, p-values can be computed via resampling methods such as permutation or the bootstrap. Unlike results we have seen in Chapter 2, results of Chung et al. (2013) are not a reason in this case to privilege permutation methods, since the structure of a properly Studentized statistic is not matched by $QT(\tau, q_A, q_B)$.

Remark: A simplified version of $QT(\tau, q_A, q_B)$ can be produced when the distribution in the two samples can be assumed to be the same, at least in the quantiles $\xi_{X_K}(1-q_K)$ and $\xi_{X_K}(q_K)$, for $K \in \{A, B\}$. Assuming that $n_A = n_B$, the simplified test statistic can be defined, under the null hypothesis, as:

$$QT(\tau, q) = \frac{\hat{\xi}_{X_A}(\tau) - \hat{\xi}_{X_B}(\tau)}{\hat{\xi}_X(1-q) - \hat{\xi}_X(q)},$$

where X is the pooled sample. The choice of pooling the samples seems reasonable under the assumption of equal distribution in the two populations. The asymptotic normality of $QT(\tau, q)$ can be proven as follows. Thanks to independence of the two samples, its numerator is asymptotically normal, if the sample sizes converge at a comparable rate, i.e. if $n_A/n_B \to c \neq 0$. In this case,

$$(\hat{\xi}_{X_A}(\tau) - \hat{\xi}_{X_B}(\tau)) - (\xi_{X_A}(\tau) - \xi_{X_B}(\tau)) \dot{\sim} N\left(0, \frac{\sigma^2_{X_A}(\tau)}{n_A} + \frac{\sigma^2_{X_B}(\tau)}{n_B}\right),$$

with $\sigma^2_{X_K}(\tau)$ defined as in (2.6), for $K \in \{A, B\}$. Moreover, the Cramér and Wold device shows that the denominator of $QT(\tau, q)$ converges in probability as follows:

$$\sqrt{n}((\hat{\xi}_X(1-q) - \hat{\xi}_X(q)) - (\xi_X(1-q) - \xi_X(q))) \xrightarrow{d} N(0, \sigma^2_X(q)),$$

where

$$\sigma^2_X(q) = \frac{q(1-q)}{f_X(\xi_X(1-q))^2} + \frac{q(1-q)}{f_X(\xi_X(q))^2} - \frac{(1-q)^2}{f_X(\xi_X(q))f_X(\xi_X(1-q))}$$

and f_X is the density function for the pooled sample. The result implies convergence in probability:

$$\hat{\xi}_X(1-q) - \hat{\xi}_X(q) \xrightarrow{p} \xi_X(1-q) - \xi_X(q).$$

Asymptotic normality of $QT(\tau, q)$ is now guaranteed by Slutsky's Theorem. Provided that $\xi_X(1-q) - \xi_X(q) \neq 0$, one has that, under (2.5):

$$QT(\tau, q) \dot{\sim} N\left(0, \omega^2\right)$$

where

$$\omega^2 = \frac{\frac{\sigma^2_{X_A}(\tau)}{n_A} + \frac{\sigma^2_{X_B}(\tau)}{n_B}}{(\xi_X(1-q) - \xi_X(q))^2}.$$

Therefore, if n_A and n_B are sufficiently large, one can use the asymptotic distribution to compute p-values for $QT(\tau, q)$. However, the above mentioned problem of density estimation would persist. In this case, a proper Studentization is not available and therefore results of Chung et al. (2013) do not apply. Moreover, the denominator of $QT(\tau, q)$ is invariant with respect to data permutation, which would make permutation p-values equal to those computed for the numerator only. An alternative option is to compute p-values via bootstrap resampling, although behaviour of unconditional type I error levels remains to be studied.

2.4.1 Asymptotic Distribution of $QT(\tau, q_A, q_B)$

We have recalled in Chapter 2 the asymptotic distribution of quantile estimators. As a direct consequence, the following asymptotic result holds:

$$QW(\tau) \dot{\sim} N\left(\xi_{X_A}(\tau) - \xi_{X_B}(\tau), \frac{\tau(1-\tau)}{n_A p^2_{X_A}(\xi_{X_A}(\tau))} + \frac{\tau(1-\tau)}{n_B p^2_{X_B}(\xi_{X_B}(\tau))}\right),$$

provided that the sample sizes converge at a comparable rate, i.e. if $n_A/n_B \to c \neq 0$ when $n_K \to \infty$, for $K \in \{A, B\}$. The denominator of $QT(\tau, q_A, q_B)$ converges in probability, as can be shown by using asymptotic joint distribution of quantiles together with Cramer and Wold device, which reduces the convergence of multivariate distribution functions to the convergence of univariate distribution functions. As a preliminary result, we show that:

$$\sqrt{n_K}(\hat{\xi}_{X_K}(1-q_K) - \hat{\xi}_{X_K}(1/2) - (\xi_{X_K}(1-q_K) - \xi_{X_K}(1/2))) \xrightarrow{d} N(0, \sigma^2_Q),$$

where

$$\sigma^2_Q = \frac{q_K(1-q_K)}{p_{X_K}(\xi_{X_K}(1-q_K))^2} + \frac{1/4}{p_{X_K}(\xi_{X_K}(1/2))^2} - \frac{(1-q_K)}{p_{X_K}(\xi_{X_K}(1-q_K))p_{X_K}(\xi_{X_K}(1/2))}.$$

Note that we are not treating q_K as an estimator, but not as a non-random approximation of ψ_K^*, for $K \in \{A, B\}$. The previous result implies convergence in probability of the quantity $\hat{\xi}_{X_K}(1-q_K) - \hat{\xi}_{X_K}(1/2)$:

$$\hat{\xi}_{X_K}(1-q_K) - \hat{\xi}_{X_K}(1/2) \xrightarrow{p} \xi_{X_K}(1-q_K) - \xi_{X_K}(1/2),$$

and therefore continuous mapping theorem guarantees that the denominator of $QT(\tau, q_A, q_B)$ converges in probability:

$$\sqrt{\frac{(\hat{\xi}_{X_A}(1-q_A) - \hat{\xi}_{X_A}(1/2))^2}{n_A} + \frac{(\hat{\xi}_{X_B}(1-q_B) - \hat{\xi}_{X_B}(1/2))^2}{n_B}} \xrightarrow{p} w,$$

where

$$w = \sqrt{\frac{(\xi_{X_A}(1-q_A) - \xi_{X_A}(1/2))^2}{n_A} + \frac{(\xi_{X_B}(1-q_B) - \xi_{X_B}(1/2))^2}{n_B}}.$$

Asymptotic normality of $QT(\tau, q_A, q_B)$ is then proven by Slutsky's Theorem. Provided that $\xi_{X_K}(1-q_K) - \xi_{X_K}(1/2) \neq 0$ for $K \in \{A, B\}$, one has that, under the null hypothesis (2.5),

$$QT(\tau, q_A, q_B) \dot{\sim} N\left(0, \frac{1}{w^2}\left(\frac{\tau(1-\tau)}{n_A p_{X_A}^2(\xi_{X_A}(\tau))} + \frac{\tau(1-\tau)}{n_B p_{X_B}^2(\xi_{X_B}(\tau))}\right)\right).$$

This result allows us to use the standardized version of $QT(\tau, q_A, q_B)$ as a pivotal quantity for hypothesis testing when n_A and n_B are sufficiently large. Unfortunately, this would require estimation of the density function at the quantile of interest, which rises problems similar to those encountered in Chapter 2. Nevertheless, we can resort to resampling methods. We will investigate the performances of $QT(\tau, q_A, q_B)$ in terms of type I error rate control and power, for the same models seen in Chapter 2, evaluating also the choices of the parameter q_K, $K \in \{A, B\}$.

2.4.2 Simulation Studies: Type I Error Rate Control

In this Section, we will perform tests on the median of the distribution, i.e., we assume $\tau = 1/2$, under the same models of Subsection 2.3.1, and we simulate two independent samples of size n_A and n_B from the following models:

1. $X_A \sim LN(0, 1)$, $X_B \sim LN(0, 5)$ (see example 2.3.1);

2. $X_A \sim LN(0,1)$, $X_B \sim Lt(0,5)$ (see example 2.3.2);

3. $X_A \sim LLog(0,1)$, $X_B \sim LU(-10,10)$ (see example 2.3.3);

4. $X_A \sim LN(-0.003,1)$, $X_B \sim MD(0,1,-3,1,0.999)$ for robustness;

5. $X_A \sim LLap(\log(2),1)$, $X_B \sim LGa(1,1)$ (see example 2.3.4), which does not satisfy the assumptions of the key case.

We choose $q_A = q_B = q$ and compute the $QT(1/2,q,q)$ statistics for $q \in \{0.10, 0.25, 0.40, 0.45\}$. In each simulation experiment, $m = 999$ permutations are used to compute p-values, except when $n_A = n_B = 5$; in that case, all available $m = 252$ permutations are used. The number of replicates in each simulation experiment is 10000. Samples are generated under the null hypothesis (2.5).

In the following, we will display the results obtained from simulation studies. Each Table refers to one of the five models of interest above mentioned, and displays permutation type I error rates. Different test statistics are displayed by row, while every column refers to a different sample size.

Table 2.11 shows Monte Carlo results for the $X_A \sim LN(0,1)$ vs. $X_B \sim LN(0,5)$ case. Type I error rate is closer to the nominal one for low sample sizes for $q \in \{0.25, 0.40, 0.45\}$ than for $q = 0.10$. As expected, best results are obtained for a value of q closer to 0.50. It is worth noting that even a little shift in the value of q has a relevant impact on the estimated type I error rate, especially for larger sample sizes. It also seems that the test statistics are more conservative for larger value of n_A and n_B.

Table 2.12 shows Monte Carlo results for the $X_A \sim LN(0,1)$ vs. $X_B \sim Lt(0,5)$ case. A more similar behaviour than in the previous caseis observed for all of the four considered test statistics. Results are also more stable for different sample sizes. This reflects the greater similarity of the distributions in the two groups compared to those of the previous case.

Table 2.13 shows Monte Carlo results for the $X_A \sim LLog(0,1)$ vs. $X_B \sim LU(-10,10)$ case. The behaviour of the four test statistics is quite different, and

		$n_A = n_B$				
		5	11	101	1001	5001
$QT(0.5, 0.10, 0.10)$	α_{perm}	0.0386	0.0358	0.0166	0.0102	0.0068
$QT(0.5, 0.25, 0.25)$	α_{perm}	0.0523	0.0541	0.0299	0.0212	0.0147
$QT(0.5, 0.40, 0.40)$	α_{perm}	0.0523	0.0576	0.0534	0.0445	0.0353
$QT(0.5, 0.45, 0.45)$	α_{perm}	0.0523	0.0576	0.0629	0.0502	0.0423

Table 2.11: Monte Carlo permutation and asymptotic type I error rates for the $LN(0,1)$ vs. $LN(0,5)$ case under $\xi_{X_A}(1/2) = \xi_{X_B}(1/2)$. Nominal type I error rate $\alpha = 0.05$.

		$n_A = n_B$				
		5	11	101	1001	5001
$QT(0.5, 0.10, 0.10)$	α_{perm}	0.0397	0.0463	0.0466	0.0461	0.0437
$QT(0.5, 0.25, 0.25)$	α_{perm}	0.0353	0.0493	0.0474	0.0507	0.0482
$QT(0.5, 0.40, 0.40)$	α_{perm}	0.0353	0.0462	0.0491	0.0516	0.0478
$QT(0.5, 0.45, 0.45)$	α_{perm}	0.0353	0.0462	0.0505	0.0519	0.0487

Table 2.12: Monte Carlo permutation and asymptotic type I error rates for the $LN(0,1)$ vs. $Lt(0,5)$ case under $\xi_{X_A}(1/2) = \xi_{X_B}(1/2)$. Nominal type I error rate $\alpha = 0.05$.

it seems that the type I error rate is not controlled by most test statistics. However, the test statistics employing $q = 0.45$ performs slightly better than the other ones for larger sample sizes. A possible explanation might be in the approximation required in the construction of $QT(\tau, q)$, which might be less accurate than for the previous cases for these specific distributions.

Table 2.14 shows Monte Carlo results for the $X_A \sim LN(-0.003, 1)$ vs. $X_B \sim MD(0, 1, -3, 1, 0.999)$ case. The four statistics all produce a type I error rate which is stable with respect with a small contamination of the data.

		$n_A = n_B$				
		5	11	101	1001	5001
$QT(0.5, 0.10, 0.10)$	α_{perm}	0.0365	0.0249	0.0002	0.0000	0.0000
$QT(0.5, 0.25, 0.25)$	α_{perm}	0.0819	0.0489	0.0022	0.0000	0.0000
$QT(0.5, 0.40, 0.40)$	α_{perm}	0.0819	0.0900	0.0555	0.0067	0.0016
$QT(0.5, 0.45, 0.45)$	α_{perm}	0.0819	0.0900	0.1032	0.0361	0.0192

Table 2.13: Monte Carlo permutation and asymptotic type I error rates for the $LLog(0, 1)$ vs. $LU(-10, 10)$ case under $\xi_{X_A}(1/2) = \xi_{X_B}(1/2)$. Nominal type I error rate $\alpha = 0.05$.

		$n_A = n_B$				
		5	11	101	1001	5001
$QT(0.5, 0.10, 0.10)$	α_{perm}	0.0409	0.0494	0.0514	0.0483	0.0504
$QT(0.5, 0.25, 0.25)$	α_{perm}	0.0361	0.0499	0.0517	0.0477	0.0504
$QT(0.5, 0.40, 0.40)$	α_{perm}	0.0361	0.0429	0.0514	0.0485	0.0501
$QT(0.5, 0.45, 0.45)$	α_{perm}	0.0409	0.0429	0.0511	0.0495	0.0510

Table 2.14: Monte Carlo permutation and asymptotic type I error rates for the $LN(-0.003, 1)$ vs $MD(0, 1, -3, 1, 0.999)$ case under $\xi_{X_A}(1/2) = \xi_{X_B}(1/2)$. Nominal type I error rate $\alpha = 0.05$.

Table 2.15 shows Monte Carlo results for the $X_A \sim LLap(\log(2), 1)$ vs. $X_B \sim LGa(1, 1)$ case. Type I error rates seem very close to the nominal one for all the considered statistics, even for moderate to low sample sizes, even when the assumption of the key case presented in Section 2.3 do not hold.

