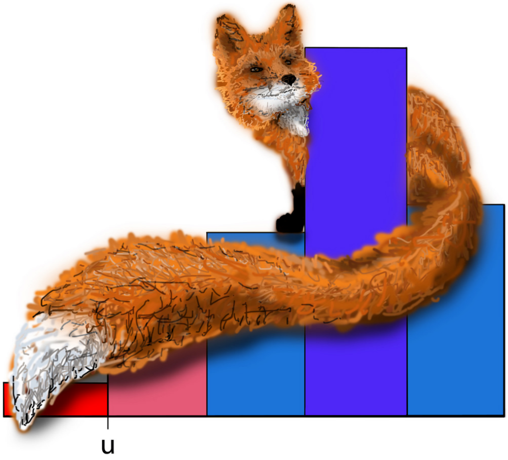
Faculty of Science

Extreme Value Analysis of Huge Datasets

Tail Estimation Methods in High-Throughput Screening and Bioinformatics



Dmitrii Zholud

Ph.D. thesis
Department of Mathematical Sciences
University of Gothenburg

UNIVERSITY OF GOTHENBURG

THESIS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

Extreme Value Analysis of Huge Datasets

Tail Estimation Methods in High-Throughput Screening and Bioinformatics

Dmitrii Zholud

CHALMERS | UNIVERSITY OF GOTHENBURG

Division of Mathematical Statistics
Department of Mathematical Sciences
Chalmers University of Technology
and Göteborg University
Göteborg, Sweden 2011

Extreme Value Analysis of Huge Datasets: Tail Estimation Methods in High-Throughput Screening and Bioinformatics

Dmitrii Zholud

ISBN 978-91-628-8354-6

Division of Mathematical Statistics
Department of Mathematical Sciences
Chalmers University of Technology and
Göteborg University
SE-412 96 Göteborg, Sweden
Telephone: +46 (0)31-772 1000

Author's e-mail: dmitrii@zholud.com
Author's homepage: www.zholud.com

Typeset with LaTeX.
Printed in Göteborg, Sweden 2011

Abstract

This thesis presents results in Extreme Value Theory with applications to High-Throughput Screening and Bioinformatics. The methods described here, however, are applicable to statistical analysis of huge datasets in general. The main results are covered in four papers.

The first paper develops novel methods to handle false rejections in High-Throughput Screening experiments where testing is done at extreme significance levels, with low degrees of freedom, and when the true null distribution may differ from the theoretical one. We introduce efficient and accurate estimators of False Discovery Rate and related quantities, and provide methods of estimation of the true null distribution resulting from data preprocessing, as well as techniques to compare it with the theoretical null distribution. Extreme Value Statistics provides a natural analysis tool: a simple polynomial model for the tail of the distribution of p-values. We exhibit the properties of the estimators of the parameters of the model, and point to model checking tools, both for independent and dependent data. The methods are tried out on two large scale genomic studies and on an fMRI brain scan experiment.

The second paper gives a strict mathematical basis for the above methods. We present asymptotic formulas for the distribution tails of probably the most commonly used statistical tests under non-normality, dependence, and non-homogeneity, and derive bounds on the absolute and relative errors of the approximations.

In papers three and four we study high-level excursions of the Shepp statistic for the Wiener process and for a Gaussian random walk. The application areas include finance and insurance, and sequence alignment scoring and database searches in Bioinformatics.

Keywords: Extreme Value Statistics, High Throughput Screening, HTS, Bioinformatics, analysis of huge datasets, quality control, correction of theoretical p-values, comparison of pre-processing methods, SmartTail, estimation of False Discovery Rates, test power, distribution tail, high level excursions, quantile estimation, multiple testing, Student $t-$test, Welch statistic, small sample sizes, $F-$test, Wiener process, Gaussian random walk, Shepp statistic, limit theorems, exotic options.

Acknowledgments

I am grateful to my supervisor Professor Holger Rootzén for his guidance and encouragement, and to my co-supervisor Professor Olle Nerman for his help and support.

Next, I would like to thank Dr. Sannikov V. F. (who engaged me in active participation in mathematical Olympiads during my school years) for instilling a love of mathematics; and my undergraduate supervisor at Lomonosov Moscow State University, Professor Piterbarg V. I., for constant support and assistance, as well as fruitful discussions on the double sum method.

Special credit goes to Valerie Fedosova for the elegant implementation of the cover drawing.

Finally, I express my sincere gratitude to my family and friends, who stayed with me in good and bad times. I would hardly be able to enjoy life the way I do, without you!

Dmitrii Zholud
Göteborg, August 2011

Where all think alike, no one thinks very much

- *Walter Lippmann* -

List of included papers

This thesis contains the following papers:

Paper I: Rootzén, H. and Zholud, D.S. (2011). Tail estimation methods for the number of false positives in high-throughput testing. *Submitted.*

Paper II: Zholud, D.S. (2011). Tail approximations for the Student $t-$, $F-$, and Welch statistics for non-normal and not necessarily i.i.d. random variables. *Submitted.*

Paper III: Zholud, D.S. (2009). Extremes of the Shepp statistic for a Gaussian random walk*. *Extremes,* **12**(1):1-17.

Paper IV: Zholud, D.S. (2008). Extremes of the Shepp statistic for the Wiener process*. *Extremes,* **11**(4):339-351.

*The thesis version differs in pagination and typographical detail. The original publication is available at www.springerlink.com

Table of contents

Abstract iii

Acknowledgments v

List of included papers ix

I Introduction 1

Background 3

1 Extreme Value Theory 3

2 High-Throughput Screening 7

3 The bridge between EVT and HTS 9

Summary of the papers 17

1 Tail estimation methods in High-Throughput Testing 17

2 Tail approximations for some common statistical tests 18

3 How big is the maximal increment of a Gaussian random walk? 19

4 How big is the maximal increment of the Wiener process? 20

5 The author's contribution 21

References 22

II The Papers 27

Paper I: Tail estimation methods for the number of false positives in high-throughput testing 29

1 Introduction 32

2 Statistical methods 39

3 Basic theory 45

4 Motivation 49

5 Examples 51

6 Discussion and conclusions 60

References 62

Paper II: Tail approximations for the Student $t-$, $F-$, and Welch statistics for non-normal and not necessarily i.i.d. random variables 67

1 Introduction 70

2 One-sample $t-$test 76

3 Two-sample $t-$test 79

4 Welch's test 83

5 $F-$test 85

6 Second and higher order approximations 89

7 Simulation study 91

Appendix A: Asymptotic behavior of an integral of a continuous function over a shrinking ball 93

Appendix B: A note on the regularity constraints and the continuity assumption 96

Appendix C: Tables and figures 100

Supplementary materials 107

References 108

Paper III: Extremes of the Shepp statistic for a Gaussian random walk 113

1 Introduction 116

2 Moderate deviations of the Shepp statistic 118

3 Very large deviations of the Shepp statistic 130

4 Limit theorems for $M_T^{(N)}$ 132

References 133

Paper IV: Extremes of the Shepp statistic for the Wiener process 135

1 Introduction 138

2 Asymptotic behavior of the distribution tail of M_T . 140

3 Limit theorem for M_T 148

References 150

Reader's notes 153

Further reading 158

Part I

Introduction

Background

In this section we give a brief introduction to Extreme Value Theory and High-Throughput Screening. The purpose is to prepare the reader for the innovative step introduced later in this thesis - a new methodology which connects the two areas.

1 Extreme Value Theory

Extreme Value Theory (EVT) is a branch of probability theory and mathematical statistics which focuses on analysis and inference about extreme events, i.e. events with very low probability of occurrence. Extreme events are of great importance in almost every field of science and technology due to the fact that they can turn out to be catastrophic[1] and thus very costly. This motivation often comes from finance and insurance - two of the most popular application areas of EVT. We now proceed with the mathematical background.

Let $\{X_i\}_{i=1}^n$ be a sequence of independent, identically distributed (i.i.d.) random variables. Further, let F be their cumulative distribution function (CDF) and consider the behavior of

$$M_n = \max\{X_1, .., X_n\}, \quad \text{as} \quad n \to \infty.$$

It then follows from elementary probability theory that

$$\mathbf{P}\left(M_n > x\right) = 1 - \left(F(x)\right)^n. \tag{1.1}$$

A naive way to estimate the tail distribution of M_n would be to take a sample of n observations from the distribution function F and then estimate $F(x)$ using the empirical CDF estimator

$$\hat{F}(x) = \frac{\#\{X_i \leq x\}}{n}$$

and substitute the result in (1.1). This however does not work in practice because small deviations from $F(x)$ induce huge inaccuracy in the estimate of $1 - (F(x))^n$ when n is large. Moreover,

[1] In this thesis extreme events will be associated with new scientific findings rather than catastrophes. These are of course important too.

the estimator $\hat{F}(x)$ is typically not reliable for large x, and it becomes clear that the estimation of $\mathbf{P}\left(M_n > x\right)$ requires a different approach.

Theorem 1.1 (Fischer and Tippett (1928), Gnedenko (1943))**.** [2] *If there exist normalizing constants $c_n > 0$ and $d_n \in \mathbb{R}$ and a non-degenerate distribution function H such that*

$$\frac{M_n - d_n}{c_n} \xrightarrow{d} H,$$

then H is a scaled and translated version of one of the following three distribution functions:

$$\textit{Fréchet}: \qquad \Phi_\alpha(x) = \begin{cases} 0, & x \leq 0 \\ \exp\left(-x^{-\alpha}\right), & x > 0 \end{cases} \quad \alpha > 0.$$

$$\textit{Weibull}: \qquad \Psi_\alpha(x) = \begin{cases} \exp\left(-(-x)^{\alpha}\right), & x \leq 0 \\ 1, & x > 0 \end{cases} \quad \alpha > 0.$$

$$\textit{Gumbel}: \qquad \Lambda(x) = \exp\left(-e^{-x}\right), \qquad x \in \mathbb{R}.$$

It is said that a distribution F belongs to the Fréchet, Weibull, or Gumbel domain of attraction if Theorem 1.1 holds with the Fréchet, Weibull or Gumbel limit distribution accordingly. Note that we now have knowledge about the distribution of the normalized maximum without knowing the true distribution.

From statistical point of view it is often convenient to parameterize the distribution H of Theorem 1.1 in the following way

$$H(x) = \exp\left(-\left[1 + \gamma\left(\frac{x - \mu}{\sigma}\right)\right]^{-1/\gamma}\right) \tag{1.2}$$

defined on $\{x : 1+\gamma(x-\mu)/\sigma > 0\}$, where $\mu, \gamma \in \mathbb{R}$ and $\sigma > 0$. The Fréchet and Weibull distributions correspond to the cases $\gamma > 0$ and $\gamma < 0$ respectively, and the Gumbel distribution is the limit case $\gamma \to 0$, which implies

$$H(x) = \exp\left(-e^{-\frac{x-\mu}{\sigma}}\right), \; x \in \mathbb{R}.$$

[2] The origins of EVT date back to the work of Fischer and Tippett (1928) on possible limits of maximum values, later formalized by Gnedenko (1943), and the book *Statistics of Extremes* by Gumbel (1958).

The distribution $H(x)$ in (1.2) is called the Generalized Extreme Value (GEV) distribution.

The Fréchet domain of attraction coincides with the class of heavy-tailed distributions; the examples are the Burr, Pareto, and Student t- distributions. Weibull domain of attraction includes light-tailed distributions with the finite right endpoint, such as the Beta and Uniform distributions. And the Gumbel domain of attraction contains a great variety of distributions ranging from moderately heavy-tailed to light-tailed, for instance the Exponential, Log-Normal and Normal distributions.

Finding the normalizing constants and determining the domain of attraction is not always easy.[3] In practice, however, the parameters of the GEV distribution (1.2) have to be estimated from data anyway, and the issue of testing for the belongingness to the domain of attraction is not a problem.[4]

Though EVT has evolved from the problem of finding the distribution of the normalized maximum, in applications it is also common to use the so called Peaks Over Threshold (POT) approach. The latter is based on the fact that for a random variable X with the distribution function F the exceedance over a threshold u is approximately Generalized Pareto (GP) distributed with parameters $\tilde{\gamma}$ and $\tilde{\sigma}(u)$, that is

$$F_u(x) = \mathbf{P}\left(X > u + x | X > u\right) \approx \left(1 + \frac{\tilde{\gamma} x}{\tilde{\sigma}(u)}\right)^{-1/\tilde{\gamma}}, \qquad (1.3)$$

defined on $\{x : x > 0 \text{ and } 1 + \tilde{\gamma} x / \tilde{\sigma}(u) > 0\}$. For $\tilde{\gamma} = 0$ we interpret the distribution as the limit as $\tilde{\gamma} \to 0$, i.e.

$$F_u(x) \approx \exp\left(-x/\tilde{\sigma}(u)\right), \; x > 0.$$

The following theorem states the relationship between the maximum and the POT approach.

Theorem 1.2 (Pickands (1975), Balkema and de Haan (1974)). *For every $\gamma \in \mathbb{R}$, F is in the corresponding domain of attraction of a Generalized Extreme Value distribution if and only if*

$$\lim_{u \uparrow x_F} \sup_{0 < x < x_F - u} |F_u(x) - G_{\tilde{\gamma}, \tilde{\sigma}(u)}(x)| = 0$$

[3] See e.g. Leadbetter et al. (1983) for different approaches.

[4] See e.g. Coles (2001) and Beirlant et al. (2004).

for some positive σ and $\mu \in \mathbb{R}$, where $G_{\tilde{\gamma},\tilde{\sigma}(u)}$ is the Generalized Pareto distribution with parameters $\tilde{\gamma} = \gamma$ and $\tilde{\sigma}(u) = \sigma + \gamma(u - \mu)$, and x_F is the upper right endpoint (possibly infinite).

Using the POT approach we can *zoom in* to the tails of a distribution. Indeed, suppose that we can model exceedances of X above a threshold u using the GP distribution. Then, for $x > u$,

$$\mathbf{P}\left(X > x | X > u\right) = \left(1 + \tilde{\gamma}\frac{x-u}{\tilde{\sigma}(u)}\right)_+^{-1/\tilde{\gamma}},$$

and

$$\mathbf{P}\left(X > x\right) = \mathbf{P}\left(X > u\right)\left(1 + \tilde{\gamma}\frac{x-u}{\tilde{\sigma}(u)}\right)_+^{-1/\tilde{\gamma}}. \tag{1.4}$$

For use in statistical analysis one wants u to be large enough so that the GP model (1.3) holds, but not too large so that one has reasonable number of observations when estimating the parameters of the model.[5] As there is no commonly agreed selection rule, the choice of u is governed by exploratory data analysis, see e.g. examples in Paper I of this thesis.

Once the threshold u is fixed we get an estimator for the probability $\mathbf{P}\left(X > x\right)$ as follows: 1) estimate the parameters $\tilde{\gamma}$ and $\tilde{\sigma}$ from (1.3) using maximum likelihood, and 2) substitute the latter in (1.4) and replace $\mathbf{P}\left(X > u\right)$ by $\hat{F}(u)$. Confidence intervals can then be obtained using the delta method.

To summarize, EVT provides a methodology for estimating the probability of events which occur very seldom or that have never occurred but may occur in the future. For a more in-depth reading we recommend Coles (2001)[6], Beirlant et al. (2004)[7], Leadbetter et al. (1983)[8] and Reiss and Thomas (2001)[9]; for a non-mathematical introduction to statistics of extremes see e.g. Brodin (2006).

[5] Such trade off between bias and variance is common in EVT and takes place also when fitting the GEV distribution using the block maximum approach, see e.g. Coles (2001).

[6] Contains the most important theoretical results without becoming tedious, and provides plenty of interesting and useful applications to real datasets.

[7] Also covers multivariate and Bayesian modeling of extremes.

[8] Focuses on EVT for stationary time series.

[9] Includes a variety of interesting case studies with applications to insurance, finance, hydrology and other fields.

2 High-Throughput Screening

High-throughput screening (HTS) is a method of experimentation which aims at accelerating scientific findings, such as e.g. identification of drug targets or genes involved in the cell cycle, by conducting thousands - or hundreds of thousands of chemical or genetic tests performed in an automated manner. HTS is a relatively recent innovation: by the beginning of the 21st century HTS has gained significant popularity and interest among scientists and industrial engineers[10], and the technology has been, and continues developing at present[11], at a tremendous pace.

HTS has been successfully applied to many areas in modern biology and has become one of the primary tools for drug development in most pharmaceutical companies. Below follow two examples of HTS studies in cell biology.

Example 1 *(quantitative phenotypic profiling in yeast):* To understand the genetic machinery of the model organism *Saccharomyces cerevisiae*[12], baker's yeast, the community of yeast researchers have constructed a collection of single gene deletion mutated haploid strains for all known protein coding genes, see Giaever et al. (2002). Some of these strains (roughly 20%) are not viable, and the challenge is to utilize the rest of the strains (approximately 5000) in a way that provides new knowledge of how genes function.

In Warringer and Blomberg (2003) and Warringer et al. (2003) the focus is on growth characteristics of different yeast strains. Omitting details, if a growth parameter of a mutant strain colony differs substantially from the corresponding growth parameter of a wild type cell, then this can be interpreted as evidence that the gene plays an important role in some intra-cellular regulation process, i.e. the gene is active, given the conditions of the experiment - growth media, temperature and etc. Of particular interest are comparisons of the growth behavior under different environmental

[10] See e.g. Bolger (1999) and Hertzberg and Pope (2000).

[11] For the most recent trends in HTS see e.g. Macarron et al. (2011), Mayr and Bojanic (2009) and Mishra et al. (2008).

[12] There is a number of advantages of a yeast cell as a model organism: it can be propagated in great numbers relatively fast, it is easy to study in a laboratory, and, being eukaryotic, it is biologically more relevant to human than bacteria.

stress conditions with the behavior under normal, unstressed condition(s). For example, a deletion strain can behave as the wild type strain in normal media, but have a severe growth defect in salt stressed environment. This can then be interpreted as a hint that the knocked out gene (or the corresponding protein) plays a key role in cellular defense or adaption processes under salt stress.

In these studies the authors recorded and analyzed the growth of colonies of the mutant strains in 5 different growth media. This gave approximately $25,000$ growth curves, and each growth curve was replicated twice, see e.g. Zholud et al. (2011) and Papers I and II of this thesis. The growth curves can be found in the PROPHECY database.

Example 2 *(large-scale mapping of genetic interactions in yeast):* Within the study of the yeast genome, let us consider the concept of synthetic lethality, which defines a relationship where two mutations, neither of which is lethal on its own, result in cell death when they are combined. Identification of synthetically lethal double mutant combinations is of basic interest because they can identify genes which are involved in the same essential biological process.

A HTS study of impressive scope is being conducted in the Boone Lab[13], where the synthetic lethality approach is used to provide a global view of functional relationships between genes and pathways, see Costanzo et al. (2010) and Baryshnikova et al. (2010). Over the period of 5 years the group has examined 5.4 million gene-gene pairs, which is approximately 30% of the total of $6000 \times 6000/2 = 18,000,000$ possible double mutants. The screening is expected to be completed by 2013.

Such huge-scale studies would not be feasible without modern advances in robotics and high-speed computer technology. Along with the ability to conduct hundreds of thousands of experiments in a short amount of time, HTS poses fundamental challenges related to experimental planning and data analysis. We discuss these challenges in the next section.

[13] The Donnelly Centre for Cellular and Biomolecular Research, University of Toronto.

3 The bridge between EVT and HTS

In this section we explain why Extreme Value Theory should be used in the analysis of High-Throughput Screening experiments. Our arguments are valid for huge datasets of any origin (and not necessarily related to HTS) as long as testing at extreme significance levels takes place.

The structure of the section is as follows. First, we give a general motivation and show that the analysis of HTS data *requires* EVT-based methods by the very definition. Then, we give two more specific notes, which, in our opinion, address fundamental challenges of HTS. Each note contains a few simple examples of how EVT can meet the challenges.

Basic motivation. Extreme values of statistical tests are of basic interest in HTS experiments for the following reasons. First, HTS uses many thousands or even millions of biochemical, genetic or pharmacological tests. In order to get a reasonable number of rejections, the significance level of the tests is often very small, say, 0.001 or lower. Second, HTS assays are often subject to numerous systematic and spatial effects and to large number of pre-processing steps. The resulting data may hence become dependent, non-normal, or non-homogeneous, which leads to deviations from the standard assumptions underlying the use of the tests. And third, under economical constraints, the number of replicates in each individual experiment that constitute a HTS study is usually as small as $2-5$, making large sample approximations, such as e.g. normal approximation, inapplicable.

This encourages the study of the asymptotic behavior of the tails of the distribution of the test statistics. Tail probabilities and accurate approximations of those can be used 1) to correct theoretical p-values when the true model is in fact different from the stated one 2) to compare different pre-processing methods 3) to study the power of statistical tests, and 4) to estimate false discovery rates and related quantities.

Thought there are other, more specific examples of the use of EVT-based methods in HTS, see e.g. Knijnenburg et al. (2009), we are not aware of any literature that describes or uses methodology or methods similar to those presented in this thesis.

Let us now consider in more detail some aspects of *data quality* and *estimation of false discovery rates.*

A note on data quality. It is a common knowledge that data quality is the primary ingredient of a successful experiment, whether it is a huge-scaled HTS study or a single test. However, even though there is a widespread opinion that advances in modern technology and methods of quality control lead to significant improvements in the accuracy of measurements involved in HTS studies, see e.g. Macarron et al. (2011), still the quality control in HTS contains many stumbling blocks.

First, it should be noted that no laboratory would be willing to publish the details on the quality of their data, if the quality of the data is poor. The author of this thesis has his own experience working with HTS data. This experience does not necessarily come from the projects mentioned here, but also from discussions with colleagues and collaborators, from attending scientific conferences, and from reading numerous articles on the subject. As long as this knowledge can be trusted without having to expose the work of others, whether such criticism would have been considered fair or not, I believe that it has become quite common practice in the literature to hide the true facts about the quality of data behind fuzzy technical details and ad-hoc implementations of data pre-processing methods. This tendency presumably comes from fear of acquiring a reputation of a ”laboratory that does bad experiments”, and is reinforced by the scale of HTS studies and the amount of time, money, and labor involved.

The second stumbling block is the lack of appropriate statistical methods that provide a proper analysis methods. We do not mean to say that existing methods fail to address the issues they are designed to cope with, e.g. various spatial effects, contaminations, outliers, systematic effects and etc., but rather that they address these issues only partially. However, the rest, that can not be fully explained or fixed, is often neglected or not mentioned. For example, such methods as removing row, column, and plate biases, see e.g. Malo et al. (2010), despite their power to reduce the impact of systematic effects, have one serious disadvantage - they introduce a bias when they are applied to data that does not contain any systematic effects. The assessment of the presence of systematic errors in a given HTS assay is a complex problem, see e.g. Dragiev et al. (2011). In biological articles, however, one rarely finds data analysis which is satisfactory from a statistical point of view.

The ultimate goal of HTS technology is of course to provide an elegant solution on the "hardware" level, that is, by means of efficient experimental design, accurate measurement systems and automated quality control. In practice, however, this is quite complicated (e.g. bacteria colonies may compete for food, or grow better on the edges rather than in the middle of the plates, and etc.), and good methods to distinguish between true positives and false positives, to choose the best pre-processing method, and to find out how the deviations from the null model affect the findings are very much needed. In Papers I and II we use Extreme Value Theory to develop such methods.

As a brief example we now show how to use EVT to compare different pre-processing methods and to correct theoretical $p-$values, even without any knowledge about the experimental design, i.e. the source of the deviations from the null model.

Let P be a generic $p-$value obtained using the Student one- or two-sample $t-$, or $F-$ test. Then, for small x, and under the assumption that there exists a continuous joint density of the vector of data[14], it follows from the results of Paper II that

$$\mathbf{P}\left(P < x\right) = Cx\left(1 + o(1)\right), \tag{3.1}$$

as $x \to 0$, where $0 < C < \infty$ is some constant. If the theoretical null hypothesis holds, then the $p-$values are uniformly distributed and $C = 1$. If, however, the true null hypothesis deviates from the theoretical one, then $C \neq 1$, and the preprocessing methods can be compared by looking at the corresponding values of the constant C. The latter can be estimated from a dataset for which the null hypothesis is known to be true[15], see Paper I for the examples of how it is done.

[14] Minor technical constraints apply.

[15] The null dataset can be obtained by conducting a separate experiment, or artificially, by e.g. alternating signs or using permutations. It may be well worth the effort to try to obtain a sample from the true null distribution anyway, both to get a better grip on risks for false positives and for general quality control purposes.

A straightforward correction of the $p-$values is as follows. Assume that one is only interested in the values of x which are small enough to make it possible to neglect the $o(1)$ term in (3.1). Then use the resulting model to estimate the constant C from a null dataset and multiply all $p-$values computed from the experimental dataset by the factor C. We, of course, assume that the experimental conditions and pre-processing methods are the same for the null and the experimental datasets.

Finally, in Paper I we present a more general EVT-based model for p-values that result from statistical tests other than the Student one- or two- sample $t-$ or $F-$ tests, and which may sometimes lead to better approximations than (3.1) for the $t-$ and $F-$ tests. It should be emphasized that it has to be checked from the data whether the model of Paper I is relevant in a concrete testing problem. The same applies, of course, to the asymptotic formula (3.1). We provide the methods for such model checking and give explicit examples of EVT-based analysis of HTS datasets in Paper I.

A note on the FDR controlling procedures. One of the primary concerns in HTS studies is to be able to control[16] or estimate[17] the proportion of false discoveries among the set of all positive findings. The false discovery rate (FDR) approach to multiple testing was introduced by Benjamini and Hochberg (1995), where the authors suggested studying the error rate vaguely described as "the proportion of false discoveries". In the notation of their paper, FDR is defined as the expectation of random variable $\mathbf{Q} = \mathbf{V}/\mathbf{R}$, the ratio of the number of erroneously rejected null hypotheses, $\mathbf{V}$, to the total number of the rejected hypotheses, $\mathbf{R}$. If $\mathbf{R} = 0$, then Benjamini and Hochberg define $\mathbf{Q}$ to be equal 0. Thus,

$$FDR = \mathbf{E}\left(\mathbf{Q}\right) = \mathbf{P}\left(\mathbf{R} > 0\right)\mathbf{E}\left(\mathbf{V}/\mathbf{R}\middle|\,\mathbf{R} > 0\right). \qquad (3.2)$$

[16] This refers to the False Discovery Rate (FDR) controlling procedure of Benjamini and Hochberg (1995), see below.

[17] This refers to the positive False Discovery Rate ($pFDR$) approach of Storey (2002). We believe that the latter is a more appropriate measure of the proportion of false discoveries than FDR.

Based on this definition Benjamini and Hochberg (1995) proposed the following multiple-testing procedure: for any given value $q^* > 0$ and m independent test statistics with the corresponding ordered p-values $P_{(1)}, P_{(2)}, .., P_{(m)}$, reject all $H_{(i)},\ i = 1, 2, .., \hat{k}$, where $\hat{k}$ is the largest i for which $P_{(i)} \leq \frac{i}{m} q^*$. This procedure ensures that $FDR \leq q^*$, that is, FDR is "controlled" at level q^*.

The critical problem with the FDR approach of Benjamini and Hochberg is that FDR, in fact, has little to do with the "expected proportion of false discoveries" when q^* is small. Small value of FDR does not necessarily mean that $\mathbf{E}\left(\mathbf{V}/\mathbf{R}|\,\mathbf{R} > 0\right)$ is small, but could instead just be caused by small value of $\mathbf{P}\left(\mathbf{R} > 0\right)$, i.e. by the probability that there are any rejections at all being small, see (3.2). It follows from the Extreme Value Theory results given in Paper II (details are given later in this section) that this is not just a theoretical possibility, but a very important practical issue. In short, the result is that for, perhaps, the most commonly used statistical tests[18], and under the assumption that the tests are independent,

$$\mathbf{E}\left(\mathbf{V}/\mathbf{R}|\,\mathbf{R} > 0\right) \to C \quad \text{as} \quad \alpha \to 0.$$

Here α is the significant level, and $C > 0$ is some positive constant.[19] Thus, under quite general conditions, $FDR \to 0$ as $q^* \to 0$ because $\mathbf{P}\left(\mathbf{R} > 0\right) \to 0$, and not because the expected proportion of false discoveries $\mathbf{E}\left(\mathbf{V}/\mathbf{R}|\,\mathbf{R} > 0\right) \to 0$.[20]

Similar criticism of the FDR approach of Benjamini and Hochberg can be found in Storey (2002). The author argues that

> *"...when controlling FDR at level q^*, and positive findings have occurred, then FDR has really only been controlled at level $q^*/\mathbf{P}\left(\mathbf{R} > 0\right)$."*

and introduces the quantity

$$pFDR = \mathbf{E}\left(\mathbf{V}/\mathbf{R}|\,\mathbf{R} > 0\right), \tag{3.3}$$

[18] For the Student one- and two- sample $t-$, and $F-$ statistics, see Theorem 1.1 of Paper II.

[19] $C = 0$ in the degenerate case when there are no null hypothesis, i.e. $\mathbf{V} \equiv 0$.

[20] The author of this thesis wonders how this approach could gain such popularity in scientific community. The arguments follow later in this section, and a separate paper that addresses the usefulness of FDR controlling procedure of Benjamini and Hochberg is in progress.

called positive False Discovery Rate (pFDR), to quantify the proportion of false positives, conditioned on the event that positive findings have occurred. For a more thorough motivation of $pFDR$ over FDR and details on the methods of estimation of $pFDR$ see e.g. Storey (2002, 2003, 2004). The $pFDR$ approach of Storey gives a more realistic picture of the expected proportion of false discoveries, with the fundamental difference that while the sequential p-value method of Benjamini and Hochberg fixes the "error rate" parameter FDR and estimates the corresponding rejection region, the Storey approach is to fix the rejection region and then estimate its corresponding error rate (without quotes!), $pFDR$.

Neither FDR nor $pFDR$, however, take advantage of Extreme Value Theory and thus do not use information from the tails of the distribution of the test statistic(s), or, alternatively, $p-$values. We now give a brief example of how EVT can contribute to the estimation of $pFDR$ and construction of confidence intervals in situations where traditional methods fail.

Consider a multiple testing problem, where there are in total m independent tests such that the null hypothesis, H_0, is true with probability π_0 and the alternative hypothesis, H_1, is true with probability $\pi_1 = 1 - \pi_0$. Let $g_0(\mathbf{x})$ and $g_1(\mathbf{x})$ be the joint densities of the data vector under H_0 and H_1 accordingly. The distribution of an arbitrary p-value then follows the semi-parametric mixture model

$$F(x) = \pi_0 F_0(x) + \pi_1 F_1(x),$$

where $F_0(x)$ and $F_1(x)$ are the CDFs of the p-values under H_0 and H_1 accordingly. Further, Theorem 1.1 of Paper II implies that if the $p-$values are obtained using the Student one- or two- sample $t-$, or $F-$ statistic, and g_0 and g_1 are continuous[21], then

$$F(x) = Cx\left(1 + o(1)\right) \quad \text{as} \quad x \to 0,$$

where $C = (\pi_0 K_{g_0} + \pi_1 K_{g_1})$ and K_{g_0} and K_{g_1} are some positive, finite constants that depend on the densities g_0 and g_1.[22] The positive False Discovery Rate parameter $pFDR$, by equation (5)

[21] Minor technical constraints apply.