Overall, it seems that control of type I error is guaranteed in all the considered settings, at least for the test statistics whose values of q are closer to 0.50, apart from the $LLog(0, 1)$ vs $LU(-10, 10)$ case. In this case, it seems that most of the test statistics provide liberal estimates of the type I error rate for small sample sizes and very conservative estimates of the type I error rate for moderate to large sample sizes. This is particularly true for $q \in \{0.10, 0.25\}$. This can be explained by the

		$n_A = n_B$				
		5	11	101	1001	5001
$QT(0.5, 0.10, 0.10)$	α_{perm}	0.0297	0.0432	0.0447	0.0488	0.0497
$QT(0.5, 0.25, 0.25)$	α_{perm}	0.0314	0.0419	0.0438	0.0478	0.0494
$QT(0.5, 0.40, 0.40)$	α_{perm}	0.0314	0.0436	0.0448	0.0476	0.0490
$QT(0.5, 0.45, 0.45)$	α_{perm}	0.0314	0.0436	0.0476	0.0472	0.0492

Table 2.15: Monte Carlo permutation and asymptotic type I error rates for the $LLap(\log(2), 1)$ vs $LGa(1, 1)$ case under $\xi_{X_A}(1/2) = \xi_{X_B}(1/2)$. Nominal type I error rate $\alpha = 0.05$.

fact that the Taylor approximation might not work well in this case. For the other settings:

1. when $q \in \{0.40, 0.45\}$, control of the type I error rate seems guaranteed across different sample sizes larger than 5, and it is usually slightly conservative for $n_A = n_B = 5$ (except for the $LN(0, 1)$ vs $LN(0, 5)$ case);

2. a similar behaviour is observed when $q \in \{0.10, 0.25\}$, but for the $LN(0, 1)$ vs $LN(0, 5)$ case test statistics seem very conservative for large sample sizes.

In the next Subsection, we will evaluate power of the $QT(\tau, q_A, q_B)$ test statistics for the same models.

2.4.3 Simulation Studies: Power

In order to assess the power of the proposed statistics, we choose the same models proposed in the analogous Subsection 2.3.2, i.e., we simulate two independent samples of size n_A and n_B from the following models:

1. $X_A \sim LN(0, 1)$, $X_B \sim LN(-1, 5)$;

2. $X_A \sim LN(0, 1)$, $X_B \sim Lt(-1, 5)$;

3. $X_A \sim LN(0, 1)$, $X_B \sim Lt(-0.5, 5)$;

4. $X_A \sim LLog(0, 1)$, $X_B \sim LU(-10, 5)$;

5. $X_A \sim LLap(2,1)$, $X_B \sim LGa(1,1)$.

We report results for the same test statistics considered in Subsection 2.4.2, i.e., for $QT(1/2, q_A, q_B)$ with $q_A = q_B = q \in \{0.10, 0.25, 0.40, 0.45\}$. In each simulation experiment, $m = 999$ permutations are used, to compute p-values, except when $n_A = n_B = 5$; in that case, all available $m = 252$ permutations are used. The number of replicates in each simulation experiment is 10000.

In the following, we report Tables containing the estimated permutation power of the test statistics for each of the above mentioned settings. In each Table, each row refers to a different test statistic, while each column refers to a different sample size. The permutation power is computed as the number of times that the null hypothesis would be rejected, with a threshold $\alpha = 0.05$.

Table 2.16 shows Monte Carlo results for the $X_A \sim LN(0,1)$ vs. $X_B \sim LN(-1,5)$ case. All of the test statistics provide power that goes to 1 for large sample sizes, with the highest power provided when $q = 0.25$ (which also led conservative type I error rates) for all sample sizes.

		$n_A = n_B$				
		5	11	21	51	101
$QT(0.5, 0.10, 0.10)$	α_{perm}	0.1532	0.2287	0.3840	0.6589	0.8999
$QT(0.5, 0.25, 0.25)$	α_{perm}	0.1473	0.2689	0.4209	0.7624	0.9582
$QT(0.5, 0.40, 0.40)$	α_{perm}	0.1473	0.2139	0.3821	0.7575	0.9453
$QT(0.5, 0.45, 0.45)$	α_{perm}	0.1473	0.2139	0.3053	0.6946	0.9362

Table 2.16: Monte Carlo permutation power for the $LN(0,1)$ vs. $LN(-1,5)$ case.

Table 2.17 shows Monte Carlo results for the $X_A \sim LN(0,1)$ vs. $X_B \sim Lt(-1,5)$ case. The power of all of the statistics is higher than in the previous case, similarly to what we noticed for the T_j, for $j = 1, \ldots, 6$ test statistics, possibly due to the clearer separations between the two distributions generating the data. Highest power is achieved at all sample sizes when $q = 0.10$.

		$n_A = n_B$				
		5	11	21	51	101
$QT(0.5, 0.10, 0.10)$	α_{perm}	0.3399	0.5401	0.8069	0.9945	1.0000
$QT(0.5, 0.25, 0.25)$	α_{perm}	0.2044	0.4606	0.7105	0.9827	1.0000
$QT(0.5, 0.40, 0.40)$	α_{perm}	0.2044	0.3175	0.5621	0.9154	0.9976
$QT(0.5, 0.45, 0.45)$	α_{perm}	0.2044	0.3175	0.4432	0.8579	0.9868

Table 2.17: Monte Carlo permutation power for the $LN(0, 1)$ vs. $Lt(-1, 5)$ case.

Table 2.18 shows Monte Carlo results for the $X_A \sim LN(0, 1)$ vs. $X_B \sim Lt(-0.5, 5)$ case. As expected, there is some loss in power with respect to the previous case, due to a smaller shift in the medians for the two groups. A power of about 0.80 or higher is obtained for all of the test statistics only for a sample size of 101. The best results are provided when $q \in \{0.10, 0.25\}$.

		$n_A = n_B$				
		5	11	21	51	101
$QT(0.5, 0.10, 0.10)$	α_{perm}	0.1455	0.2308	0.3693	0.6898	0.9112
$QT(0.5, 0.25, 0.25)$	α_{perm}	0.1025	0.2084	0.3267	0.6602	0.8945
$QT(0.5, 0.40, 0.40)$	α_{perm}	0.1025	0.1578	0.2660	0.5722	0.8466
$QT(0.5, 0.45, 0.45)$	α_{perm}	0.1025	0.1578	0.2211	0.5051	0.7930

Table 2.18: Monte Carlo permutation power for the $LN(0, 1)$ vs. $Lt(-0.5, 5)$ case.

Table 2.19 shows Monte Carlo results for the $X_A \sim LLog(0, 1)$ vs. $X_B \sim LU(-10, 5)$ case. We report them for completeness, even though in this case it seemed from the previous Subsection that control of type I error rate is not achieved. Power seems to achieve 1 for large sample sizes for all of the test statistics, with $QT(0.5, 0.40, 0.40)$ providing the highest power for most sample sizes.

Table 2.20 shows Monte Carlo results for the $X_A \sim LLap(2, 1)$ vs. $X_B \sim LGa(1, 1)$ case. The permutation power for all of the test statistics approaches 1

		$n_A = n_B$				
		5	11	21	51	101
$QT(0.5, 0.10, 0.10)$	α_{perm}	0.1484	0.2083	0.3194	0.5728	0.8546
$QT(0.5, 0.25, 0.25)$	α_{perm}	0.1812	0.2644	0.3948	0.7011	0.8950
$QT(0.5, 0.40, 0.40)$	α_{perm}	0.1812	0.2455	0.4157	0.7929	0.9554
$QT(0.5, 0.45, 0.45)$	α_{perm}	0.1812	0.2455	0.3476	0.7394	0.9607

Table 2.19: Monte Carlo permutation power for the $LLog(0, 1)$ vs $LU(-10, 5)$ case.

for large sample sizes, with $QT(0.5, 0.10, 0.10)$ providing the highest power almost always across different sample sizes.

		$n_A = n_B$				
		5	11	21	51	101
$QT(0.5, 0.10, 0.10)$	α_{perm}	0.3662	0.5651	0.8069	0.9844	1.0000
$QT(0.5, 0.25, 0.25)$	α_{perm}	0.2398	0.5238	0.7967	0.9891	1.0000
$QT(0.5, 0.40, 0.40)$	α_{perm}	0.2398	0.4034	0.6804	0.9643	1.0000
$QT(0.5, 0.45, 0.45)$	α_{perm}	0.2398	0.4034	0.5459	0.9308	0.9988

Table 2.20: Monte Carlo permutation power for the $LLap(2, 1)$ vs $LGa(1, 1)$ case.

Overall, it seems that all of the test statistics are consistent, i.e., that their power approaches 1 when the sample sizes increase. The sample sizes required in order to obtain a reasonably good power (say over 0.80) vary considerably for the different models. For the $LLap(2, 1)$ vs $LGa(1, 1)$ case good power results are achieved for most statistics already when $n_A = n_B = 21$, while for the $LN(0, 1)$ vs. $Lt(-0.5, 5)$ case possibly a sample size of 51 or even 101 is needed. Overall, it seems that the test statistics employing $q \in \{0.10, 0.25\}$ provide higher power than the other ones for most cases.

2.5 Final Remarks

In this Chapter, we have built Studentized statistics for the comparison of two groups. After having defined them for a very general case, we have derived simplified versions for a specific key case, i.e., when both samples can be transformed by the same function in order to obtain a distribution for which $\xi_{X_K}(\tau) = \mu_{Y_K}$ holds. Under these assumptions, we have assessed performances of the test statistics in terms of type I error control rate and power. Our main findings may be summarized as follows:

- it seems that all the proposed statistics perform well in most of the considered statistical models, both in terms of type I error control rate and power, even if with several differences between them;
- statistics T_4, T_5 and T_6 achieve the best type I error rate control. This can be due to the fact that their variance estimators take into account the fact that data are generated under the null hypothesis. They also provide the same permutation p-values, as noted in Subsection 2.3.1;
- statistics T_1, T_2 and T_3 achieve the best power for most of the considered models.

Afterwards, we have defined a t-test-like Pseudo-Studentization, entirely based on quantiles on the X scale. Preliminary results show that also these test statistics perform well for most proposed models, both in terms of control of the type I error rate, for which a small value of the parameters q_A and q_B seems necessary, and in terms of power.

Appendix

In this Appendix we will report, with minor modifications, Theorem 2.2 in Chung et al. (2013), and prove that the main conditions hold for the statistics T to T_6, taking T as an example.

Permutation Distribution of S

Assume $Y_{K1}, \ldots, Y_{Kn_K}$ are i.i.d. with distribution function P_{Y_K}, for $K \in \{A, B\}$, independent from each other. Assume also that the vector of all observations $Y = (Y_{A1}, \ldots, Y_{An_A}, Y_{B1}, \ldots, Y_{Bn_B})$ can be interpreted as a vector of i.i.d. random variables from the mixture population, i.e. with distribution function $\overline{P}_Y = qP_{Y_A} + (1-q)P_{Y_B}$, with $q \in (0,1)$, where $V(Y) < \infty$. Consider testing $H_0 : \theta_A = \theta_B$ for some parameter of interest θ_K. For this purpose, consider $\hat{\theta}_K$, an asymptotically linear estimator for θ_K, i.e, an estimator such that there exist functions $l_K(\cdot)$ such that:

$$n_K^{1/2}(\hat{\theta}_K - \theta_K) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n_K} l_K(Y_{Ki}) + o_P(1),$$

under P_{Y_K}, and the same holds for the mixture population $Y \sim \overline{P}_Y$. The functions $l_K(\cdot)$ must be such that $E(l_K(Y_{Ki})) = 0$ and $E(l_K^2(Y_{Ki})) < \infty$, both under P_{Y_K} and under $\overline{P}_Y$. A Studentized statistic is then defined as:

$$S = \frac{n^{1/2}(\hat{\theta}_A - \hat{\theta}_B)}{\sqrt{\frac{n}{n_A}\hat{\sigma}^2_{Y_A} - \frac{n}{n_B}\hat{\sigma}^2_{Y_B}}},$$

where $\hat{\sigma}^2_{Y_K}$ is a consistent estimator of $\sigma^2_{Y_K}$, and $\hat{\sigma}^2_{Y_K}$ is consistent also for σ^2_Y when $Y_1, \ldots, Y_n$ are i.i.d. from $\overline{P}_Y$. Let $n_A \to \infty$, $n_B \to \infty$, with $n_A/n \to q \in (0,1)$ and $n_A/n - q = O(n^{-1/2})$. Define the permutation distribution of S as

$$\hat{R}^S(s) = \frac{1}{n!} \sum_{\pi \in G_n} \mathbf{I}\left\{S\left(Y_{\pi(1)}, \ldots, Y_{\pi(n)}\right) \leq s\right\}$$

where $\{\pi(1), \ldots, \pi(n)\}$ is any permutation of $\{1, \ldots, n\}$, and G_n is the set of all the permutations π of $\{1, \ldots, n\}$. Then it holds that

$$\sup_s |\hat{R}^S(s) - \Phi(s)| \xrightarrow{p} 0.$$

Therefore the permutation distribution of S is asymptotically standard Normal, as is its true unconditional limiting distribution.

Computation of S permutation p-value

An exact level α permutation test can be constructed as follows: for each permutation $\pi \in G_n$, compute the statistic $S(Y)$, and let their ordered values be:

$$S^{(1)}(Y), \ldots, S^{(n!)}(Y)$$

Fix a nominal level $\alpha \in (0, 1)$ and let $k = n! - [\alpha n!]$, where $[a]$ denotes the largest integer less or equal to a. Let $M^+(Y) = \#\{S^{(j)}(Y) : S^{(j)}(Y) > S^{(k)}(Y), j = 1 \ldots, n!\}$ and $M^0(Y) = \#\{S^{(j)}(Y) : S^{(j)}(Y) = S^{(k)}(Y), j = 1 \ldots, n!\}$, where $\#\{A\}$ is the cardinality of set A. The randomization test function is defined as:

$$\phi(Y) = \begin{cases} 1 & \text{if } S > S^{(k)}(Y) \\ \dfrac{\alpha n! - M^+(Y)}{M^0(Y)} & \text{if } S = S^{(k)}(Y) \\ 0 & \text{if } S < S^{(k)}(Y) \end{cases}$$

and $E(\phi(Y)) = \alpha$ exactly.