[22] The constants K_{g_0} and K_{g_1} depend on the test statistic as well.

of Theorem 1 in Storey (2002) and the above equality[23], is

$$pFDR(x) = \frac{\pi_0 K_{g_0}}{C} + o(x). \tag{3.4}$$

It follows, perhaps in contradiction to what could be expected, that $pFDR$ does not become smaller as one goes further out in the tail. Equation (3.4) shows that for the Student one- and two- sample $t-$ and $F-$ tests, and under quite general assumptions on the data, $pFDR$ can not be "controlled", no matter what the value of π_0 is!

We propose the following EVT-based estimator of $pFDR$: estimate the unknown parameter π_0 using the estimator $\hat{\pi}_0$ of Storey (2002), and K_{g_0} and C using methods of EVT[24] - these methods and related exploratory and model checking analysis are studied in detail in Paper I. We can then use the delta method to obtain confidence intervals on $pFDR(x)$ in (3.4), see Zholud (2011b).

As a final comment, Storey (2002) estimates $pFDR$ using the empirical CDF as estimator of $F(x)$. The EVT-based methods, however, lead to substantial gains in accuracy and efficiency, as indicated by the examples in Paper I. Further, we are not aware of any other methods that make it possible to estimate $pFDR(x)$ for x-es which are smaller than the smallest observed $p-$value. This can sometimes be of interest, e.g. as input to the planning of the next experiment.

Paper I develops methods to estimate $pFDR$ under more general model for the observed p-values, i.e. not necessarily for the tests mentioned above. This model is motivated by general asymptotic arguments from EVT, and also by extensive practical experience from Extreme Value Statistics.

Conclusion. In this thesis we make an attempt to start a new trend for the use of EVT-based methods in the analysis of huge datasets. A brief summary of our contribution is given in the next chapter.

[23] The equation (5) of Storey (2002) is based on the assumption that the $p-$values are uniformly distributed under H_0. In this case, of course, $K_{g_0} = 1$. We though consider a more general case when the true null distribution may differ from the theoretical one - this happens quite often in HTS experiments.

[24] Here K_{g_0} and C are estimated from the null and experimental datasets accordingly.

Summary of the papers

In this chapter we give a short summary of the papers included in the thesis. Papers I and II develop new methodology for statistical analysis of huge datasets; in particular, datasets that arise in High-Throughput Screening experiments are in focus. Papers III and IV take their origin from the problem of sequence alignment and database searches in Bioinformatics, but the results may find suitable application in other areas as well. The last section contains a note on the author's contribution to the papers.

1 Tail estimation methods in High-Throughput Testing

In Paper I, *Tail estimation methods for the number of false positives in high-throughput testing*, we develop novel methods to handle false rejections in HTS experiments. The innovative step emphasized in the paper is to use Extreme Value Theory to study the tails of the distribution of p-values associated with these highly-multiple testing setups.

Let P denote a generic p-value, and, as usual, write H_0 for the null hypothesis and H_1 for the alternative hypothesis. Our basic model is that there are positive constants $c_0, \gamma_0, c_1, \gamma_1$ such that

$$F_0(x) = \Pr(P \leq x \mid H_0) = c_0 x^{1/\gamma_0}(1 + o(1))$$

and

$$F_1(x) = \Pr(P \leq x \mid H_1) = c_1 x^{1/\gamma_1}(1 + o(1)),$$

as $x \to 0$. This model is motivated by general asymptotic arguments from EVT, and also by extensive practical experience from Extreme Value Statistics. We give both theoretical and data-based motivation, and also introduce and illustrate a number of model checking tools.

If the testing procedure is reasonable, then small p-values should be more likely under H_1 than under H_0, so it is typically reasonable to expect that $\gamma_0 \leq \gamma_1$.[1]

[1] However, for the Student one- and two- sample $t-$ and $F-$ tests the model is expected to hold with $\gamma_0 = \gamma_1 = 1$, see Paper II.

If one further assumes that tests are independent, and that H_0 is true with probability π_0, and H_1 is true with probability $\pi_1 = 1 - \pi_0$, then we get the following extreme tail mixture form for the distribution

$$F(x) = \Pr(P \leq x) = \big(\pi_0 c_0 x^{1/\gamma_0} + \pi_1 c_1 x^{1/\gamma_1}\big)(1 + o(1)) \qquad (1.1)$$

of the observed p-values.

We show that the conditional distribution of the number of false positives, given that there is in all r positives, approximately has a binomial distribution, and use extreme tail model to develop efficient and accurate methods to estimate its success probability parameter. Furthermore, we provide efficient and accurate methods to estimate the true null distribution resulting from a preprocessing method, and techniques to compare it with the theoretical null distribution.

The methods are tried out on two large scale genomic studies and on an fMRI brain scan experiment. A software implementation of the methods, SmartTail, will appear at www.smarttail.se.

2 Tail approximations for some common statistical tests

In Paper II, *Tail approximations for the Student $t-$, $F-$, and Welch statistics for non-normal and not necessarily i.i.d. random variables*, we study the tails of the distribution of these, perhaps, most commonly used test statistics under non-standard conditions. If the null hypothesis

$$H_0 : \mathbf{X} \sim MVN(\mathbf{0}, \sigma^2 \mathbf{I}_n)$$

holds, where $\mathbf{X} \in \mathbb{R}^n,\ n \geq 2$, is the vector of data, $\sigma^2 > 0$, and $\mathbf{I}_n$ is the identity matrix, then the distribution of the Student one- or two- sample $t-$ or $F-$ statistic, denoted further by $T_n = T_n(\mathbf{X})$, is of course well known. The challenge is to find the *asymptotic* behavior of the probability of high-level excursion of T_n under the *alternative hypothesis*; and preferably for *small sample sizes*, i.e. to study

$$\mathbf{P}\left(T_n > u | H_1\right), \quad \text{as} \quad u \to \infty \quad \text{for } n \text{ fixed.}$$

The motivation for such problem statement comes from HTS experiments, where testing is done at extreme significance levels, and often with very low degrees of freedom.

The multivariate joint density g of the vector X under H_1 does not have to be Gaussian, and may allow dependence and non-homogeneity of the elements of the sample. We show that if the density g is continuous[2], then

$$\mathbf{P}\left(T_n > u | H_1\right) = K_g t(u)(1 + o(1)) \quad \text{as} \quad u \to \infty, \qquad (2.1)$$

where $0 < K_g < \infty$ is some constant that depends on g, and $t(u)$ is the tail of the distribution of T_n under H_0. We give exact algebraic expressions for K_g, compute the theoretical value of the constant for a number of particular cases[3], and present MATLAB (2010) scripts to evaluate K_g numerically for an arbitrary g.

We also derive bounds for the absolute and relative errors of the asymptotic approximations, and study the rate of convergence, both theoretically and using simulations.

The paper provides explicit conditions on the multivariate density g which ensure that the asymptotic formulas are valid, as well as simpler conditions that can be easily checked in many situations.

Finally, we study high-level excursions of the Welch statistic, and suggest an alternative to the Welch-Satterthwaite approximation for the case when sample sizes are small, and when testing is done at extreme significance levels.

The results of Paper II give a basis for new methods to correct theoretical $p-$values, to compare different pre-processing methods, and to estimate False Discovery Rates and related quantities in HTS experiments. These methods have been developed in Paper I, under the more general model (1.1) for the tails of the distribution of the $p-$values. For the Student one- and two- sample $t-$ and $F-$ tests, however, the model (1.1) is expected to hold with $\gamma_0 = \gamma_1 = 1$, which reduces the number of parameters and therefore the variability of the corresponding estimators.

Another important consequence is that for these test statistics the positive False Discovery Rate of Storey (2002), $pFDR(x)$, is asymptotically constant as $x \to 0$.

[2] Minor regularity constraints apply.

[3] Including multivariate Gaussian and non-normal i.i.d. cases.

3 How big is the maximal increment of a Gaussian random walk?

The motivation for Paper III, *Extremes of the Shepp statistic for a Gaussian random walk*, comes from the problem of finding similarities between long biological sequences. In order to detect similarities that indicate real biological kinship it is important to study similarities between random Bernoulli sequences.

A well understood application of Extreme Value Theory to sequence alignment is *gapless alignment* as implemented, e.g., in BLAST, see Altschul et al. (1990).[4] In this case it has been rigorously proven, see Karlin and Dembo (1992), that the distribution of the maximal score of the alignment of two random sequences belongs to the Gumbel domain of attraction, and explicit formulas for the parameters have been given.

In this paper we study mathematically closely related matching problem[5] and derive a theoretical result on the limiting distribution of the maximal increment of a Gaussian random walk.

Let $(\xi_i, i \geq 1)$ be a sequence of independent standard normal random variables and let $S_k = \sum_{i=1}^{k} \xi_i$ be the corresponding random walk. We study the re-normalized Shepp statistic

$$M_T^{(N)} = \frac{1}{\sqrt{N}} \max_{1\leq k \leq TN} \max_{1 \leq L \leq N} (S_{k+L-1} - S_{k-1})$$

and determine asymptotic expressions for

$$\mathbf{P}\left(M_T^{(N)} > u\right) \quad \text{when} \quad u, N \quad \text{and} \quad T \to \infty$$

in a synchronized way. There are three types of relations between u and N that give different asymptotic behavior. For these three cases we establish the limiting Gumbel distribution of $M_T^{(N)}$ when $T, N \to \infty$ and present the corresponding normalization sequences.

[4] Gapless alignments are most commonly used in database searches.

[5] That is as if the distribution of the substitution weights was standard normal.

4 How big is the maximal increment of the Wiener process?

In Paper IV, *Extremes of the Shepp statistic for the Wiener process*, we study the tail of the distribution of

$$M_T = \max_{0 \le t \le T} \max_{0 \le s \le 1} W(t+s) - W(t),$$

where $W(\cdot)$ is the standard Wiener process. By analogy with the discrete case studied in Paper III, M_T is referred to as the Shepp statistic for the Wiener process.

We determine an asymptotic expression for the probability of high-level excursion of M_T,

$$\mathbf{P}\left(M_T > u\right) \quad \text{when} \quad u \to \infty, \tag{4.1}$$

establish the limiting Gumbel distribution of M_T as $T \to \infty$, and present the corresponding normalization functions.

The results of Paper IV are essential for the proof of Theorem 1.2 of Paper III. We note, however, that though the motivation to study (4.1) had purely *scientific* origin, the solution may be useful in *practice*: a quick example comes from finance and insurance, where the tail distribution of the maximal increment M_T of, say, a stock price may be used to model prices of some exotic options.

5 The author's contribution

Paper I is a joint work with Professor Holger Rootzén. The authors have contributed equally to the paper.

In Paper II, special credit goes to Professor Holger Rootzén for careful reading and useful comments, and Professor Olle Nerman for pointing out that the proof does not use the assumptions of independence and identical distribution, as stated in a very early version of the manuscript.

The problem statement in Papers III and IV comes from Professor V.I. Piterbarg.

References

S.F. Altschul, W. Gish, W. Miller, E.W. Myers, and D.J. Lipman. Basic alignment search tool. *J. Mol. Biol.*, 215:403–410, 1990. 20

B.B. Balkema and L. de Haan. Residual lifetime of great age. *Annals of Probability*, 2:792–804, 1974. 5

A. Baryshnikova, M. Costanzo, Y. Kim, H. Ding, J. Koh, K. Toufighi, J.-Y. Youn, J. Ou, B.-J. San Luis, S. Bandyopadhyay, M. Hibbs, D. Hess, A.-C. Gingras, G.D. Bader, O.G. Troyanskaya, G.W. Brown, B. Andrews, C. Boone, and C.L. Myers. Quantitative analysis of fitness and genetic interactions in yeast on a genome scale. *Nature Methods*, 7(12):1017–1024, 2010. 8

J. Beirlant, Y. Goegebeur, J. Segers, and J. Teugels. *Statistics of Extremes, Theory and Applications.* Wiley, Chichester, 2004. 5, 6

Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: A practical and powerful aproach to multiple testing. *J. R. Statist. Soc. B*, 57(1):289–300, 1995. 12, 13, 14

R. Bolger. High-throughput screening: new frontiers for the 21st century. *Drug Discovery Today*, 4(6):251–253, 1999. 7

Boone Lab, 2011. *url:* www.utoronto.ca/boonelab/ - global mapping of genetic interactions in yeast (*Saccharomyces cerevisiae*). 8

E. Brodin. A non-mathematical introduction to statistics of extremes. *Scandinavian Insurance Quarterly*, 3:247–252, 2006. 6

S.G. Coles. *An Introduction to Statistical Modeling of Extreme Values.* Springer, London, 2001. 5, 6

M. Costanzo, A. Baryshnikova, J. Bellay, Y. Kim, E.D. Spear, C.S. Sevier, H. Ding, J.L.Y. Koh, K. Toufighi, S. Mostafavi, J. Prinz, R.P.St. Onge, B. VanderSluis, T. Makhnevych, F.J. Vizeacoumar, S. Alizadeh, S. Bahr, R.L. Brost, Y. Chen, M. Cokol, R. Deshpande, Z. Li, Z.-Y. Lin, W. Liang, M. Marback, J. Paw, B.-J. San Luis, E. Shuteriqi, A.H.Y. Tong, N. van Dyk, I. M.

Wallace, J.A. Whitney, M.T. Weirauch, G. Zhong, H. Zhu, W.A. Houry, M. Brudno, S. Ragibizadeh, B. Papp, C. Pál, F.P. Roth, G. Giaever, C. Nislow, O.G. Troyanskaya, H. Bussey, G.D. Bader, A.-C. Gingras, Q.D. Morris, P.M. Kim, C.A. Kaiser, C.L. Myers, B.J. Andrews, and C. Boone. The genetic landscape of a cell. *Science*, 327(5964):425–431, 2010. 8

P. Dragiev, R. Nadon, and V. Makarenkov. Systematic error detection in experimental high-throughput screening. *BMC Bioinformatics*, 12(1):25, 2011. 11

R.A. Fischer and L.H.C. Tippett. Limiting forms of the frequency distribution of the largest or smallest member of a sample. *Proceedings of the Cambridge Philosophical Society*, 24:180–190, 1928. 4

G. Giaever, A.M. Chu, L. Ni, C. Connelly, L. Riles, S. Véronneau, S. Dow, A. Lucau-Danila, K. Anderson, B. André, A.P. Arkin, A. Astromoff, M. El-Bakkoury, R. Bangham, R. Benito, S. Brachat, S. Campanaro, M. Curtiss, K. Davis, A. Deutschbauer, K.D. Entian, P. Flaherty, F. Foury, D.J. Garfinkel, M. Gerstein, D. Gotte, U. Güldener, J.H. Hegemann, S. Hempel, Z. Herman, D.F. Jaramillo, D.E. Kelly, S. L. Kelly, P. Kötter, D. LaBonte, D.C. Lamb, N. Lan, H. Liang, H. Liao, L. Liu, C. Luo, M. Lussier, R. Mao, P. Menard, S.L.L. Ooi, J.L. Revuelta, C.J. Roberts, M. Rose, P. Ross-Macdonald, B. Scherens, G. Schimmack, B. Shafer, D.D. Shoemaker, S. Sookhai-Mahadeo, R.K. Storms, J.N. Strathern, G. Valle, M. Voet, G. Volckaert, C.Y. Wang, T.R. Ward, J. Wilhelmy, E.A. Winzeler, Y. Yang, G. Yen, E. Youngman, K. Yu, H. Bussey, J.D. Boeke, M. Snyder, P. Philippsen, R.W. Davis, and M. Johnston. Functional profiling of the saccharomyces cerevisiae genome. *Nature*, 418(6896):387–391, 2002. 7

B.V. Gnedenko. Sur la distribution limit'e du term d'une série aléatoire. *Annals of Mathematics*, 44:423–453, 1943. 4

E.J. Gumbel. *Statistics of Extremes.* Columbia University Press, New York, 1958. 4

R.P. Hertzberg and A.J. Pope. High-throughput screening: new technology for the 21st century. *Current Opinion in Chemical Biology*, 4(4):445–451, 2000. 7

S. Karlin and A. Dembo. Limit distributions of maximal segmental score among Markov-dependent partial sums. *Adv. Appl. Prob.*, 24:113–140, 1992. 20

T.A. Knijnenburg, L.F.A. Wessels, J.T.M. Reinders, and I. Shmulevich. Fewer permutations, more accurate P-values. *Bioinformatics*, 25(12):161–168, 2009. 9

M.R. Leadbetter, G. Lindgren, and H. Rootzén. *Extremes and related properties of random sequences and processes.* Springer series in statistics. Springer-Verlag, 1983. 5, 6

R. Macarron, M.N. Banks, D. Bojanic, D.J. Burns, D.A. Cirovic, T. Garyantes, D.V.S. Green, R.P. Hertzberg, W.P. Janzen, J.W. Paslay, U. Schopfer, and G.S. Sittampalam. Impact of high-throughput screening in biomedical research. *Nature Reviews Drug Discovery*, 10(3):188–195, 2011. 7, 10

N. Malo, J.A. Hanley, G. Carlile, J. Liu, J. Pelletier, D. Thomas, and R. Nadon. Experimental design and statistical methods for improved hit detection in high-throughput screening. *Journal of Biomolecular Screening*, 15(8):990–1000, 2010. 10

MATLAB. *Version 7.10.0 (R2010a).* The MathWorks, Inc., Natick, Massachusetts, 2010. 19

L.M. Mayr and D. Bojanic. Novel trends in high-throughput screening. *Current Opinion in Pharmacology*, 9(5):580 – 588, 2009. 7

K.P. Mishra, L. Ganju, M. Sairam, P.K. Banerjee, and R.C. Sawhney. A review of high throughput technology for the screening of natural products. *Biomedicine & Pharmacotherapy*, 62(2):94 – 98, 2008. 7

J. Pickands. Statistical inference using extreme order statistics. *Annals of Statistics*, 3:119–131, 1975. 5

PROPHECY, 2011. *url:* prophecy.lundberg.gu.se - quantitative information about phenotypes for the complete collection of deletion strains in yeast (*Saccharomyces cerevisiae*). 8

R.-D. Reiss and M. Thomas. *Statistical Analysis of Extreme Values. With Applications to Insurance, Finance, Hydrology and Other Fields.* Birkhäuser, Basel, 2nd edition, 2001. 6

H. Rootzen and D.S. Zholud. Tail estimation methods for the number of false positives in high-throughput testing. *Submitted*, 2011.

SmartTail, 2011. *url:* www.smarttail.se - software for the analysis of false discovery rates in high-throughput screening experiments. 18

J.D. Storey. A direct approach to false discovery rates. *J. R. Statist. Soc. B*, 64(3):479–498, 2002. 12, 13, 14, 15, 19

J.D. Storey. The positive false discovery rate: a bayesian interpretation and the q-value. *The annals of Statistics*, 31(6):2013–2035, 2003. 14

J.D. Storey. Srong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. *J. R. Statist. Soc. B*, 66(1):187–205, 2004. 14

J. Warringer and A. Blomberg. Automated screening in environmental arrays allows analysis of quantitative phenotypic profiles in saccharomyces cerevisiae. *Yeast*, 20:53–67, 2003. 7

J. Warringer, E. Ericson, L. Fernandez, O. Nerman, and A. Blomberg. High-resolution yeast phenomics resolves different physiological features in the saline response. *Proc Natl Acad Sci USA*, 100(26):15724–15729, 2003. 7

D.S. Zholud. Extremes of the Shepp statistic for the Wiener process. *Extremes*, 11(4):339–351, 2008.

D.S. Zholud. Extremes of the Shepp statistic for a Gaussian random walk. *Extremes*, 12(1):1–17, 2009.

D.S. Zholud. Tail approximations for the Student $t-$, $F-$, and Welch statistics for non-normal and not necessarily i.i.d. random variables. *Submitted*, 2011a.

D.S. Zholud. Confidence intervals for the False Discovery Rate - the SmartTail method. *Work in progress*, 2011b. 15

D.S. Zholud, H. Rootzèn, O. Nerman, and A. Blomberg. Positional effects in biological array experiments and their impact on False Discovery Rate. *Work in progress*, 2011. 8

Part II

The Papers

PAPER I

PAPER I

Tail estimation methods for the number of false positives in high-throughput testing

Holger Rootzén and Dmitrii Zholud *

Abstract

This paper develops methods to handle false rejections in high-throughput screening experiments. The setting is very highly multiple testing problems where testing is done at extreme significance levels and with low degrees of freedom, and where the true null distribution may differ from the theoretical one. We show that the conditional distribution of the number of false positives, given that there is in all r positives, approximately has a binomial distribution, and develop efficient and accurate methods to estimate its success probability parameter. Furthermore we provide efficient and accurate methods for estimation of the true null distribution resulting from a preprocessing method, and techniques to compare it with the theoretical null distribution. Extreme Value Statistics provides the natural analysis tools, a simple polynomial model for the tail of the distribution of p-values. We provide asymptotics which motivate this model, exhibit properties of estimators of the parameters of the model, and point to model checking tools, both for independent data and for dependent data. The methods are tried out on two large scale genomic studies and on an fMRI brain scan experiment. A software implementation, SmartTail, may be downloaded from the web.

Keywords high-throughput screening, false discovery rate, FDR, SmartTail, correction of theoretical p-values, extreme value statistics, multiple testing, Bioscreen MBR, MRI brain scan, microarray data analysis.

AMS 2000 Subject Classifications: $\frac{\text{Primary-62G32;}}{\text{Secondary-62P10, 60G70;}}$.

Department of Mathematical Statistics
Chalmers University of Technology and University of Göteborg, Sweden.
E-mails: hrootzen@chalmers.se and dmitrii@zholud.com

1 Introduction

Setting: High-throughput measurements and screenings in modern bioscience differ from classical statistical testing in several ways. First, it involves testing thousands - or hundreds of thousands - of hypotheses. Second, to get a manageable amount of rejected null hypotheses, testing is typically done at extreme significance levels, e.g. with $\alpha = 0.001$ or smaller. Third, each of the individual tests are often based on very few observations, so that the degrees of freedom for $t-$ or $F-$tests may be as low as 1 or 2, and degrees of freedom less than 10 are common. Fourth, in such large and complicated experiments the real null distribution of test statistics and p-values frequently deviates from the theoretical one.

In Section 5 we consider three such high-throughput investigations: testing of gene interaction in yeast using a Bioscreen C Analyzer robot Warringer et al. (2003), Zholud et al. (2011); a Genome Wide association scan *Arabidopsis* microarray experiment Zhao et al. (2007); and a fMRI brain imaging experiment, see Dehaene-Lambertz et al. (2006) and Taylor and Worsley (2006). The number of tests for the different data sets in these investigations varied between 1,700 and 35,000 and the typical significance levels were 0.001 or less. The degrees of freedom was 1, the lowest possible, in the Bioscreen experiment. It seemed clear that for all three investigations the real null distribution differed from the theoretical one. Property four, that the real null distribution often is different from the hypothetical one has been widely observed and discussed in the literature. For a few examples see Efron et al. (2001), Efron (2004, 2008), Jin and Cai (2007), Zhao et al. (2007), and Schwartzman (2008). Cope et al. (2004) developed a standardized set of graphical tools to evaluate high-throughput testing data.

False positives: The aim of this paper is to develop methods to understand and handle false rejections in high-throughput screening experiments. Thus we study very highly multiple testing problems where testing is performed at extreme significance levels and with low degrees of freedom, and where the true null distribution may differ from the theoretical one. Our illustrations come from biology, but the same problems appear in many other areas, too.

Our point of view is the following: the tests use a very small α and hence false rejections are determined by the extreme tails of the distributions of test statistics and p-values, and the central parts of the distributions are largely irrelevant. For this reason we throughout use the powerful methods for tail estimation which have been developed in Extreme Value Statistics.

Our main results are answers to the questions *"how many of the positive test results are false?"* and *"how should one judge if one preprocessing method makes the true null distribution closer to the theoretical one than another method?"* for such testing problems.

Our answer to the first question is i) the conditional distribution of the number of false positives given that there are in all r positives is approximately binomial, and ii) efficient and accurate methods to estimate the success probability parameter of this binomial distribution.

As answer to the second question we provide efficient and accurate methods for estimation of the true null distribution resulting from a preprocessing method, and techniques to compare it with the theoretical null distribution.

Perhaps the words *efficient* and *accurate* above should be emphasized. Existing approaches use either fully parametric models for the distributions of test quantities or p-values, or else use the empirical distribution function as estimator. Our approach instead is semi-parametric: we use a parametric model, but only for the tails of the distributions. The meaning of "efficient" then is that the random variation in our estimates is substantially smaller than for the empirical distribution function. With "accurate" we mean that we do not make the very strong assumptions that models like normal or beta distributions can be trusted far out in the tails of the distribution. Instead we only model the tail, and let data determine the part of the tail for which the model fits well. For the details of this, see Sections 2 below, and for some concrete numerical results see Section 5.

A third contribution of this paper is that it provides an accurate estimator of Efron's local false discovery rate $\mathrm{fdr}(x)$, see later in this section. Note that the empirical distribution does not provide any estimate of fdr(x) at all, and hence one can not talk about our estimator's "efficiency" relative to the empirical estimator (this is why here we write "accurate", not "efficient and accurate").

There is an enormous and rapidly expanding literature on multiple testing. The monograph of Dudoit and van der Laan (2008) is one entrance point into this literature. Kerr (2009) gives a recent useful review of the area. Noble (2009) is directed at practitioners. Below we discuss some specific parts of the recent literature on multiple testing more in detail. The recent paper Knijnenburg et al. (2009) suggests using Generalized Pareto approximations to improve efficiency of permutation test, in particular in bioinformatics. We are not aware of any other papers which connect EVS and high-throughput testing.

Estimation, not error control: The aim of this paper is estimation of the distribution of the number of false positives and of related quantities, and not "error control". The False Discovery Rate (FDR) error control procedure, Benjamini and Hochberg (1995), has had an enormous influence on the field of multiple testing, and has seen extensive further development. However, in screening studies the aim isn't final confirmation of an effect. It instead is to select a number of interesting cases for further study. In such situations, error control may be less natural. The estimation approach to multiple testing of course already has attracted significant interest in the literature. E.g., this is the point of view in Storey (2002, 2003, 2004), Efron et al. (2001), Efron (2004, 2008), Ruppert et al. (2007), and Jin and Cai (2007).

By way of further comment, in high-throughput testing and for many standard tests, such as $t-$ or $F-$tests, the error control provided by the Benjamini-Hochberg method is different from what one naively could be led to expect. The reason is that for such tests the ratio {probability of false rejection}/{probability of rejection} converges to a constant as the significance level tends to zero (see e.g. Zholud (2011a)). Then, if the desired FDR is less than this constant, the Benjamini-Hochberg method simply makes it highly probable that there are no rejections at all. Since FDR is defined to be zero if there are no rejections this in turn makes the achieved FDR small. However, this may be more a formality than anything else. For high-throughput screening it seems to be more useful to have good estimates of the probability of false rejection and of the distribution of the number of false rejections, rather than FDR control. These issues are discussed in further in e.g. Storey (2002, 2003), Kerr (2009), and Zholud (2011a).

Tail model: Our methods can equivalently be presented in terms of test statistics or in terms of p-values. We have found the latter formulation convenient and use it throughout.

Let P denote a generic p-value, and, as usual, write H_0 for the null hypothesis and H_1 for the alternative hypothesis. Our basic model is that there are positive constants $c_0, \gamma_0, c_1, \gamma_1$ such that

$$F_0(x) = \Pr(P \leq x \mid H_0) = c_0 x^{1/\gamma_0}(1 + o(1)) \tag{1.1}$$

and

$$F_1(x) = \Pr(P \leq x \mid H_1) = c_1 x^{1/\gamma_1}(1 + o(1)), \tag{1.2}$$

as $x \to 0$. This model is motivated by general asymptotic arguments from extreme value theory, and also by extensive practical experience from extreme value statistics. The theoretical motivation and references are given in Section 4. Section 5 gives a data-based motivation using three studies from biology, and also introduces and illustrates a number of model checking tools, cf. Figure 2.2 in Section 5.

If one further assumes that tests are independent, that H_0 is true with probability π_0, and H_1 is true with probability $\pi_1 = 1 - \pi_0$, then we get the following extreme tail mixture form for the distribution

$$F(x) = \Pr(P \leq x) = \big(\pi_0 c_0 x^{1/\gamma_0} + \pi_1 c_1 x^{1/\gamma_1}\big)(1 + o(1)), \tag{1.3}$$

for the observed p-values. If the testing procedure is reasonable, then small p-values should be more likely under H_1 than under H_0, so it is typically reasonable to expect that $\gamma_0 \leq \gamma_1$.

Mixture models are of course very widely used in many areas, and in particular in multiple testing in bioinformatics. For a long list of references to this, see Kerr (2009); in particular Allison et al. (2002) discusses such models under the heading Mixture Model Methods, or MMM-s. We find these models natural and useful. However the results of this paper continue to hold also for models where the numbers of true and false null hypotheses are considered as fixed numbers, see below.

Asymptotics and the tail model: The situation described above is in an "asymptotics formulation" described as

$$n \,\text{fixed}, m \to \infty, \ \alpha \to 0, \tag{1.4}$$

where n is the number of observations used in the individual tests, m is the number of tests, and α is the significance level.

The first two assumptions, n fixed, $m \to \infty$, delineate the class of high-throughput testing situations which are studied in this paper. As for the third one, $\alpha \to 0$, suppose one chooses to reject the null hypotheses for all tests that give a p-value less than some critical value α. For example, in a Bonferroni procedure one would choose $\alpha = \eta/m$ (cf. Gordon et al. (2007)), with η "a fixed number". In theory the choice of η, and hence of α, would be guided by the fact that $\pi_0\eta$ is the expected number of false rejections (provided the null distribution one uses is, in fact, the true one). In practice, the choice of α is often based on beliefs of how often the null hypothesis is violated, and on available capacity for further study of rejected hypotheses. Since the number of p-values, m, is assumed to be very large, α in the end typically in the situations we consider is chosen quite small, also because one wants to get a manageable number of rejections. Hence, the assumption $\alpha \to 0$.

There of course exists very large literature on central limit type approximations for the case $n \to \infty$. E.g., sharp results for uniformity in approximate $t-$distributions when $n \to \infty$ and $m \to \infty$ simultaneously and a literature review is given by Fan et al. (2007). However, this is not the case of interest here, and it is of course well known that for low values of n approximations of t or $F-$distributions can be quite inaccurate. In another set of literature it is instead proven that if the underlying observations deviate from normality or independence then, under very general conditions, tails of one- and two-sample t-statistic and of $F-$statistic are not the same as if the observations really were normal, see Hotelling (1961), Zholud (2011a), and references in the latter paper. However the latter paper also shows that the deviation is of a simple kind: under the asymptotics (1.4) the tail probabilities under non-normality are proportional to the tails of the relevant t or $F-$distributions, see Zholud (2011a). It in particular follows that (1.1) - (1.3) are satisfied, since these equations are known to hold for $t-$ and $F-$distributions.