To compute the permutation p-value of S, we refer to the general theory of permutation tests, keeping in mind the observations of Smyth & Phipson (2011) to make sure that the permutation p-values are never null. We will assume that the Y_{Ki} are all distinct, so that the test statistic can assume $n!$ possible distinct values, with equal probability. The original idea of permutation tests was that all possible permutations would be enumerated (Fisher, 1935). However $n!$ can be very large even for moderate sample sizes. In this case, it is common practice to examine a random subset of the possible permutations, of size $m \leq n!$. Then the exact permutation p-value is:

$$p_{\text{perm}} = \frac{b + 1}{m},$$

where b is the number of permutation $S^{(j)}(Y)$ which are equally or further from the null hypothesis than the observed S computed on the original sample. The quantity p_{perm} converges to the asymptotic p-value for large n.

Asymptotic Linearity of $\tilde{\xi}_{X_K}(\tau)$

We will now prove that the quantity $\tilde{\xi}_{X_K}(\tau) = f_K(\overline{Y}_K)$ is asymptotically linear in the sense of Theorem 2.2 in Chung et al. (2013). From a Taylor expansion, centered in μ_{Y_K} and computed in $\overline{Y}_K$, we have that:

$$f_K(\overline{Y}_K) = f_K(\mu_{Y_K}) + (\overline{Y}_K - \mu_{Y_K}) f'_K(\mu_{Y_K}) + \sum_{j=2}^{\infty} \frac{\partial^j f_K(t)}{\partial t^j}\Big|_{t=\mu_{Y_K}} \frac{(\overline{Y}_K - \mu_{Y_K})^j}{j!}.$$

This can be rewritten as:

$$f_K(\overline{Y}_K) - f_K(\mu_{Y_K}) = \sum_{i=1}^{n_K} \frac{Y_{Ki} - \mu_{Y_K}}{n_K} f'_K(\mu_{Y_K}) + \sum_{j=2}^{\infty} \frac{\partial^j f_K(t)}{\partial t^j}\Big|_{t=\mu_{Y_K}} \frac{(\overline{Y}_K - \mu_{Y_K})^j}{j!},$$

and as:

$$n_K^{1/2}(f_K(\overline{Y}_K) - f_K(\mu_{Y_K})) = n_K^{-1/2} \sum_{i=1}^{n_K} (Y_{Ki} - \mu_{Y_K}) f'_K(\mu_{Y_K}) + n_K^{1/2} \sum_{j=2}^{\infty} \frac{\partial^j f_K(t)}{\partial t^j}\Big|_{t=\mu_{Y_K}} \frac{(\overline{Y}_K - \mu_{Y_K})^j}{j!}.$$

Then we could define $l_K(Y_{Ki}) = (Y_{Ki} - \mu_{Y_K}) f'_K(\mu_{Y_K})$. We still have to prove that

$$n_K^{1/2} \sum_{j=2}^{\infty} \frac{\partial^j f_K(t)}{\partial t^j}\Big|_{t=\mu_{Y_K}} \frac{(\overline{Y}_K - \mu_{Y_K})^j}{j!} = o_P(1).$$

Because of the weak law of large numbers,

$$\overline{Y}_K - \mu_{Y_K} = o_P(1),$$

and therefore

$$(\overline{Y}_K - \mu_{Y_K})^j = o_P(1).$$

Since the other quantities are non-random, it also holds that:

$$\frac{\partial^j f_K(t)}{\partial t^j}\Big|_{t=\mu_{Y_K}} \frac{(\overline{Y}_K - \mu_{Y_K})^j}{j!} = o_P(1).$$

The sum of the quantities above is dominated in probability by the $j = 2$ term, i.e.,

$$\sum_{j=2}^{\infty} \frac{\partial^j f_K(t)}{\partial t^j}\Big|_{t=\mu_{Y_K}} (\overline{Y}_K - \mu_{Y_K})^j = o_P(1)$$

And finally, since $n_K^{-1/2} = o(1)$,

$$\frac{1}{\sqrt{n_K}} \sum_{j=2}^{\infty} \frac{\partial^j f_K(t)}{\partial t^j}\Big|_{t=\mu_{Y_K}} (\overline{Y}_K - \mu_{Y_K})^j = o_P(1).$$

Therefore, $\tilde{\xi}_{X_K}(\tau)$ is an asymptotically linear estimator.

Additional Simulation Studies

While performing initial simulation studies, we also generated data under the hypothesis of equal distribution of the two populations, $p_{X_A} = p_{X_B}$. In these cases, permutation p-values not only approach asymptotically the actual ones, but are also exact for finite sample sizes (Chung et al., 2013). This was done for some of the models considered in Subsection 2.3.1, i.e.,

1. $X_K \sim LN(0,1)$, for $K \in \{A, B\}$;
2. $X_K \sim LN(0,5)$, for $K \in \{A, B\}$;
3. $X_K \sim Lt(0,5)$, for $K \in \{A, B\}$;
4. $X_K \sim MD(0,1,-3,1,0.999)$, for $K \in \{A, B\}$;
5. $X_K \sim LGa(1,1)$, for $K \in \{A, B\}$.

In each simulation experiment, $m = 999$ permutations are used to compute p-values, except when $n_A = n_B = 5$; in that case, all available $m = 252$ permutations are used. The number of replicates in each simulation experiment is 1000.

In the following, we will display the results obtained from simulation studies. Each Table refers to one of the five models of interest above mentioned, and displays permutation and asymptotic type I error rates. Different test statistics are displayed by row, while every column refers to a different sample size. It seems that, for most models, the type I error is closer to the nominal one even for low sample sizes, as expected.

Table 2.21 shows Monte Carlo results for the $X_K \sim LN(0,1)$ case, for $K \in \{A, B\}$. As expected, permutation type I error rates seem much closer to the nominal type I error rate even for low sample sizes.

		$n_A = n_B$				
		5	11	101	1001	5001
T_1	α_{perm}	0.047	0.044	0.038	0.049	0.052
	α_{asymp}	0.057	0.037	0.040	0.048	0.049
T_2	α_{perm}	0.049	0.051	0.036	0.050	0.052
	α_{asymp}	0.066	0.049	0.041	0.047	0.049
T_3	α_{perm}	0.054	0.051	0.036	0.055	0.052
	α_{asymp}	0.066	0.049	0.041	0.047	0.049
T_4	α_{perm}	0.047	0.041	0.038	0.049	0.052
	α_{asymp}	0.083	0.052	0.044	0.048	0.049
T_5	α_{perm}	0.047	0.041	0.038	0.049	0.052
	α_{asymp}	0.084	0.060	0.045	0.048	0.049
T_6	α_{perm}	0.047	0.041	0.038	0.049	0.052
	α_{asymp}	0.058	0.041	0.042	0.048	0.049

Table 2.21: Monte Carlo permutation and asymptotic type I error rates for the $LN(0,1)$ case under $p_{X_A} = p_{X_B}$. Nominal type I error rate $\alpha = 0.05$.

Table 2.22 shows Monte Carlo results for the $X_K \sim LN(0,5)$ case, for $K \in \{A, B\}$. Also in this case, as expected, permutation type I error rates seem much closer to the nominal type I error rate even for low sample sizes.

		$n_A = n_B$				
		5	11	101	1001	5001
T_1	α_{perm}	0.049	0.052	0.038	0.050	0.052
	α_{asymp}	0.022	0.008	0.033	0.046	0.049
T_2	α_{perm}	0.052	0.059	0.040	0.051	0.052
	α_{asymp}	0.062	0.049	0.035	0.048	0.049
T_3	α_{perm}	0.058	0.059	0.040	0.051	0.052
	α_{asymp}	0.062	0.049	0.035	0.048	0.049
T_4	α_{perm}	0.047	0.041	0.038	0.049	0.052
	α_{asymp}	0.106	0.068	0.046	0.049	0.049
T_5	α_{perm}	0.047	0.041	0.038	0.049	0.052
	α_{asymp}	0.127	0.081	0.045	0.050	0.050
T_6	α_{perm}	0.047	0.041	0.038	0.049	0.052
	α_{asymp}	0.022	0.021	0.037	0.048	0.049

Table 2.22: Monte Carlo permutation and asymptotic type I error rates for the $LN(0,5)$ case under $p_{X_A} = p_{X_B}$. Nominal type I error rate $\alpha = 0.05$.

Table 2.23 shows Monte Carlo results for the $X_K \sim Lt(0, 5)$ case, for $K \in \{A, B\}$. Also in this case, as expected, permutation type I error rates seem much closer to the nominal type I error rate even for low sample sizes.

		$n_A = n_B$				
		5	11	101	1001	5001
T_1	α_{perm}	0.041	0.053	0.057	0.041	0.054
	α_{asymp}	0.042	0.036	0.057	0.038	0.053
T_2	α_{perm}	0.037	0.051	0.058	0.039	0.054
	α_{asymp}	0.057	0.046	0.058	0.040	0.054
T_3	α_{perm}	0.040	0.051	0.058	0.039	0.054
	α_{asymp}	0.057	0.046	0.058	0.040	0.054
T_4	α_{perm}	0.046	0.054	0.053	0.041	0.054
	α_{asymp}	0.076	0.062	0.059	0.038	0.053
T_5	α_{perm}	0.046	0.054	0.053	0.041	0.054
	α_{asymp}	0.085	0.066	0.059	0.040	0.054
T_6	α_{perm}	0.046	0.054	0.053	0.041	0.054
	α_{asymp}	0.051	0.050	0.057	0.038	0.053

Table 2.23: Monte Carlo permutation and asymptotic type I error rates for the $Lt(0, 5)$ case under $p_{X_A} = p_{X_B}$. Nominal type I error rate $\alpha = 0.05$.

Table 2.24 shows Monte Carlo results for the $X_K \sim MD(0, 1, -3, 1, 0.999)$ case, i.e., for the small contamination scenario, for $K \in \{A, B\}$. The property of exactness for finite samples does not seem affected by a small contamination in the data.

		$n_A = n_B$				
		5	11	101	1001	5001
T_1	α_{perm}	0.057	0.049	0.038	0.053	0.043
	α_{asymp}	0.045	0.046	0.039	0.056	0.044
T_2	α_{perm}	0.050	0.047	0.038	0.052	0.043
	α_{asymp}	0.054	0.042	0.041	0.055	0.044
T_3	α_{perm}	0.052	0.047	0.038	0.052	0.043
	α_{asymp}	0.054	0.042	0.041	0.055	0.044
T_4	α_{perm}	0.058	0.052	0.039	0.053	0.043
	α_{asymp}	0.081	0.059	0.042	0.056	0.044
T_5	α_{perm}	0.058	0.052	0.039	0.053	0.043
	α_{asymp}	0.087	0.063	0.043	0.055	0.044
T_6	α_{perm}	0.058	0.052	0.039	0.053	0.043
	α_{asymp}	0.064	0.053	0.040	0.056	0.044

Table 2.24: Monte Carlo permutation and asymptotic type I error rates for the $MD(0, 1, -3, 1, 0.999)$ case under $p_{X_A} = p_{X_B}$. Nominal type I error rate $\alpha = 0.05$.

Table 2.25 shows Monte Carlo results for the $X_K \sim LGa(1,1)$ case, for $K \in \{A, B\}$. The property of exactness for finite samples seem to hold well even for this setting, that does not satisfy the assumptions of the key case presented in Section 2.3.

		$n_A = n_B$				
		5	11	101	1001	5001
T_1	α_{perm}	0.047	0.068	0.048	0.046	0.045
	α_{asymp}	0.016	0.033	0.047	0.047	0.046
T_2	α_{perm}	0.046	0.065	0.046	0.047	0.045
	α_{asymp}	0.065	0.099	0.110	0.123	0.114
T_3	α_{perm}	0.049	0.065	0.046	0.047	0.045
	α_{asymp}	0.065	0.099	0.115	0.123	0.114
T_4	α_{perm}	0.043	0.067	0.047	0.047	0.045
	α_{asymp}	0.071	0.077	0.052	0.048	0.046
T_5	α_{perm}	0.043	0.067	0.047	0.047	0.045
	α_{asymp}	0.128	0.127	0.118	0.123	0.113
T_6	α_{perm}	0.043	0.067	0.047	0.047	0.045
	α_{asymp}	0.050	0.064	0.051	0.048	0.046

Table 2.25: Monte Carlo permutation and asymptotic type I error rates for the $LGa(1,1)$ case under $p_{X_A} = p_{X_B}$. Nominal type I error rate $\alpha = 0.05$.

Chapter 3

Application to Microarray Data

3.1 An Introduction to Gene Differential Expression in Microarray Data

A gene is currently defined as "a locatable region of genomic sequence, corresponding to a unit of inheritance, which is associated with regulatory regions, transcribed regions, and or other functional sequence regions" Pearson (2006). Gene expression is the process by which information from a gene is used in the synthesis of a functional gene product, such as RNA or proteins. Depending on the specific cell they belong to, genes' expression is regulated in a different way, so that specific needs of the cell are satisfied. However, differences in expression might be due also, for example, to mutations in the genome or degenerative processes. The consequence of abnormally high (or low) gene expressions can result in pathologies (as in the case of deregulation of insuline expression) and in different responses to drugs and treatments. Therefore, gene expression studies are a crucial tool for understanding genes' participation in biological processes (Alberts et al., 2013), and to develop appropriate response to abnormal behaviours.

From a statistical point of view, a way to identify genes of potential interest (i.e., genes that express differently between two conditions) is to collect data regarding several genes from different samples and identify which ones are consistently either up- or down-regulated in the two groups. The groups might refer, for example, to mutant and wild-type genes, healthy and sick tissues, or different arms

of a clinical trial. DNA microarrays have been one of the technologies of choice for this task. After early attempts in the 1980s, a focused use for gene profiling is first reported in Schena et al. (1996). Since then, microarrays have been a major source of information about gene expression, providing in the last two decades crucial insights in important areas of biological investigation. Tailor-suited statistical tools have grown together with the development and diffusion of data from microarray. In fact, one of the most critical features of microarray experiments is the capability of measuring the expression of a large number of genes simultaneously, while the number of biological or technical replicates is usually kept quite small due to practical reasons and economic constraints. Situations where tens of thousands of genes are sampled over a few tens of biological or technical replicates are not infrequent. In this setting, the experimenter's goal is usually to identify a small subset of genes that show differential expression between the two groups of interest. The selected genes will then be the object of further biological analysis, impractical or impossible to be done over the original whole set of genes. Therefore, statistical methods are necessary to identify genes that exhibit a "significantly" different behaviour in the two conditions.

The so-called "small n, large p" context, where n is the number of replicates and p is the number of genes, together with the difficulty to find a shared statistical and biological definition of the complex concept of "differential expression", makes the development of suitable statistical tools a delicate matter. In fact, practice has shown that different methods, or even different versions of the same method, can lead to quite different results, i.e., to the identification of different sets of "differentially expressed" genes. As a consequence, statistical contributions have been numerous and diverse. The earliest and simplest methods used to deem a gene to be differentially expressed if its fold-change (the log ratio of the means in the two groups) exceeded a prefixed threshold such as 2 (Schena et al., 1996). Classical t-statistics and many subsequent modifications have improved the basic fold-change methods by incorporating statistical variability in the process, therefore providing inference and introducing the concept of "statistical significance" to the difference of gene expression between two groups of interest. Modifications of the basic t-test have often aimed at "moderating" the value of the statistic by artificially increasing the variability of the genes, i.e., by increasing the denominator

of the test statistic. The aim of this moderation is to reduce the significance of the genes which might report a low variability only by chance, and therefore might be identified as false positives in the search for differentially expressed genes. Among the most notable examples, the SAM procedure introduced a stabilisation factor for experimental variability (Tusher et al., 2001), later developed in the context of an empirical Bayesian model by Efron et al. (2001). The Bayesian perspective has been the instrument of choice also for Smyth et al. (2004), that developed an alternative Bayesian framework for the moderation problem. The research on the topic is still very active, and both modifications of the above methods and original solutions have been proposed. For a comparison of different alternatives we refer, for example, to Kooperberg et al. (2005), Jeffery et al. (2006) and Pan (2002).

However, we feel that some aspects of commonly used statistical procedures have not been fully explored: possibly the need to quickly obtain biologically meaningful results has sometimes made some of the properties of the test methods not sufficiently investigated. Moreover, as we have mentioned in Chapter 2, transformation of the data, which is often a routine step of these procedures, has often the consequence of transforming also the statistical hypotheses under investigation, but this phenomenon is not always acknowledged by researchers. Therefore, we would like to propose the statistical test statistics developed in Chapter 2 as alternatives to existing methods, capable of providing sound statistical inference, robustness of interpretation with respect to data transformation, and not less importantly, interpretable results. As a reference method, we will take the widespread SAM statistic (Tusher et al., 2001), one of the earliest and still most used tools. The choice is motivated by its extreme popularity in current literature: although it was proposed in 2001 and many other statistical procedures for microarray data analysis have been introduced since, the original SAM article has been cited more than 9000 times since its publication, and more than 400 times in 2014 alone, according to Google Scholar indexes at the moment of the writing. However, we underline a major difference between the SAM procedure and our proposed methods, i.e., the former one is specifically designed for multiple hypothesis testing, while the latter ones are univariate procedures. In the next Section, we provide an overview of the SAM procedure and point out some aspects of interest; in the following Sections, we compare the performance of some of our proposed test statistics to those of

SAM in simulated and real-data experiments.

3.2 Overview of the SAM procedure

Assume the setting of Chapter 2, i.e., let $X_{K1}, \ldots, X_{Kn_K}$ be two simple random samples of size n_K, independent from each other, from the density function p_{X_K}, for $K \in \{A, B\}$, and let $n = n_A + n_B$ be the total sample size. Let $Y_{K1}, \ldots, Y_{Kn_K}$ be the transformed samples on the Y scale, i.e., $Y_{Ki} = g_K(X_{Ki})$, with density function p_{Y_K}, mean μ_{Y_K} and variance $\sigma^2_{Y_K} < \infty$, for $i \in \{1, \ldots, n_K\}$ and $K \in \{A, B\}$. Moreover, let p be the number of genes under investigation. The SAM statistic is defined for the gene j as:

$$SAM_j = \frac{\overline{Y}_{Aj} - \overline{Y}_{Bj}}{S_j + S_0}, \qquad j = 1, \ldots, p$$

where

$$S_j = \sqrt{\frac{1}{n_A} + \frac{1}{n_B}} \times \sqrt{\frac{(n_A - 1)S^2_{Y_{Aj}} + (n_B - 1)S^2_{Y_{Bj}}}{n_A + n_B - 2}}, \qquad j = 1, \ldots, p$$

and S_0 is the so-called "fudge factor", added for the aim of moderation. In the expressions above, $\overline{Y}_{Kj}$ and $S^2_{Y_{Kj}}$ are the sample mean and the unbiased sample variance estimators for gene j, $j = 1, \ldots, p$, and group K, $K \in \{A, B\}$, respectively. The fudge factor is computed as a function of the empirical distribution of the S_j values, often in a numerical way. In the original paper, the fudge factor is chosen in such a way that minimises the coefficient of variation of the SAM_j statistics, but several different methods have been proposed, see for example Broberg (2002). The aim of the fudge factor is to artificially increase the variance of genes with very low variability, which might have happened only by chance due to low sample sizes. In fact, unstable estimates of the variance might provide an excessively high level of significance for some genes, in some cases producing a "false positive", i.e., a gene which is deemed by the procedure to be differentially expressed while it is actually not.

Of course, an explicit expression for the distribution of SAM_j is not available, neither for finite samples nor asymptotically. The Authors, therefore, resort to permutation methods. According to the proposed procedure, a gene is deemed to be differentially expressed if the difference between the observed value of SAM_j and its expected value computed over a number B of permuted samples is larger than some moving threshold Δ. For each value of Δ, the False Discovery Rate (FDR) proposed by Benjamini & Hochberg (1995) is computed, and the set of differentially expressed genes is chosen according to an "acceptable" value of the FDR. Permutation p-values can also be produced as a measure of statistical significance, and genes that show an (adjusted) p-value lower than a prefixed threshold such as $\alpha = 0.05$ are deemed to be differentially expressed.

We argue that the procedure arises few controversial points. Firstly, the null hypothesis under investigation is not explicitly stated, nor is it specified if data undergo some kind of transformation before hypothesis testing. Therefore, it is not immediate to compare the procedure to other methods testing equivalent hypotheses. In the following, we assume that the SAM procedure tests equality of the means after a logarithmic transformation, which seems a reasonable choice since the test statistic is based on a t-test statistic. Under this assumption, results obtained in the following for the SAM procedure are comparable to those provided in Chapter 2 of this Thesis. Although a permutation procedure requires the distribution in the two samples to be the same, the so-called randomisation hypothesis in Lehmann & Romano (2006), to guarantee control of type I error rates, the SAM procedure proves to be very robust with respect to deviations from this hypothesis. Its power depends on the underlying models, but in all cases approaches 1 for large enough sample sizes. We report results concerning type I error rates and power in Table 3.1 and Table 3.2, respectively, for the same models considered in Chapter 2.

A second point of concern is that the fudge factor is computed over the whole set of the genes. This is done with the aim of having as a moderation factor a quantity produced by taking into account all the genes present in the experiments, but can be also be a source of irreproducibility of results. In fact, a different data set might produce different value of the SAM_j statistic even if the observed data for the gene j were exactly the same, due to the presence of different genes in the

	$n_A = n_B$				
	5	11	101	1001	5001
$LN(0,1)$ vs $LN(0,5)$	0.0555	0.0498	0.0480	0.0499	0.0464
$LN(0,1)$ vs $Lt(0,5)$	0.0522	0.0511	0.0501	0.0485	0.0466
$LLog(0,1)$ vs $LU(-10,10)$	0.0617	0.0538	0.0529	0.0484	0.0532
$LN(0,1)$ vs $MD(0,1,-3,1,0.999)$	0.0548	0.0497	0.0528	0.0463	0.0498
$LLap(\log(2),1)$ vs $LGa(\log(2),1)$	0.0572	0.0551	0.0529	0.0505	0.0488

Table 3.1: Monte Carlo permutation type I error rates for the SAM procedure for the models of Chapter 2 under $\xi_{X_A}(1/2) = \xi_{X_B}(1/2)$. Nominal type I error rate $\alpha = 0.05$.

	$n_A = n_B$				
	5	11	101	1001	5001
$LN(0,1)$ vs $LN(-1,5)$	0.1521	0.2601	0.4594	0.8184	0.9827
$LN(0,1)$ vs $Lt(-1,5)$	0.2682	0.5570	0.8806	0.9980	1.0000
$LN(0,1)$ vs $Lt(-0.5,5)$	0.1143	0.2122	0.3761	0.7183	0.9556
$LLog(0,1)$ vs $LU(-10,5)$	0.2020	0.3868	0.6478	0.9616	0.9996
$LLap(2,1)$ vs $LGa(\log(2),1)$	0.2921	0.4964	0.7261	0.9733	0.9997

Table 3.2: Monte Carlo permutation power for the SAM procedure for the models of Chapter 2 under $\xi_{X_A}(1/2) = \xi_{X_B}(1/2)$.

experiment. It is possible to imagine cases where the set of shared genes between two experiments is quite limited: in that case, comparison of the statistics (and of their level of significance) runs the risk of producing ambiguous results.

One last, more extensive point, regards the choice of permutation methods used together with variance moderation in the sense proposed by the SAM procedure. In fact, although many different options have been proposed for the computation of S_0 (which might be a point of concern in itself, since, to the best of our knowledge,

shared guidelines do not seem to exist in the literature), they all share a potential issue. This is due to the fact that re-computation of S_0 in the permutation process is likely to yield quite similar values for each permutation, and therefore to produce levels of statistical significance quite close to those provided, by permutation, by a classical Student's t-test, or even by a plain mean difference estimator. To see why this is true, let us first prove that a permutation p-value obtained from a Student's t-test coincides with the permutation p-value that would be obtained from the absolute value of the mean difference statistic $|\overline{Y}_A - \overline{Y}_B|$. In this context, we refer to quantities computed over the original sample with the superscript or subscript *oss*, and to quantities computed over the permuted sample without any superscript or subscript: for example, T_{oss} denotes the observed value of the test statistic, and T the test statistic computed over the permuted sample. Therefore, we will prove that, for $T > T_{oss}$ to hold, it is sufficient that $|\overline{Y}_A - \overline{Y}_B| > |\overline{Y}_A^{oss} - \overline{Y}_B^{oss}|$ holds. Let us denote with BSS the sum of squares between groups and with WSS the sum of squares within groups, i.e.,

$$BSS = \sum_{K \in \{A,B\}} n_K (\overline{Y}_K - \overline{Y})^2 = \frac{n_A n_B}{n} (\overline{Y}_A - \overline{Y}_B)^2,$$

where $\overline{Y}$ is the mean of the pooled sample Y, and

$$WSS = \sum_{K \in \{A,B\}} \sum_{i=1}^{n_K} (Y_{Ki} - \overline{Y}_K)^2 = \frac{n_A n_B}{n} S^2.$$

Let us denote the total sum of squares with TSS, where $TSS = BSS + WSS$. Then, it is possible to write the squared T statistic as

$$T^2 = \left(\frac{\overline{Y}_A - \overline{Y}_B}{S} \right)^2 = \frac{BSS}{WSS},$$

which is the F-statistic interpretation in the context of the analysis of variance for two groups comparison. As a preliminary observation, we note that the total sum of squares is constant across permutation, i.e., $TSS = TSS_{oss} = c$. We will use this result in the following. Therefore, we prove that

$$T^2 > T_{oss}^2 \Longleftrightarrow BSS > BSS_{oss},$$

which is equivalent to prove that

$$T^2 > T_{oss}^2 \Longleftrightarrow |\overline{Y}_A - \overline{Y}_B| > |\overline{Y}_A^{oss} - \overline{Y}_B^{oss}|.$$

Proof:

$$
\begin{aligned}
& T^2 > T_{oss}^2 \\
\Longleftrightarrow & \frac{BSS}{WSS} > \frac{BSS_{oss}}{WSS_{oss}} \\
\Longleftrightarrow & \frac{BSS}{TSS - BSS} > \frac{BSS_{oss}}{TSS_{oss} - BSS_{oss}} \\
\Longleftrightarrow & \frac{BSS}{c - BSS} > \frac{BSS_{oss}}{c - BSS_{oss}} \\
\Longleftrightarrow & BSS(c - BSS_{oss}) > BSS_{oss}(c - BSS) \\
\Longleftrightarrow & cBSS - BSS_{oss}BSS > cBSS_{oss} - BSS_{oss}BSS \\
\Longleftrightarrow & cBSS > cBSS_{oss} \\
\Longleftrightarrow & BSS > BSS_{oss} \qquad \square
\end{aligned}
$$

We will prove that an analogous result holds when a positive constant, such as the fudge factor S_0, is added to the denominator of the T statistic. For the moment, we will assume that S_0 is constant across permutation, i.e., $S_0 = S_0^{oss}$, and we will prove that

$$SAM^2 > SAM_{oss}^2 \Longleftrightarrow BSS > BSS_{oss},$$

which is equivalent to prove that

$$SAM^2 > SAM_{oss}^2 \Longleftrightarrow |\overline{Y}_A - \overline{Y}_B| > |\overline{Y}_A^{oss} - \overline{Y}_B^{oss}|.$$