The Extreme Tail Mixture Model: It is reasonable to neglect the $o(1)$-terms in (1.1) - (1.3) if x is not too far from α, and α is small. This leads to the following Extreme Tail Mixture Model

$$F_0(x) = \Pr(P \leq x \mid H_0) = c_0 x^{1/\gamma_0}, \tag{1.5}$$

$$F_1(x) = \Pr(P \leq x \mid H_1) = c_1 x^{1/\gamma_1}, \tag{1.6}$$

and

$$F(x) = \Pr(P \leq x) = \pi_0 c_0 x^{1/\gamma_0} + \pi_1 c_1 x^{1/\gamma_1}. \tag{1.7}$$

If this model holds, tests are made at the "fixed" level α, and assuming independence between tests, then (see Section 3 below) the conditional distribution of the number of false rejections, given that there has been $r > 0$ rejections in total, has approximately a binomial distribution with "number of trials" parameter r and success probability parameter

$$\text{pFDR} = \frac{\pi_0 c_0 \alpha^{1/\gamma_0}}{\pi_0 c_0 \alpha^{1/\gamma_0} + \pi_1 c_1 \alpha^{1/\gamma_1}} = \frac{\pi_0 F_0(\alpha)}{F(\alpha)}. \tag{1.8}$$

Here we use the notation pFDR for this parameter because it coincides with the pFDR of Storey (2002), see Section 3. (The leftover case $r = 0$ is not interesting - if there are no rejections, then one knows for sure that there are no false rejections either!)

These results apply also for the more ad hoc method, which is presumably often used in practice, where one chooses to reject the null hypothesis for the r tests which gave the smallest p-values, with "r a fixed number". Again, the choice of r is typically influenced by beliefs about the frequency of null hypothesis violations, and on available capacity for study of rejected hypotheses.
Further, the local false discovery rate (Efron et al. (2001)), which measures the "a posteriori likelihood of false rejection" of a hypothesis with p-value $x \leq \alpha$ is then

$$\begin{aligned} \text{fdr}(x) = \Pr(H_0 \text{ true} \mid P = x) &= \frac{\pi_0 c_0 \gamma_0^{-1} x^{1/\gamma_0}}{\pi_0 c_0 \gamma_0^{-1} x^{1/\gamma_0} + \pi_1 c_1 \gamma_1^{-1} x^{1/\gamma_1}} \\ &= \frac{\pi_0 \frac{d}{dx} F_0(x)}{\frac{d}{dx} F(x)}. \end{aligned} \tag{1.9}$$

The empirical distribution function doesn't provide any estimate of fdr. An alternative nonparametric possibility could be to use some kernel type estimator. However such estimators are well known to provide erratic tail estimators.

The simplest situation is when the theoretical null distribution is in fact the true null distribution, so that F_0 is the uniform distribution and $c_0 = \gamma_0 = 1$. However our focus (cf. discussion above) is situations where this doesn't hold, but we instead have an additional sample where the null hypothesis is known to be true. It is typically quite worth the effort to acquire such a sample. It could be achieved by performing an extra experiment, as in the Bioscreen example in Section 5, or, in the brain imaging example, by finding regions where it seems likely there are no effects, or by randomly changing signs in contrasts, or in other ways, see e.g. in Efron et al. (2001) and Taylor and Worsley (2006).

In this paper we show how Extreme Value Statistics can be used to get *efficient* and *accurate* estimates of the distributions (1.5) and (1.7), and of their derivatives, for the small x which are the values of interest in the present situation. EVS in addition provides confidence intervals and goodness-of-fit tests and focusses analysis and graphics on what is at the heart of the problem. The remaining parameter π_0 in (1.8) and (1.9) has to be estimated through other means, e.g. by a variant of the method in Storey (2002) – this is the only point where the entire range of p-values comes into play. Alternatively, π_0 can be conservatively estimated by setting it equal to 1. The loss of accuracy in the conservative approach is small in the situations we consider, and is quantified by (1.8) and (1.9). The estimates of $F_0(x)$, $F(x)$, and π_0 directly lead to estimates of the success probability parameter pFDR in the binomial distribution, and of fdr(x).

Dependence: Complexity and preprocessing in high-throughput testing can introduce dependence between tests. The effects of time series dependence on extremes has been extensively studied in the extreme value literature, and many of the issues are well understood. In particular, extremes may be asymptotically independent even if typical observations are dependent. For one instance of this, in a paper inspired by high-throughput testing in biology, see Clarke and Hall (2009). However, also for the opposite case when extremes are "asymptotically dependent", there exist good methods to deal with time dependence. Further, even if less is rigorously proven for the more complicated "spatial" dependence which often is of interest, e.g. in gene expression experiments, it is typically relatively clear how the known time series results extend to such situations. In the following sections we provide some more details on this.

Overview: In Section 2 we develop the estimation methods, and discuss how dependence influences estimation. Section 3 derives the conditional binomial distribution for the number of false positives. It also provides estimates of error control parameters such as the Benjamini-Hochberg False Detection Rate and the Family-Wise ERror. Section 4 gives the motivation for the models (1.1) and (1.2). In Section 5 we use our methods to analyze the two data sets from genomics, and the brain imaging data set discussed above. Section 6 contains a concluding discussion.

A statistical software tool, SmartTail, for performing the analyzes described in this paper may be downloaded from the web, see www.smarttail.se, and details on technical implementation of the methods can be found in Zholud (2011c,b).

2 Statistical methods

In this section we first discuss how to estimate $F_0(x)$, and how to test and produce confidence intervals. This discussion is for the case of independent p-values. A natural simplification of the mixture model (1.3) then makes it possible to use the same methods to estimate $F(x)$. For $t-$ and $F-$tests with low degrees of freedom, it may sometimes be reasonable to replace γ_0 and γ_1 by 1, and, if the theoretical null distribution is the true one, then $c_0 = \gamma_0 = 1$.

We further discuss how the method of Storey (2002) to estimate π_0 translates to the present setting. Together this provides all the ingredients needed for estimation of (1.8) and (1.9). Finally, if there is dependence between observations the estimators still are consistent and asymptotically normal, but if there is clustering of extremely small p-values, then the standard deviations of the estimators may be inflated.

Estimation of $F_0(x)$: To estimate F_0 we assume that it has been possible to obtain a (perhaps approximate) sample of m_0 p-values, $p_1^0, \ldots, p_{m_0}^0$, from the true null distribution, cf. the discussion in the introduction. Our EVS procedure for estimation of the parameters of (1.1) is then as follows:

The first step is to choose a threshold $u > 0$ which is small enough to make it possible to neglect the $o(1)$ term in (1.1) for $x \leq u$ and hence to use the Extreme Tail Model (1.5). This u then plays the central role that the statistical analysis only uses those of the observations (i.e. the p_i^0-s) which are less than u, and that the analysis only tries to estimate values of $F_0(x)$ for $x \leq u$.

This choice of the threshold u is a compromise between bias and variance (or, in the terminology of the introduction, between accuracy and efficiency) and is similar to the choice of bandwidth in kernel density estimation: a small u leads to less "model error", and hence less bias, but also to fewer observations to base estimation on, and hence more variance. In practice, the choice of u is guided by goodness-of-fit test and plots; see Coles (2001), Beirlant et al. (2004), and the analysis of the examples in Section 5.

For the next step, write P^0 for a random variable which has the (true) null distribution of the p-values. From (1.5) it follows that

$$\Pr\big(-\log(P^0/u) \geq x \mid P^0 \leq u \big) = \frac{c_0(ue^{-x})^{1/\gamma_0}}{c_0 u^{1/\gamma_0}} = e^{-x/\gamma_0}, \quad (2.1)$$

for x positive. Thus, the variable $-\log(P^0/u)$ conditionally on $P^0 \leq u$ has an exponential distribution with mean γ_0. Let $N = \#\{1 \leq i \leq m_0;\ p_i^0 \leq u\}$ be the number of the $p_1^0, \ldots, p_{m_0}^0$ that are less than u. Since the sample mean of the observations is the natural estimator of the mean of an exponential distribution, the natural estimator of γ_0 is

$$\hat{\gamma}_0 := \frac{1}{N} \sum_{1 \leq i \leq m_0;\, p_i^0 \leq u} -\log(p_i^0/u). \quad (2.2)$$

This is just the ubiquitous Hill estimator in a somewhat different guise, cf. Beirlant et al. (2004), Section 4.2. Further, for $0 \leq x \leq u$, we have that $F_0(x) = \Pr\big(P^0 \leq x\big) = \Pr\big(P^0 \leq u\big)\Pr\big(P^0 \leq x \mid P^0 \leq u\big) = \Pr\big(P^0 \leq u\big)\, c_0 x^{1/\gamma_0}/(c_0 u^{1/\gamma_0})$. Since N/m_0 is the nonparametric estimator of $\Pr\big(P^0 \leq u\big)$ we get the semiparametric estimator

$$\hat{F}_0(x) = \frac{N}{m_0} \left(\frac{x}{u}\right)^{1/\hat{\gamma}_0} \quad (2.3)$$

of $F_0(x)$, for $0 \leq x \leq u$. An estimate of $\frac{d}{dx}F_0(x)$ is obtained by differentiating (2.3).

The important point here is the following. We only trust the model $F_0(x) = c_0 x^{1/\gamma_0}$ to be sufficiently accurate for "small" values of x, i.e for $x \leq u$, where u is "small". However, still this threshold u often can be chosen much larger than the critical value $x = \alpha$ used to decide if a test rejects or not, and hence the estimate $\hat{F}_0(\alpha)$ is based on many more observations - and accordingly is much more efficient - than the standard empirical distribution function estimator of $F_0(\alpha)$. Quantitative examples of this are given below, and for any specific data set the efficiency gain can be obtained from the SmartTail software.

If the observations $p_1^0, \ldots, p_{m_0}^0$ are independent, then the variance of N/m_0 is estimated by $\frac{N}{m_0}\left(1 - \frac{N}{m_0}\right)/m_0$. Since the variance of an exponential distribution is equal to the mean we have that conditionally on N the variance of $\hat{\gamma}_0$ is γ_0/N. Hence the variance of $\hat{\gamma}_0$ may be estimated by $\hat{\gamma}_0/N$. Further the parameter estimators N/m_0 and $\hat{\gamma}_0$ are asymptotically uncorrelated and asymptotically normally distributed. Thus, asymptotic confidence intervals, e.g. for $\hat{F}_0(\alpha)$, can be computed using the delta method, see Zholud (2011b).

Estimation of $F(x)$: The straightforward way to estimate parameters in the mixture density (1.3) would be to write down the joint conditional likelihoods of the sample from the null distribution and of the observed p-values $p_1, \ldots, p_m$ that are less than the threshold u and then maximize numerically to find the parameters $c_0, c_1, \gamma_0, \gamma_1, \pi_0$. However, if $\gamma_0 = \gamma_1 =: \gamma$ then the model collapses to

$$F(x) = cx^{1/\gamma}, \tag{2.4}$$

with $c = \pi_0 c_0 + \pi_1 c_1$, and the parameters become unidentifiable. This would presumably also make it difficult to estimate parameters if γ_0 and γ_1 are similar, even if they are not exactly equal. This identifiability problem is further compounded by the fact that in typical situations where the test works as desired, the first term in (1.7) would be substantially smaller than the second one – if not there would be too many false rejections.

However, turning this around, one can often expect that (2.4) in fact would model the observed p-values quite well. We hence propose the following procedure: First estimate γ in the model (2.4) from the observed p-values $p_1, \ldots, p_m$ in precisely the same way as γ_0 was estimated from $p_1^0, \ldots, p_{m_0}^0$. If this estimate is reasonably close to the estimate of γ_0 - just use the model (2.4) for the distribution of p-values in the experiment and estimate $F(x)$ in the same way as $F_0(x)$. Confidence intervals for $F(x)$ are also obtained in the same way as for $F_0(x)$.

If the estimated γ_0 and γ are substantially different, then one might try to complement with the maximum likelihood approach outlined above, perhaps with π_0 estimated ”externally” by just guessing, or by the Storey (2002) method which we discuss below. The final decision on whether to use (1.7) or (2.4) can then be based on the extent to which the fitted distributions differ, and on the relative sizes of the two terms in (1.3).

As a further comment, the preceding results of course are valid not only for the mixture model, but also if one assumes fixed numbers of true and false null hypotheses.

As mentioned above, there are important cases when some of the parameters are known from theory. In particular, if the p-values have been produced by one- or two-sample $t-$tests or $F-$tests, then $\gamma_0 = \gamma_1 = 1$, see Zholud (2011a), and in particular (2.4) is satisfied. If the theoretical null distribution is in fact equal to the true one, then $\gamma_0 = c_0 = 1$. In such cases one may of course use these known values instead of the estimates – but it still may be a good idea to check if they agree with the estimates.

Estimation of π_0: Storey (2002) proposed the following conservative estimator $\hat{\pi}_0 = (\#\{p_i > \lambda\}/m)/(1-\lambda)$ for the proportion of cases where the null hypothesis is true. Here $\lambda \in (0,1)$ is a suitably chosen (not ”small”) number. The idea behind the estimator is that ”most” of the large p-values come from the null distribution, so that the numerator is an estimate of $\pi_0 \Pr(P^0 > \lambda)$, while the denominator is an estimate of $\Pr(P^0 > \lambda)$, provided the p-values in fact are uniformly distributed under the null hypothesis. In the present situation where this last assumption may not be true one can instead use the estimator

$$\hat{\pi}_0 = \frac{\#\{p_i > \lambda\}/m}{\#\{p_i^0 > \lambda\}/m_0}.$$

The choice of λ is discussed in Storey's paper.

Efficiency and accuracy : As one quantification of the gain in efficiency from using the Extreme Tail Model instead of the empirical distribution function, assume that the model (2.4) holds exactly for $x \leq u$, and let $\hat{F}_E(x)$ be the empirical distribution function estimator of $F(x)$. Straightforward but lengthy calculations show that then, for $x \leq u$ and $F(x) \leq 0.01$,

$$\frac{\mathrm{Var}(\hat{F}(x))}{\mathrm{Var}(\hat{F}_E(x))} \leq \left(\frac{x}{u}\right)^{1/\gamma}\left(1+\frac{1.02}{\gamma^2}\left(\log\left(\frac{x}{u}\right)\right)^2\right),$$

and that this bound is quite precise. For details see Zholud (2011b). In practical use often x/u would be of the order of 0.1 - 0.01 and the value of γ would be around 1. The resulting efficiency gain is illustrated in Figure 2.1 for three values of γ.

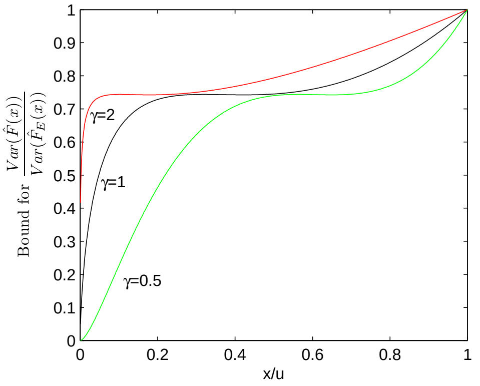


Figure 2.1: Efficiency gain from using the Extreme Tail Model as compared to the standard empirical CDF.

The same results of course apply to $\hat{F}_0(x)$. The examples in Section 5 contain some further numerical examples of the increased efficiency which can be obtained by using the Extreme Tail Model. As a final comment, for very small values of x, which sometimes can be of interest, the empirical distribution function is just not possible to use as an estimator.

As for accuracy, e.g. the papers Efron (2004, 2008, 2010) first uses the theoretical null distribution to transform test statistics to a N(0, 1) distribution and then accounts for deviations from this theoretical null distribution by fitting a N(μ, σ) distribution to these values, and finally uses the result as a tail approximation.

Schwartzman (2008) instead assumes an exponential family of distributions. These papers emphasize the risks for wrong inferences which can result from the theoretical null distribution being different from the theoretical one, but don't seem to account for the risk that the assumed distribution doesn't fit in the tails. In Allison et al. (2002) tail approximations are instead obtained by fitting a beta distribution to observed p-values. Tang et al. (2007) increases the sophistication of this approach by adding a Dirichlet prior. In these approaches, data from the center of the distributions determine the fit, and thus the statistical variability of the resulting estimators is small. Instead the fitted models are assumed to hold in the extreme tails. However, for the case considered here where each test is based on very few observations, such an assumption is not backed by theory, and in many similar situations has been observed to lead to bad modeling of tails (also in cases where the parametric model fitted very well in the center of the data). Thus there is substantial risk that these methods could lead to bad accuracy and, in our opinion, if they nevertheless are used they should at least be checked through comparison with methods such a those in this paper. In particular, it is wise to employ appropriate graphical tools which concentrate on tail fit, rather than on overall fit of distributions.

Dependence: The estimators discussed above are consistent and asymptotically normal also for dependent observations, in quite general circumstances. For the standard Hill estimator and for the "time series case" Rootzén et al. (1991) and Hsing (1991) give the precise conditions for this, using strong mixing as the dependence condition. Via (2.1) and (2.2) this directly provides corresponding results for the present situation.

Dependence can potentially inflate the variances of the estimators. However this only happens if there is clustering of small p-values, i.e. if the small p-values appear in clusters located closely together. Thus, if there isn't any clustering one can just ignore dependence and proceed as if the p-values were independent. How to check for clustering is discussed in the next section.

If extreme values in fact do cluster, it is nevertheless still possible, see Rootzén et al. (1991), to estimate the asymptotic variance by "blocking". In this method the p_i^0-s are first grouped in equal sized blocks of neighboring observations, and then estimates of γ in the blocks are used to estimate the variance, in the following way: Let ℓ be the lengths of the blocks, and for simplicity assume that $n = \ell \times k$ for k integer (if this is not satisfied, one may discard the last few observations). The standard deviation of $\hat{\gamma}_0$ is then estimated by the standard deviation of the k γ-estimates computed in the blocks, divided by $\sqrt{k}$. The standard deviation of $\hat{\gamma}$ is estimated in the same way.

The choice of ℓ is again a compromise between bias and variance: too small an ℓ might cut to many clusters into two and lead to an underestimated standard deviation, while too large an ℓ gives too few block averages to use for estimation, and hence very variable estimates of the standard deviation. In practice one typically would compute the estimates for a range of values of ℓ and use this together with visual information about clustering to make the final choice of ℓ. This method of course is closely related to the block bootstrap methods used in time series analysis.

In addition to γ_0 the estimator (2.3) of $\hat{F}_0(x)$ also contains the factor N/m_0 where $N = \#\{1 \leq i \leq m_0;\, p_i^0 \leq u\}$. The discussion above of the influence of dependence on γ_0 carries directly over to the variance of N and the covariance between γ_0 and N, see Rootzén et al. (1991).

In biological applications it is common that there is no natural linear ordering of the observations. However, it may still be possible to group the observations into equal-sized blocks such that there may be considerable dependence inside the blocks, while block averages are substantially independent. If this is possible, the method described above can still be used to estimate the standard deviations of $\hat{\gamma}_0$ and $\hat{\gamma}$.

3 Basic theory

The basic binomial conditional distribution for the number of false positives in the introduction follows from completely elementary reasoning. However, for completeness we still give a derivation of it for three cases: the basic model (1.3) where H_0 is true with probability π_0 and H_1 is true with probability π_1; the case when $m_0 = \#\{\text{false null hypotheses}\}$ is thought of as non-random; and a case when the critical level α is a random variable, e.g. when it is equal to the k-th largest p-value for "k non-random".

Although this is not the main thrust of this paper we also briefly illustrate how the estimates from the previous section may be used to give efficient and accurate estimates of some other standard error control parameters.

Approximate binomial distribution of the number of false positives: With notation from above we have that

$$\begin{aligned} m_0 &= \text{\# true null hypotheses} \\ m_1 &= \text{\# false null hypotheses} \\ m &= m_0 + m_1 = \text{total number of tests} \\ r &= \text{\# rejections} \\ \alpha &= \alpha_m = \text{critical level of the tests.} \end{aligned}$$

Further, in the derivations below, if nothing is said to the contrary, we assume that the observed p-values are mutually independent.

Now, suppose m_0 and m_1 are non-random and tend to infinity and that α_m tends to zero, in a coordinated way so that $m_0 F_0(\alpha_m)$, and $m_1 F_1(\alpha_m)$ are bounded. Accordingly, also the expected total number of rejections is bounded. Then, by the standard Poisson limit theorem for the binomial distribution, the number of false rejections is approximately Poisson distributed with parameter $m_0 F_0(\alpha_m)$ and the number of correct rejections is approximately Poisson distributed with parameter $m_1 F_1(\alpha_m)$. It then follows at once that the conditional distribution of the number of false rejections, given that there are in total r rejections, is approximately Binomial with number of trials parameter r and success probability $m_0 F_0(\alpha_m)/(m_0 F_0(\alpha_m) + m_1 F_1(\alpha_m))$. This is the same as (1.8) if $\pi_0 = m_0/m, \pi_1 = m_1/m$.

If instead the mixture model (1.3) is assumed to hold, then $m_0/m \xrightarrow{P} \pi_0$ as $m \to \infty$. Thus, for any $\epsilon > 0$ we have that for sufficiently large m the number of false rejections is less than the number of rejections of a sample of size $(\pi_0 + \epsilon)m$ so that the number of false rejections is stochastically smaller than a binomial variable with $m_0 + \epsilon m$ trials and success probability $F_0(\alpha_m)$. Similarly one gets an upper bound for the number of correct rejections, and also corresponding lower bounds. Using monotonicity and the Poisson limit distribution of binomial distributions, as above one obtains an approximate conditional binomial distribution with success probability (1.8) for the number of false rejections.

Next, if instead α_m is random and it is assumed that there exists a non-random sequence $\tilde{\alpha}_m$ such that $m|F(\alpha_m) - F(\tilde{\alpha}_m)| \xrightarrow{P} 0$, then a simple argument similar to the previous one shows that asymptotically there are no p-values in the interval with endpoints α_m and $\tilde{\alpha}_m$. It then follows that the conditional distribution of the number of false rejections is the same as if the rejection level was $\tilde{\alpha}_m$, and hence the results above again apply. In particular this is the case if α_m is the r-th smallest of the p-values, for some fixed r.

Dependence and clustering of small p-values: So far we have assumed that the p-values were independent. If they instead are dependent, then the results above continue to hold if Leadbetter's conditions $\mathrm{D}'(\alpha_m)$ and $\mathrm{D}(\alpha_m)$ are satisfied; see Leadbetter (1974) and Leadbetter and Rootzén (1998). Here $\mathrm{D}(\alpha_m)$ is a quite weak restriction on dependence at long distances, and can be expected to hold very generally. Instead $\mathrm{D}'(\alpha_m)$ restricts dependence between neighboring variables. It may be violated in circumstances where small p-values occur in clusters, and typically holds otherwise.

Clustering of p-values which could make $\mathrm{D}'(\alpha_m)$ invalid can be investigated informally by inspection of the samples, and there is also a large literature on formal estimation of the amount of clustering, as measured by the so-called Extremal Index, see e.g. Beirlant et al. (2004), Section 10.3.2. However, the issue is somewhat delicate: clustering caused by local dependence will violate the asymptotic Poisson distribution, but clusters of very small p-values may also be caused by non-null experiments occurring at neighboring locations, and this would then not contradict an asymptotic Poisson distribution. The latter situation, for example, is expected to occur in the brain scan experiment discussed in Section 5 below.

Estimation of error control parameters: With standard notation in multiple testing, let the random variables V and R be the number of false positives, and the total number of rejections, respectively. We now list number of common error control quantities, and how they may be estimated using the results from Section 2 (it is assumed that α and x are less than the threshold u). The second and third one have already been discussed above, but are included for completeness. For comprehensive listing and discussion of such parameters we refer to Dudoit and van der Laan (2008). Motivation of the estimators comes after the table. In each case a conservative estimate is obtained by setting $\hat{\pi}_0 = 1$, and the degree of conservatism can be judged directly from the formulas for the estimators.

Parameter	Estimate
The Benjamini and Hochberg (1995) False Detection Rate:	
FDR:= $E\left(\frac{V}{R} \mid R > 0\right) \Pr(R > 0)$	$\frac{\hat{\pi}_0 \hat{F}_0(\alpha)}{\hat{F}(\alpha)}(1 - e^{-m\hat{F}(\alpha)})$.
The Storey (2002) False Detection Rate:	
pFDR:= $E\left(\frac{V}{R} \mid R > 0\right)$	$\frac{\hat{\pi}_0 \hat{F}_0(\alpha)}{\hat{F}(\alpha)}$.
The Efron et al. (2001) Local False Detection Rate:	
fdr(x):= $\frac{\pi_0 \frac{d}{dx} F_0(x)}{\frac{d}{dx} F(x)}$	$\frac{\hat{\pi}_0 \frac{d}{dx} \hat{F}_0(x)}{\frac{d}{dx} \hat{F}(x)}$.
The FamilyWise ERror:	
FWER:= $\Pr(V \neq 0)$	$1 - e^{-m\hat{\pi}_0 \hat{F}_0(\alpha)}$.
The k-FamilyWise ERror:	
k-FWER:= $\Pr(V \geq k)$	$\sum_{i=k}^{\infty} \frac{(m\hat{\pi}_0 \hat{F}_0(\alpha))^i}{i!} e^{-m\hat{\pi}_0 \hat{F}_0(\alpha)}$.

The estimate of pFDR is a consequence of the conditional binomial distribution of the number of false positives. Specifically, conditional on $R = r > 0$ the number of false positives has a binomial distribution with parameters r and $\pi_0 F_0(\alpha)/F(\alpha)$ so that $E\big(V/R \mid R = r) = \pi_0 F_0(\alpha)/F(\alpha)$. Hence we also have that $E\big(V/R \mid R > 0) = \pi_0 F_0(\alpha)/F(\alpha)$. The estimate of FDR is obtained by using the pFDR estimate for the first factor and the asymptotic Poisson distribution of the number of false positives to estimate the second factor. The FWER and k-FWER estimates use the asymptotic Poisson distribution of the number of false positives. Since they also involve $\hat{\pi}_0$ they may be harder to use in practice.

It may be noted that our estimates of pFDR and FDR are slightly different from the Storey (2002) estimates: translating Storey's estimates to the present situation, Storey's estimate of FDR is the same as our estimate of pFDR, and Storey's estimate of pFDR is obtained by dividing our pFDR estimate by $1-(1-F_0(\alpha))^m$.

The pFDR estimate also works quite generally for dependent p-values, see Section 2. The other three estimates require that "local dependence" of p-values is negligible for small p-values so that there is no clustering of extreme values, cf. the discussion of the condition $\mathrm{D}'(\alpha_n)$ at the end of Section 3.

4 Motivation

In this section we give the theoretical motivation for the models (1.1) and (1.2) by showing that, very generally, they are asymptotically valid when m is large and one is far out into the tails – basically the models apply if tails of test statistics have a natural "asymptotic stability property", or, equivalently, if the distribution of test statistics is in the domain of attraction of an extreme value distribution. This motivation of course comes from general mathematical arguments (as does the motivation for the use of normal distribution), and not from, say, specifics of biology.

Practical motivation is given by the analysis of three examples in Section 5 below and, of course, more generally from very extensive experience in using extreme value statistics in many areas of science.

For simplicity, in this section we phrase the discussion in terms of large values of the test statistic being "significant" i.e. leading to small p-values. Let T_h, T_0, and T_1 be random variables which have the distribution of the test statistic under the hypothetical (=theoretical) null distribution, the true null distribution, and the distribution under the alternative hypothesis, respectively, and let G_h, G_0, and G_1 be the corresponding distribution functions. Further, throughout let $\bar{G}=1-G$ denote the tail (or "survival") function associated with a distribution function (d.f.) G.

The simplest motivation is as follows. Suppose that there are constants $C_h>0$ and $\tilde{\gamma}_h>0$ such that

$$\bar{G}_h(x)\sim C_h\frac{1}{x^{1/\tilde{\gamma}_h}}, \quad \text{as } x\to\infty, \tag{4.1}$$

which e.g. holds for the Student one- and two- sample $t-$statistics, and for $F-$statistics. Further suppose that $\bar{G}_0$ and $\bar{G}_1$ satisfy corresponding expressions. For one- and two-sample $t-$statistics this again holds very generally indeed, with $\tilde{\gamma}_h = \tilde{\gamma}_0 = \tilde{\gamma}_1 = f$, where f is the degrees of freedom; see Zholud (2011a). The distribution F_0 of the p-values under the true null distribution is $G_0(G_h^{\leftarrow})$, where $G^{\leftarrow}$ denotes is the right continuous inverse of a d.f. G. Thus,

$$F_0(x) = G_0(G_h^{\leftarrow}(x)) \sim C_0 \frac{1}{\left((x/C_h)^{-\tilde{\gamma}_h}\right)^{1/\tilde{\gamma}_0}} = c_0 x^{1/\gamma_0}, \text{ as } x \to 0,$$

with $c_0 = C_0/C_h^{\tilde{\gamma}_h/\tilde{\gamma}_0}$ and $\gamma_0 = \tilde{\gamma}_h/\tilde{\gamma}_0$, so that (1.1) holds. Similarly it follows that (1.2) holds with $c_1 = C_1/C_h^{\tilde{\gamma}_h/\tilde{\gamma}_1}$ and $\gamma_1 = \tilde{\gamma}_h/\tilde{\gamma}_1$.

However, (1.1) and (1.2) hold much more generally. Let T be a random variable with d.f. G and suppose that G satisfies either one of the following two equivalent conditions: a) G belongs to the domain of attraction of an extreme value distribution, i.e. the distribution of linearly normalized maxima of i.i.d. variables with d.f. G converges, or b) the tail of G is asymptotically stable, i.e. the distribution of a scale normalized exceedance of a level u converges as u tends to the right hand endpoint of the distribution G. Then there are constants $\sigma = \sigma_u > 0$ and γ such that

$$\Pr\left(\frac{T-u}{\sigma} > x \mid T > u\right) \approx \left(1 + \frac{\gamma}{\sigma}x\right)_+^{-1/\gamma}, \tag{4.2}$$

for u close to the right endpoint of G. Here the $+$ signifies that the expression in parentheses should be replaced by zero if it is negative, and the right hand side is the tail function of a Generalized Pareto distribution. The parameter γ can be positive, zero, or negative. For $\gamma = 0$ the last term in (4.2) is interpreted as its limit as $\gamma \to 0$, i.e. it is $e^{-x/\sigma}$. Writing $v = \Pr(T > u)$ we get that

$$\bar{G}(x) \approx v\left(1 + \frac{\gamma}{\sigma}(x-u)\right)_+^{-1/\gamma}, \text{ for } x > u,$$

and

$$\bar{G}^{\leftarrow}(y) \approx u + \frac{\sigma}{\gamma}\left(\left(\frac{v}{y}\right)^{\gamma} - 1\right), \text{ for } y \le v.$$

Suppose now that G_h and G_0 satisfy (4.2). Then, repeating the calculations above (with the same u for both distributions) we get, with self-explanatory notation, that for $\gamma_h, \gamma_0 > 0$

$$F_0(x) \approx v_0 \left(1 - \frac{\gamma_0 \sigma_h}{\gamma_h \sigma_0} + \frac{\gamma_0 \sigma_h}{\gamma_h \sigma_0} \left(\frac{v_h}{x}\right)^{\gamma_h}\right)^{-1/\gamma_0}$$

$$\approx v_0 \left(\frac{\gamma_h \sigma_0}{\gamma_0 \sigma_h}\right)^{\gamma_0} \; v_h^{-\gamma_h/\gamma_0} x^{\gamma_h/\gamma_0},$$

for small x, so that (1.1) again holds.
If instead $\gamma_h = \gamma_0 = 0$ one obtains

$$F_0(x) \approx v_0 x^{\sigma_h/\sigma_0},$$

which again is of the form (1.1). Cases where one γ is > 0 and the other is ≤ 0, or where one γ is ≥ 0 and the other is < 0 are not interesting, since one of the tails then completely dominates the other. Calculations become more complex if both γ-s are negative, so that the corresponding generalized Pareto distributions have a finite upper endpoint. However such cases are not expected to occur in practice, either. The motivation for (1.2) is the same as for (1.1).

5 Examples

In this section the methods introduced here are illustrated by analyses of three different data sets. All analyses were made using the SmartTail tool, see also Zholud (2011c) and Zholud (2011b).

Example 1: *Yeast genome screening, Warringer et al. (2003) and Zholud et al. (2011).* The data sets in this example come from a long sequence of Genome Wide screening experiments for detecting differential growth under different conditions. The experiments use *Saccharomyces cerevisiae*, baker's yeast, a model organism for advancing understanding of genetics. The experiments were run on a Bioscreen Microbiology Reader (also known as Bioscreen C Analyzer). In an experiment different yeast strains are grown on two 100-well (10 × 10) honeycomb agar plates. The output is 200 growth curves, each representing a time series of optical density measurements from a well. Here we only consider one of the parameters extracted from these curves, the so-called *logarithmic doubling time.*

In a typical experiment a mutant yeast strain with one gene knocked out is grown in the same position in each of the two plates. A reference wild type strain without any gene knockout is grown in four wells in each plate, one well in each quadrant of the plate. Differential growth caused by a mutant is measured separately for each of the two plates by subtracting the average of the logarithmic doubling times of the four reference strains from the logarithmic doubling times for the mutant strain. This gives one value for each plate. Differential growth is then tested by comparing these two values with zero in a one-sample $t-$test with 1 degree of freedom.

We consider three data sets: A *Wild Type Data Set* with 1,728 observed p-values, a *Genome Wide Data Set* with 4,896 observed p-values, where the single knockout mutants and reference strains were grown under normal conditions, and a *Salt Stress Data Set* with 5,280 observed p-values, where all single knockout mutants and reference strains were grown in a salt stress environment. The Wild Type data set was obtained for quality control purposes, and hence was analyzed in exactly the same way as the Genome Wide scans. However for this data set one knows that there are no real effects, so it in fact is a sample from the true null distribution. These data sets are available from the PROPHECY database, see the list of references, and Fernandez-Ricaud et al. (2006).

As a theoretical background, from Zholud (2011a) follows that for one sample $t-$tests the models (1.1) and (1.2) are expected to hold, with $\gamma_0 = \gamma_1 = 1$. It then follows that also (2.4) holds, with $\gamma = 1$ and $c = \pi_0 c_0 + \pi_1 c_1$. However, non-asymptotically, other values of the γ-s may give a reasonable fit.

Figure 2.2 illustrates the results of the analysis of the wildtype data set (recall that this data set is a sample from the true null distribution). From the top right and middle panels it is clear that the true null distribution is different from the uniform distribution, and also that the model (1.5) fits quite well. A formal likelihood ratio test of the hypothesis $\gamma_0 = 1$ gave the p-value 0.67, and setting $\gamma_0 = 1$ doesn't change the estimates much (the estimated value of γ_0 is 0.96). The remaining three plots in Figure 2.2 illustrate different ways of checking the model assumption (1.5).

The top right panel shows the Kolmogorov-Smirnov 95% confidence limits for the exponential distribution of $-\log(p/u)|p \leq u$ transformed to the uniform scale, see Lilliefors (1969) and Schafer et al. (1972). The test doesn't reject the model (1.5) (p-value 0.49). The bottom right panel shows that the estimate of $\hat{F}_0$ is quite insensitive to the choice of the threshold u. The individual estimates of $1/\gamma_0$ (bottom left panel) and c_0 (not shown) change slightly more when u is changed, but are still quite stable. The two bottom plots are customarily used to guide the choice of u and to check the model fit.

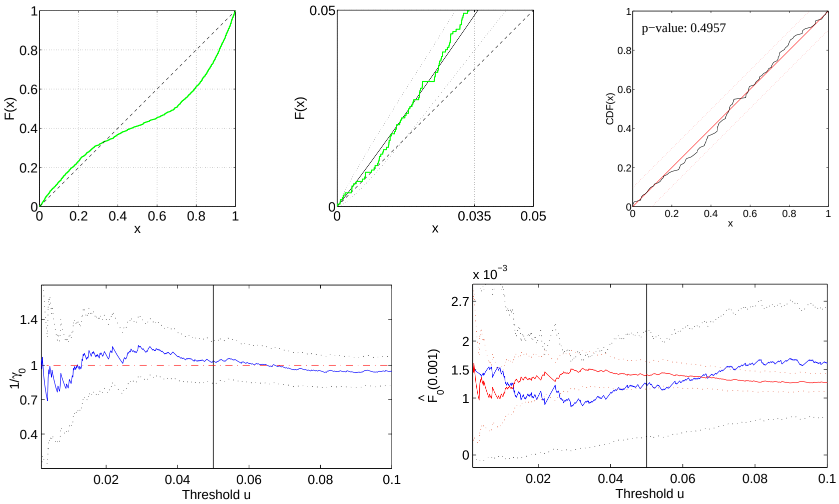


Figure 2.2: *The Wild Type Data Set*

Top left: Empirical distribution function. Dashed line is the uniform distribution. *Top middle:* Empirical distribution function for $p \leq 0.05$ (122 values). Solid line is (1.5) estimated using $u = 0.05$; dashed line is the uniform distribution. Dotted lines are 95% pointwise confidence intervals. *Top right:* Empirical conditional distribution function of $-\log(p/0.05)|p \leq 0.05$ transformed to the uniform scale and Kolmogorov-Smirnov 95% goodness of fit limits. *Bottom left/right:* Estimated $1/\gamma_0$ and $\hat{F}_0(0.001)$ as function of the threshold u. Red line is the same function but with γ_0 set to 1. Dotted lines are 95% pointwise confidence intervals.

The left panel of Figure 2.3 indicates that the model (2.4) fits the Genome Wide Data Set well (the Kolmogorov-Smirnov p-value was 0.38). The estimates of pFDR at $\alpha = 0.001$ are somewhat higher for small values of the threshold u. This behavior is reversed if γ is set to 1. The model checking plots corresponding to the four last panels in Figure 2.2 where slightly less stable than those for the wildtype data set.

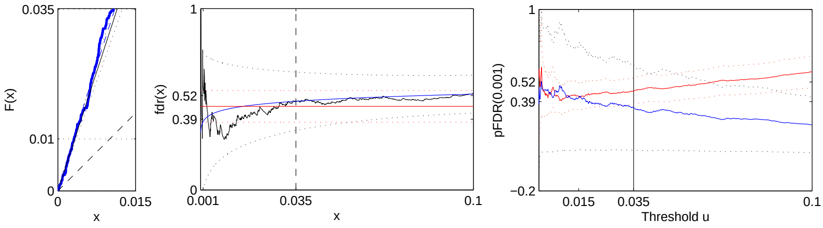


Figure 2.3: *The Genome Wide Data Set*

Left: Empirical distribution function for $p \leq 0.035$ (525 values); dashed line is the uniform distribution. Solid line is (2.4) estimated using $u = 0.035$; dot-dashed line is the same function for γ_0 set to 1. *Middle:* $\mathrm{fdr}(x)$ (smooth curve) and empirical FDR (edgy curve) for $u = 0.035, \pi_0 = 1$. Horizontal line is $\mathrm{fdr}(x)$ with γ_0 set to 1. *Right:* pFDR at $\alpha = 0.001$ as function of the threshold u, for $\pi_0 = 1$. Dot-dashed line is the same function with γ_0 set to 1. Dotted lines are the corresponding 95% pointwise confidence intervals.

The estimate of pFDR at $\alpha = 0.001$ is 0.34 (here $\gamma_0 = 0.96$ and $\gamma = 1.03$ are estimated from the data using $u = 0.05$ and $u = 0.035$ accordingly). Since there were 14 p-values less than 0.001 we hence estimate the expected number of false positives at this level to be 4.76. Using the binomial approximation to the number of false positives, we further estimate that with probability greater than 95% the number of false positives was at most 8.

If we instead had believed in a uniform distribution under the null hypothesis, we would have estimated the mean number of false positives to be 4 and that the number of false positives with probability greater than 95% was less than 7 - a somewhat too positive picture of experimental precision.

A further practically important question is "Which out of the 14 rejections are the true positives?". Sometimes one meets the idea that one should make an ordered list of the p-values corresponding to rejected null hypotheses and make further investigation starting with the smallest p-value, then go to the next smallest one, and so on, in the hope that the smaller the p-value, the more likely it is to correspond to true positives, see e.g. Noble (2009). For $t-$ and $F-$tests with low degrees of freedom, and far out in tails theory suggests that this hope often is unfounded, see Zholud (2011a). Nevertheless, for less extreme situation Efron's fdr(x) can be used to measure how likely it is that a rejection is a false positive. The fdr(x) plot in middle panel of Figure 2.3 decreases as x tends to zero, but the decrease is small. This indicates that it is slightly (but only slightly) more probable that it is the rejections with the smallest p-values which are the true positives. However, it is still quite likely that also some of the tests with the smallest p-values are false positives.

Theoretically, that fdr(x) is almost constant for small x of course is a consequence of the asymptotic tail behavior of the $t-$statistics discussed above. For the present data set this theory is also borne out by the empirical results.

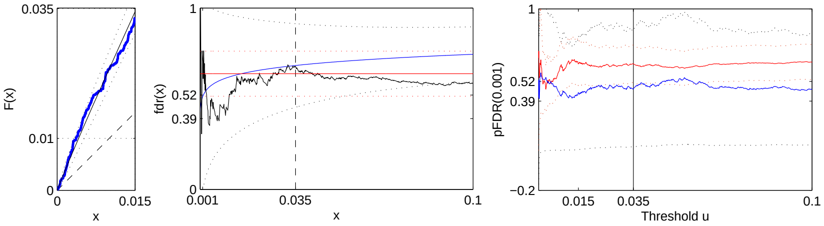


Figure 2.4: *The Salt Stress Data Set*
Left: Empirical distribution function for $p \leq 0.035$ (409 values); dashed line is the uniform distribution. Solid line is (2.4) estimated using $u = 0.035$; dot-dashed line is the same function for γ_0 set to 1. *Middle:* fdr(x) (smooth curve) and empirical FDR (edgy curve) for $u = 0.035, \pi_0 = 1$. Horizontal line is fdr(x) with γ_0 set to 1. *Right:* pFDR at $\alpha = 0.001$ as function of the threshold u, for $\pi_0 = 1$. Red line is the same function with γ_0 set to 1. Dotted lines are the corresponding 95% pointwise confidence intervals.

The Salt Stress Data Set by and large behaved in the same way as the Genome Wide Data Set, see Figure 2.4. A difference was that model checking plots were more stable, and in fact model fit seemed even better than for the wild type data.

To illustrate the gain in efficiency from using the estimates from Section 2 instead of the empirical distribution function estimator we set $x = 0.001$ and SmartTail estimated that the ratio $\mathrm{Var}(\hat{F}_0(x))/\mathrm{Var}(\hat{F}_E(x)) = 0.46$ for the Wild Type Data Set where $u = 0.05$, and for $u = 0.035$ that $\mathrm{Var}(\hat{F}(x))/\mathrm{Var}(\hat{F}_E(x))$ was 0.7 for the Genome Wide Data Set and 0.69 for the Salt Stress Data Set. Variance estimates for the version of pFDR which uses the empirical distribution functions don't seem to be available, and hence we haven't compared the variance of the pFDR estimates from Section 2 with the variance of the empirically estimated pFDR.

At this point it should perhaps be recalled that the estimates of fdr and pFDR are somewhat biased upwards as we have set $\pi_0 = 1$. However, for the situations we are interested in, the true π_0 should be close to 1, and this bias accordingly is insignificant. The same comment applies also to all the plots which follow.

Example 2: *Association mapping in Arabidopsis, Zhao et al. (2007).* This data set comes from 95 *Arabidopsis Thaliana* samples, with measurements of flowering-related phenotypes together with genotypes in the form of over 900 short sequenced fragments, distributed throughout the genome. The goal was association mapping, i.e. identification of regions of the genome where individuals who are phenotypically similar are also unusually closely genetically related. A problem is that spurious correlations may arise if the population is structured so that members of a subgroup, say samples obtained from a specific geographical area, tend to be closely related. One of the main thrusts of the paper was to evaluate 9 different statistical methods to remove such spurious correlations. But of course an ultimate aim is to identify interesting genes.

Here we only consider the SNP (Single Nucleotide Polymorphism) data, and one phenotype, the one called JIC4W, which we choose since it was of special interest in the paper. Further, we only display results for two of the statistical methods, the KW method which just consisted in making Kruskal-Wallis tests without correction for population structure, and a method called Q+K which may have been the most successful of the 9 methods studied. The number of tests was 3745.

Figures 2.5 and 2.6 show that the model (2.4) fits both the Kruskal-Wallis and the Q+K p-values well for the values $u = 0.001$ and $u = 0.01$ of the threshold accordingly. The p-values for the Kolmogorov-Smirnov test were 0.43 and 0.38, respectively. The estimate of pFDR at $\alpha = 0.0001$ for the Kruskal-Wallis test was 0.013, and for the Q+K the estimate was s0.1, i.e. more than 7 times bigger. Both these numbers assume that the true null distribution is the uniform distribution. Zhao et al. (2007) argue that most of the Kruskal-Wallis p-values are spurious. We also performed the same analysis for the other test methods proposed in their paper. For most, but not all, of them, the model (2.4) gave a good fit. Of course the quality of the fit also depended on the choice of u.

Led by some speculation in the paper, we tried to use chromosomes 2 and 3 as a surrogate null distribution sample. However, the tail distribution of p-values in those chromosomes were in fact, if anything, heavier than for those in chromosomes 1, 4, 5, although the differences were well within the range of random statistical variation. Thus if there indeed were no effects present in chromosomes 2 and 3, then also most of the positives in the other genes might be false, as also is discussed in Zhao et al. (2007).

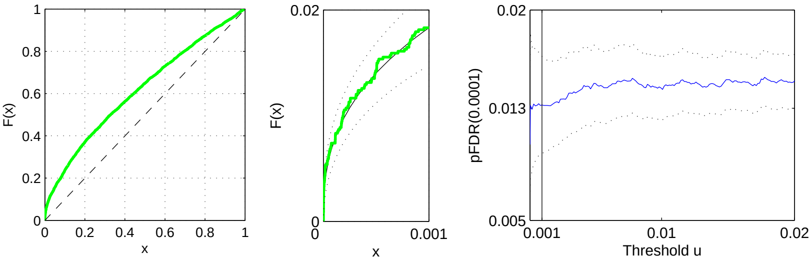


Figure 2.5: *The KW analysis of the JIC4W data set*
Left: Empirical distribution function. Dashed line is the uniform distribution. *Middle:* Empirical distribution function for $p \leq 0.001$ (99 values). Solid line is (2.4) estimated using $u = 0.001$. Dotted lines are 95% pointwise confidence intervals. OBS scale: x-axis is stretched 10 times. *Right:* pFDR at $\alpha = 0.0001$ as function of the threshold u, for $\pi_0 = 1$. Dotted lines are 95% pointwise confidence intervals.

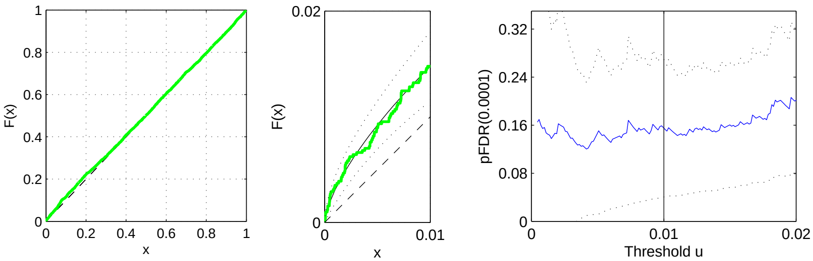


Figure 2.6: *The K+Q analysis of the JIC4W data set*
Left: Empirical distribution function. Dashed line is the uniform distribution. *Middle:* Empirical distribution function for $p \leq 0.01$ (76 values). Solid line is (2.4) estimated using $u = 0.01$. Dotted lines are 95% pointwise confidence intervals. *Right:* p-FDR at $\alpha = 0.0001$ as function of the threshold u, for $\pi_0 = 1$. Dotted lines are 95% pointwise confidence intervals.

Again to illustrate the gain in efficiency from using the estimates from Section 2, we set $x = 0.0001$ and SmartTail for $u = 0.001$ estimated that $\mathrm{Var}(\hat{F}(x))/\mathrm{Var}(\hat{F}_E(x))$ was 0.53 for the KW method and 0.85 for the Q+K method and $u = 0.01$. Note that the values of γ for these two sets of p-values were 2.6 and 1.5 accordingly.

Example 3: *fMRI brain scans, Taylor and Worsley (2006).* The Functional Image Analysis Contest (FIAC) data set contains results from an fMRI experiment aimed at exploring functional organization of the language network in the human brain. The part we use here is "the Block Experiment", see Dehaene-Lambertz et al. (2006). In this experiment 16 subjects were instructed to lie still in a scanner with eyes closed and to attentively listen to blocks of 6 sentences, either different ones or the same sentence, and either read by the same speaker or by different speakers. Each subject was asked to participate in two "runs", with 16 blocks presented in each run. In Taylor and Worsley (2006), for each run and each voxel in the brain scans, the data was used to study the significance of two contrasts, "different minus same sentence" and "different minus same speaker" and the interaction between these two. Roughly 35,000 voxels per subject were used. For each voxel in each subject and each run quite sophisticated preprocessing was used to construct the corresponding 3 $t-$test quantities. One subject dropped out of the experiment, and one only completed one run, so the end results was $(15 \times 2 + 1) \times 3 = 93$ sets of roughly $35,000$ $t-$test quantities.

To study the fit of Equation (2.4) we transformed these $t-$values to p-values using a $t-$distribution with 40 degrees of freedom (this was the approximate degrees of freedom according to Taylor and Worsley (2006) - it can in fact be seen that to check model fit the precise number is not important). For each of the 93 resulting data sets of about 35,000 p-values we performed a Kolmogorov-Smirnov goodness-of-fit test of the fit of the model (2.4) for the p-values which were smaller than the threshold $u = 0.01$. In these 93 data sets the smallest number of p-values less than 0.01 was 117, and the largest number was 973. Figure 2.7 shows that the distribution of the 93 goodness-of-fit p-values are somewhat skewed towards smaller values, as compared with the uniform distribution. However, this deviation from uniformity is small, and the overall impression is that Equation (2.4) fits the Block Experiment FIAC data well.

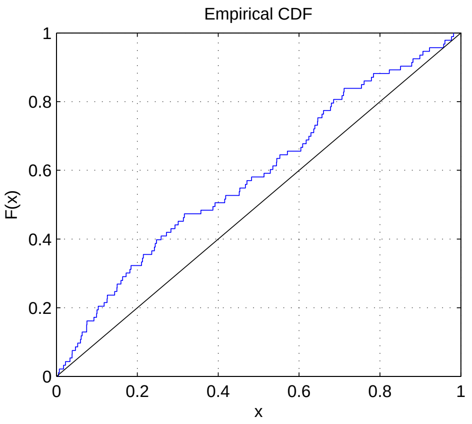


Figure 2.7: Empirical d.f. of the Kolmogorov-Smirnov goodness-of-fit p-values from the 93 sets of p-values in the fMRI brain scan data set.

In fact, even for the two data sets where (2.4) was clearly rejected (the Kolmogorov-Smirnov p-values were 0.005 and 0.007), the Kolmogorov-Smirnov plots showed that the deviations from the model were still quite moderate, and, as expected, even smaller for thresholds u lower than 0.01 (Kolmogorov-Smirnov p-values 0.32 and 0.29 accordingly for $u = 0.005$).

This FIAC experiment opens possibilities for substantial further analysis. For example, Taylor and Worsley (2006) suggest randomly changing signs in contrasts to get surrogate observations from the true null distribution. A null sample might also be obtained from areas in the brain which are not involved in language processing. However, we stop at this point.

6 Discussion and conclusions

This paper is about high-throughput screening – experiments where very many individual statistical tests are performed, but where one expects that it is only possible to detect a real effect in a small or moderate number of the tests, so that testing is done at quite extreme significance levels. High-throughput testing typically involves considerable preprocessing of data before the tests are made. This, and the complexity of the experiments often cause the true null distribution to be different from the theoretical null distribution. We believe that if one suspects this is the case it may be well worth the effort to try to obtain a sample from the true null distribution, both to get a better grip on risks for false positives and for general quality control purposes. Examples of how this can be done are mentioned above.

This paper gives answers to the two questions from the introduction: "How many of the positive test results are false?" and "How should one judge if one preprocessing method makes the true null distribution closer to the theoretical one than another method?". The questions concern tails of distributions, with the central part of the distributions being largely irrelevant. We accordingly use Extreme Value Statistics in the answers. Our answer to the first question is that the conditional distribution of the number of false positives is approximately binomial, and efficient and accurate methods to estimate the success probability parameter of this binomial distribution. The answer rests on assuming a simple polynomial model for the lower tail of the distribution of p-values (cf. (1.5) and (1.6)). In Section 4 this assumption is shown to be quite generally asymptotically valid. However, of course, whether these asymptotics are relevant in a concrete testing problem has to be checked from data. We also provide methods for such model checking (see the analyzes in Section 5, in particular Figure 2.2).

Our answer to the second question is to compare the estimates of the true null distribution with the theoretical uniform distribution. This can be done informally from plots, or by a formal test of the hypothesis that the parameters in the null distribution (1.1) satisfy $c_0 = \gamma_0 = 1$. Again it is useful to complement this analysis with model checking.

A third basic question is "Which of the rejections are caused by real effects?". The answer one might hope for is that the smallest of the p-values which lead to rejections are those which correspond to real effects. Our $\mathrm{fdr}(x)$ plots can be used to judge if this in fact is the case. However, both from asymptotic theory and from our experience with data analysis, the answer might be disappointing: often the real effects are fairly randomly spread out amongst the rejections.

The p-values obtained from high-throughput screening sometimes are dependent. However, not unusually this dependence affects the extreme tails less than the centers of distributions - whether this is the case or not depends on the amount of clustering of small p-values. This is discussed in Sections 2 and 3. A comforting message is that even in cases where dependence persists into the extreme tails, the estimates of basic quantities, such as pFDR, still under very wide conditions are consistent and asymptotically normal. There exists a very extensive literature about dependent extremes for the case when observations are ordered "in time". However less is proven for the much more complicate "spatial" dependence patterns which may occur in high-throughput testing, and more research is needed.

We have applied the methods developed in this paper to data from two genomics experiments, a Bioscreen yeast experiment, and an *Arabidopsis* study, and to a fMRI brain scan experiment. For all three data sets our analysis methods seem to fit the data well, and to provide useful information. In particular, they proved that for the yeast data the real null distribution was different from the uniform distribution, and quantified the rather low specificity of the tests. For the *Arabidopsis* data the methods put numbers on the differences between alternative statistical processing methods and indicated that even for the best test method, specificity may not have been all that good.

Finally, the aim of this paper is not just technical development. It is also to deliver a message: *If you are concerned with false positives in high-throughput testing, then it is the tails (and not the centers) of distributions which matter!* And, Extreme Value Statistics is *the* instrument for looking at tails. Further, already in the near future, screening experiments will become even much larger, and testing will be done at even more extreme significance levels - so the issues raised in this paper will become even more important than they now are.

Acknowledgements We are very grateful to Simon Tavaré for constructive and helpful reading of a version of this paper, and for pointing us to the Microarray data set. We thank Olle Nerman for inspiring comments, and for starting us on this research project. We also thank Anders Blomberg for providing access to the Bioscreen data and Jonathan Taylor for the brain imagining data. Research supported in part by the Swedish Foundation for Strategic Research.

References

D.B. Allison, G.L. Gadbury, M. Heo, J.R. Fernández, C.-K. Lee, T.A. Prolla, and R. Weindruch. A mixture model approach for the analysis of microarray gene expression data. *Comput. Stat. Data Anal.*, 39(1):1–20, 2002. 35, 44

J. Beirlant, Y. Goegebeur, J. Segers, and J. Teugels. *Statistics of Extremes, Theory and Applications.* Wiley, Chichester, 2004. 40, 47

Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: A practical and powerful aproach to multiple testing. *J. R. Statist. Soc. B*, 57(1):289–300, 1995. 34, 48

S. Clarke and P. Hall. Robustness of multiple testing procedures against dependence. *Ann. Statist.*, 37(1):332–358, 2009. 38

S.G. Coles. *An Introduction to Statistical Modeling of Extreme Values.* Springer, London, 2001. 40

L.M. Cope, R.A. Irizarry, H.A. Jaffee, Z. Wu, and T.P. Speed. A benchmark for Affymetrix GeneChip expression measures. *Bioinformatics*, 20(3):323–331, 2004. 32

G. Dehaene-Lambertz, S. Dehaene, J.-L. Anton, A. Campagne, A. Jobert, D. LeBihan, M. Sigman, C. Pallier, and J.-B. Poline. Functional segregation of cortical language areas by sentence repetition. *Human Brain Mapping*, 27(5):360–371, 2006. 32, 58

S. Dudoit and M.J. van der Laan. *Mulitple Testing Procedures with Applications in Genomics.* Wiley, New York, 2008. 34, 48

B. Efron. Large-scale simultaneous hypothesis testing: the choice of a null hypothesis. *J. Amer. Statist. Ass.*, 99(465):96–104, 2004. 32, 34, 43

B. Efron. Microarrays, empirical Bayes and the two-groups model. *Statist. Sci.*, 23(1):1–22, 2008. 32, 34, 43

B. Efron. Correlated z-values and the accuracy of large-scale statistical estimates. *J. Amer. Statist. Assoc.*, 105(491):1042–1055, 2010. 43

B. Efron, R. Tibshirani, J.D. Storey, and V. Tusher. Empirical Bayes analysis of a microarray experiment. *Journal of the American Statistical Association*, 96(456):1151–1160, 2001. 32, 34, 37, 38, 48

J. Fan, P. Hall, and Q. Yao. To how many simultaneous hypothesis tests can normal, student's t or bootstrap calibration be applied? *J. Amer. Statist. Assoc.*, 102(480):1282–1288, 2007. 36

L. Fernandez-Ricaud, J. Warringer, E. Ericson, K. Glaab, P. Davidsson, F. Nilsson, G.J. Kemp, O. Nerman, and A. Blomberg. PROPHECY-a yeast phenome database, update 2006. *Nucleic Acids Res*, 35:D463–D467, 2006. 52

A. Gordon, G. Glazko, X. Qui, and A. Yakovlev. Control of the mean number of false discoveries, Bonferroni and stability of multiple testing. *Ann. Appl. Statist.*, 1(1):179–190, 2007. 36

H. Hotelling. The behavior of some standard statistical tests under non-standard conditions. *Proceedings of the Fourth Berkeley*

Symposium on Mathematical Statistics and Probability, 1:319–360, 1961. 36

T. Hsing. On tail index estimation using dependent data. *Ann. Statist.*, 19(3):1547–1569, 1991. 44

J. Jin and T.T. Cai. Estimating the null and the proportion of nonnull effects in large-scale multiple comparisons. *J. Amer. Statist. Assoc.*, 102(478):495–506, 2007. 32, 34

K.F. Kerr. Comments on the analysis of unbalanced microarray data. *Bioinformatics*, 25(16):2035–2041, 2009. 34, 35

T.A. Knijnenburg, L.F.A. Wessels, J.T.M. Reinders, and I. Shmulevich. Fewer permutations, more accurate P-values. *Bioinformatics*, 25(12):161–168, 2009. 34

M.R. Leadbetter. On extreme values in stationary sequences. *Probability Theory and Related Fields*, 28(4):289–303, 1974. 47

M.R. Leadbetter and H. Rootzén. On extreme values in stationary random fields. In I. Karatzas, B.S. Rajput, and M.S. Taqqu, editors, *Stochastic Processes and Related Topics, in Memory of Stamatis Cambanis, 1943–1995*, pages 275–285. Birkhäuser, Boston, 1998. 47

H.W. Lilliefors. On the Kolmogorov-Smirnov test for the exponential distribution with mean unknown. *J. Amer. Statist. Assoc.*, 63(325):387–389, 1969. 52

W.S. Noble. How does multiple testing correction work? *Nature Biotechnology*, 27(12):1135–1137, 2009. 34, 55

PROPHECY, 2011. *url:* prophecy.lundberg.gu.se - quantitative information about phenotypes for the complete collection of deletion strains in yeast (*Saccharomyces cerevisiae*). 52

H. Rootzén, L. de Haan, and M.R. Leadbetter. Tail and quantile estimation for strongly mixing stationary sequences. Technical report, dept of Statistics,University of North Carolina, 1991. 44, 45

D. Ruppert, D. Nettleton, and J.T.G. Hwang. Exploring the Information in p-Values for the Analysis and Planning of Multiple-Test Experiments. *Biometrics*, 63(2):483–495, 2007. 34

R.E. Schafer, J.M. Finkelstein, and J. Collins. On a goodness-of-fit test for the exponential distribution with mean unknown. *Biometrika*, 59(1):222–224, 1972. 52

A. Schwartzman. Empirical null and false discovery rate inference for exponential families. *Ann. Appl. Statist.*, 2:1332–1359, 2008. 32, 44

SmartTail, 2011. *url:* www.smarttail.se - software for the analysis of false discovery rates in high-throughput screening experiments. 41, 51, 58

J.D. Storey. A direct approach to false discovery rates. *J. R. Statist. Soc. B*, 64(3):479–498, 2002. 34, 37, 38, 39, 42, 48, 49

J.D. Storey. The positive false discovery rate: a bayesian interpretation and the q-value. *The annals of Statistics*, 31(6):2013–2035, 2003. 34

J.D. Storey. Srong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. *J. R. Statist. Soc. B*, 66(1):187–205, 2004. 34, 42

Y. Tang, S. Ghosal, and A. Roy. Nonparametric Bayesian Estimation of Positive False Discovery Rates. *Biometrics*, 63(4): 1126–1134, 2007. 44

J.E. Taylor and K.J. Worsley. Inference for magnitudes and delays of response in the FIAC data using BRAINSTAT/FMRISTAT. *Human Brain Mapping*, 27:434–441, 2006. 32, 38, 58, 59, 60

J. Warringer, E. Ericson, L. Fernandez, O. Nerman, and A. Blomberg. High-resolution yeast phenomics resolves different physiological features in the saline response. *Proc Natl Acad Sci USA*, 100(26):15724–15729, 2003. 32, 51

K. Zhao, M.J. Aranzana, S. Kim, C. Lister, C. Shindo, C. Tang, C. Toomajian, H. Zheng, C. Dean, P. Marjoram, and M. Nordborg. An *Arabidopsis* example of association mapping in structured samples. *PLoS Genet*, 3(1):71–82, 2007. 32, 56, 57

D.S. Zholud. Tail approximations for the Student $t-$, $F-$, and Welch statistics for non-normal and not necessarily i.i.d. random variables. *Submitted*, 2011a. 34, 36, 42, 50, 52, 55

D.S. Zholud. Confidence intervals for the False Discovery Rate - the SmartTail method. *Work in progress*, 2011b. 39, 41, 43, 51

D.S. Zholud. SmartTail - software for the analysis of False Discovery Rates in High-Throuput Screening experiments. *Work in progress*, 2011c. 39, 51

D.S. Zholud, H. Rootzèn, O. Nerman, and A. Blomberg. Positional effects in biological array experiments and their impact on False Discovery Rate. *Work in progress*, 2011. 32, 51

PAPER II

Tail approximations for the Student $t-$, $F-$, and Welch statistics for non-normal and not necessarily i.i.d. random variables

Dmitrii Zholud [*]

Abstract

Let T be the Student one- or two-sample $t-$, $F-$, or Welch statistic. Now release the underlying assumptions of normality, independence and identical distribution and consider a more general case where one only assumes that the vector of data has a continuous joint density. We determine asymptotic expressions and rates of convergence for $\mathbf{P}(T > u)$ as $u \to \infty$ for this case. The approximations are particularly useful for small sample sizes and are aimed at analysis of High-Throughput Screening experiments, where the number of replicates can be as low as two to five and where often extremely high significance levels are used. A particular conclusion is that the Welch-Satterthwaite approximation to the distribution of the Welch statistic is inaccurate in such situations.