Proof:

$$
\begin{aligned}
& SAM^2 > SAM_{oss}^2 \\
\Longleftrightarrow & \frac{\frac{n}{n_A n_B} BSS}{\left(\sqrt{\frac{n}{n_A n_B} WSS} + S_0\right)^2} > \frac{\frac{n}{n_A n_B} BSS_{oss}}{\left(\sqrt{\frac{n}{n_A n_B} WSS_{oss}} + S_0\right)^2} \\
\Longleftrightarrow & \frac{\frac{n}{n_A n_B} BSS}{\frac{n}{n_A n_B} WSS + S_0^2 + 2S_0\sqrt{\frac{n}{n_A n_B}}\sqrt{WSS}} \\
& > \frac{\frac{n}{n_A n_B} BSS_{oss}}{\frac{n}{n_A n_B} WSS_{oss} + S_0^2 + 2S_0\sqrt{\frac{n}{n_A n_B}}\sqrt{WSS_{oss}}} \\
\Longleftrightarrow & \frac{BSS}{WSS + \frac{n_A n_B}{n} S_0^2 + 2S_0\sqrt{\frac{n_A n_B}{n}}\sqrt{WSS}} \\
& > \frac{BSS_{oss}}{WSS_{oss} + \frac{n_A n_B}{n} S_0^2 + 2S_0\sqrt{\frac{n_A n_B}{n}}\sqrt{WSS_{oss}}}
\end{aligned}
$$

$$
\begin{aligned}
&\iff WSS_{oss}BSS + \frac{n_A n_B}{n} S_0^2 BSS + 2S_0\sqrt{\frac{n_A n_B}{n}}\sqrt{WSS_{oss}}BSS \\
&\quad > WSSBSS_{oss} + \frac{n_A n_B}{n} S_0^2 BSS_{oss} + 2S_0\sqrt{\frac{n_A n_B}{n}}\sqrt{WSS}BSS_{oss} \\
&\iff cBSS - BSS_{oss}BSS \\
&\quad + \frac{n_A n_B}{n} S_0^2 BSS + 2S_0\sqrt{\frac{n_A n_B}{n}}\sqrt{WSS_{oss}}BSS \\
&\quad > cBSS_{oss} - BSSBSS_{oss} + \frac{n_A n_B}{n} S_0^2 BSS_{oss} \\
&\quad + 2S_0\sqrt{\frac{n_A n_B}{n}}\sqrt{WSS}BSS_{oss} \\
&\iff BSS\left(c + \frac{n_A n_B}{n} S_0^2 + 2S_0\sqrt{\frac{n_A n_B}{n}}\sqrt{WSS_{oss}}\right) \\
&\quad > BSS_{oss}\left(c + \frac{n_A n_B}{n} S_0^2 + 2S_0\sqrt{\frac{n_A n_B}{n}}\sqrt{WSS}\right) \\
&\iff \left(c + \frac{n_A n_B}{n} S_0^2\right)(BSS - BSS_{oss}) \\
&\quad + \left(2S_0\sqrt{\frac{n_A n_B}{n}}\right)(BSS\sqrt{WSS_{oss}} - BSS_{oss}\sqrt{WSS}) > 0.
\end{aligned}
$$

The last inequality implies that $BSS > BSS_{oss}$. We can prove this by contradiction. In fact, if it held that $BSS \leq BSS_{oss}$, then both addends in the last inequality would be non positive, therefore making the inequality not hold, and proving the initial statement by contradiction. In detail, the first term would be non positive as the quantity $c + \frac{n_A n_B}{n} S_0^2$ is a sum of positive quantities, and $BSS - BSS_{oss} \leq 0$ by assumption. The second term would also be non positive, since the quantity $2S_0\sqrt{\frac{n_A n_B}{n}} S_0^2$ is the product of positive quantities and the quantity $BSS\sqrt{WSS_{oss}} - BSS_{oss}\sqrt{WSS}$ is non positive. In fact, we have that $BSS\sqrt{WSS_{oss}} \leq BSS_{oss}\sqrt{WSS}$, under the assumption that $BSS - BSS_{oss} \leq 0$, which implies also $WSS_{oss} - WSS \leq 0$. Therefore, the above inequality holds, and we can write

$$SAM^2 > SAM_{oss}^2 \iff BSS > BSS_{oss}. \qquad \square$$

Therefore, adding a constant to the denominator has no effect on the permutation p-value of a Student's t-test (if this value is kept fixed across permutations), which remains equal to the permutation p-value of a raw mean difference test.

So far, we have assumed that S_0 was constant across permutations. This seems a reasonable approximation of what happens in real life experiments, where p is usually high and there is little reason to believe that a function of the distribution of the standard deviations changes substantially across permutations.

These considerations should make researchers cautious with employing the SAM procedure. Notwithstanding the good results obtained over a variety of different models generating the data, when the randomisation hypothesis does not hold, there is no formal guarantee that, unconditionally, the actual type I error rate provided by the SAM procedure is close to the nominal one even asymptotically. This is instead true for the T_j statistics, $j = 1, \ldots, 6$ introduced in Chapter 2 of the present Thesis, because of the results provided by Chung et al. (2013).

In the next Section, we will compare the test statistics introduced in Chapter 2, i.e., T_j, for $j = 1, \ldots, 6$, and $QT(\tau, q_A, q_B)$ for different values of the parameters, to the SAM procedure for simulated data that mimic actual microarray data. We will compare the different methods in terms of type I error rate control and power. We will also propose useful measures for evaluating the ranking of the genes.

3.3 Comparison of Procedures via Simulation Studies

We simulate data from a model that recreates with good approximation actual microarray gene expression data, following the directions of Kendziorski et al. (2003). Data are simulated according to a two-step Log-Normal-Normal hierarchical model. The first step generates the means of the different genes according to a Normal distribution with pre-specified parameters, therefore allowing for different genes to have different means, which is expected to happen with real data. The second step generates the actual expression data according to a Log-Normal distribution with the means generated at the previous step and pre-fixed variances. Formally, the first step generates the gene means according to:

$$\mu_{Y_K^j}|(\mu_0, \tau_0^2) \sim N(\mu_0, \tau_0^2), \qquad j = 1, \ldots, p, \qquad K \in \{A, B\}$$

and the second step, conditioned on the first one, generates the actual expression data according to:

$$Y_{Ai}^j|(\mu_{Y_A^j}, \sigma_0^2) \sim LN(\mu_{Y_A^j}, \sigma_0^2), \qquad i = 1, \ldots, n_A, \qquad j = 1, \ldots, p,$$
$$Y_{Bi}^j|(\mu_{Y_B^j}, \sigma_0^2) \sim LN(\mu_{Y_B^j}, \sigma_0^2), \qquad i = 1, \ldots, n_B, \qquad j = 1, \ldots, p.$$

Values for hyperparameters are fixed by following Chiogna et al. (2009), who chose based on real-life observations, and are defined as $\sigma_0^2 = 0.164$, $\tau_0^2 = 0.895$ and $\mu_0 = 7.9$. In this setting, a gene is set to be differentially expressed with probability $\pi \in (0, 1)$. In this case, the means in the two groups are generated independently from each other, i.e., $\mu_{Y_A^j} \neq \mu_{Y_B^j}$. For equivalently expressed genes, instead, the means are the same in the two groups, i.e., $\mu_{Y_A^j} = \mu_{Y_B^j}$.

3.3.1 Type I Error Rate Control

Following the setting of Chapter 2, we generate 10000 gene values for different sample sizes, and compute asymptotic and permutation p-values for the test statistics introduced in Chapter 2 and, for reference, for the SAM statistic. The permutation p-values are computed on the basis of $m = 999$ permuted samples. Data are generated assuming $\pi = 0$, i.e., that the null hypothesis holds for all the genes. The sample size settings are chosen as follows:

1. small sample sizes setting, i.e., $n_A = n_B = 11$;
2. strongly unbalanced sample sizes, i.e., $n_A = 101$; $n_B = 11$;
3. large sample sizes setting, i.e., $n_A = n_B = 101$.

Results are reported in Table 3.3. It seems that the type I error rate is consistently close to the nominal one for all the statistics, across the different sample sizes. Results based on asymptotic distributions seem to generally produce larger type I error rates than permutation ones, in particular for the low sample size setting and the strongly unbalanced setting. Among the $QT(1/2, q, q)$ statistics, as expected, the best results are provided when q is closer to 1/2. Results seem in general comparable with those obtained by SAM.

	n_A	11	101	101
	n_B	11	11	101
T_1	α_{perm}	0.0514	0.0523	0.0504
	α_{asymp}	0.0590	0.0653	0.0518
T_2	α_{perm}	0.0512	0.0518	0.0498
	α_{asymp}	0.0598	0.0656	0.0518
T_3	α_{perm}	0.0512	0.0518	0.0498
	α_{asymp}	0.0598	0.0656	0.0518
T_4	α_{perm}	0.0515	0.0438	0.0502
	α_{asymp}	0.0597	0.0556	0.0516
T_5	α_{perm}	0.0517	0.0434	0.0502
	α_{asymp}	0.0597	0.0553	0.0518
T_6	α_{perm}	0.0517	0.0434	0.0502
	α_{asymp}	0.0597	0.0553	0.0518
$QT(1/2, 0.10, 0.10)$	α_{perm}	0.0388	0.0383	0.0309
$QT(1/2, 0.25, 0.25)$	α_{perm}	0.0436	0.0447	0.0403
$QT(1/2, 0.40, 0.40)$	α_{perm}	0.0514	0.0457	0.0439
$QT(1/2, 0.45, 0.45)$	α_{perm}	0.0514	0.0475	0.0473
SAM	α_{perm}	0.0498	0.0496	0.0493

Table 3.3: Monte Carlo permutation and asymptotic type I error rates for different test statistics for the Log-Normal-Normal model with $\pi = 0$. Nominal type I error rate $\alpha = 0.05$.

3.3.2 Power

To investigate power, we choose $\pi = 1$, under the constraint that $\mu_{Y_A^j} > \mu_{Y_B^j}$, for $j = 1, \ldots, p$, in order to generate data under a unilateral alternative hypothesis. Before proceeding, it is worth noting that, given the structure of the model generating the data, the difference in the means between the two groups could be very little up to the point where they are undistinguishable from each other. This allows the setting to include differentially expressed genes that are harder than others to spot, which mimics well actual data. Let us call a true positive an actual differentially expressed gene which is deemed by the test statistic to be differentially expressed, and a true negative an actual equivalently expressed gene which is deemed by the test statistic to be equivalently expressed. Power of the statistics is then measured by the proportion of true positives among all the genes which are claimed to be differentially expressed by the test statistic.

As in the previous Subsection, we generate 10000 samples for different sample sizes, and compute asymptotic and permutation p-values for the test statistics introduced in Chapter 2 and, for reference, for the SAM statistic. The permutation p-values are computed on the basis of $m = 999$ permuted samples. The sample size settings are chosen as follows:

1. small sample sizes setting, i.e., $n_A = n_B = 11$;
2. strongly unbalanced sample sizes, i.e., $n_A = 101$; $n_B = 11$;
3. large sample sizes setting, i.e., $n_A = n_B = 101$.

Results are reported in Table 3.4. It seems that the power of the test statistics is very large, and quite similar to that of SAM for most settings. There do not seem to exist notable differences between the T_j statistics, for $j = 1, \ldots, 6$. On the contrary, $QT(1/2, q, q)$ statistics display a small power for low sample sizes when q is closer to 1/2.

	n_A	11	101	101
	n_B	11	11	101
T_1	α_{perm}	0.9230	0.9470	0.9741
	α_{asymp}	0.9277	0.9511	0.9739
T_2	α_{perm}	0.9229	0.9465	0.9742
	α_{asymp}	0.9273	0.9510	0.9741
T_3	α_{perm}	0.9229	0.9465	0.9742
	α_{asymp}	0.9273	0.9510	0.9741
T_4	α_{perm}	0.9229	0.9420	0.9741
	α_{asymp}	0.9275	0.9476	0.9739
T_5	α_{perm}	0.9229	0.9420	0.9741
	α_{asymp}	0.9272	0.9478	0.9742
T_6	α_{perm}	0.9229	0.9420	0.9741
	α_{asymp}	0.9272	0.9478	0.9742
$QT(1/2, 0.10, 0.10)$	α_{perm}	0.9011	0.9344	0.9666
$QT(1/2, 0.25, 0.25)$	α_{perm}	0.8953	0.9339	0.9673
$QT(1/2, 0.40, 0.40)$	α_{perm}	0.6009	0.9024	0.9660
$QT(1/2, 0.45, 0.45)$	α_{perm}	0.6009	0.8996	0.9627
SAM	α_{perm}	0.9120	0.8841	0.9563

Table 3.4: Monte Carlo permutation and asymptotic power for different test statistics for the Log-Normal-Normal model with $\pi = 1$. Threshold for significance $\alpha = 0.05$.

3.3.3 Ranking

Several considerations can be done regarding the use of p-values in the context of identifying differentially expressed genes. The first concerns the issue of multiple testing. It is well-known that the overall type I error rate, when considering testing simultaneously several null hypothesis, is not equal in general to the nominal level α. Many methods for dealing with this problem have been developed, from simple Bonferroni correction to control of the False Discovery Rate proposed by Benjamini & Hochberg (1995) and employed by the SAM procedure, to the min-p and max-T procedure proposed by Westfall & Young (1993) and implemented in the context of microarray data analysis by Dudoit et al. (2002). Although we do not face directly the issue of multiple testing in this Thesis, a starting point for extending our results could be given by Chung & Romano (2013), that provide useful results about Studentized statistics in the context of multiple testing. We also notice that two other specific issues arise in the context of permutation testing. The first one is that the lower bound of the p-value depends on the number m of permutations used by the procedure, i.e., is equal to $1/m$. Therefore, in order to obtain p-values small enough for the desired threshold, a very large number of permutations might be needed, which is not always computationally feasible. The second one is that there is a non-null probability that two genes share the exact same p-value, due to the fact that permutation p-value can take only one out of m possible values and generally $p \gg m$.

As a last aspect, let us consider another key specific feature of microarray data analysis. As we have mentioned before, usually researchers are interested in identifying a subset of differentially expressed genes. From this point of view, in addition to statistical significance, ranking on the genes might be also of interest. The SAM procedure itself incorporates the idea of ranking (Tusher et al., 2001), introducing a tuning parameter, Δ, which accounts for the number of false positives that are expected in the set of genes deemed to be differentially expressed. Starting from this idea, we propose some simple measures for the analysis of the ranking produced by the test statistics introduced in Chapter 2. First of all, we notice that a ranking based on the p-values is basically the same as a ranking based on the values of the test statistic themselves. In fact, if the p-value is computed based on a pivotal

quantity, then it is a strictly monotone transformation of the observed test statistic. Otherwise, if the p-value is computed according to a resampling procedure, it still can be seen as a monotone transformation of the observed test statistic, for a large enough m. Then, in both cases, the ranking of the genes produced by the p-values is equal to the one produced by ranking the test statistics themselves. This brings an alternative to p-values computation for obtaining quick results. We also notice that use of "meaningful" values as ranking measures, i.e., values on the scale of the data, rather than transformation with a mere statistical interpretation, such as the p-value, can provide also some benefits. In fact, for some genes the change in expression might be statistically significant, but not large enough, from a biological point of view, to be of actual interest for further investigation, while "biological significance" is easier to assess on the scale of the data. Moreover, problems of ties in the ranking, i.e., of genes with the same ranking score, are reduced if using the values of the test statistic as opposed to permutation p-values, for the reasons stated above. Of course, a ranking approach does not provide measures of statistical significance, but it could still be useful as an integrative tool for microarray data analysis (Boulesteix & Slawski, 2009). An example of integration of statistical and biological significance is provided by the volcano plots proposed by Cui & Churchill (2003).

In the following, we report some ranking-based measures that could be of interest for comparing different procedures. We refer, in particular, to the lowest, highest and average position held in the ranking by true differentially expressed gene, and average number of true and false positives (expressed as a ROC curve). Of course, these measures are valid only if the set of differentially expressed genes is known in advance, therefore is most suitable for simulation studies. We choose the same generating mechanism of the previous Subsection to simulate the data, choosing $\pi = 0.05$, so that a small percentage of the genes is differentially expressed, which is what is usually expected in real life experiments. Since, of course, a crucial aspect is choice of the threshold, which is as arbitrary as the choice of the significance level for a p-value, we report results for different thresholds. In real life experiments, the threshold can be defined a priori based on the number of differentially expressed genes that the investigator expects to find or is able to analyse afterwards. As in the previous Subsection, we generate 10000 samples for different

sample sizes, and compute the observed values of the test statistics introduced in Chapter 2 and, for reference, of the SAM statistic. We repeat the experiment 100 times and average the results, and compute average measures over the 100 replications. We notice that, since the ranking is computed according to the magnitude of the test statistic, a gene with a larger value of the test statistic will have a lower ranking than one with a smaller value of the test statistic. The sample size settings are chosen as follows:

1. small sample sizes setting, i.e., $n_A = n_B = 11$;
2. strongly unbalanced sample sizes, i.e., $n_A = 101$; $n_B = 11$;
3. large sample sizes setting, i.e., $n_A = n_B = 101$.

Table 3.5 contains the average over the 100 replications of the lowest rank among differentially expressed genes. All of the test statistics almost always put a truly differentially expressed gene in the first position of the ranking.

average lowest rank	n_A	11	101	101
	n_B	11	11	101
T_1		1.00	1.00	1.00
T_2		1.00	1.00	1.00
T_3		1.00	1.00	1.00
T_4		1.00	1.00	1.00
T_5		1.00	1.00	1.00
T_6		1.00	1.00	1.00
$QT(1/2, 0.10, 0.10)$		1.00	1.00	1.00
$QT(1/2, 0.25, 0.25)$		1.00	1.00	1.00
$QT(1/2, 0.40, 0.40)$		1.03	1.00	1.00
$QT(1/2, 0.45, 0.45)$		1.03	1.00	1.00
SAM		1.00	1.00	1.00

Table 3.5: Monte Carlo average lowest rank of differentially expressed genes for different test statistics for the Log-Normal-Normal model with $\pi = 0.05$.

Table 3.6 contains the average over the 100 replications of the median rank among differentially expressed genes. All of the test statistics, apart from the $QT(\tau, q, q)$ statistic with $q \in \{0.40, 0.45\}$ for the low sample size scenario, display a value very close to 250, meaning that, on average, half of the differentially expressed genes are in the top 250 positions of the ranking. Since the expected total number of differentially expressed genes is equal to $\pi p = 500$, it seems that the top 250 positions of the rank are on average held by truly differentially expressed genes. The effect of the sample size is almost null for most statistics, but for example for $QT(1/2, 0.45, 0.45)$ it can be seen that the median indicator approaches 250 with larger sample sizes. It is worth noting that the variability of the indicator across different test statistics is barely perceivable both for Studentized and for pseudo-Studentized statistics. For the latter, best results seem to be obtained for low values of q_A and q_B, as it happened in the power context.

average median rank	n_A	11	101	101
	n_B	11	11	101
T_1		250.57	250.94	250.14
T_2		250.57	250.94	250.14
T_3		250.57	250.94	250.14
T_4		250.57	250.94	250.14
T_5		250.57	250.94	250.14
T_6		250.57	250.94	250.14
$QT(1/2, 0.10, 0.10)$		250.92	250.94	250.14
$QT(1/2, 0.25, 0.25)$		251.94	250.94	250.14
$QT(1/2, 0.40, 0.40)$		324.76	250.98	250.14
$QT(1/2, 0.45, 0.45)$		324.76	251.71	250.14
SAM		250.57	250.94	250.10

Table 3.6: Monte Carlo average median rank of differentially expressed genes for different test statistics for the Log-Normal-Normal model with $\pi = 0.05$.

Table 3.7 contains the average over the 100 replications of the highest rank

among differentially expressed genes. Results are slightly more variable than for the previous measures, but still quite similar across different statistics, with worse results provided by the $QT(\tau, q, q)$ statistic with $q \in \{0.40, 0.45\}$ for the low sample size scenario. The average highest rank is in general very large (over 8000), meaning that at least one differentially expressed gene has a quite small difference in means between the two groups and is therefore hard to find by all of the test statistics, SAM included. As the sample sizes increase, as expected, it seems that the average highest rank of differentially expressed genes tends to decrease, i.e., even the genes with highest ranking tend to lower their position.

average highest rank	n_A	11	101	101
	n_B	11	11	101
T_1		9157.33	8871.85	8046.76
T_2		9160.61	8860.68	8045.75
T_3		9160.61	8860.68	8045.75
T_4		9158.24	8868.49	8046.91
T_5		9156.72	8869.17	8047.32
T_6		9156.69	8869.17	8047.32
$QT(1/2, 0.10, 0.10)$		9200.37	9171.45	8441.58
$QT(1/2, 0.25, 0.25)$		9236.62	9162.57	8460.13
$QT(1/2, 0.40, 0.40)$		9436.39	9138.12	8424.39
$QT(1/2, 0.45, 0.45)$		9436.39	9149.28	8453.56
SAM		9144.70	8843.05	8044.71

Table 3.7: Monte Carlo average highest rank of differentially expressed genes for different test statistics for the Log-Normal-Normal model with $\pi = 0.05$.

In addition to the above presented measures, in Figure 3.1 we report Receiver Operating Characteristic (ROC) curves for the three different sample sizes settings. ROC curves display the number of true and false positives identified by the statistics for different thresholds. All of the curves are overlapping for most values, meaning that the ranking that they provide is quite similar. Moreover, all the curves are all quite steep, i.e., the number of false positives decrease very slowly when the number of true positives increases. This seems to suggest that the ranking procedure is quite effective for identifying differentially expressed genes for the proposed model.

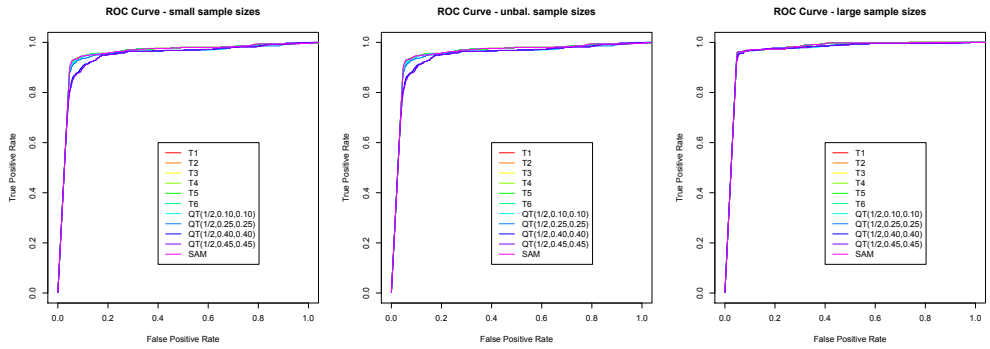

Figure 3.1: Left to right: ROC curves relative to the $n_A = n_B = 11$ setting, to the $n_A = 101$, $n_B = 11$ setting, and to the $n_A = n_B = 101$ setting.

It seems that ranking the genes according to the magnitude of the test statistics produces quick and accurate results, and it could be an interesting tool for comparing different methods for identifying differentially expressed data. Of course, for an evaluation of this kind to be possible, the set of truly differentially expressed genes should be known a priori. In the next Section, we will provide application of Studentized and pseudo-Studentized test statistics to real data from microarray.

3.4 Application to Real Data from Microarray

In this Section, we compare the capacity to identify well-known differentially expressed genes by the Studentized test statistics T_j, for $j = 1, \ldots, 6$, by pseudo-Studentized test statistics $QT(\tau, q_A, q_B)$ and, for comparison, by the SAM procedure, by means of a real microarray dataset. The chosen dataset was published by Chiaretti et al. (2005) and contains gene expression from microarray experiments conducted on sample cells of patients with acute lymphocytic leukemia (ALL), which are associated with known genotypic abnormalities in adult patients. Working expression data appropriately are already normalised according to robust multiarray analysis and quantile normalisation (Martini et al., 2013). The dataset contains 37 observations from patient with the so-called "chimeric" BCR/ABL gene rearrangement, linked with ALL, and 41 observations from patients without the rearrangement, over 8384 genes in total. Our aim is to check if, according to our test statistic, the BCR/ABL gene is identified as differentially expressed between the two groups of patients, as expected, and to compare the overall set of differentially expressed genes identified by our proposed statistics and by SAM.

Table 3.8 contains the observed p-value for the "chimera" (gene BCR/ABL), and its ranking across the whole set of genes, for the statistics of interest. Permutation p-values are computed over $m = 999$ permutations. All of the proposed methods identify the chimera as statistically significant, and the majority of them place it first in the ranking computed according to the magnitude of the test statistic. Test statistics T_5, T_6 and $QT(1/2, q)$, for $q \in \{0.25, 0.40, 0.45\}$ rank the gene in a top position, however not in the first one. Results are very similar to those obtained by the SAM procedure both in terms of statistical significance and of ranking (the statistical significance computed by SAM for the chimeric gene is 0 up to the third decimal place).

Since a larger list of true differentially expressed genes is not available in this context, we focus on comparison of results provided by our test statistics with those provided by SAM, in terms of ranking, i.e., we compute the percentage of overlap in the top-ranking provided by our test statistics and the SAM procedure. Figure 3.2 contains the results of this comparison. On the horizontal axis, thresholds

	permutation p-value	rank
T_1	0.001	1
T_2	0.001	1
T_3	0.001	1
T_4	0.001	1
T_5	0.001	2
T_6	0.001	2
$QT(1/2, 0.10, 0.10)$	0.001	1
$QT(1/2, 0.25, 0.25)$	0.001	3
$QT(1/2, 0.40, 0.40)$	0.007	15
$QT(1/2, 0.45, 0.45)$	0.025	8
SAM	0.000	1

Table 3.8: Permutation p-values and rank position for the chimeric gene for different test statistics.

for the identification of differentially expressed genes are reported, ranging from 1 to 100. On the vertical axis, the length of the intersection of the top rankings (rankings above the threshold) is reported as a fraction of the length of the top ranking, for the considered thresholds. The dashed line, drawn as a reference, indicates complete overlap at all thresholds, i.e., identical overall ranking. Solid lines represent results for the different test statistics. Results show how the Studentized test statistics produce a ranking more similar to the ranking provided by SAM, than the pseudo-Studentized ones for most thresholds, with T_4 resulting the most similar to SAM. However, we remark that the true set of differentially expressed genes is not known, so that Figure 3.2 is only a way to compare our statistics to SAM, not to investigate their efficacy in general. This implies, for example, that pseudo-Studentized statistics produce a ranking which is quite different than the one produced by SAM, which means that they propose different genes as candidates for differential expression. Since simulation studies reported good results also for the pseudo-Studentized statistics, it could be worth to consider inclusion also of these genes (which would not be identified by SAM) in further analysis.

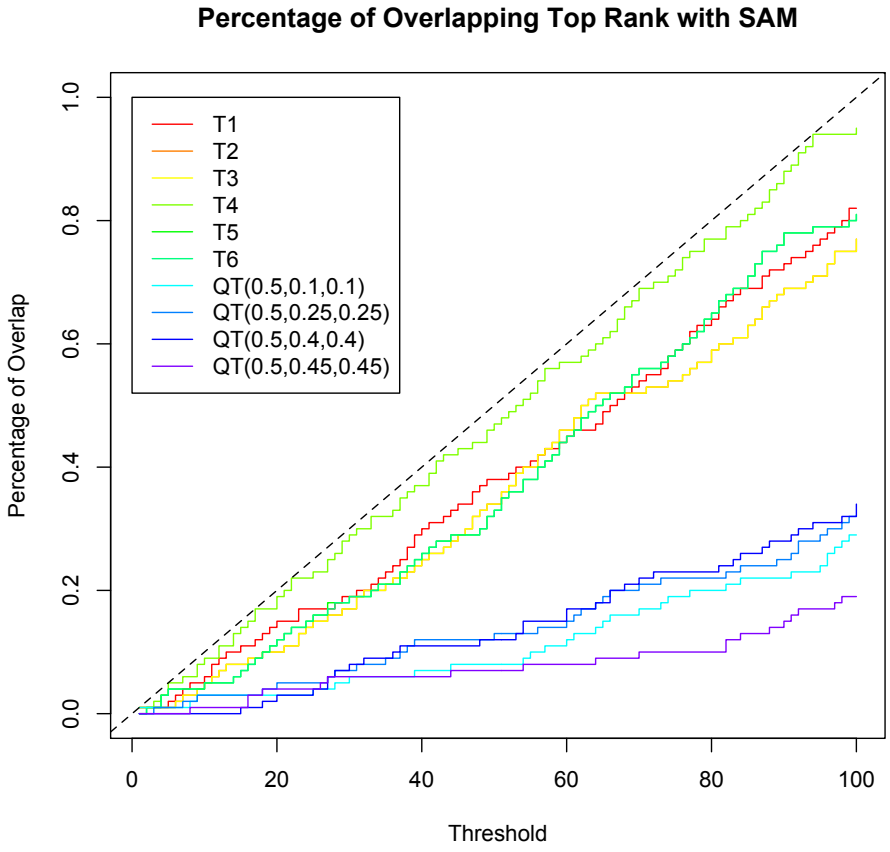

Figure 3.2: Percentage of overlap of top ranking for different thresholds for different test statistics compared to the SAM procedure.

3.5 Final Remarks

In this Chapter, we have applied our novel Studentized and pseudo-Studentized test statistics introduced in Chapter 2 to simulated data and real data from microarray, and compared them with the performance of the SAM procedure. The following observations seem to have emerged:

1. simulation experiments from a Log-Normal-Normal model that mimics well data from microarray show that both Studentized and pseudo-Studentized test statistics give satisfactory results (comparable to those of SAM) both in terms of control of type I error rate and power for different sample sizes, including strongly unbalanced ones;

2. synthesis measures based on the ranking of the genes have been proposed, that show how ranking the genes according to the magnitude of the test statistics is a procedure which is able to discriminate well differentially expressed genes from the rest, both for Studentized and most pseudo-Studentized statistics;

3. application to real data has shown the efficacy of the proposed statistics in identifying a specific gene, known to be differentially expressed between the two conditions considered, and overall agreement with the SAM procedure on the top ranking of differentially expressed genes for different thresholds (more for the Studentized than for the pseudo-Studentized statistics).