Keywords Student's one- and two-sample $t-$test, $F-$test, Welch statistic, small sample size, high-throughput screening, test power, non-normal population distribution, systematic effects, outliers, non-homogeneous data, dependent random variables.

AMS 2010 Subject Classifications: $\frac{\text{Primary-62G32;}}{\text{Secondary-62P10, 60F10;}}$.

[*] *Department of Mathematical Statistics*
Chalmers University of Technology and University of Göteborg, Sweden.
E-mail: dmitrii@zholud.com

1 Introduction

This article extends and fills out early results of Bradley and Hotelling on the tails of the distributions of, probably the most popular and frequently used statistical tests under arbitrary distributional assumptions. We present asymptotic results which quantify the effect of non-normality, dependence, and non-homogeneity of data on distribution tails of the Student one- and two-sample $t-$, $F-$ and Welch statistics. The approximations are valid for samples of any size, but may be most useful for very small sample sizes when standard central limit based approximations are inaccurate. This problem has gained significant new importance through the explosive increase of high-throughput testing, where sample sizes are often as small as two to four, but when instead thousands - or millions - of tests are performed, at extremely high significance level. Below we briefly illustrate this point by describing a biological experiment which in fact was the motivation for the present paper.

Let $\mathbf{X} \in \mathbb{R}^n,\ n \geq 2$, be a random vector and $T_n = T_n(\mathbf{X})$ be a Student's one- or two-sample $t-$test or an $F-$test (an $F-$test for comparison of variances, an $F-$test used in one-way ANOVA analysis, an $F-$test in full/fractional factorial experimental designs, a lack-of-fit sum of squares test, or an $F-$test for comparison of two nested linear models in regression analysis). Now let $MVN(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denote the multivariate normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, and define

$$t(u) = \mathbf{P}\left(T_n > u | H_0\right),$$

the distribution tail of T_n under the "standard" null hypothesis $H_0 : \mathbf{X} \sim MVN(\mathbf{0}, \sigma^2 \mathbf{I}_n)$, where $\sigma^2 > 0$ and $\mathbf{I}_n$ is the identity matrix. Here we study the *asymptotic* behavior of the tail distribution of T_n under the *alternative hypothesis* H_1 for *small and fixed* sample sizes, that is,

$$\mathbf{P}\left(T_n > u | H_1\right), \quad \text{as} \quad u \to \infty \quad \text{for } n \text{ fixed.}$$

Further, let $g_0(\mathbf{x})$ and $g_1(\mathbf{x})$ be the joint densities of the data vector $\mathbf{X}$ under some null and alternative hypotheses H_0 and H_1 respectively, and let $\mathcal{G}$ be a set of continuous densities that satisfy the regularity constraint of Theorem 2.1, 3.1 or 5.1 below for the three tests. Our main result is as follows.

Theorem 1.1. *There exists a functional $K : \mathcal{G} \to R^+$ such that $\forall g_0, g_1 \in \mathcal{G}$ and fixed n the limit expression*

$$\frac{\boldsymbol{P}(T_n > u|H_1)}{\boldsymbol{P}(T_n > u|H_0)} = \frac{K_{g_1}}{K_{g_0}} + o(1) \quad as \quad u \to \infty \tag{1.1}$$

holds with constants $0 < K_{g_0} = K(g_0) < \infty$ and $0 < K_{g_1} = K(g_1) < \infty$. Further, $MVN(\mathbf{0}, \mathbf{I}_n) \in \mathcal{G}$ and $K(MVN(\mathbf{0}, \mathbf{I}_n)) = 1$. The exact forms of K_g for arbitrary densities g are given in (2.3), (3.3) and (5.3) below.

Note that if T_n is a Z-test and $g \sim MVN(\mu \times \mathbf{1}, \mathbf{I}_n)$, then the right-hand side of (1.1) is either 0 , 1 or ∞ for $\mu < 0$, $\mu = 0$ and $\mu > 0$, respectively. This shows that small deviations from $g_0(x)$ do not necessarily induce small changes in the tails of the distribution of the test statistic.

The proofs of (1.1) for the different test statistics all follow a common path: through suitably chosen changes of variables the problem is reduced to using a lemma, down in Appendix A, which describes the behavior of integrals over small balls around zero in $\mathbb{R}^k$, $k \geq 1$.

Figure 3.1 below was the original motivation for writing this article. It comes from a paper Zholud et al. (2011) which studies systematic errors in a particular kind of biological experiments, so called Bioscreen array experiments, see Warringer and Blomberg (2003) and Warringer et al. (2003), and their impact on false positive and false discovery rates. Omitting details, the parameter of interest, called LSC, was assumed to be normally distributed with mean $\mu = 0$ if the null hypothesis H_0 were true.

However, a histogram of the LSC values in the wildtype dataset for which H_0 is known to be true, see Figure 3.1, left panel, showed clear deviations from normality. We therefore plotted the empirical cumulative distribution function (CDF) for the 1728 p-values $p_i = t(T_2(\mathbf{X}_i))$ computed in the wildtype experiment. Each one-sample $t-$test $T_2(\mathbf{X}_i)$, $i = 1, 2, .., 1728$, is a function of a pair $\mathbf{X}_i = (LSC_{i,1}, LSC_{i,2})$ of LSC replicates, and $t(\cdot)$ is the tail of Student's $t-$distribution with one degree of freedom.

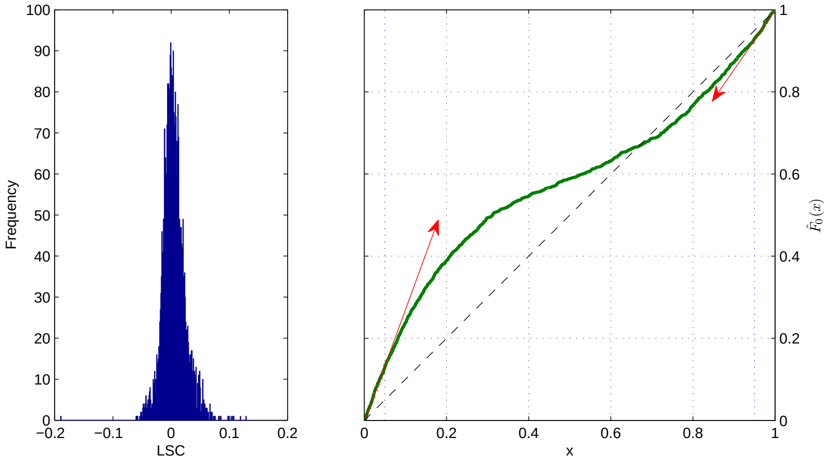


Figure 3.1: *The Wild Type Data Set*
A histogram of 3456 LSC values from the wildtype dataset (*left*) and empirical CDF based on the corresponding 1728 p-values (*right*).

The empirical CDF plot in Figure 3.1 showed clear deviation from the theoretical uniform distribution. Subsequent analysis revealed spatial systematic effects, see Zholud et al. (2011), which, in part, explained such deviations. The nature of these systematic effects are not yet fully understood. However, it was possible to use the p-values obtained in wildtype experiment to improve estimates of false discovery rates in other Bioscreen studies. The key observation, from the right panel of Figure 3.1, was that both lower and upper tails of the plot approached straight lines, as indicated by the two arrows.

The connection between the straight tails observed in Bioscreen HTS data and (1.1) is as follows. Let $g_1(\mathbf{x})$ be the joint density of the data vector $\mathbf{X}$ and let $F_1(x)$ be the CDF of the p-value $p = t(T_2(\mathbf{X}))$ that corresponds to the right-sided $t-$test. Assuming (1.1) holds with $K_{g_0} = 1$ it follows that $F_1(0) = 0$ and the right-sided derivative of F_1 at zero is equal to

$$\lim_{x\to 0+} \frac{F_1(x) - F_1(0)}{x} = \lim_{u\to\infty} \frac{F_1(t(u))}{t(u)} = \lim_{u\to\infty} \frac{\mathbf{P}\left(t(T_2(\mathbf{X})) \leq t(u)\right)}{t(u)}$$

$$= \lim_{u\to\infty} \frac{\mathbf{P}\left(T_2(\mathbf{X}) \geq u\right)}{t(u)} = \lim_{u\to\infty} \frac{\mathbf{P}\left(T_2(\mathbf{X}) \geq u\right)}{\mathbf{P}\left(T_2(\mathbf{X}) \geq u | H_0\right)} = K_{g_1}.$$

The constant K_{g_1}, the slope of the line tangent to the graph of $F_1(x)$ at zero, can be estimated from data using technique similar to peak-over-threshold method in Extreme Value Theory, see Rootzen and Zholud (2011). If, on the other hand, the density g_1 were known, then the exact value of K_{g_1} could be computed according to formula (2.3) of Theorem 2.1.

There is an extensive literature on the behavior of the Student $t-$ and $F-$ statistic under deviation from normality. Below we focus mostly on the Student one-sample $t-$statistic, with references to the Student two-sample $t-$, $F-$, and Welch statistics given as needed.

A brief introduction to the Student one-sample $t-$test can be found in Zabell (2008), and Cressie (1980) is a review with emphasis on understanding of the behavior of the test statistic under non-normality.

A main theme in the literature is the *Normal Approximation*, which is commonly stated as follows: "if the sample size is large enough and the population distribution is in the domain of attraction of the normal law, then the Student one-sample $t-$statistic is approximately $N(0, 1)$ distributed", see e.g. Giné et al. (1997). The non-central $t-$statistic is discussed in Bentkus et al. (2007). However, the Normal Approximation is inaccurate for small sample sizes, which are the center of interest in this paper.

Additional accuracy in the Normal approximation can be obtained by using the first few terms of a Gram-Charlier series, Geary (1936), Bartlett (1935), or *Edgeworth expansion*, see e.g. Field and Ronchetti (1990), Hall (1987) and Gaen (1949, 1950). The Edgeworth expansion improves the Normal approximation and performs better for smaller sample sizes, but still is inaccurate in the extreme tail area.

A different approach is to use *Saddlepoint approximations* to the distribution of the test statistics, see e.g. Zhou and Jing (2006), Jing et al. (2004) and Daniels and Young (1991). The simulation study of Jing et al. (2004) showed that the Saddlepoint approximation for Student's $t-$statistic is very accurate in the tail area, however it it is yet not shown whether the relative error of such approximation tends to zero as one goes far out in the tail. The formulas for the Saddlepoint approximation do not give explicit formulas for the tail behavior of the test statistic and are computationally complex. In particular, Saddlepoint approximations require knowledge of the population density, which limits their use in statistical inference. Details on Saddlepoint approximations can be found in Kolassa (2006), Jensen (1995), Reid (1988) and Lugannani and Rice (1980).

The Student one-sample $t-$test is closely related to the so-called self-normalized sum, see e.g. Shao (2004), Shao (1997) and Logan et al. (1973). The exact distribution of $t-$statistic for some special cases is discussed in Eden and Yates (1933), Rider (1929), Perlo (1933) and Laderman (1939).

Fundamental results on the Student two-sample $t-$test and $F-$test were formulated by Fisher (1924, 1925, 1935a,b) and the behavior of these tests under deviation from standard assumptions was thoroughly studied. One particular case is known as the two-means Behrens-Fisher problem, see e.g. Sawilowsky (2002) and Kim and Cohen (1998) and the list of references in these papers, and addresses the question of using the Student two-sample $t-$test when the variances of the two populations are unequal. A common approach to the Behrens-Fisher problem is to use the Welch-Satterthwaite approximation, see Aspin and Welch (1949), Welch (1937, 1947), and Satterthwaite (1941, 1946). Exact distribution for the Welch $t-$test for odd sample sizes is given in Ray and Pitman (1961). The formulas in the latter paper can be easily modified to hold for the Student two-sample t-test as well.

The effect of non-normality on the $F-$test was considered by e.g. Box (1953, 1954), Gaen (1950) and David and Johnson (1951).

Finally, the papers which are closest to the results of our research are Bradley (1952a,b) which describe the tail behavior of the Student one- and two- sample $t-$test and $F-$test under various deviations from the standard assumptions.

Bradley covers the Student one-sample $t-$test for i.i.d. non-normal observations, and also makes a somewhat less complete study of the corresponding cases for the Student two- sample $t-$test and the $F-$test of equality of variances. Bradley (1952b) derives the constant K_g from geometrical considerations, but does not state any assumptions on the underlying population density which ensure that the results hold. Bradley (1952a), on the other hand, gives assumption on the population density for the formulas to hold. However, he overlooked that applying the Leibnitz's rule for differentiation under the integral sign requires that the integration is taken over a set of finite measure. Consequently, the assumptions may not be sufficient.

Hotelling (1961) studies the Student one-sample $t-$test for an "arbitrary" joint density of the data vector. Hotelling derives the constant K_g assuming that the limit in the left-hand side of (1.1) exists and that the function

$$D_n(\xi) = \int_0^\infty r^{n-1} g(r\xi_1, \cdots, r\xi_n) dr$$

is continuous for the two population densities g_0 and g_1. When it comes to the examples, however, the existence of the limit in (1.1) is taken for granted and the assumption of continuity of $D_n(\xi)$ is never verified.

Our results cover the general case where neither independence nor identical distribution is assumed, and hold for both the Student one- and two- sample $t-$tests and $F-$test of the equality of variances. We also derive the asymptotic formula for the Welch $t-$test and offer a somewhat different way of proof which may be applied to a wider class of statistical tests: in particular, $F-$tests in their most general meaning, and, perhaps, other statistical tests that have χ^2-distributed denominator under the null hypothesis, see Appendix A.

We also provide general (and correct) conditions which ensure that the asymptotic formulas are valid, and simpler versions of those which can be easily checked in many situations.

The structure of this paper is as follows. Sections 2 - 5 contain the main theorems and examples, and Section 6 is a brief note on the speed of convergence and on second and higher order terms in the asymptotic expression (1.1). Section 7 complements the analytical results with a simulation study. The key lemma used in the proofs is given in Appendix A, and a discussion of the regularity conditions can be found in Appendix B. Appendix C contains tables related to Sections 2 and 3 and figures from the simulation section.

2 One-sample $t-$test

The aim of the current section is to establish asymptotic expression for the probability of high level excursions of the Student one-sample $t-$statistic. The setting is that the population distribution may be non-normal, and standard assumptions of independence and homogeneity are relaxed as well.

Let $\mathbf{X} = (X_1, X_2, .., X_n)$, $n \geq 2$, be a random vector with joint density g and define

$$T_n = \sqrt{n}\frac{\overline{\mathbf{X}}}{\sqrt{S^2}},$$

where $\overline{\mathbf{X}}$ and S^2 denote the sample mean and sample variance of the vector $\mathbf{X}$. Introduce the unit vector $\mathbf{I}_d = (1/\sqrt{n}, 1/\sqrt{n}, .., 1/\sqrt{n})$, and assume that

$$g(x\mathbf{I}_d) > 0 \text{ for some } x \geq 0 \tag{2.1}$$

and that

$$\int_0^\infty r^{n-1} \sup_{\substack{\|\boldsymbol{\xi}\|<\varepsilon, \\ \boldsymbol{\xi}\in L^\perp}} g\Big(r\left(\mathbf{I}_d + \boldsymbol{\xi}\right)\Big)\, dr < \infty \tag{2.2}$$

for some $\varepsilon > 0$, where L is a linear subspace of $\mathbb{R}^n$ spanned by the vector $\mathbf{I}_d$ and $L^\perp$ is its orthogonal complement. Finally, let

$$K_g = 2\frac{\pi^{\frac{n}{2}}}{\Gamma(\frac{n}{2})}\int_0^\infty r^{n-1} g\left(r\mathbf{I}_d\right) dr \tag{2.3}$$

be a constant determined by the density g and sample size n.

Theorem 2.1. *If g is continuous and satisfies (2.1) and (2.2), then*

$$\frac{\boldsymbol{P}\left(T_n > u\right)}{t_{n-1}(u)} = K_g + o(1) \quad \textit{as} \quad u \to \infty, \tag{2.4}$$

where $t_{n-1}(u)$ is the tail of the $t-$distribution with $n-1$ degrees of freedom and $0 < K_g = K(g) < \infty$.

Proof. The starting point of the proof is the equality

$$\mathbf{P}\left(T_n > u\right) = \int\limits_{D_1} g(\mathbf{x})d\mathbf{x},$$

where $D_1 = \{\mathbf{x} : T_n > u\}$ and $d\mathbf{x}$ is the notation for $dx_1 dx_2..dx_n$. We now conduct a series of variable changes. Let $\mathbf{e_1}, \mathbf{e_2}, .., \mathbf{e_n}$ be the standard basis in $\mathbb{R}^n$ and A be an orthogonal linear operator which satisfies

$$A\mathbf{e_n} = \mathbf{I}_d. \tag{2.5}$$

First, changing coordinate system $\mathbf{x} = A\mathbf{y}$ we have

$$\overline{X} = \frac{1}{\sqrt{n}} y_n \quad \text{and} \quad S^2 = \frac{1}{n-1}\sum_{i=1}^{n-1} y_i^2.$$

Therefore

$$\mathbf{P}\left(T_n > u\right) = \int\limits_{D_2} g(A\mathbf{y})d\mathbf{y},$$

where $D_2 = \left\{ \mathbf{y} : \frac{y_n}{\sqrt{\frac{1}{n-1}\sum\limits_{i=1}^{n-1} y_i^2}} > u \right\}.$

Next, the variable change

$$y_i = (n-1)^{1/2} r t_i \ \text{ for } \ i \le n-1,$$

$$y_n = r, \quad r > 0,$$

and applying Fubini's theorem and recalling (2.5) leads to

$$\mathbf{P}\left(T_n > u\right) = \int\limits_{\sum t_i^2 < u^{-2}} \cdots \int G(\mathbf{t})d\mathbf{t}, \tag{2.6}$$

where

$$G(\mathbf{t}) = (n-1)^{\frac{n-1}{2}} \int\limits_0^\infty r^{n-1} g\Big(r\left(\mathbf{I}_d + A\mathbf{v}(\mathbf{t})\right)\Big)\, dr,$$

and

$$\mathbf{v}(\mathbf{t}) = (n-1)^{1/2}\left(t_1, t_2, .., t_{n-1}, 0\right).$$

Continuity of g and (2.2) ensure that G is continuous at zero, by the dominated convergence theorem, and Corollary 7.1.1 in Appendix A completes the proof. □

Assumption (2.1) ensures $K_g > 0$ and the condition (2.2) holds if, for example, $K_g < \infty$ and g is continuous and has the asymptotic monotonicity property, see Lemma 7.2 in Appendix B.

Now consider the case when one of the assumptions (2.2) or (2.1) is violated. If (2.2) holds and (2.1) is violated, then (2.4) holds with $K_g = 0$, that is, the right tail of the distribution of T_n is "strictly lighter" than the tail of the $t-$distribution with $n-1$ degrees of freedom. If, instead, (2.1) holds and (2.2) is violated, then, Theorem 7.2 in Appendix B shows that the right tail of the distribution of T_n is "at least as heavy as" $K_g t(u)$, provided $K_g < \infty$, and "strictly heavier" than $t_{n-1}(u)$ if $K_g = \infty$.

Next, two corollaries. The first one concerns dependent Gaussian vectors.

Corollary 2.1.1 (Gaussian zero-mean case)**.** *If $X \sim MVN(\mathbf{0}, \mathbf{\Sigma})$, where $\mathbf{\Sigma}$ is a strictly positive-definite covariance matrix, then (2.4) holds with*

$$K_g = \frac{\left(\mathbf{I}_d \mathbf{\Sigma} \mathbf{I}_d^T\right)^{n/2}}{|\mathbf{\Sigma}|^{1/2}}.$$

Proof. Deriving the expression for K_g in (2.3) is straightforward. Furthermore, $K_g < \infty$ since $\mathbf{\Sigma}$ is non-degenerate, and $MVN(\mathbf{0}, \mathbf{\Sigma})$ has the asymptotic monotonicity property defined in Appendix B. From Lemma 7.2 it then follows that the regularity constraint (2.2) holds, and then so does (2.4). □

Now consider the effect of non-normality. Assume that the elements X_i of the vector $\mathbf{X}$ are independent and identically distributed and let $h(x)$ be their common marginal density, that is, $g(\mathbf{x}) = h(x_1)h(x_2)\cdots h(x_n)$.

Corollary 2.1.2 (i.i.d. case)**.** *If $h(x)$ is continuous and monotone on $[L, \infty)$ for some finite constant L, then (2.4) holds with*

$$K_g = 2\frac{(\pi n)^{\frac{n}{2}}}{\Gamma(\frac{n}{2})} \int\limits_0^\infty r^{n-1} h\left(r\right)^n dr < \infty.$$

Proof. The monotonicity of $h(x)$ on $[L, \infty)$ implies that $g(\mathbf{x})$ has the asymptotic monotonicity property, see Appendix B, and the regularity assumption (2.2) hence follows from finiteness of K_g and

Lemma 7.2. The finiteness of K_g, in turn, follows if we show that $rh(r) \to 0$ as $r \to \infty$.

Indeed, assume to the contrary that $\limsup rh(r) > 0$. Then there exists $\delta > 0$ and a sequence $\{r_k\}_{k=0}^{\infty}$ with $r_0 = L+1$ and such that $r_{k+1} > 2r_k$ and $r_k h(r_k) > \delta$ for any $k > 0$. Now the monotonicity of $h(x)$ on $[L, \infty)$ gives

$$\int\limits_{L+1}^{\infty} h(r)dr \geq \sum_{k=1}^{\infty} (r_k - r_{k-1})\, h(r_k) > \delta \sum_{k=1}^{\infty} \frac{r_k - r_{k-1}}{r_k} = \infty,$$

contradicting that $h(x)$ is a density. □

The constants K_g for some common densities $h(x)$ are given in Table 3.1 in Appendix C.

3 Two-sample $t-$test

In this section we study the tail of the distribution of the Student two-sample $t-$statistic under non-standard conditions. However, we first, consider a more general case. For $n_1 \geq 2,\ n_2 \geq 2$, set $n = n_1 + n_2$ and let $\mathbf{X} = (X_1, X_2, .., X_n)$ be a random vector with multivariate joint density g. Now let S_1 and S_2 denote the sample variances of the vectors $(X_1, X_2, .., X_{n_1})$ and $(X_{n_1+1}, X_{n_1+2}, .., X_n)$ and define

$$T_n = \frac{\frac{1}{n_1}\sum\limits_{i=1}^{n_1} X_i - \frac{1}{n_2}\sum\limits_{i=n_1+1}^{n} X_i}{\sqrt{\alpha S_1^2 + \beta S_2^2}},$$

where α and β are some positive constants.

Our aim is to establish asymptotic expression for the probability of high-level excursions $\mathbf{P}\,(T_n > u)$ as $u \to \infty$. The constants α and β will be specified later. Introduce the unit vectors

$$\mathbf{I}_{d1} = (1/\sqrt{n_1}, 1/\sqrt{n_1}, .., 1/\sqrt{n_1}, 0, 0, .., 0)$$

and

$$\mathbf{I}_{d2} = (0, 0, .., 0, 1/\sqrt{n_2}, 1/\sqrt{n_2}, .., 1/\sqrt{n_2}),$$

let $\omega_0 = \arccos\left(\sqrt{\frac{n_2}{n}}\right)$, and assume that

$$g\Big(r\,\big(\cos(\omega - \omega_0)\mathbf{I}_{d1} + \sin(\omega - \omega_0)\mathbf{I}_{d2}\big)\Big) > 0 \qquad (3.1)$$

for some $r \geq 0$ and $\omega \in [-\pi/2,\ \pi/2]$, and that

$$\int_{-\pi/2}^{\pi/2} \cos(\omega)^{n-2} \int_0^{\infty} r^{n-1} \times$$

$$\times \sup_{\substack{\|\boldsymbol{\xi}\|<\varepsilon \\ \boldsymbol{\xi}\in L^{\perp}}} g\Big(r\,\big(\cos(\omega-\omega_0)\mathbf{I}_{d1} + \sin(\omega-\omega_0)\mathbf{I}_{d2} + \boldsymbol{\xi}\big)\Big)\, dr d\omega \tag{3.2}$$

is finite for some $\varepsilon > 0$, where L is a linear subspace of $\mathbb{R}^n$ spanned by the vectors $\mathbf{I}_{d1}$ and $\mathbf{I}_{d2}$ and $L^{\perp}$ is its orthogonal complement. Next, define

$$K_g = C(n_1, n_2, \alpha, \beta) \int_{-\pi/2}^{\pi/2} \cos(\omega)^{n-2} \times \tag{3.3}$$

$$\times \int_0^{\infty} r^{n-1} g\Big(r\,\big(\cos(\omega-\omega_0)\mathbf{I}_{d1} + \sin(\omega-\omega_0)\mathbf{I}_{d2}\big)\Big)\, dr d\omega,$$

where the constant $C(n_1, n_2, \alpha, \beta)$ is given by

$$C(n_1, n_2, \alpha, \beta) = \frac{2\pi^{\frac{n-1}{2}} \left(\frac{n_1-1}{\alpha}\right)^{\frac{n_1-1}{2}} \left(\frac{n_2-1}{\beta}\right)^{\frac{n_2-1}{2}} \left(\frac{1}{n_1} + \frac{1}{n_2}\right)^{\frac{n-2}{2}}}{\Gamma\left(\frac{n-1}{2}\right)(n-2)^{\frac{n-2}{2}}}.$$

Theorem 3.1. *If g is continuous and satisfies (3.1) and (3.2), then*

$$\frac{\boldsymbol{P}(T_n > u)}{t_{n-2}(u)} = K_g + o(1) \quad as \quad u \to \infty, \tag{3.4}$$

where $t_{n-2}(u)$ is the tail of the $t-$distribution with $n-2$ degrees of freedom and $0 < K_g = K(g) < \infty$.

Proof. The proof is similar to the proof of Theorem 2.1. We start with

$$\mathbf{P}(T_n > u) = \int_{D_1} g(\mathbf{x}) d\mathbf{x},$$

where $D_1 = \{\mathbf{x} : T_n > u\}$. Let A be an orthogonal linear operator such that

$$A\mathbf{e}_{\mathbf{n_1}} = \mathbf{I}_{d1} \text{ and } A\mathbf{e}_{\mathbf{n}} = \mathbf{I}_{d2}. \tag{3.5}$$

Changing coordinate system $\mathbf{x} = A\mathbf{y}$ gives

$$\frac{1}{n_1}\sum_{i=1}^{n_1} X_i = \frac{1}{\sqrt{n_1}} y_{n_1}, \quad \frac{1}{n_2}\sum_{i=n_1+1}^{n} X_i = \frac{1}{\sqrt{n_2}} y_n,$$

$$S_1^2 = \frac{1}{n_1-1}\sum_{i=1}^{n_1-1} y_i^2 \quad \text{and} \quad S_2^2 = \frac{1}{n_2-1}\sum_{i=n_1+1}^{n-1} y_i^2$$

and therefore

$$\mathbf{P}\left(T_n > u\right) = \int\limits_{D_2} g(A\mathbf{y}) d\mathbf{y},$$

where $D_2 = \left\{ \mathbf{y} : \frac{\frac{1}{\sqrt{n_1}} y_{n_1} - \frac{1}{\sqrt{n_2}} y_n}{\left(\frac{\alpha}{n_1-1}\sum\limits_{i=1}^{n_1-1} y_i^2 + \frac{\beta}{n_2-1}\sum\limits_{i=n_1+1}^{n-1} y_i^2\right)^{1/2}} > u \right\}.$

Next, define $c_1(\omega)$ and $c_2(\omega)$ by

$$\frac{c_1(\omega)}{\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} = \sqrt{\frac{n_1-1}{\alpha}}\cos(\omega) \quad \text{and} \quad \frac{c_2(\omega)}{\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} = \sqrt{\frac{n_2-1}{\beta}}\cos(\omega),$$

and introduce new variables $t_1, t_2, .., t_{n-2}, r, \omega$ such that

$$\begin{aligned} y_i &= rc_1(\omega)t_i \quad \text{for } i = 1, 2, .., n_1 - 1, \\ y_i &= rc_2(\omega)t_{i-1} \quad \text{for } i = n_1+1, n_1+2, .., n-1, \\ y_{n_1} &= r\cos(\omega - \omega_0) \quad \text{and} \quad y_n = r\sin(\omega-\omega_0), \quad r > 0. \end{aligned}$$

The identity

$$\frac{1}{\sqrt{n_1}}\cos(\omega-\omega_0) - \frac{1}{\sqrt{n_2}}\sin(\omega-\omega_0) = \sqrt{\frac{1}{n_1}+\frac{1}{n_2}}\cos(\omega)$$

and Fubini's theorem and (3.5) then give

$$\mathbf{P}\left(T_n > u\right) = \int\limits_{\sum\limits_{i=1}^{n-2} t_i^2 < u^{-2}} \cdots \int G(\mathbf{t}) d\mathbf{t}, \tag{3.6}$$

where

$$G(\mathbf{t}) = M \int\limits_{-\pi/2}^{\pi/2} \cos(\omega)^{n-2} \int\limits_{0}^{\infty} r^{n-1} \times$$

$$\times g\Big(r\left(\cos(\omega-\omega_0)\mathbf{I}_{d1}+\sin(\omega-\omega_0)\mathbf{I}_{d2}+A\mathbf{v}(\mathbf{t},\omega-\omega_0)\right)\Big)\,dr d\omega$$

with

$$\mathbf{v}(\mathbf{t},\omega)=\big(c_1(\omega)t_1,..,c_1(\omega)t_{n_1-1},0,c_2(\omega)t_{n_1},..,c_2(\omega)t_{n-2},0\big)$$

and

$$M=\left(\frac{n_1-1}{\alpha}\right)^{\frac{n_1-1}{2}}\left(\frac{n_2-1}{\beta}\right)^{\frac{n_2-1}{2}}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)^{\frac{n-2}{2}}.$$

The finiteness of the integral in (3.2) and continuity of g imply the continuity of G at zero by the dominated convergence theorem, and Corollary 7.1.1 of Appendix A gives the asymptotic expression (3.4) with the constant K_g defined in (3.3). □

The assumption (3.1) ensures that $K_g > 0$, and the regularity constraint (3.2) can be verified directly, or using simpler criteria, see Appendix B.