In the next Chapter, we will propose application of our pseudo-Studentized statistics to a different kind of data, i.e., data from sequencing following chromatin-immunoprecipitation (ChIP-Seq). The basic idea is to briefly illustrate an example of different application of the pseudo-Studentized statistics, to show how they are not strictly confined to the context of microarray data analysis.

Chapter 4

Application to ChIP-Seq Data

4.1 An Introduction to ChIP-Seq Data

The so-called "ChIP-Seq" technology takes its name from the process of Chromatin-ImmunoPrecipitation followed by Sequencing (Barski et al., 2007; Johnson et al., 2007). One of the main aims of the procedure is to identify binding sites of transcription factors of interest, i.e., how a transcription factor is deployed across the genome for a given cell. In order to investigate this, the ChIP-Seq goes through the following steps to produce data: the first step includes using an immune reagent specific for a DNA binding factor to enrich target DNA sites. After immunoprecipitation, a large number of "short reads", i.e., short fragments of genetic material, are collected, and mapped to the reference genome. Then, for each chromosome, the number of reads mapping to every genomic location is counted. In this way, the raw data for identifying transcription sites are the counts of how many reads map to any single location in the genome. Usually, the interest is not in a single genomic position, instead counts are considered together for a continuous genomic region with at least one read at each position, called "island". After having pooled reads into islands, biologists are often interested in identifying "peaks", i.e., islands where the number of mapped reads is particularly large. These regions are usually a few kilobase-pairs long, but in some cases, like for histone modifications analysis, can also be much longer. In order to identify peaks along the chromosome, which might correspond to location of binding of the transcription factor, a reference sample is usually needed. This is often a control sample that did not

undergo immunoprecipitation, or that was immunoprecipitated with a generic antibody. Comparison between the ChIP sample and the reference makes it possible to evaluate if an island is significantly enriched in the ChIP sample with respect to the reference. After identification of significant peaks, annotation procedures link them to biological areas of interest, so that knowledge of gene functions and biological pathways can be updated. Of course, this setting gives wide opportunities for application and development of statistical methods, in particular concerning the search for statistically significant peaks.

In the last years, many statistical approaches have been developed in order to identify peaks and compute their statistical significance. A review of methods for peak calling, most of which are based on modifications of Poisson models, is provided by Wilbanks & Facciotti (2010b). An interesting aspect brought to light by the Authors is the very poor agreement of different methods in identifying peaks, which underlines the complexity of the problem and the lack of shared guidelines in such a new and challenging topic. Among the most popular methods, it is worth mentioning the CisGenome algorithm (Ji et al., 2011), the MACS procedure (Zhang et al., 2008), and the spp model (Kharchenko et al., 2008). A more complete discussion about challenges posed by analysis of ChIP-Seq data might be found in Park (2009). In the following, we will briefly overview some of the statistical aspects involved with the identification of peaks in ChIP-Seq data. We will limit the description to some key aspects that will be employed in the next Section, dedicated to application of our pseudo-Studentized statistics to real data from ChIP-Seq. In this context, when the research hypothesis can be phrased in terms of counts comparison between two groups, and therefore be translated into a statistical hypothesis concerning two distributions, our pseudo-Studentized test statistics can be suitably applied. We remark the fact that, in this context, our proposed Studentized statistics like T_j, for $j \in \{1, \dots, 6\}$ do not seem appropriate, since no result is available concerning relationship between parameters of the data (such as symmetry) or of some transformation of them, similarly to what happened with microarray data. We stress again that this Chapter aims only at showing a possible different field of application for the pseudo-Studentized test statistics we introduced in Chapter 2, with data different from microarray. Therefore we will not go too far in the detail, and leave some consideration to the last Section of

this Chapter. Before showing an example of an application of pseudo-Studentized statistics in the next Section, we point out some aspects of interest specific of ChIP-Seq data and related to the test statistics we want to apply:

1. a general point of interest is how to compare two islands that are not entirely overlapping between the sample of interest and the reference, which is in general expected to happen. Several solutions have been proposed, including use of the union or the intersection of the islands from both samples. We feel that using the intersection might be an appropriate solution in order not to have two possibly very unbalanced sample sizes; however, in the following we also report results obtained for the union of the two islands (on non-overlapping regions, and therefore for samples of different sizes). We remark that the union is intended in the sense of considering the specific island for each sample, not in the sense of extending the shorter island to the region of the wider one. In this way, we avoid the introduction of many zero counts in the sample with a shorter island;

2. data are counts. Therefore, before applying our pseudo-Studentized statistics, they must undergo some transformation such that they are on a continuous scale. For this task, several solutions exist. We choose to take the jittering approach proposed by Machado & Silva (2005), which consists in constructing a continuous random variable whose quantiles have a one-to-one relation with those of X_K, for $K \in \{A, B\}$. Such a variable can be built by adding to the raw counts, X_A and X_B, standard Uniform random variables, say U_A and U_B. The uniform distribution is chosen for simplicity, even if any continuous distribution with support on the interval $(0, 1)$ could be used instead. The test statistics are then computed on the transformed variables X'_A and X'_B;

3. data have a strong spatial correlation across the genome. In the following, we will not take this correlation into account, and treat the data from each sample as independent realisations from an underlying generating model. This is clearly a simplification, but it is worth noting that many other methods consider only the sum of the counts for a specific island in the process of peak calling, therefore not taking into account spatial correlation either.

Having said this, we will show how the test statistic $QT(\tau, q_A, q_B)$ can be computed, for an island of interest for different values of $\tau \in (0, 1)$ and of $q_K \in \{0, 1/2\}$, for $K \in \{A, B\}$, and its significance computed via permutation. In the next Section, we will report a brief example of application.

4.2 Analysis of Data from ChIP-Seq

We apply the pseudo-Studentized statistics introduced in Chapter 2 to a real dataset, made public by Chen et al. (2008) and partially included in the `R` library `chipseq`, containing ChIP-Seq data for the mouse genome. The `R` dataset contains a subset of the original data, over chromosomes 10, 11 and 12, for a sample of interest (`ctcf`) and a reference sample (`gfp`). For illustration purposes, we will focus on comparison of two specific islands on chromosome 11. The islands we take into account range from position 3045392 to 3047368, therefore with a width of 1977, for the sample of interest, and from position 3045752 to 3046466, therefore with a width of 715, for the reference sample. Hence the second island is a subset of the first one, and we can either work on the union or on the intersection of the two islands. In the latter case, we obtain two samples of equal size 715. In order to apply our test statistics, the next step is applying the jittering procedure that we have mentioned in the previous Section. Figure 4.1 shows how the jittering does not alter substantially the distribution of the data; Figure 4.2 shows how the same holds also for data over the union of the islands.

In the following, we report statistical significance obtained via permutation for the pseudo-Studentized test statistics computed on the two samples (on the jittered data). For comparison, we report also the significance that would be obtained by a classic t-test on the jittered data, and by a Poisson and a Negative Binomial model for two groups comparison on the count data. Pseudo-Studentized statistics are computed for $q_A = q_B \in \{0.10, 0.25, 0.40, 0.45\}$, in order to evaluate the influence of the q_A and q_B parameters on the results, and for values of

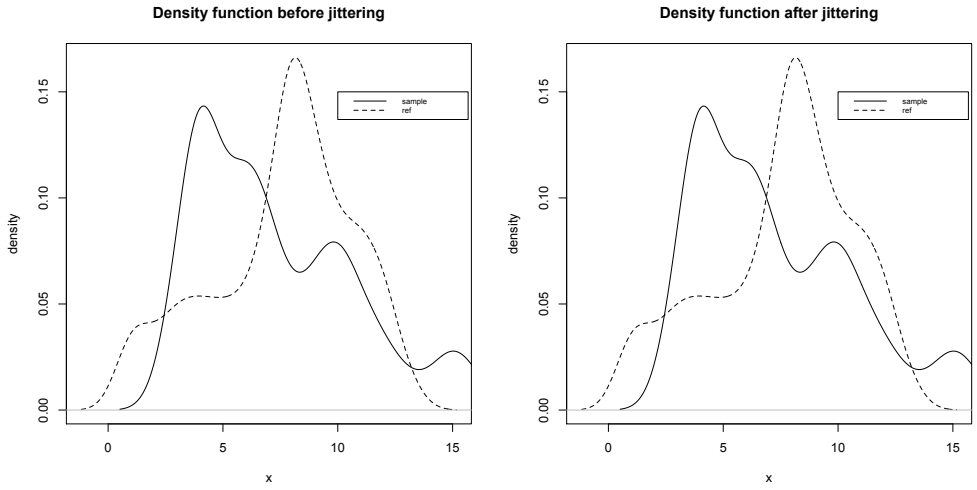

Figure 4.1: Probability density function before (left) and after (right) jittering for the sample of interest (solid line) and the reference (dashed line) for the data on the intersection of the islands.

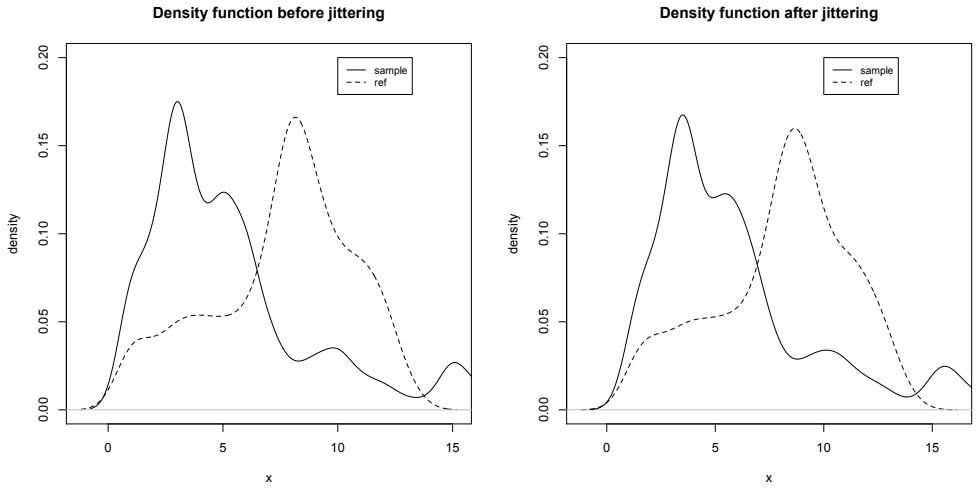

Figure 4.2: Probability density function before (left) and after (right) jittering for the sample of interest (solid line) and the reference (dashed line) for the data on the union of the islands.

$\tau \in \{0.10, 0.25, 0.50, 0.75, 0.90\}$, to investigate differences at different points of the distributions. Moreover, results are reported both for the union ($n_A = 1977$;

$n_B = 715$) and the intersection ($n_A = n_B = 715$) of the islands. The alternative hypothesis is unilateral, as we expect the counts to be higher for the sample of interest than in the reference sample. Table 4.1 contains the results. The effect of q_A and q_B seems almost non-existent in this case, while the statistics computed at different values of τ have very different significance levels. In fact, for the data based on the union of the islands, only differences at the 0.1- and at the 0.9-level quantiles are significant, while the ones at the quartiles and at the median are not significant. It is interesting to note how these differences would not be spotted by a classic t-test, or via a Poisson or Negative Binomial model, being located in an area of the distribution which is far from the mean. The same computation over the data based on the intersection of the islands data, for which the sample sizes are heavily unbalanced between the two samples, brings very similar results for all the considered quantiles except for the 0.1-level, where the test statistic becomes non significant possibly due to the fact that the difference between the islands, excluded from the intersection, refers to the tails of the peak, i.e., to counts that are mostly 1, which influences mainly the lower quantiles. Since a large number of counts equal to 1 is expected for any island, one could consider excluding those counts from computation in further analysis.

4.3 Final Remarks

In this Chapter, we have briefly introduced ChIP-Seq data analysis, with a focus on identifying differences in peaks between two groups. Although the one presented is a very simple example, it seems that there is actually space for application of our pseudo-Studentized statistics in the field of ChIP-Seq data; in particular, in application on the mouse data, pseudo-Studentized statistics were capable of identify a statistically significant difference on the 0.90 quantile of the two distributions, that a classic test on the means of the two groups would fail to notice. This brings support to the idea that use of the proposed statistic can contribute to the analysis of data from different sources, and it is by no means limited to the context of analysis of data from microarray.

test statistic	p-value data from union	p-value data from intersection
$QT(0.10, 0.10, 0.10)$	0.001	1.000
$QT(0.10, 0.25, 0.25)$	0.001	1.000
$QT(0.10, 0.40, 0.40)$	0.001	1.000
$QT(0.10, 0.45, 0.45)$	0.001	1.000
$QT(0.25, 0.10, 0.10)$	1.000	1.000
$QT(0.25, 0.25, 0.25)$	1.000	1.000
$QT(0.25, 0.40, 0.40)$	1.000	1.000
$QT(0.25, 0.45, 0.45)$	1.000	1.000
$QT(0.50, 0.10, 0.10)$	1.000	1.000
$QT(0.50, 0.25, 0.25)$	1.000	1.000
$QT(0.50, 0.40, 0.40)$	1.000	1.000
$QT(0.50, 0.45, 0.45)$	1.000	1.000
$QT(0.75, 0.10, 0.10)$	0.520	1.000
$QT(0.75, 0.25, 0.25)$	0.520	1.000
$QT(0.75, 0.40, 0.40)$	0.520	1.000
$QT(0.75, 0.45, 0.45)$	0.528	1.000
$QT(0.90, 0.10, 0.10)$	0.001	0.020
$QT(0.90, 0.25, 0.25)$	0.001	0.002
$QT(0.90, 0.40, 0.40)$	0.001	0.001
$QT(0.90, 0.45, 0.45)$	0.001	0.001
t-test	0.911	1.000
Poisson model	0.933	1.000
Negative Binomial model	0.893	1.000

Table 4.1: p-values for pseudo-Studentized test statistics for jittered data from ChIP-Seq.