Corollary 3.1.1 (Gaussian zero-mean case). *If $X \sim MVN(\mathbf{0},\mathbf{\Sigma})$, where $\mathbf{\Sigma}$ is a strictly positive-definite covariance matrix, then (3.4) holds with*

$$K_g = C(n_1,n_2,\alpha,\beta)\frac{\Gamma\left(\frac{n}{2}\right)}{2\pi^{\frac{n}{2}}|\mathbf{\Sigma}|^{1/2}}\int_{-\pi/2}^{\pi/2}\frac{\cos(\omega)^{n-2}}{\left(\mathbf{v}(\omega)\mathbf{\Sigma}^{-1}\mathbf{v}(\omega)^T\right)^{n/2}}d\,\omega, \tag{3.7}$$

where

$$\mathbf{v}(\omega)=\cos(\omega-\omega_0)\mathbf{I}_{d1}+\sin(\omega-\omega_0)\mathbf{I}_{d2}.$$

Proof. Let λ be the smallest eigenvalue of $\mathbf{\Sigma}^{-1}$. Note that $\lambda > 0$, which implies that

$$g(\mathbf{x}) \le \frac{1}{(2\pi)^{n/2}|\mathbf{\Sigma}|^{1/2}}e^{-\frac{\lambda}{2}\|\mathbf{x}\|^2} < \frac{1}{\|\mathbf{x}\|^{n+1}}$$

for $\|\mathbf{x}\|$ large enough. The above and Lemma 7.1 in Appendix B ensure that (3.2) holds, and deriving K_g is a calculus exercise. □

The asymptotic expression for the probability of high level excursions for the Student two-sample $t-$test is obtained by setting

$$\alpha=\frac{n_1-1}{n-2}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)\quad\text{and}\quad\beta=\frac{n_2-1}{n-2}\left(\frac{1}{n_1}+\frac{1}{n_2}\right).$$

For the Gaussian zero-mean case the expression (3.7) then reduces to

$$\frac{\Gamma(\frac{n}{2})}{\Gamma(\frac{n-1}{2})\sqrt{\pi}|\mathbf{\Sigma}|^{1/2}}\int_{-\pi/2}^{\pi/2}\frac{\cos(\omega)^{n-2}}{\left(\mathbf{v}(\omega)\mathbf{\Sigma}^{-1}\mathbf{v}(\omega)^T\right)^{n/2}}d\,\omega. \qquad (3.8)$$

As expected, if $\mathbf{\Sigma} = \sigma^2\mathbf{I}_n$ (recall, $\mathbf{I}_n$ is the identity matrix) and $\sigma^2 > 0$, then direct calculation shows that $K_g = 1$. Now consider the case when the population variances are unequal. Substituting the diagonal matrix

$$\mathbf{\Sigma} = diag\{\underbrace{\sigma_1^2, .., \sigma_1^2}_{n_1}, \underbrace{\sigma_2^2, .., \sigma_2^2}_{n_2}\}$$

into (3.8), the latter, after some lengthy algebraic manipulations, takes form

$$\frac{\Gamma(\frac{n}{2})n_1^{\frac{n}{2}-1}k^{n_2}}{n^{\frac{n-1}{2}}\Gamma(\frac{n-1}{2})\sqrt{\pi}}\left[\int_{-\infty}^{1}\frac{(1-x)^{n-2}}{(1+ck^2x^2)^{n/2}}dx + \int_{1}^{\infty}\frac{(x-1)^{n-2}}{(1+ck^2x^2)^{n/2}}dx\right],$$

where $k = \sigma_1/\sigma_2$ and $c = n_2/n_1$. The integrals can be computed by resolving the corresponding rational functions into partial fractions (n is even) or by expanding brackets in the numerator and integrating by parts (n is odd). We hence computed K_g for sample sizes up to 5, see Table 3.2 in Appendix C.

Unfortunately there is no closed form expressions for (3.7) or (3.8) for an arbitrary covariance matrix $\mathbf{\Sigma}$. The same applies to the i.i.d case, that is, when the vector $\mathbf{X}$ consists of the i.i.d. elements X_i having a common continuous marginal density $h(x)$. However, in both cases the constant K_g may be computed numerically, see Supplementary materials. Note also that for odd sample sizes the exact distribution of the Student two-sample $t-$statistic is known, see Ray and Pitman (1961).

4 Welch's test

In this section we consider the tails of the Welch statistic T_n and discuss the accuracy of the approximation of Corollary 3.1.1 for $\mathbf{P}\left(T_n > u\right)$ as $u \to \infty$. For the rest of the section we assume that the data is a Gaussian zero-mean vector, and the components are independent, and identically distributed within the two samples $(X_1, X_2, .., X_{n_1})$ and $(X_{n_1+1}, X_{n_2+2}, .., X_n)$ having population variances σ_1^2 and σ_2^2 respectively.

Under the above assumptions Ray and Pitman (1961) showed that for odd sample sizes n_1 and n_2 the exact distribution of the Welch statistic T_n (also known as Fisher-Behrens-Welch or Welch-Aspin statistic) is given by the weighted sum of t-distributions. More precisely, formula (4.1) on page 380 of the cited reference (we use the notation k instead of n) says that

$$\rho_{T_n}(v) = (2\pi)^{-\frac{1}{2}} \left\{ \sum_{r=0}^{m-1} \alpha_r \Gamma\left(r+\frac{3}{2}\right)\left(\frac{1}{2a}+\frac{v^2}{2}\right)^{-(r+3/2)} + \right. \tag{4.1}$$

$$\left. + \sum_{r=0}^{k-1} \beta_r \Gamma\left(r+\frac{3}{2}\right)\left(\frac{1}{2b}+\frac{v^2}{2}\right)^{-(r+3/2)} \right\},$$

where $\rho_{T_n}(v)$ is the p.d.f. of the Welch statistic T_n with odd sample sizes $n_1 = 2m+1$ and $n_2 = 2k+1$. The constants a, b, α_r and β_r depend on m, k, σ_1 and σ_2, and the constants α_r and β_r depend also on r.

In the light of the asymptotic representation of Theorem 3.1, the expression (4.1) for the exact density of T_n for odd sample sizes may look suspicious, since $\rho_{T_n}(v)$ is a mixture of the Student $t-$densities that have tails heavier than the tail of the Student $t-$density $t(u)$ in Theorem 3.1. A more detailed investigation, however, revealed that $\alpha_0 + \beta_0 = 0$, and the main terms in the asymptotic expansion for the heaviest densities in (4.1) cancel out, bringing in summands of higher order, which, in turn, may cancel out the main terms of the asymptotic expansion of the next most heavy summands (i.e. the ones that correspond to $r = 1$), and so on.

We tested our theory for the case $n_1 = 3$, $n_2 = 5$ and $\sigma_1^2 = \sigma_2^2$ by computing the left-hand side of (3.4) using (4.1), see Figure 3.2, and also for other choices of n_1 and n_2 and for $\sigma_1 \neq \sigma_2$, see Supplementary materials.

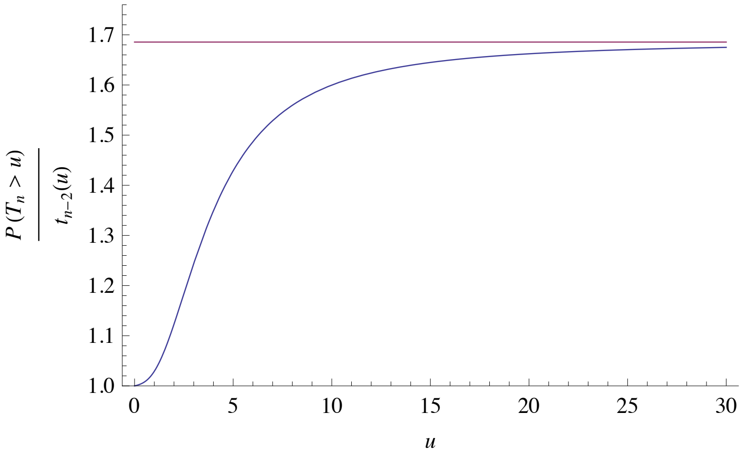


Figure 3.2: *Convergence in (3.4) for the Welch $t-$test with sample sizes $n_1 = 3$ and $n_2 = 5$, variances equal.*
Blue curve is the left-hand side of (3.4) computed using the exact density of Welch's statistic, see (4.1). Horizontal line is the constant $K_g \approx 1.68$ computed from (3.7).

The figure above shows that the left-hand side of (3.4) converges to the constant K_g. Recall that the latter is obtained by substituting $\alpha = 1/n_1$ and $\beta = 1/n_2$ in (3.7). We also computed the first 10 terms in the Taylor series expansion for $\rho_{T_n}(u)$ and $K_g t(u)$ as $u \to \infty$, and found that the two expansions are in agreement with each other, see Supplementary materials.

The asymptotic expressions for the Welch test under non-normality and dependence are obtained using the procedure similar to the one employed in the derivation of the asymptotic expressions for the Student two-sample $t-$ test, see Section 3. We also study the accuracy of the asymptotic approximations of Corollary 3.1.1 for the Student two-sample $t-$test and Welch test using simulations, see Section 7.

5 $F-$test

In this section we study the tails of an F-statistic for non-normal, dependent, and/or non-stationary data. For the sake of simplicity we proceed with the F-test of the equality of variances. The results hold also for an F-test used in one-way ANOVA, lack-of-fit sum of squares, and when comparing two nested linear models in regression analysis. The exact form of the constant factors in the asymptotic expressions for these tests can be derived in a similar way. The formulation of the problem is follows.

Let $\mathbf{X} = (X_1, X_2, .., X_{n_1})$ and $\mathbf{Y} = (Y_1, Y_2, .., Y_{n_2})$, $n_1 \geq 2$ and $n_2 \geq 2$, be random vectors, and let $g(\mathbf{x}, \mathbf{y})$ be the joint density of the vector $(\mathbf{X}, \mathbf{Y})$. Next, set $n = n_1 + n_2$ and define

$$T_n = \frac{S_1^2}{S_2^2},$$

where S_1 and S_2 are the sample variances of $\mathbf{X}$ and $\mathbf{Y}$ respectively. Let $s_1(\mathbf{x})$ denote the sample standard deviation of the vector $\mathbf{x} \in \mathbb{R}^{n_1}$ and define the unit vector $\mathbf{I}_d = \left(1/\sqrt{n_2}, 1/\sqrt{n_2}, .., 1/\sqrt{n_2}\right)$. We assume that

$$s_1(\mathbf{x})\, g\left(\mathbf{x}, r\mathbf{I}_d\right) > 0 \tag{5.1}$$

for some $\mathbf{x}$ and r, and that the integral

$$\int_{\mathbb{R}^{n_1}} \cdots \int s_1(\mathbf{x})^{n_2-1} \int_{-\infty}^{\infty} \max_{\substack{\|\boldsymbol{\xi}\|<\varepsilon,\\ \boldsymbol{\xi}\in L^{\perp}}} g\left(\mathbf{x}, r\mathbf{I}_d + s_1(\mathbf{x})\boldsymbol{\xi}\right) dr d\mathbf{x} \tag{5.2}$$

is finite for some $\varepsilon > 0$, where L is a linear subspace spanned by vector $\mathbf{I}_d$ and $L^{\perp}$ is its orthogonal complement. Finally, define the constant

$$K_g = \frac{\Gamma\left(\frac{n_1-1}{2}\right)\left(\pi(n_1-1)\right)^{\frac{n_2-1}{2}}}{\Gamma\left(\frac{n-2}{2}\right)} \int_{\mathbb{R}^{n_1}} \cdots \int s_1(\mathbf{x})^{n_2-1} \int_{-\infty}^{\infty} g\left(\mathbf{x}, r\mathbf{I}_d\right) dr d\mathbf{x}. \tag{5.3}$$

Theorem 5.1. *If g is continuous and satisfies (5.1) and (5.2), then*

$$\frac{\boldsymbol{P}\left(T_n > u\right)}{F_{n_1-1,n_2-1}(u)} = K_g + o(1) \quad \text{as} \quad u \to \infty, \tag{5.4}$$

where $F_{n_1-1,n_2-1}(u)$ is the tail of the F-distribution with parameters $n_1 - 1$ and $n_2 - 1$ and $0 < K_g = K(g) < \infty$.

Proof. The first part of the proof repeats the proofs of Theorems 2.1 and 3.1. Let A be an orthogonal linear operator defined on $\mathbb{R}^{n_2}$ such that $A\mathbf{e_{n_2}} = \mathbf{I}_d$. We have

$$\mathbf{P}\left(T_n > u\right) = \int\limits_{D} g(\mathbf{x}, A\mathbf{y})d\mathbf{x}d\mathbf{y},$$

where $D = \left\{ (\mathbf{x}, \mathbf{y}) : \frac{s_1^2(\mathbf{x})}{\sum\limits_{i=1}^{n_2-1} y_i^2} > \frac{1}{n_2-1}u \right\}$.

Changing variables $y_i = (n_2 - 1)^{1/2} r s_1(\mathbf{x}) t_i$ for $1 \leq i < n_2$ and $y_{n_2} = r$ (though, formally, we should have considered the case $r > 0$ and $r < 0$ separately), we write

$$\mathbf{P}\left(T_n > u\right) = \int\limits_{\sum\limits_{i=1}^{n_2-1} t_i^2 < u^{-1}} \cdots \int G(\mathbf{t})d\mathbf{t}, \tag{5.5}$$

where

$$G(\mathbf{t}) = (n_2 - 1)^{\frac{n_2-1}{2}} \int\limits_{\mathbb{R}^{n_1}} \cdots \int s_1(\mathbf{x})^{n_2-1} \int\limits_{-\infty}^{\infty} g\left(\mathbf{x}, r\mathbf{I}_d + s_1(\mathbf{x}) A\mathbf{v}(\mathbf{t})\right) dr d\mathbf{x}$$

and

$$\mathbf{v}(\mathbf{t}) = (n_2 - 1)^{1/2} \left(t_1, t_2, .., t_{n_2-1}, 0\right).$$

Continuity of G at zero follows from the finiteness of integral (5.2) and continuity of g by the dominated convergence theorem, and Lemma 7.1 (A) of Appendix A implies that $\mathbf{P}\left(T_n > u\right)$ is asymptotically proportional to $t_{n_2-1}(\sqrt{u})$. It can be shown that

$$\frac{t_{n_2-1}(\sqrt{u})}{F_{n_1-1,n_2-1}(u)} = \frac{\Gamma\left(\frac{n_2}{2}\right)\Gamma\left(\frac{n_1-1}{2}\right)(n_1 - 1)^{\frac{n_2-1}{2}}}{2\sqrt{\pi}\Gamma\left(\frac{n-2}{2}\right)} + o(1) \quad \text{as} \quad u \to \infty,$$

and (5.4) and the expression for K_g in (5.3) follow. □

Corollary 5.1.1 (Gaussian zero-mean case, independent samples)**.** *If X and Y are independent zero-mean Gaussian random vectors* *with strictly non-degenerate covariance matrices $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$, then (5.4) holds with*

$$K_g = C \underset{\mathbb{R}^{n_1}}{\int \cdots \int} \frac{s_1(\mathbf{x})^{n_2-1}}{\left(1+\mathbf{x}\mathbf{\Sigma}_1^{-1}\mathbf{x}^T\right)^{n/2}} d\mathbf{x}, \tag{5.6}$$

where the constant C is given by

$$C = \frac{(n-2)(n_1-1)^{\frac{n_2-1}{2}}\Gamma\left(\frac{n_1-1}{2}\right)|\mathbf{I}_d\mathbf{\Sigma}_2\mathbf{I}_d^T|^{1/2}}{2\pi^{\frac{n_1+1}{2}}|\mathbf{\Sigma}_1|^{1/2}|\mathbf{\Sigma}_2|^{1/2}}.$$

Proof. The regularity assumption (5.2) follows from Lemma 7.1 and the derivation of the constant K_g is a calculus exercise. □

An immediate consequence of the above is the expression for the asymptotic power of the $F-$statistic.

Corollary 5.1.2 (Asymptotic Power)**.** *If X and Y are independent zero-mean Gaussian random vectors with strictly non-degenerate covariance matrices $\sigma_1^2\mathbf{I}_{n_1}$ and $\sigma_2^2\mathbf{I}_{n_2}$, then*

$$\lim_{u\to\infty} \frac{\boldsymbol{P}(T_n > u)}{F_{n_1-1,n_2-1}(u)} = \left(\frac{\sigma_1}{\sigma_2}\right)^{n_2-1}. \tag{5.7}$$

Proof. Changing variables $\mathbf{x} = \sigma_1 B\mathbf{y}$, where B is an orthogonal operator such that $B\mathbf{e}_{\mathbf{n_1}} = \left(1/\sqrt{n_1}, 1/\sqrt{n_1}, .., 1/\sqrt{n_1}\right)$, the integral on the right-hand side of (5.6) takes form

$$\sigma_1^{n-1}\left(\frac{1}{n_1-1}\right)^{\frac{n_2-1}{2}} \underset{\mathbb{R}^{n_1}}{\int \cdots \int} \frac{\left(\|\mathbf{y}\|^2 - y_{n_1}^2\right)^{\frac{n_2-1}{2}}}{(1+\|\mathbf{y}\|^2)^{n/2}} d\mathbf{y},$$

which can be evaluated by passing to spherical coordinates. □

A careful reader may note that (5.7) follows from the equality

$$\mathbf{P}\left(T_n > u\right) = F_{n_1-1,n_2-1}\left((\sigma_2/\sigma_1)^2 u\right)$$

and the asymptotic expansion for the right hand side of the above in terms of $F_{n_1-1,n_2-1}(u)$. Our aim, however, was to show that despite the complexity of the expression (5.3), the constant K_g can be evaluated directly for some densities g.

It is often possible to compute K_g numerically, see the MATLAB (2010) scripts in the Supplementary materials section.

6 Second and higher order approximations

In this section we give a brief note on the speed of convergence in Theorem 1.1. First, consider the case $H_0 : g \sim MVN(\mathbf{0}, \mathbf{I}_n)$, that is, when the classical normal i.i.d. assumption holds and $K_{g_0} = 1$.

Let T_n be the Student one- or two- sample $t-$test or an F-test of the equality of variances, and let $t_k(u)$ be the Student $t-$distribution tail with k degrees of freedom and $F_{m,k}(u)$ be the $F-$distribution tail with parameters m and k. We assume that the regularity assumptions (2.2), (3.2) and (5.2) hold, and let K_g be the constant in (2.3), (3.3) and (5.3) for the three tests accordingly. The function $G(\mathbf{t})$ is given by (2.6), (3.6) and (5.5). Finally, with the standard notation ∇f for the gradient of a scalar function f, and a parameter α which can take values 1 or 2, define

$$d_{\alpha,m,k}(u) = \frac{1}{u^{\frac{\alpha(k+1)}{2}}} \left[C_1 \sup_{\|\mathbf{x}\| \le u^{-\frac{\alpha}{2}}} \left\| \nabla G(\mathbf{x}) \right\| + C_2 \frac{K_g}{\alpha} \frac{1}{u^{\frac{\alpha}{2}}} \right] \tag{6.1}$$

where the constants C_1, C_2 (which depend on m and k) are given in Lemma 7.1 (B) below.

Lemma 6.1 (Absolute error bound)**.** *If $G(\mathbf{t})$ is differentiable in some neighborhood of zero, then for any $u > 0$ the following inequalities hold*

Student's one-sample $t-$test:

$$|\boldsymbol{P}(T_n > u) - K_g\, t_{n-1}(u)| \le d_{2,1,n-1}(u),$$

Student's two-sample $t-$test:

$$|\boldsymbol{P}(T_n > u) - K_g\, t_{n-2}(u)| \le d_{2,1,n-2}(u),$$

$F-$test:

$$|\boldsymbol{P}(T_n > u) - K_g\, F_{n_1-1,n_2-1}(u)| \le d_{1,n_1-1,n_2-1}(u).$$

Proof. For the $F-$test the statement follows from (5.5) and Lemma 7.1 (B) where we set $\alpha = 1$ and replace u by $\sqrt{u}$. For the Student one- and two- sample $t-$test we use the representations (2.4) and (3.4) and Corollary 7.1.1. □

Next, we recall the notation $t(u)$ for the distribution tail of T_n under $H_0 : g \sim MVN(\mathbf{0}, \mathbf{I}_n)$ and derive the asymptotic formula for the relative error.

Lemma 6.2 (Relative error decrease rate)**.** *If $G(\mathbf{t})$ is twice differentiable in some neighborhood of zero, then*

$$\frac{\boldsymbol{P}(T_n > u) - K_g t(u)}{\boldsymbol{P}(T_n > u)} = \frac{C_3}{u^\alpha}(1 + o(1)),$$

where

$$C_3 = \frac{\alpha\, k\, B\left(\frac{m}{2}, \frac{k}{2}\right)}{2\left(\frac{k}{m}\right)^{k/2}} \frac{L_{G_\alpha}}{K_g},$$

the triple (α, m, k) is set to $(2, 1, n-1)$, $(2, 1, n-2)$ and $(1, n_1, n_2)$ for the Student one- and two- sample $t-$ and $F-$ tests respectively, and the constant L_G is defined in Lemma 7.1 C).

Proof. The result is follows from the representations (2.4), (3.4) and (5.4) for $\mathbf{P}(T_n > u)$, Lemma 7.1 C), and formula (7.7). □

The bounds and asymptotic expressions for the case of an arbitrary null hypothesis H_0 follow from basic calculus. Indeed,

$$\mathbf{P}(T_n > u|H_1) - \frac{K_{g_1}}{K_{g_0}}\mathbf{P}(T_n > u|H_0) = \Big(\mathbf{P}(T_n > u|H_1) - K_{g_1}t(u)\Big)$$
$$-\frac{K_{g_1}}{K_{g_0}}\Big(\mathbf{P}(T_n > u|H_0) - K_{g_0}t(u)\Big),$$

and the absolute error of the approximation (1.1) is thus bounded by the linear combination of the absolute errors, and we can now apply Lemma 6.1. For the relative error - simply replace the probabilities $\mathbf{P}(T_n > u|H_1)$ and $\mathbf{P}(T_n > u|H_0)$ by their second order expansions given by Lemma 7.1 C) and use (7.7).

Note that Lemma 6.2 expresses the rate of decrease of the relative error $RE(u)$ as $u \to \infty$. In practice, however, it might be desirable to look on the p-value scale. Take, for example, the Student one-sample $t-$test based on $n = 2$ observations. Omitting the $o(1)$ term in Lemma 6.2 this means that $RE(u) < 0.01C_3$ for $u > 10$, or, equivalently, for $t(u) < t_1(10) \approx 0.03$. But if, instead, one had $n = 6$ observations, and assuming that the constant C_3 (which depends on n) did not change, similar precision would hold for p-values of at most $t_5(10) \approx 0.00008$. In this hypothetical example, the conclusion is that the asymptotic formulas are more accurate for small sample sizes. Such behavior was also observed in the simulation study, see Section 7.

Lemma 7.1 can be generalized to obtain a series expansion for the probability $\mathbf{P}\,(T_n > u)$ as $u \to \infty$. However, formulating the assumptions which would allow to interchange differentiation and integration in the expressions for the coefficients is yet an open problem, see Section 7.

7 Simulation study

In this section we study the accuracy of the asymptotic formulas of Sections 2, 3, 4 and 5 using simulations. With the notation T_n for the Student one- or two- sample $t-$ or $F-$ test, $t(u)$ for the distribution tail of T_n under $H_0 : g \sim MVN(\mathbf{0}, \mathbf{I}_n)$, and K_g for the corresponding asymptotic constant, consider the following procedure:

1. Choose n_1 and n_2 (or simply n for the Student one-sample $t-$test).
2. Specify the density $g(\mathbf{x})$ of the data vector X.
3. Set the values of the integer parameters r and res (description follows) and simulate $N = r \times res$ random vectors $\mathbf{X} \sim g$.
4. For each vector $\mathbf{X}$ compute $t^* = T_n(\mathbf{X})$, the value of the test statistic T_n, and two p-values $p^R = t(t^*)$ and $p^C = K_g\, t(t^*)$.
5. Plot the empirical CDF of p^R and p^C over the range $[0, 1/r]$.

The *Zoom Factor* (Z.F.) parameter r determines the tail region of interest: only p-values that fall in the interval $I(r) = [0, 1/r]$ are kept. The parameter res determines the desired number of p-values out of N to fall in the interval $I(r)$ as if the p-values were uniformly distributed. For the present, we set res to $10,000$ - the latter gives high accuracy in the approximation of the tails of the distribution of p^R and p^C by the empirical CDF.

The letters "R" and "C" in the notation for the simulated p-values stand for "Raw" (computed using $t(u)$) and "Corrected" (computed using $K_g t(u)$) accordingly. For the Welch test "Raw" p-values are computed using the Welch-Satterthwaite approximation and the notation is p^{WS}.

Within the scope of the current study we consider only the i.i.d case. Further, let $h(x)$ be the marginal density of the vector $\mathbf{X}$. From now on $h(x)$ is one of the following densities: $Uniform(-1, 1)$, *Standard normal*, *Centered exponential*, *Cauchy* or $t-$ density with 2 or 5 degrees of freedom, and the values of the constant K_g are computed numerically using MATLAB (2010). The scripts can be found in the Supplementary materials section.

The empirical CDF plots for the p-values p^R and p^C for the Student one- and two- sample $t-$ and $F-$ tests, with different sample sizes n, n_1, and n_2, and Zoom Factor r ranging from 20 to $1,000,000$ are shown in Figures 3.3, 3.4 and 3.5 for the three tests accordingly, see Appendix C.

The plots show that the asymptotic approximations perform uniformly better in the tail regions for all the three tests and for all the densities $h(x)$ and sample sizes n, n_1 and n_2 considered in the study.

The speed of convergence varies depending on the choice of the sample size(s) and population density $h(x)$. The two observations are: 1) the convergence is faster for small sample sizes, and 2) the convergence is faster for the case when the deviation of the distribution tail of p^R from the theoretical uniform distribution is small (i.e. when K_g is not "too far" from 1). Both observations are in agreement with the bounds for the absolute error in Lemma 6.1, see Section 6.

The only plots which showed slow convergence speed were those that correspond to the case of the *Uniform* population distribution for the Student one-sample $t-$test of sample size 5, and the *Uniform* and *Centered exponential* densities for the Student two-sample $t-$test of sample sizes $n_1 = 3$ and $n_2 = 5$. Note also that for these cases the sample sizes and the values of the constant K_g are larger compared to the other cases.

The convergence speed for the $F-$test of the equality of variances is, in fact, faster than for Student's one- and two- sample $t-$tests.

For the Welch test we compared the asymptotic approximation (3.7) of Corollary 3.1.1 with the Welch-Satterthwaite approximation. The latter suggests that the distribution of the Welch statistic T_n should be approximated by the Student $t-$distribution with

$$\nu = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{S_1^4}{n_1^2(n_1-1)} + \frac{S_2^4}{n_2^2(n_2-1)}}$$

degrees of freedom, where ν is estimated from data.

Based on the full scale plots of Figure 3.6, the top row, one may get an impression that the empirical CDF of the p-values p^{WS} is in agreement with the theoretical uniform distribution over the whole range of values between 0 and 1. Such impression is, however, misleading - this becomes clear as one "zooms in" to the tails of the distribution of the p-values: the plots in the middle row of Figure 3.6 correspond to the tail region $[0, 0.001]$, and one can see that the p-values obtained using the Welch-Satterthwaite approximation deviate significantly from the theoretical uniform distribution. The plot of the empirical CDF for the p-values p^C, on the other hand, almost coincides with the diagonal line. The advantage of using the tail approximation is fully convincing at Zoom Factor 100 000, see the bottom row of Figure 3.6 in Appendix C.

The simulation study can be extended to cover the case of dependence and non-stationarity of the data. All the scripts used in this section are available as a Supplementary material.

Appendix A: Asymptotic behavior of an integral of a continuous function over a shrinking ball

In this Section we introduce the key lemma which is repeatedly used in Sections 2, 3 and 5. It was shown that the probability of high-level excursions for Student's one-sample $t-$test, Student's two-sample $t-$test, Welch's test and $F-$test is determined by the asymptotic behavior of an integral of some function (different for each of the tests) over a shrinking ball.

Let $G(\mathbf{x}),\ \mathbf{x} \in \mathbb{R}^k$ be some real-valued function and consider the asymptotic behavior of

$$F(u) = \underset{\sum x_i^2 < u^{-2}}{\int \cdots \int} G(\mathbf{x})\, d\mathbf{x}. \tag{7.1}$$

The value of k is fixed and $u \to \infty$.

Lemma 7.1. *Set $f(u) = \alpha^{-1} F_{m,k}(u^2)$, where $F_{m,k}(\cdot)$ is the tail of the $F-$distribution with $m \geq 1$ and $k \geq 2$ degrees of freedom, and let $\mathrm{Vol}(B_k)$ be the volume of the unit k-ball B_k and $B(x, y)$ be the Beta function. The parameter α will be set later. With the above notation we have:*

(A) If G is continuous at zero, then

$$\frac{F(u)}{f(u)} = K_{G,\alpha} + o(1) \quad as \quad u \to \infty, \tag{7.2}$$

where

$$K_{G,\alpha} = \frac{\alpha\, k\, B\left(\frac{m}{2}, \frac{k}{2}\right)}{2\left(\frac{k}{m}\right)^{k/2}} \mathrm{Vol}(B_k) G(\mathbf{0}). \tag{7.3}$$

(B) If G is differentiable in some neighborhood of zero, then for any $u > 0$

$$|F(u) - K_{G,\alpha}\, f(u)| \leq \frac{C_1}{u^{k+1}} \sup_{\|\mathbf{x}\| \leq u^{-1}} \left\|\nabla G(\mathbf{x})\right\| + C_2 \frac{K_{G,\alpha}}{\alpha} \frac{1}{u^{k+2}}, \tag{7.4}$$

where

$$C_1 = \mathrm{Vol}(B_k) \quad and \quad C_2 = \frac{k(k+m)}{m(k+2)} \frac{\left(\frac{k}{m}\right)^{k/2}}{B\left(\frac{m}{2}, \frac{k}{2}\right)}, \tag{7.5}$$

and $\nabla G(\mathbf{x})$ is a gradient of G evaluated at point $\mathbf{x}$.

(C) If G is twice differentiable in some neighborhood of zero, then

$$u^{k+2} \left(F(u) - K_{G,\alpha}\, f(u)\right) = L_{G,\alpha} + o(1) \quad as \quad u \to \infty, \tag{7.6}$$

where

$$L_{G,\alpha} = C_1 \frac{tr\left(Hess\left(G(\mathbf{0})\right)\right)}{2(k+2)} - C_2 \frac{K_{G,\alpha}}{\alpha},$$

$tr(A)$ is the trace of square matrix A, and $Hess\,(G(\mathbf{x}))$ is the Hessian matrix of G at point $\mathbf{x}$. Constants C_1 and C_2 are given in (7.5).

Proof. The first statement follows from the asymptotic expansion for the $F-$distribution tail (derivation is straightforward)

$$f(u) = \frac{2\left(\frac{k}{m}\right)^{k/2}}{\alpha\, k\, B\left(\frac{m}{2}, \frac{k}{2}\right)} \left[\frac{1}{u^k} - \frac{k^2(k+m)}{2m(k+2)}\frac{1}{u^{k+2}}\right] + o\left(\frac{1}{u^{k+2}}\right). \quad (7.7)$$

Indeed, changing variables $\mathbf{x} = \mathbf{y}/u$ we write

$$F(u) = \underset{\sum x_i^2 < u^{-2}}{\int \cdots \int} G(\mathbf{x})d\mathbf{x} = \frac{1}{u^k} \underset{B_k}{\int \cdots \int} G(\mathbf{y}/u)d\mathbf{y}. \quad (7.8)$$

Continuity of G at zero implies uniform convergence $G(\mathbf{y}/u) \overset{B_k}{\rightrightarrows} G(\mathbf{0})$, and thus

$$F(u) = \mathrm{Vol}\,(B_k)\, G(\mathbf{0})\frac{1}{u^k}\,(1 + o(1)). \quad (7.9)$$

Dividing (7.9) by (7.7) we get that the value of K_G in (7.2) coincides with (7.3).

Now assume G is differentiable in some neighborhood of zero and consider the Lagrange form of the Tailor expansion of $G(\mathbf{y}/u)$. The latter and (7.8) give

$$\begin{aligned} |F(u) - K_{G,\alpha}\, f(u)| \;&\le\; \frac{1}{u^k}\left|\mathrm{Vol}\,(B_k)\, G(\mathbf{0}) - u^k K_{G,\alpha}\, f(u)\right| \\ &+\; \frac{1}{u^{k+1}}\left|\underset{B_k}{\int \cdots \int} \nabla G\left(\xi(\mathbf{y})\mathbf{y}\right)\mathbf{y}^T d\mathbf{y}\right|, \end{aligned} \quad (7.10)$$

where $0 \le \xi(\mathbf{y}) \le 1/u$. The second summand in the right-hand side of the above inequality is bounded by

$$\frac{1}{u^{k+1}}\mathrm{Vol}\,(B_k) \sup_{B_k} \|\nabla G(\mathbf{x}/u)\|,$$

and the bound for the remaining summand follows from (7.7), where we note that $f(u)$ is bounded by the two successive partial sums in its alternated series (7.7) and that the factors before $\mathrm{Vol}\,(B_k)\,G(\mathbf{0})$ in the expression for $K_{G,\alpha}$ and before the square brackets in (7.7) cancel out. The last step is to use formulas (7.3) and (7.5) to express $\mathrm{Vol}\,(B_k)\,G(\mathbf{0})$ in terms of $K_{G,\alpha}$ and C_2.

We move on to the proof of (7.6). Taylor expansion for $G(\mathbf{y}/u)$ yields

$$F(u) = \frac{1}{u^k}\mathrm{Vol}\,(B_k)\,G(\mathbf{0}) + \\ + \frac{1}{u^{k+2}}\underset{B_k}{\int\cdots\int}\frac{\mathbf{y}Hess\,(G\,(\mathbf{0}))\,\mathbf{y}^T}{2}d\mathbf{y} + o\left(\frac{1}{u^{k+2}}\right),$$

where we took into account that the integral of the odd function $\nabla G(\mathbf{0})\mathbf{y}$ over the ball B_k is zero. Neglecting odd terms in $\mathbf{y}Hess\,(G\,(\mathbf{0}))\,\mathbf{y}^T$ we have

$$\underset{B_k}{\int\cdots\int}\mathbf{y}Hess\,(G\,(\mathbf{0}))\,\mathbf{y}^T d\mathbf{y} = \sum\underset{B_k}{\int\cdots\int}\frac{\partial^2 G(\mathbf{0})}{\partial^2 y_i}y_i^2 d\mathbf{y}$$

$$= \left(\sum\frac{\partial^2 G(\mathbf{0})}{\partial^2 y_i}\right)\underset{B_k}{\int\cdots\int}\frac{\sum y_i^2}{k}d\mathbf{y} = \mathrm{Vol}\,(B_k)\,\frac{tr(Hess(G(\mathbf{0}))}{k+2},$$

where the last integral was computed using spherical coordinates. Substituting the second order Taylor expansion for $F(u)$ and expression for $f(u)$ in (7.7) into the left-hand side of (7.6) we get the constant $L_{G,\alpha}$. □

Note that the expression $\alpha^{-1}K_{G,\alpha}$ does not depend on α and thus the right-hand side of (7.4) and (7.6) depends only on the integrand G in (7.1) and parameters m and k.

Corollary 7.1.1. *Let $t_k(u)$ be the Student $t-$distribution tail with k degrees of freedom. If G is continuous at zero, then*

$$\frac{F(u)}{t_k(u)} = K_{G,2} + o(1) \quad as \quad u \to \infty,$$

where $K_{G,2}$ is given by (7.3) with $m = 1$. Statements (B) and (C) also hold for $f(u) = t_k(u)$, provided $m = 1$ and $\alpha = 2$.

Proof. Note that $t_k(u) = \frac{1}{2}F_{1,k}(u^2)$ and apply Lemma 7.1. □

Appendix B: A note on the regularity constraints and the continuity assumption

The aim of this section is to replace the technical constraints (2.2), (3.2) and (5.2) of Theorems 2.1, 3.1 and 5.1 by simpler criteria, and to weaken the assumption of continuity of the multivariate density g of the vector of data.

The key observation that provides better understanding of the nature of the regularity constraints (2.2), (3.2) and (5.2) is that the proofs of the theorems share a common part, which is to apply Lemma 7.1 (A) or Corollary 7.1.1 to the representations (2.6), (3.6) and (5.5), and then use dominated convergence theorem to show that the corresponding function $G(\mathbf{t})$ is continuous at zero. The only purpose of the regularity constraints is to ensure that the limiting and integration operations are interchangeable, and that the limit constant K_g is finite. Note, however, that omitting the regularity assumptions (2.2), (3.2) and (5.2) we immediately obtain

Theorem 7.2 ("liminf" analogue of Theorems 2.1, 3.1 and 5.1). *Let T_n be the Student one- or two-sample $t-$test or an $F-$test and let $t(u)$ be the distribution tail of T_n under the null hypothesis $H_0 : g \sim MVN(\mathbf{0}, \sigma^2\mathbf{I}_n)$, where $\sigma^2 > 0$ and $\mathbf{I}_n$ is the identity matrix. If g is continuous, then*

$$\liminf_{u\to\infty} \frac{\boldsymbol{P}(T_n > u)}{t(u)} \geq K_g,$$

where the constant K_g is given by (2.3), (3.3) and (5.3) accordingly, though it may be infinite.

Proof. The statement is a consequence of representations (2.6), (3.6) and (5.5), Fatou's lemma, and inequality analog of Lemma 7.1 (A) and Corollary 7.1.1. □

We now give the sufficient conditions for the regularity constraints of Theorems 2.1, 3.1 and 5.1 to hold. These conditions, however, are not necessary. One may expect that the statements of Theorems 2.1, 3.1 and 5.1 hold when g is continuous and K_g is finite. The proof or disproof of the latter claim is an open problem.

Lemma 7.1. *If $g(\mathbf{x})$ is bounded and there exist positive constants R, C and δ such that*

$$g(\mathbf{x}) \leq \frac{C}{\|\mathbf{x}\|^{n+\delta}} \quad \textit{for} \quad \|\mathbf{x}\| > R, \tag{7.11}$$

then the assumptions (2.2), (3.2) and (5.2) of Theorems 2.1, 3.1 and 5.1 hold.

Proof. The integrals in (2.2), (3.2) and (5.2) will be estimated by partitioning the integration domain into several disjoint parts D_i and D_j^* and considering the integrals over these sets separately. For non-compact domains D_j^* the integrand will be estimated from above using the bound (7.11). The key task is of course to show that these bounds are integrable. The integrability over the compact domains D_i follows from the fact that $g(\mathbf{x})$ is bounded.

In the notation below let $G(r)$, $G(\omega, r)$ and $G(\mathbf{x}, r)$ be the integrands in (2.2), (3.2) and (5.2) accordingly.

Student's one-sample t−test: Set $D_1 = [0, R]$ and $D_1^* = [R, \infty]$. Since $\mathbf{I}_d$ and $\boldsymbol{\xi}$ are orthogonal and taking into account that $\|\mathbf{I}_d\| = 1$ we have

$$\|r\,(\mathbf{I}_d + \boldsymbol{\xi})\,\|^2 = r^2(1 + \|\boldsymbol{\xi}\|^2) \geq r^2,$$

and the bound (7.11) gives

$$\int\limits_{D_1^*} G(r)dr < \int\limits_{R}^{\infty} \frac{C}{r^{1+\delta}} dr < \infty.$$

Student's two-sample t−test: Setting

$$D_1 = [-\pi/2, \pi/2] \times [0, R] \quad \text{and} \quad D_1^* = [-\pi/2, \pi/2] \times [R, \infty]$$

and noting that $\mathbf{I}_{d_1}$, $\mathbf{I}_{d_2}$ and $\boldsymbol{\xi}$ are mutually orthogonal we get

$$\|r\,\big(\cos(\omega - \omega_0)\mathbf{I}_{d1} + \sin(\omega - \omega_0)\mathbf{I}_{d2} + \boldsymbol{\xi}\big)\,\|^2 = r^2(1 + \|\boldsymbol{\xi}\|^2) \geq r^2,$$

where we used the fact that $\|\mathbf{I}_{d_1}\| = \|\mathbf{I}_{d_2}\| = 1$. Now the bound (7.11) implies

$$\int\limits_{D_1^*} G(\omega, r)dr < \int\limits_{-\pi/2}^{\pi/2} \cos(\omega)^{n-2} d\omega \times \int\limits_{R}^{\infty} \frac{C}{r^{1+\delta}} < \infty.$$

F−test: Consider the following partition of $\mathbb{R}^{n_1+1}$:

$$\begin{aligned} D_1 &= \Big\{(\mathbf{x}, r) : \|\mathbf{x}\| \le R \ , \ |r| \le R\Big\}, \\ D_1^* &= \Big\{(\mathbf{x}, r) : \|\mathbf{x}\| \le R \ , \ |r| > R\Big\} \\ D_2^* &= \Big\{(\mathbf{x}, r) : \|\mathbf{x}\| > R\Big\}. \end{aligned}$$

Since $\mathbf{I}_d$ and $\boldsymbol{\xi}$ are orthogonal and $\|\mathbf{I}_d\| = 1$ we have

$$\| \, (\mathbf{x}, r\mathbf{I}_d + s_1(\mathbf{x})\boldsymbol{\xi}) \, \|^2 = \|\mathbf{x}\|^2 + r^2 + s_1(\mathbf{x})^2\|\boldsymbol{\xi}\|^2 \ge \|\mathbf{x}\|^2 + r^2.$$

Then

$$\int\limits_{D_1^*}\!\cdots\!\int G(\mathbf{x}, r) dr d\mathbf{x} < \int\limits_{\|\mathbf{x}\|\le R}\!\cdots\!\int s_1(\mathbf{x})^{n_2-1} d\mathbf{x} \ \times \int\limits_{|r|>R} \frac{C}{|r|^{n+\delta}} dr < \infty$$

and

$$\int\limits_{D_2^*}\!\cdots\!\int G(\mathbf{x}, r) dr d\mathbf{x} < \int\limits_{\|\mathbf{x}\|>R}\!\cdots\!\int \int\limits_{-\infty}^{\infty} \frac{s_1(\mathbf{x})^{n_2-1}}{(\|\mathbf{x}\|^2 + r^2)^{\frac{n+\delta}{2}}} dr d\mathbf{x} <$$

$$< \int\limits_{\|\mathbf{x}\|>R}\!\cdots\!\int \frac{s_1(\mathbf{x})^{n_2-1}}{\|\mathbf{x}\|^{n-1+\delta}} d\mathbf{x} \ \times \int\limits_{-\infty}^{\infty} \frac{1}{(1 + r^2)^{n/2}} dr < \infty,$$

where the multidimensional integral in the last inequality is computed by means of passing to spherical coordinates. □

Note that in the i.i.d case (7.11) is equivalent to the existence of the $n-1+\delta$ moment of the marginal density $h(x)$. For the Student one-sample $t-$test, however, the criterium of Lemma 7.1 is ”too strict”.

Definition. We say that the multivariate density $g(\mathbf{x})$ has the asymptotic monotonicity property if there exists a constant M such that for any $1 \le i \le n$ and constants c_j, $j \ne i$, the function $f(x) = g(c_1, c_2, .., c_{i-1}, x, c_{i+1}, .., c_n)$ is monotone on $[M \ , \ \infty)$.

Lemma 7.2. *If K_g is finite and $g(\mathbf{x})$ is bounded and has the asymptotic monotonicity property, then the assumption (2.2) holds.*

Proof. Setting ε equal to $(2\sqrt{n})^{-1}$ and using asymptotic monotonicity property we get that the integral in (2.2) is bounded by

$$\int\limits_0^{2M\sqrt{n}} r^{n-1} \sup_{\|\boldsymbol{\xi}\|<\frac{1}{2\sqrt{n}}} g\Big(r\left(\mathbf{I}+\boldsymbol{\xi}\right)\Big)\,dr + \int\limits_{2M\sqrt{n}}^{\infty} r^{n-1} g\left(r\frac{\mathbf{I}}{2}\right) dr < \infty.$$

The first summand is finite owing to the boundness of g and the finiteness of the second summand is equivalent to the finiteness of K_g. □

Now consider for a moment the i.i.d case of the Student one-sample $t-$test. Lemma 7.2 shows that the statement of Theorem 2.1 holds for any continuous marginal density $h(\mathbf{x})$ that has monotone tails and such that $K_g < \infty$. The latter assumption, however, is weaker than the assumption of existence of the first moment and holds even for such heavy tailed densities as Cauchy. As we can see, asymptotic monotonicity and finiteness of K_g is a very mild constraint, at least from the practical point of view.

Unfortunately, due to the geometrical considerations, there is no asymptotic monotonicity criterium analogue for the case of the Student two-sample $t-$test and $F-$test, and the constant K_g in (3.3) and (5.3) may be infinite for some heavy-tailed densities.

Finally, note that in the proofs of Theorems 2.1, 3.1 and 5.1 we may have used the "almost everywhere" version of the dominated convergence theorem. For the Student one-sample $t-$test this means that the assumption of continuity of g in the corresponding theorems can be replaced by the assumption that $g(\mathbf{x})$ is continuous a.e. on the set of points

$$\mathbf{x} = r\mathbf{I}_d, \quad r > 0,$$

for the Student two-sample $t-$test - on the set of points

$$\mathbf{x} = r\left(\cos(\omega-\omega_0)\mathbf{I}_{d_1} + \sin(\omega-\omega_0)\mathbf{I}_{d_2} + \mathbf{z}\right),$$

where $r > 0$ and $\omega \in [-\pi/2\ ,\ \pi/2]$, and for the $F-$test - on the set of points

$$\mathbf{x} = \mathbb{R}^{n_1} \times r\mathbf{I}_d,\ r \in \mathbb{R}.$$

Here a.e. means almost everywhere with respect to the Lebesque measure induced by the d-dimensional Lebesque measure of the linear space L in (2.2), (3.2) and (5.2), where $d = dim(L)$. Continuous means continuous as a function of $\mathbf{x} \in \mathbb{R}^n$.

Appendix C: Tables and figures

Table 3.1: The constants K_g for the i.i.d case of the Student one-sample $t-$test. Here $\Gamma(x)$, $B(x)$ and $M(a,b,x)$ are the Gamma, Beta and Kummer confluent hypergeometric function, see e.g. Polyanin and Manzhirov (2008) and Hayek (2001).

Density / Constant K_g
Normal with mean $\mu \neq 0$ and standard deviation $\sigma > 0$
$M\left(\frac{1-n}{2},\frac{1}{2},-\frac{n\mu^2}{2\sigma^2}\right)+\frac{\mu}{\sigma}\frac{\sqrt{2n}\Gamma\left(\frac{1+n}{2}\right)}{\Gamma\left(\frac{n}{2}\right)}M\left(1-\frac{n}{2},\frac{3}{2},-\frac{n\mu^2}{2\sigma^2}\right)$
Half-normal, and log-normal derived from a $N(\mu,\sigma^2)$
2^n and $\frac{n^{\frac{n-1}{2}}\sqrt{\pi}}{2^{\frac{n-3}{2}}\sigma^{n-1}\Gamma\left(\frac{n}{2}\right)}$
χ with $\nu > 0$, and χ^2 (and its inverse) with $\nu \geq 2$ d.f.
$\frac{2^n\pi^{n/2}\Gamma\left(\frac{n\nu}{2}\right)}{n^{\frac{n}{2}(\nu-1)}\Gamma\left(\frac{\nu}{2}\right)^n\Gamma\left(\frac{n}{2}\right)}$ and $\frac{2\pi^{n/2}\Gamma\left(\frac{n\nu}{2}\right)}{n^{\frac{n}{2}(\nu-1)}\Gamma\left(\frac{\nu}{2}\right)^n\Gamma\left(\frac{n}{2}\right)}$
F with $\mu > 0$ and $\nu > 0$ degrees of freedom
$\frac{2(\pi n)^{n/2}\Gamma\left(\frac{\mu n}{2}\right)\Gamma\left(\frac{\nu n}{2}\right)\Gamma\left(\frac{\mu+\nu}{2}\right)^n}{\Gamma\left(\frac{n}{2}\right)\left[\Gamma\left(\frac{\mu}{2}\right)\Gamma\left(\frac{\nu}{2}\right)\right]^n\Gamma\left(\frac{\mu+\nu}{2}n\right)}$
T with $\nu > 0$ d.f. and Cauchy
$\frac{n^{n/2}\Gamma\left(\frac{\nu n}{2}\right)}{\Gamma\left(\frac{(\nu+1)n}{2}\right)}\left(\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)}\right)^n$ and $\frac{n^{n/2}}{2^{n-1}\pi^{\frac{n-1}{2}}\Gamma\left(\frac{n+1}{2}\right)}$
Beta with shape parameters $\alpha > 1$ and $\beta > 1$
$\frac{2(\pi n)^{n/2}\Gamma(\alpha n)\Gamma\left(1+(\beta-1)n\right)}{B(\alpha,\beta)^n\Gamma\left(\frac{n}{2}\right)\Gamma\left(1+(\alpha+\beta-1)n\right)}$

Gamma (and its inverse) with shape $\alpha > 1$
$\frac{2n^{\frac{n}{2}(1-2\alpha)}\pi^{n/2}\Gamma(\alpha n)}{\Gamma(\alpha)^n\Gamma\left(\frac{n}{2}\right)}$
Uniform on interval $[a,b], b > 0$
$\frac{(\pi n)^{\frac{n}{2}}}{\Gamma\left(\frac{n}{2}+1\right)}\begin{cases}\left(\frac{b}{b-a}\right)^n & 0 \in [a,b] \\ \frac{b^n-a^n}{(b-a)^n} & [a,b] \subset [0,\infty)\end{cases}$
Centered exponential and exponential
$\frac{2\left(\frac{\pi}{n}\right)^{n/2}\Gamma(n)}{e^n\Gamma\left(\frac{n}{2}\right)}$ and $\frac{2\left(\frac{\pi}{n}\right)^{n/2}\Gamma(n)}{\Gamma\left(\frac{n}{2}\right)}$
Maxwell, and Pareto with $k > 0$ and scale $\alpha > 0$
$\frac{\left(\frac{4}{n}\right)^n\Gamma\left(\frac{3n}{2}\right)}{\Gamma\left(\frac{n}{2}\right)}$ and $\frac{(\pi n)^{n/2}\alpha^{n-1}}{\Gamma\left(\frac{n}{2}+1\right)}$

Table 3.2: Constants K_g for the Student two-sample $t-$test, variances unequal.

$n_2 \backslash n_1$	$n_1 = 2$	$n_1 = 3$	$n_1 = 4$	$n_1 = 5$
$n_2 = 2$	$\frac{k^2+1}{2k}$	$\frac{\left(2k^2+3\right)^{3/2}}{5\sqrt{5}k^2}$	$\frac{\left(k^2+2\right)^2}{9k^3}$	$\frac{\left(2k^2+5\right)^{5/2}}{49\sqrt{7}k^4}$
$n_2 = 3$	$\frac{\left(3k^2+2\right)^{3/2}}{5\sqrt{5}k}$	$\frac{\left(k^2+1\right)^2}{4k^2}$	$\frac{\left(3k^2+4\right)^{5/2}}{49\sqrt{7}k^3}$	$\frac{\left(3k^2+5\right)^3}{512k^4}$
$n_2 = 4$	$\frac{\left(2k^2+1\right)^2}{9k}$	$\frac{\left(4k^2+3\right)^{5/2}}{49\sqrt{7}k^2}$	$\frac{\left(k^2+1\right)^3}{8k^3}$	$\frac{\left(4k^2+5\right)^{7/2}}{2187k^4}$
$n_2 = 5$	$\frac{\left(5k^2+2\right)^{5/2}}{49\sqrt{7}k}$	$\frac{\left(5k^2+3\right)^3}{512k^2}$	$\frac{\left(5k^2+4\right)^{7/2}}{2187k^3}$	$\frac{\left(k^2+1\right)^4}{16k^4}$

Student's one-sample t−test

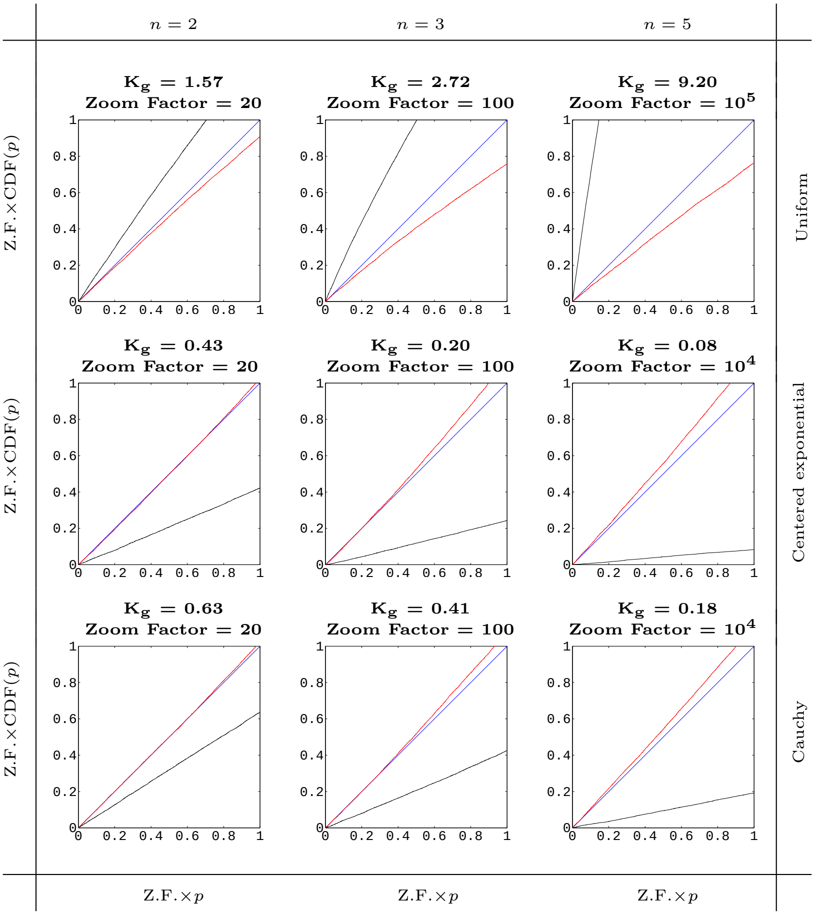


Figure 3.3: *The eCDF of the p-values for the Student one-sample t−test.*

The empirical CDFs of the raw and corrected p-values p^R and p^C are shown in black and red accordingly. The top, middle and bottom rows correspond to the $Uniform(-1, 1)$, *Centered exponential* and *Cauchy* densities, and left, middle and right columns correspond to sample sizes $n = 2$, $n = 3$ and $n = 5$. The blue diagonal line is the theoretical uniform distribution. The axes are scaled according to the Zoom Factor (Z.F.) parameter r in the title of the graphs.

Student's two-sample t−test

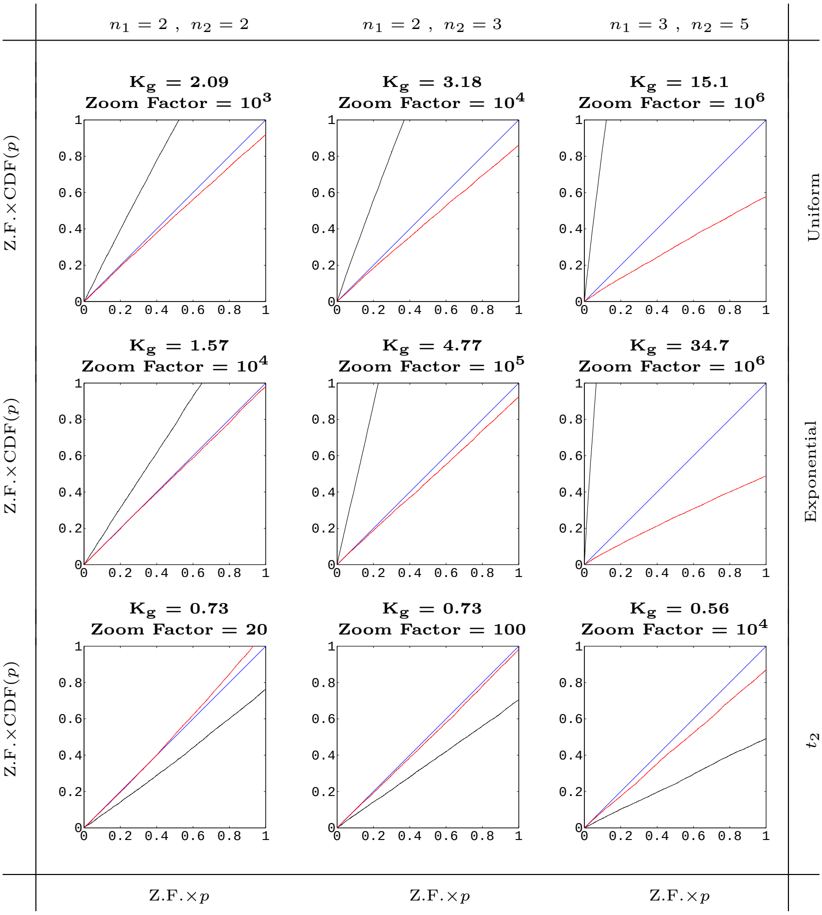


Figure 3.4: *The eCDF of the p-values for the Student two-sample t−test.*

The empirical CDFs of the raw and corrected p-values p^R and p^C are shown in black and red accordingly. The top, middle and bottom rows correspond to the *Uniform*$(-1, 1)$, *Exponential* and t_2 densities, and left, middle and right columns correspond to sample sizes $(n_1 = 2, n_2 = 2)$, $(n_1 = 2, n_2 = 3)$ and $(n_1 = 3, n_2 = 5)$. The blue diagonal line is the theoretical uniform distribution. The axes are scaled according to the Zoom Factor (Z.F.) parameter r in the title of the graphs.

$F-test$

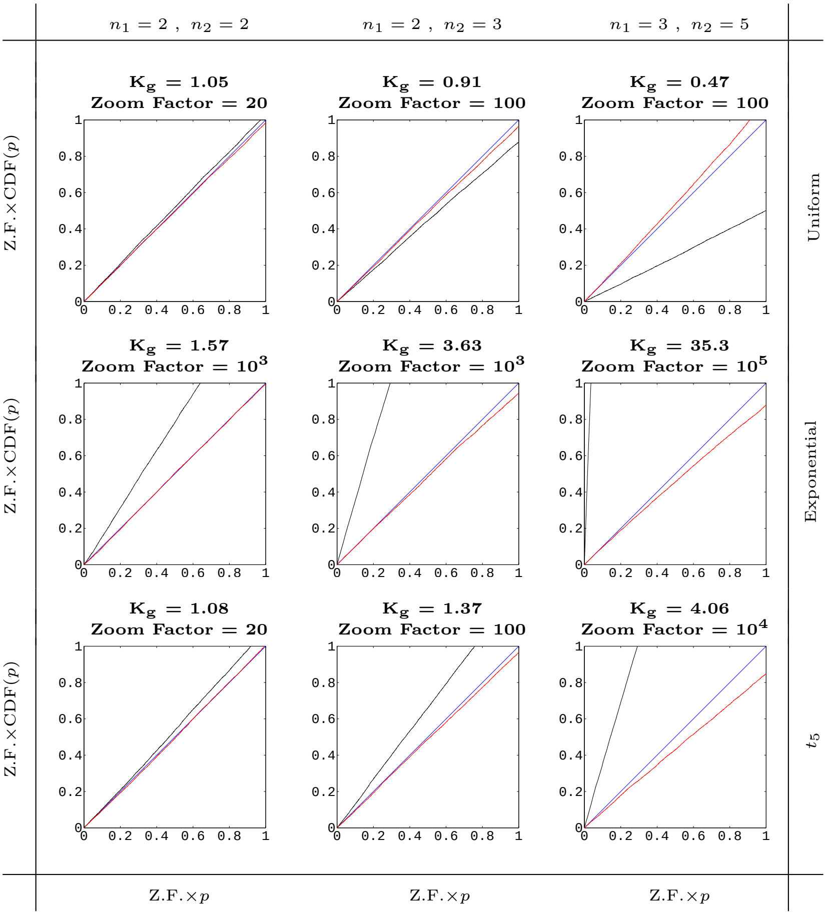


Figure 3.5: *The eCDF of the p-values for an $F-$test of the equality of variances.*

The empirical CDFs of the raw and corrected p-values p^R and p^C are shown in black and red accordingly. The top, middle and bottom rows correspond to the $Uniform(-1,1)$, $Exponential$ and t_5 densities, and left, middle and right columns correspond to sample sizes $(n_1 = 2, n_2 = 2)$, $(n_1 = 2, n_2 = 3)$ and $(n_1 = 3, n_2 = 5)$. The blue diagonal line is the theoretical uniform distribution. The axes are scaled according to the Zoom Factor (Z.F.) parameter r in the title of the graphs.