Chapter 5

Conclusions

In this Thesis, we have explored some possible approaches to inference on quantiles in the context of genomic data analysis. In Chapter 2, we have developed novel statistical tools for testing statistical hypothesis when relationships exist between different parameters of the distributions, therefore making it possible to exploit the mean estimator to test hypothesis on quantiles of the distributions. Following the approach of Chung et al. (2013), we have developed Studentized statistics for quantile comparison, which do not require estimation of the density function at the quantile of interest, like classic test statistics on quantiles would. The permutation p-value of the proposed statistics coincides asymptotically with the true unconditional one, though retaining the exactness property for finite samples when the two samples have the same distribution. We have also proposed pseudo-Studentized test statistics, which are an approximation of the above mentioned ones, with a simple structure and easy interpretation. Although the test statistics are developed in the context of analysis of genomic data, their use is quite general, and could be easily extended to the analysis of different kinds of data. Further development of the methodology could include comparison of different groups, or extensions to regression models. Moreover, extensions to the context of multiple testing could be an area of further investigation, possibly based on the results provided by Chung & Romano (2013). In Chapter 3, we have applied the proposed statistics to data from microarray, both in simulation experiments and with a real dataset, and their performances in terms of identification of correct identification of differentially expressed genes have proven at least comparable with those of the most popular

method in this context, the SAM procedure proposed by Tusher et al. (2001). In Chapter 4, we have applied pseudo-Studentized statistics to data from ChIP-Seq, obtaining interesting results when comparing peaks for different samples. In this context, the ability of quantile-based statistics to investigate different aspects of the distribution of the data seems a desirable feature to discover differences not detected by traditional methods devoted to comparison of the means of the distributions. Overall, the results obtained in this Thesis seem promising, and further investigation both from a methodological point of view and in application to genomic data and other kinds of data might widen the result and incorporate them in larger statistical frameworks of analysis.

References

ALBERTS, B., BRAY, D., HOPKIN, K., JOHNSON, A., LEWIS, J., RAFF, M., ROBERTS, K. & WALTER, P. (2013). *Essential cell biology*. Garland Science.

BAHADUR, R. R. (1966). A note on quantiles in large samples. *Annals of Mathematical Statistics* 37 577–580.

BARSKI, A., CUDDAPAH, S., CUI, K., ROH, T.-Y., SCHONES, D. E., WANG, Z., WEI, G., CHEPELEV, I. & ZHAO, K. (2007). High-resolution profiling of histone methylations in the human genome. *Cell* 129 823–837.

BENJAMINI, Y. & HOCHBERG, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)* 289–300.

BOULESTEIX, A. & SLAWSKI, M. (2009). Stability and aggregation of ranked gene lists. *Briefings in bioinformatics* 10 556–568.

BROBERG, P. (2002). Ranking genes with respect to differential expression. *Genome Biology* 3 1–0007.

CHEN, X., XU, H., YUAN, P., FANG, F., HUSS, M., VEGA, V. B., WONG, E., ORLOV, Y. L., ZHANG, W., JIANG, J. ET AL. (2008). Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. *Cell* 133 1106–1117.

CHIARETTI, S., LI, X., GENTLEMAN, R., VITALE, A., WANG, K. S., MANDELLI, F., FOÀ, R. & RITZ, J. (2005). Gene expression profiles of b-lineage adult acute lymphocytic leukemia reveal genetic patterns that identify lineage derivation and distinct mechanisms of transformation. *Clinical cancer research* 11 7209–7219.

CHIOGNA, M., MASSA, M. S., RISSO, D. & ROMUALDI, C. (2009). A comparison on effects of normalisations in the detection of differentially expressed

genes. *BMC bioinformatics* 10 61.

CHUNG, E. & ROMANO, J. P. (2013). Multivariate and multiple permutation tests. Tech. rep., Technical report, Stanford University.

CHUNG, E., ROMANO, J. P. ET AL. (2013). Exact and asymptotically robust permutation tests. *The Annals of Statistics* 41 484–507.

CUI, X. & CHURCHILL, G. A. (2003). Statistical tests for differential expression in cdna microarray experiments. *Genome Biol* 4 210.

CURRY, E. W., STRONACH, E. A., RAMA, N. R., WANG, Y. Y., GABRA, H. & EL-BAHRAWY, M. A. (2013). Molecular subtypes of serous borderline ovarian tumor show distinct expression patterns of benign tumor and malignant tumor-associated signatures. *Modern pathology : an official journal of the United States and Canadian Academy of Pathology, Inc* 1–10URL `http://www.ncbi.nlm.nih.gov/pubmed/23948749`.

DASGUPTA, A. & HAFF, L. (2006). Asymptotic values and expansions for the correlation between different measures of spread. *Journal of statistical planning and inference* 136 2197–2212.

DELAIGLE, A., HALL, P. & JIN, J. (2011). Robustness and accuracy of methods for high dimensional data analysis based on Student's t-statistic. *J. R. Stat. Soc. Ser. B Stat. Methodol.* 73 283–301. URL `http://dx.doi.org/10.1111/j.1467-9868.2010.00761.x`.

DRAPER, N. & COX, D. (1969). On distributions and their transformation to normality. *Journal of the Royal Statistical Society. Series B (* ... 31 472–476. URL `http://www.jstor.org/stable/2984350`.

DUDOIT, S., YANG, Y. H., CALLOW, M. J. & SPEED, T. P. (2002). Statistical methods for identifying differentially expressed genes in replicated cdna microarray experiments. *Statistica sinica* 12 111–140.

EFRON, B. (1979). Bootstrap methods: another look at the jackknife. *The annals of Statistics* 7 1–26.

EFRON, B., TIBSHIRANI, R., STOREY, J. & TUSHER, V. (2001). Empirical bayes analysis of a microarray experiment. *Journal of the American Statistical Association* 96 1151–1160.

HALL, P., DICICCIO, T. J. & ROMANO, J. P. (1989). On smoothing and the bootstrap. *The Annals of Statistics* 17 692–704.

JEFFERY, I., HIGGINS, D. & CULHANE, A. (2006). Comparison and evaluation of

methods for generating differentially expressed gene lists from microarray data. *Bmc Bioinformatics* 7 359.

JI, H., JIANG, H., MA, W. & WONG, W. H. (2011). Using cisgenome to analyze chip-chip and chip-seq data. *Current Protocols in Bioinformatics* 2–13.

JOHNSON, D. S., MORTAZAVI, A., MYERS, R. M. & WOLD, B. (2007). Genome-wide mapping of in vivo protein-dna interactions. *Science* 316 1497–1502.

KENDZIORSKI, C., NEWTON, M., LAN, H. & GOULD, M. (2003). On parametric empirical bayes methods for comparing multiple groups using replicated gene expression profiles. *Statistics in medicine* 22 3899–3914.

KHARCHENKO, P. V., TOLSTORUKOV, M. Y. & PARK, P. J. (2008). Design and analysis of chip-seq experiments for dna-binding proteins. *Nature biotechnology* 26 1351–1359.

KOENKER, R. & BASSETT JR, G. (1978). Regression quantiles. *Econometrica: journal of the Econometric Society* 33–50.

KOOPERBERG, C., ARAGAKI, A., STRAND, A. & OLSON, J. (2005). Significance testing for small microarray experiments. *Statistics in medicine* 24 2281–2298.

LAMBERT, S. R., WITT, H., HOVESTADT, V., ZUCKNICK, M., KOOL, M., PEARSON, D. M., KORSHUNOV, A., RYZHOVA, M., ICHIMURA, K., JABADO, N., FONTEBASSO, A. M., LICHTER, P., PFISTER, S. M., COLLINS, V. P. & JONES, D. T. W. (2013). Differential expression and methylation of brain developmental genes define location-specific subsets of pilocytic astrocytoma. *Acta neuropathologica* 126 291–301. URL `http://www.ncbi.nlm.nih.gov/pubmed/23660940`.

LEHMANN, E. L. & ROMANO, J. P. (2006). *Testing statistical hypotheses*. springer.

LIU, Z. & YIN, Y. (1994). Asymptotic representations for quantiles of pooled samples. *Statistics & Probability Letters* 19 299–305.

MACHADO, J. A. F. & SILVA, J. S. (2005). Quantiles for counts. *Journal of the American Statistical Association* 100 1226–1237.

MANN, H. & WALD, A. (1943). On stochastic limit and order relationships. *Ann. Math. Statist.* 14 205–226.

MARTINI, P., SALES, G., MASSA, M. S., CHIOGNA, M. & ROMUALDI, C.

(2013). Along signal paths: an empirical gene set approach exploiting pathway topology. *Nucleic acids research* 41 e19–e19.

PAN, W. (2002). A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments. *Bioinformatics* 18 546–554.

PARK, P. J. (2009). Chip–seq: advantages and challenges of a maturing technology. *Nature Reviews Genetics* 10 669–680.

PEARSON, H. (2006). Genetics: what is a gene? *Nature* 441 398–401.

SCHENA, M., SHALON, D., DAVIS, R. W. & BROWN, P. O. (1995). Quantitative monitoring of gene expression patterns with a complementary dna microarray. *Science* 270 467–470.

SCHENA, M., SHALON, D., HELLER, R., CHAI, A., BROWN, P. & DAVIS, R. (1996). Parallel human genome analysis: microarray-based expression monitoring of 1000 genes. *Proceedings of the National Academy of Sciences* 93 10614.

SMYTH, G. ET AL. (2004). Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical applications in genetics and molecular biology* 3 3.

SMYTH, G. K. & PHIPSON, B. (2011). Permutation P -values Should Never Be Zero : Calculating Exact P -values When Permutations Are Randomly Drawn ? *Statistical Applications in Genetics and Molecular Biology* 9 1–12.

STUDENT (1908). The probable error of a mean. *Biometrika* 6 1–25.

TAI, Y. C., SPEED, T. P. ET AL. (2006). A multivariate empirical bayes statistic for replicated microarray time course data. *The Annals of Statistics* 34 2387–2412.

TUSHER, V., TIBSHIRANI, R. & CHU, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences* 98 5116.

WESTFALL, P. & YOUNG, S. (1993). Resampling-based multiple testing: examples and methods for p-value adjustment. *Wiley series in probability and mathematical statistics (Applied probability and statistics)* .

WILBANKS, E. G. & FACCIOTTI, M. T. (2010a). Evaluation of Algorithm Performance in ChIP-Seq Peak Detection. *PLoS ONE* 5.

WILBANKS, E. G. & FACCIOTTI, M. T. (2010b). Evaluation of algorithm performance in chip-seq peak detection. *PloS one* 5 e11471.

ZHANG, Y., LIU, T., MEYER, C. A., EECKHOUTE, J., JOHNSON, D. S., BERNSTEIN, B. E., NUSBAUM, C., MYERS, R. M., BROWN, M., LI, W. ET AL. (2008). Model-based analysis of chip-seq (macs). *Genome Biol* 9 R137.

Lorenzo Maragoni

Curriculum Vitae

Contact Information

University of Padua
Department of Statistical Sciences
Via C. Battisti, 241
35121 Padova, Italy
tel: +39 049 827 4111
e-mail: maragoni@stat.unipd.it

Current Position

Since January, 2011 (expected completion: July, 2014)
PhD Student in Statistical Sciences
University of Padua, Italy
Thesis title: Quantile Inference in Genomic Studies
Supervisor: Prof. Monica Chiogna

Research Interests

Statistical methods for genomic data.
Quantile inference.

Hypothesis testing.
Statistical analysis of rankings.

Education

July 2006 - July 2008
Master degree in Statistical, Social and Demographical Sciences
Faculty of Statistical Sciences
University of Padua, Italy
Thesis: "Combining graphical models for the analysis of biological networks"
Supervisor: Prof. Monica Chiogna
Final mark: 110 (out of 110) with honors

September 2003 - July 2006
Bachelor degree in Statistics, Population and Society
Faculty of Statistical Sciences
University of Padua, Italy
Thesis: "Confidence intervals for discrete distributions: a comparison of methods by simulation techniques"
Supervisor: Prof. Alessandra Salvan
Final mark: 110 (out of 110) with honors

Visiting Periods

March - June 2013
Visiting Researcher at the Department of Statistics
University of California at Berkeley, US.
Supervisor: Prof. Sandrine Dudoit

March - June 2011

Visiting Researcher at the Institute of Environmental Medicine
Karolinska Institutet, Stockholm, Sweden.
Supervisor: Prof. Matteo Bottai

August - December 2006
LLP - Erasmus Programme at the School of Social and Behavioral Sciences
Universiteit van Tilburg, The Netherlands.

Scholarships

January 2011 - December 2013
PhD Scholarship
University of Padua, Italy.

September - December 2008
Scholarship at the Deparment of Statistical Sciences
University of Padua, Italy.
Project: "Graphical modelling of microarray data"
Supervisor: Prof. Gianfranco Adimari

September - December 2007
Scholarship at the Deparment of Statistical Sciences
University of Padua, Italy.
Project: "Multivariate analysis of socio-demographic data and textual data"
Supervisor: Prof. Luigi Fabbris

Computer Skills

Operating Systems: MAC OS X; MS Windows
Programming Languages: Python, shell script (basic skills)
Statistical Software: S-plus/R, SAS, SPSS
Other: LaTeX

Language Skills

Italian: native
English: fluent
German: basic

Talks

June 2014
47th Scientific Meeting of the Italian Statistical Society
University of Cagliari, Italy
Talk title: "A Quantile-based Test for Detecting Differential Expression in Microarray Data"
June 2013

Deparment of Statistics
University of California at Berkeley, US
Talk title: "Methods for ChIP-Seq: finding expression peaks in human epidermal keratinocytes samples"
February 2013 Department of Statistical Sciences

University of Padua, Italy
??Talk title: "Quantile Inference for Identification of Differential Expression in Microarray Studies" *June 2012* Institute of Environmental Medicine

Karolinska Institutet, Stockholm, Sweden
??Talk title: “Quantile Comparison for Gene Ranking”

Teaching Activity

October 2013 - July 2014
Tutor for the Department of Statistical Sciences
University of Padua, Italy
Teaching task: tutoring undergraduate students on the subjects of Calculus, Probability and Statistics, 100 hours
Supervisor: Prof. Laura Ventura, Prof. Stefano Mazzuco *October - December 2010*
Course name: Multivariate Statistics
Degree: Master Degree in Natural Sciences
Teaching task: computer labs, 25 hours
Institution: University of Padua, Italy
Instructor: Prof. Giovanna Boccuzzo

Other Interests

May, 2010
Acting diploma at Teatro Stabile del Veneto drama school.

References

Monica Chiogna
Associate Professor
Department of Statistical Sciences
University of Padua
via C. Battisti, 241
35121, Padova, Italy

tel: +39 049 827 4183
e-mail: monica@stat.unipd.it