Welch's test

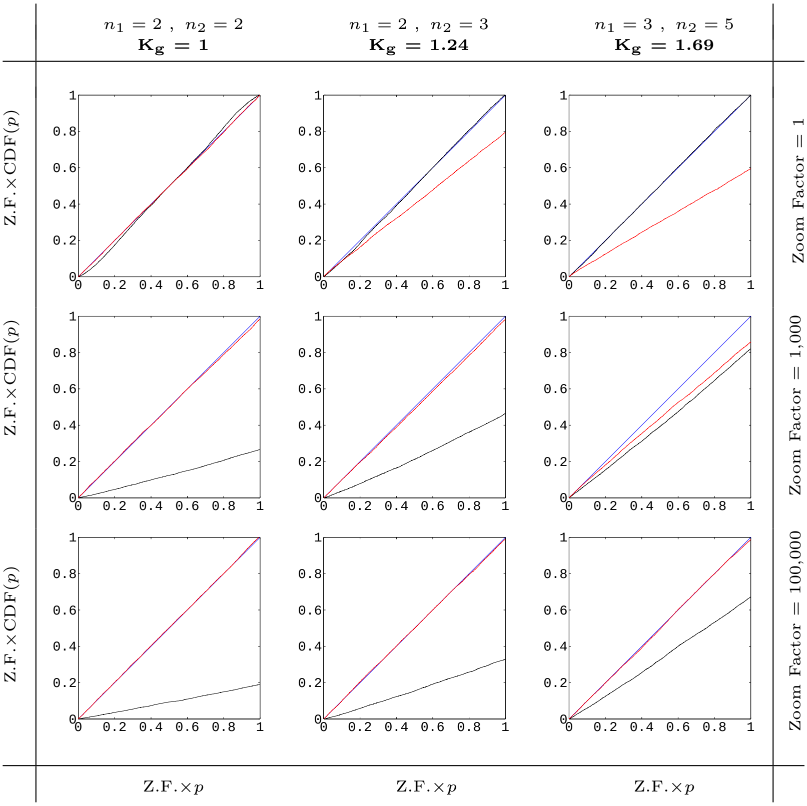


Figure 3.6: *The distribution tails of the p-values for the Welch test.*

The empirical CDFs of the raw (Welch-Satterthwaite) and corrected p-values p^R and p^{WS} for the *Standard normal* density are shown in black and red accordingly. The top, middle and bottom rows correspond to the different values of the Zoom Factor (Z.F.) parameter r shown on the right, and the axes are scaled accordingly. The left, middle and right columns correspond to sample sizes $(n_1 = 2, n_2 = 2)$, $(n_1 = 2, n_2 = 3)$ and $(n_1 = 3, n_2 = 5)$. The blue diagonal line is the theoretical uniform distribution.

Supplementary Materials

MATLAB scripts

[*OST*/*TST*/*WELCH*/*F*]+*ComputeKg.m* - compute K_g numerically for the Student one- and two- sample $t-$, Welch, and $F-$statistics using adaptive Simpson or Lobatto quadratures. Here g is an arbitrary multivariate density.[1]

[*OST*/*TST*/*WELCH*/*F*]+*ComputeKg*+[IS/IID]+.m - the same as above but for the case of independent samples and i.i.d. random variables accordingly.[2]

Wolfram Mathematica scripts

[*OST*/*TST*/*WELCH*/*F*]+*ComputeKg.nb* - compute the exact expression for K_g for an arbitrary multivariate density g and given sample size(s). The scripts include a number of instructive examples, including evaluation of K_g for the zero-mean Gaussian case with an arbitrary covariance matrix $\mathbf{\Sigma}$ (and the "unequal variances" case for the Student two-sample $t-$ and Welch tests) and for the densities considered in the simulation study of Section 7.

OSTComputeKgIID.nb - verifies the constants in Table 3.1 for the i.i.d. case of the Student one-sample $t-$statistic.

TSTExactPDF.nb and *WELCHExactPDF.nb* - the exact distribution for the Student two-sample $t-$ and Welch statistics for odd sample sizes, see Ray and Pitman (1961), and Figure 3.2 of Section 4 for the case $\sigma_1^2 \neq \sigma_2^2$.

The scripts are available at www.zholud.com

[1] For the $F-$test we use Monte Carlo integration.

[2] For the $F-$test and $n_1 > 3$ we use Monte Carlo integration.

Acknowledgements The author gives his deep thanks to Professor Holger Rootzén for constant assistance and fruitful discussions, and wishes to thank Professors Olle Nerman and Anders Blomberg for the collaboration on the BioScreen project and providing the data which became the motivation for the present article.

References

A.A. Aspin and B.L. Welch. Tables for use in comparisons whose accuracy involves two variances, separately estimated. *Biometrika*, 36(3):290–296, 1949. 74

M.S. Bartlett. The effects of non-normality on the t-distribution. *Proc. Cambridge Philos. Soc.*, 31:223–231, 1935. 73

V. Bentkus, B.-Y. Jing, Q.M. Shao, and W. Zhou. Limiting distributions of the non-central t-statistic and their applications to the power of t-tests under non-normality. *Bernoulli*, 13(2):346–364, 2007. 73

G.E.P. Box. Non-normality and tests on variances. *Biometrika*, 40:318–335, 1953. 74

G.E.P. Box. Some theorems on quadratic forms applied in the study of analysis of variance problems, i. effect of inequality of variance in the one-way classification. *The Annals of Mathematical Statistics*, 25(2):290–302, 1954. 74

R.A. Bradley. The distribution of the t and f statistics for a class of non-normal populations. *Virginia J. Sci.*, 3:1–32, 1952a. 70, 74

R.A. Bradley. Corrections for non-normality in the use of the two-sample t- and f-tests at high significance levels. *Ann. Math. Statist.*, 23:103–113, 1952b. 74

Noel Cressie. Relaxing assumptions in the one sample t-test. *Australian Journal of Statistics*, 22(2):143–153, 1980. 73

H.E. Daniels and G.A. Young. Saddlepoint approximation for the studentized mean, with an application to the bootstrap. *Biometrika*, 78:169–179, 1991. 73

F.N. David and N.L. Johnson. The effect of non-normality on the power function of the f-test in the analysis of variance. *Biometrika*, 38:43–57, 1951. 74

T. Eden and F. Yates. On the validity of fisher's z test when applied to an actual example of non-normal data. (with five text-figures.). *The Journal of Agricultural Science*, 23(01):6–17, 1933. 74

C. Field and E. Ronchetti. *Small Sample Asymptotics*. IMS, Hayward, CA, 1990. 73

R.A. Fisher. On a distribution yielding the error functions of several well known statistics. *Proceeding of the International Mathematical Congress, Toronto*, 2:805–813, 1924. 74

R.A. Fisher. Applications of "student's" distribution. *Metron*, 5: 90–104, 1925. 74

R.A. Fisher. The mathematical distributions used in the common tests of significance. *Econometrica*, 3(4):353–365, 1935a. 74

R.A. Fisher. The fiducial argument in statistical inference. *Annals of Eugenics*, 8:391398, 1935b. 74

A.K. Gaen. The distribution of 'student's' t in random samples of any size drawn from non-normal universes. *Biometrika*, 36: 353–369, 1949. 73

A.K. Gaen. The distribution of the variance ratio in random samples of any size drawn from non-normal universes. *Biometrika*, 37:236–255, 1950. 73, 74

R.C. Geary. The distribution of student's ratio for non-normal samples. *J. Roy. Statist. Soc., Suppl.*, 3:178–184, 1936. 73

E. Giné, F. Götze, and D.M. Mason. When is the student t - statistic asymptotically standard normal? *Ann. Probab.*, 25: 1514–1531, 1997. 73

P. Hall. Edgeworth expansion for students t statistic under minimal moment conditions. *Ann. Probab.*, 15:920–931, 1987. 73

S.I. Hayek. *Advanced Mathematical Methods in Science and Engineering.* Marcel Dekker, New York, NY, 2001. 101

H. Hotelling. The behavior of some standard statistical tests under non-standard conditions. *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, 1:319–360, 1961. 70, 75

J.L. Jensen. *Saddlepoint Approximations.* Oxford Univ. Press., 1995. 74

B.-Y. Jing, Q.M. Shao, and W. Zhou. Saddlepoint approximation for students t-statistic with no moment conditions. *Ann. Statist.*, 32(6):2679–2711, 2004. 73

S.-H. Kim and A.S. Cohen. On the behrens-fisher problem: A review. *Journal of Educational and Behavioral Statistics*, 23(4): 356–377, 1998. 74

J.E. Kolassa. *Series Approximation Methods in Statistics*, volume 88 of *Lecture Notes in Statistics.* Springer, New York., 3 edition, 2006. 74

Jack Laderman. The distribution of "student's" ratio for samples of two items drawn from non-normal universes. *The Annals of Mathematical Statistics*, 10(4):376–379, 1939. 74

B.F. Logan, C.L. Mallows, S.O. Rice, and L.A. Shepp. Limit distributions of self-normalized sums. *Ann. Probab.*, 1:788–809, 1973. 74

R. Lugannani and S. Rice. Saddlepoint approximation for the distribution of the sum of independent random variables. *Adv. in Appl. Probab.*, 12:475–490, 1980. 74

Mathematica. *Version 8.0.* Wolfram Research, Inc., Champaign, IL, 2010. 107

MATLAB. *Version 7.10.0 (R2010a).* The MathWorks, Inc., Natick, Massachusetts, 2010. 88, 92, 107

V. Perlo. On the distribution of the "student's" ratio for samples of three drawn from rectangular distribution. *Biometrika*, 25: 203–204, 1933. 74

A.D. Polyanin and Manzhirov. *Handbook of Integral Equations*. Chapman & Hall/CRC Press, Boca Raton, FL, second edition, 2008. 101

W.D. Ray and A.E.N.T. Pitman. An exact distribution of the fisher-behrens-welch statistic for testing the difference between the means of two normal populations with unknown variance. *Journal of the Royal Statistical Society. Series B (Methodological)*, 23(2):377–384, 1961. 74, 83, 84, 107

N. Reid. Saddlepoint methods and statistical inference (with discussion). *Statist. Sci.*, 3:213–238, 1988. 74

P.R. Rider. On the distribution of the ratio of mean to standard deviation in small samples from non-normal universes. *Biometrika*, 21(1):124–143, 1929. 74

H. Rootzen and D.S. Zholud. Tail estimation methods for the number of false positives in high-throughput testing. *Submitted*, 2011. 73

F.E. Satterthwaite. Synthesis of variance. *Psychometrika*, 6(5): 309–316, 1941. 74

F.E. Satterthwaite. An approximate distribution of estimates of variance components. *Biometrics Bulletin*, 2(6):110–114, 1946. 74

S.S. Sawilowsky. Fermat, schubert, einstein, and behrensfisher: The probable difference between two means when $\sigma_1^2 \neq \sigma_2^2$. *Journal of Modern Applied Statistical Methods*, 1(2):461–472, 2002. 74

Q.-M. Shao. Self-normalized large deviations. *Ann. Probab.*, 25: 285–328, 1997. 74

Q.-M. Shao. Recent progress on self-normalized limit theorems. *Probability, Finance and Insurance*, pages 50–68, 2004. 74

SmartTail, 2011. *url:* www.smarttail.se - software for the analysis of false discovery rates in high-throughput screening experiments.

J. Warringer and A. Blomberg. Automated screening in environmental arrays allows analysis of quantitative phenotypic profiles in saccharomyces cerevisiae. *Yeast*, 20:53–67, 2003. 71

J. Warringer, E. Ericson, L. Fernandez, O. Nerman, and A. Blomberg. High-resolution yeast phenomics resolves different physiological features in the saline response. *Proc Natl Acad Sci USA*, 100(26):15724–15729, 2003. 71

B.L. Welch. The significance of the difference between two means when the population variances are unequal. *Biometrika*, 29(3-4): 350–362, 1937. 74

B.L. Welch. The generalization of Student's problem when several different population variances are involved. *Biometrika*, 34(1-2): 28–35, 1947. 74

S.L. Zabell. On student's 1908 article. *Journal of the American Statistical Association*, 103:1–7, 2008. 73

D.S. Zholud. SmartTail - software for the analysis of False Discovery Rates in High-Throuput Screening experiments. *Work in progress*, 2011.

D.S. Zholud, H. Rootzèn, O. Nerman, and A. Blomberg. Positional effects in biological array experiments and their impact on False Discovery Rate. *Work in progress*, 2011. 71, 72

W. Zhou and B.-Y. Jing. Tail probability approximations for student's t-statistics. *Probability Theory and Related Fields*, 136(4): 541–559, 2006. 73

PAPER III

PAPER III

Extremes
2009, Vol. 12, No. 1, pp. 1-17.

Extremes of the Shepp statistic for a Gaussian random walk

Dmitrii Zholud [*]

Abstract

Let $(\xi_i, i \geq 1)$ be a sequence of independent standard normal random variables and let $S_k = \sum_{i=1}^{k} \xi_i$ be the corresponding random walk. We study the renormalized Shepp statistic $M_T^{(N)} = \frac{1}{\sqrt{N}} \max_{1 \leq k \leq TN} \max_{1 \leq L \leq N} (S_{k+L-1} - S_{k-1})$ and determine asymptotic expressions for the probability $\mathbf{P}\left(M_T^{(N)} > u\right)$ when u, N and $T \to \infty$ in a synchronized way. There are three types of relations between u and N that give different asymptotic behavior. For these three cases we establish the limiting Gumbel distribution of $M_T^{(N)}$ when $T, N \to \infty$ and present corresponding normalization sequences.

Keywords Gaussian random walk, increments, maximum, extreme values, high excursions, large deviations, moderate deviations, asymptotic behavior, Shepp statistic, distribution tail, Gumbel law, limit theorems, weak theorems, Wiener Process.

AMS 2000 Subject Classifications: $\frac{\text{Primary-60G70;}}{\text{Secondary-62P10, 60F10;}}$.

[*] *Department of Mathematical Statistics*
Chalmers University of Technology and University of Göteborg, Sweden.
E-mail: dmitrii@zholud.com

1 Introduction

Let $(\xi_i, i \geq 1)$ be a sequence of independent standard normal random variables, and let $S_k = \sum\limits_{i=1}^{k} \xi_i$, with $S_0 = 0$, be the corresponding random walk. Introduce the Shepp and the closely related Erd$\acute{o}$s-R$\acute{e}$nyi statistics

$$W_{N,L} = \max_{1 \leq l \leq L} T_{N,l} \;\text{ and }\; T_{N,L} = \max_{1 \leq k \leq N} S_{k+L-1} - S_{k-1},$$

and define

$$\zeta_L^{(N)}(k) = \frac{1}{\sqrt{N}}\left(S_{k+L-1} - S_{k-1}\right) = \frac{1}{\sqrt{N}} \sum_{i=k}^{k+L-1} \xi_i.$$

We study the asymptotic behavior of the probability

$$\mathbf{P}\left(\max_{\substack{0<k\leq TN\\ 0<L\leq N}} \zeta_L^{(N)}(k) > u\right) \tag{1.1}$$

when $u \to \infty$, $N \to \infty$ in a coordinated way. In fact, (1.1) is the probability of exceeding the level $u\sqrt{N}$ by the Shepp statistic $W_{TN,N}$. Related problems were described in Erdös and Rényi (1970), Piterbarg (1991), Kozlov (2004) and Piterbarg and Kozlov (2002). Paper Piterbarg (1991) presents the asymptotic behavior of the probability of moderate deviations for the Erd$\ddot{o}$s-R$\acute{e}$nyi statistic under the assumption of sub-gaussian distribution of random walk step and papers Kozlov (2004) and Piterbarg and Kozlov (2002) study large deviations of the Erd$\ddot{o}$s-R$\acute{e}$nyi and Shepp statistics for Cramér random walk. To get the full picture of all possible cases of asymptotic behavior of (1.1) we reformulate the result obtained by A.M. Kozlov in Kozlov (2004). Let $\psi(u) = \frac{1}{\sqrt{2\pi}} \int\limits_{u}^{\infty} e^{-x^2/2} dx$ be the tail of standard normal distribution and introduce the finite positive constant

$$J_\theta = \lim_{l \to \infty} \frac{1}{\theta l} \mathbf{E} exp\left\{\theta \max_{0 \leq n < l} (\sqrt{2} S_n - \theta n)\right\}.$$

Theorem 1.1 (A.M. Kozlov)**.** *Assume* $u \to \infty$, $N \to \infty$, $\frac{u}{\sqrt{N}} \to \theta$, *where* $0 < \theta < \infty$*. If* $Tu^2\psi(u) \to 0$ *and* $Tu^2 \to \infty$*, then*

$$\boldsymbol{P}\left(\max_{\substack{0<k\leq TN\\ 0<L\leq N}} \zeta_L^{(N)}(k) > u\right) \sim J_\theta T u^2 \psi(u).$$

The present paper extends this result to moderate and excessively large deviations. For comparison and ease of reference we also restate the main result of Zholud (2008) which deals with the continuous time case and is crucial in proving the asymptotic formula for the case of moderate deviations. Let $W(\cdot)$ be the standard Brownian motion.

Theorem 1.2 (D. Zholud)**.** *Assume $u \to \infty$. If $Tu^2 \to \infty$ and $Tu^2\psi(u) \to 0$, then*

$$\boldsymbol{P}\left(\max_{\substack{0\le t\le T\\0\le s\le 1}} (W(t+s)-W(t)) > u\right) = HTu^2\psi(u)(1+o(1)),$$

where the constant

$$H = \lim_{B\to\infty}\lim_{A\to\infty} A^{-1}e^{-\frac{A+B}{2}}\mathbf{E}\exp\left(\max_{\substack{0\le t\le A\\0\le s\le B}}(W(t+s+A)-W(t))\right)$$

is finite and positive.

The case of moderate deviations (i.e. $\frac{u}{\sqrt{N}} \to 0$ when $u \to \infty$) is intermediate between Theorem 1.1 and Theorem 1.2.

Theorem 1.3. *Assume $u \to \infty$, $N \to \infty$, $\frac{u}{\sqrt{N}} \to 0$. If $Tu^2 \to \infty$ and $Tu^2\psi(u) \to 0$, then*

$$\boldsymbol{P}\left(\max_{\substack{0<k\le TN\\0<L\le N}} \zeta_L^{(N)}(k) > u\right) \sim HTu^2\psi(u).$$

Indeed, this asymptotic behavior is different from the one in Theorem 1.1 in the constant multiplier and coincides with the asymptotic behavior for the case of continuous time, Theorem 1.2. The proof of Theorem 1.3 can be found in Section 2.

A further comment is that if $N \to \infty$ and u is fixed, then we could apply weak convergence of a random walk to the Wiener process, and the probabilities in Theorem 1.2 and Theorem 1.3 would coincide. However Section 3 shows that just applying weak convergence under the probability sign leads to incorrect results, and that the rigorous and somewhat technical proof of Theorem 1.3 is indeed needed. The main result of this section is as follows.

Theorem 1.4. *Assume $u \to \infty$, $N \to \infty$, $\frac{u}{\sqrt{N}} \to \infty$. If $TN \geq 1$ and $TN\psi(u) \to 0$, then*

$$\boldsymbol{P}\left(\max_{\substack{0<k\leq TN\\ 0<L\leq N}} \zeta_L^{(N)}(k) > u\right) \sim [TN]\psi(u).$$

This theorem completes full description of the possible asymptotic behavior of (1.1) under various relations between u and N.

Finally, using the results of Sections 2 and 3 we obtain limit Gumbel distribution for $M_T^{(N)}$ when $T, N \to \infty$. If one of the following relations holds,

$$1)\frac{2\ln T}{N} \to 0. \quad 2)\frac{2\ln T}{N} \to \theta^2 > 0. \quad 3)\frac{2\ln T}{N} \to \infty,$$

then, there exist functions a_T and b_T such that for any fixed x

$$\mathbf{P}\left(\max_{\substack{0<k\leq TN\\ 0<L\leq N}} a_T(\zeta_L^{(N)}(k) - b_T) \leq x\right) = e^{-e^{-x}} + o(1).$$

The corresponding theorems and normalizing constants can be found in Section 4. A similar result for standardized increments of Gaussian random walk is obtained in Kabluchko (2007).

There is also extensive literature on a.s. convergence of related quantities, see e.g. Shepp (1964), Erdös and Rényi (1970) and Frolov (2004).

2 Moderate deviations of the Shepp statistic

In this section we prove Theorem 1.3. That is we find the asymptotic behavior of the probability (1.1) when $u \to \infty$ and $u/\sqrt{N} \to 0$. It will be shown that it coincides with the asymptotic behavior for continuous time case, given in Theorem 1.2. The idea of the proof is similar to Zholud (2008) and we divide it into two main parts.

First, for any positive constant B we will focus on the asymptotic behavior of maximum of $\zeta_L^{(N)}(k)$ over a thin layer

$$\{(k, L) : 0 < k \leq TN, (1 - Bu^{-2})N < L \leq N\}.$$

Within this area and for large u, $\zeta_L^{(N)}(k)$ behaves approximately like $\zeta_N^{(N)}(k)$, and it will be shown that it is possible to determine the asymptotic behavior using similar techniques as used for stationary process in Piterbarg (1991).

Second, knowing the asymptotic behavior of maximum of the random variable $\zeta_L^{(N)}(k)$ over the area of its maximum variance, we will show that the maximum over the complementary set $\{(k,L) : 0 < k \le TN, 0 < L \le (1 - Bu^{-2})N\}$ gives a neglible contribution to the probability in (1.1).

The proof of the first part is based on the Double Sum Method. The lemma below is the analog of Lemma 2.1 in Zholud (2008). Let A and B be positive constants and set $p = Au^{-2}$, $q = Bu^{-2}$. By $A_0(u)$ we refer to the set of pairs $(k,L) \in [0,pN] \times ((1-q)N, N]$, where k and L are positive integers.

Lemma 2.1. *Let $u \to \infty$. Then*

$$\boldsymbol{P}\left(\max_{A_0(u)} \zeta_L^{(N)}(k) > u\right) = H_A^B \frac{1}{\sqrt{2\pi}u} e^{-\frac{u^2}{2}}(1+o(1)), \qquad (2.1)$$

where

$$H_A^B = e^{-\frac{A+B}{2}} \mathbf{E} \exp\left(\max_{\substack{0\le t\le A\\ 0\le s\le B}} W(t+s+A) - W(t)\right).$$

Proof. Let $[x]$ denote the integer part of x. We have

$$\max_{A_0(u)} \zeta_L^{(N)}(k) = \max_{A_0(u)} \tfrac{1}{\sqrt{N}} \sum_{i=k}^{[k+L-1]} \xi_i$$

$$= \tfrac{1}{\sqrt{N}} \sum_{i=[pN]+1}^{[(1-q)N]} \xi_i + \max_{A_0(u)} \tfrac{1}{\sqrt{N}} \left(\sum_{i=k}^{[pN]} \xi_i + \sum_{i=[(1-q)N]+1}^{k+L-1} \xi_i \right).$$

The $L + [pN] - [(1-q)N]$ random variables in the sums inside the "max" sign are independent of the variables in the sum outside the "max" sign. We renumber the variables inside the maximum sign and denote them by $\xi_i^{'}$. Thus,

$$\mathbf{P}\left(\max_{A_0(u)} \zeta_L^{(N)}(k) > u\right)$$

$$= \mathbf{P}\left(\frac{1}{\sqrt{N}}\sum_{i=[pN]+1}^{[(1-q)N]}\xi_i + \max_{\substack{0<k\leq pN\\ 0<L\leq qN}}\frac{1}{\sqrt{N}}\sum_{i=k}^{k+L+[pN]-1}\xi_i' > u\right)$$

$$= \int_{-\infty}^{\infty}\frac{e^{-\frac{v^2}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}}\mathbf{P}\left(\max_{\substack{0<k\leq pN\\ 0<L\leq qN}}\frac{1}{\sqrt{N}}(S'_{k+L+[pN]-1} - S'_{k-1}) > u - v\right)dv,$$

where $\sigma^2 = \frac{[(1-q)N]-[pN]}{N}$ and $S'_k = \sum_{i=1}^{k}\xi_i'$ with $S'_0 = 0$.

For the sake of briefness introduce

$$M(k,L) = \max_{\substack{0<k\leq pN\\ 0<L\leq qN}}\frac{1}{\sqrt{pN}}(S'_{k+L+[pN]-1} - S'_{k-1}).$$

Using the change of variables $v = u - \frac{\sqrt{A}w}{u}$, and recalling that $u\sqrt{p} = \sqrt{A}$, the probability in question is seen to equal to

$$\frac{\sqrt{A}}{\sqrt{2\pi\sigma^2}u}\int_{-\infty}^{\infty} e^{-\frac{(u-\sqrt{A}w/u)^2}{2\sigma^2}}\mathbf{P}\left(M(k,L) > w\right)ds$$

$$= \frac{\sqrt{A}}{\sqrt{2\pi\sigma^2}u}e^{-\frac{u^2}{2\sigma^2}}\int_{-\infty}^{\infty} e^{-\frac{Aw^2/u^2}{2\sigma^2}}e^{\frac{\sqrt{A}w}{\sigma^2}}\mathbf{P}\left(M(k,L) > w\right)dw. \quad (2.2)$$

By weak convergence of a random walk to the Wiener process, for any w,

$$\lim_{pN\to\infty}\mathbf{P}\left(M(k,L) > w\right) = \mathbf{P}\left(\max_{\substack{0\leq t\leq 1\\ 0\leq s\leq B/A}} W(t+s+1) - W(t) > w\right),$$

where $W(\cdot)$ is the standard Wiener process; using Lemma 1 Piterbarg (1991) it is straightforward to show that

$$\mathbf{P}\left(M(k,L) > w\right) \leq 2e^{-\frac{w^2}{24}}.$$

Thus, by dominated convergence

$$\int_{-\infty}^{\infty} e^{-\frac{Aw^2/u^2}{2\sigma^2}}e^{\frac{\sqrt{A}w}{\sigma^2}}\mathbf{P}\left(M(k,L) > w\right)dw$$

$$= \int_{-\infty}^{\infty} e^{\sqrt{A}w} \mathbf{P}\left(\max_{\substack{0\le t\le 1\\ 0\le s\le B/A}} (W(t+s+1)-W(t)) > w \right) dw + o(1)$$

$$= \tfrac{1}{\sqrt{A}} \mathbf{E} \exp\left(\max_{\substack{0\le t\le A\\ 0\le s\le B}} W(t+s+A) - W(t) \right) + o(1).$$

Finally, since $\sigma^2 = 1-p-q+o(u^{-2})$ the factor in front of the integral (2.2) is equal to

$$\frac{\sqrt{A}}{\sqrt{2\pi}u} e^{-\frac{u^2}{2}(1+p+q+o(u^{-2}))}(1+o(1)) = \frac{1}{\sqrt{2\pi}u} e^{-\frac{u^2}{2}} \sqrt{A} e^{-\frac{A+B}{2}}(1+o(1))$$

and we obtain (2.1). □

Our next aim is to consider the layer $[0,TN]\times((1-q)N,N]$. We use Lemma 2.1 and the Bonferroni inequality to obtain estimates of the probability of high level excursions of the maximum of $\zeta_L^{(N)}(k)$. Then we will show that estimates from below and from above are asymptotically equivalent.

Define $\Delta_k(u) = \{kpN+1,...,(k+1)pN\}\times\{(1-q)N+1,...,N\}$. For ease of notation we suppress dependence on u and assume that pN and qN are integers. Using stationarity of $\zeta_L^{(N)}(k)$ with respect to k, we obtain that

$$(Tp^{-1}+1)\mathbf{P}\left(\max_{\Delta_0} \zeta_L^{(N)}(k) > u\right) \ge \mathbf{P}\left(\max_{\substack{0<k\le TN\\ (1-q)N<L\le N}} \zeta_L^{(N)}(k) > u\right)$$

$$\ge (Tp^{-1}-1)\mathbf{P}\left(\max_{\Delta_0} \zeta_L^{(N)}(k) > u\right)$$

$$- \sum_{\substack{0\le l,m\le Tp^{-1}+1\\ l\ne m}} \mathbf{P}\left(\max_{\Delta_l} \zeta_L^{(N)}(k) > u, \max_{\Delta_m} \zeta_L^{(N)}(k) > u\right).$$

Let $p_{l,m}$ denote the summands in the last sum. This sum, due to stationarity, does not exceed

$$2(Tp^{-1}+1)\sum_{n=1}^{Tp^{-1}+1} p_{0,n}.$$

Estimating the probabilities $p_{0,n}$ from above we will show that the double sum is negligible, and thus the upper and lower estimates in the Bonferroni inequality are asymptotically equivalent. The estimates of $p_{0,n}$ are obtained in the same manner as in Piterbarg (1991). The proof will be divided into four parts.

Case 1.1: $1 \le n < n_0$. The value of n_0 will be chosen later. We have:

$$p_{0,n} \le \mathbf{P}\left(\max_{\substack{0<k\le pN(n+1)/2 \\ (1-q)N<L\le N}} \zeta_L^{(N)}(k) > u, \max_{\substack{pN(n+1)/2<k\le pN(n+1) \\ (1-q)N<L\le N}} \zeta_L^{(N)}(k) > u \right)$$

$$= 2\mathbf{P}\left(\max_{\substack{0<k\le pN(n+1)/2 \\ (1-q)N<L\le N}} \zeta_L^{(N)}(k) > u \right) - \mathbf{P}\left(\max_{\substack{0<k\le pN(n+1) \\ (1-q)N<L\le N}} \zeta_L^{(N)}(k) > u \right).$$

Applying Lemma 2.1 we obtain that

$$p_{0,n} \le \frac{1}{\sqrt{2\pi} u} (2H^B_{A(n+1)/2} - H^B_{A(n+1)}) e^{-\frac{u^2}{2}} (1 + g_n(u, N)), \quad (2.3)$$

where $g_n(u, N) \to 0$.

Case 1.2: $n_0 \le n \le \varepsilon p^{-1} - 1$. The value of ε will be chosen later. First, introduce random variables

$$\eta = \tfrac{1}{\sqrt{N}} \sum_{i=(n+1)pN+1}^{(1-q)N} \xi_i, \quad \zeta_1 = \tfrac{1}{\sqrt{N}} \sum_{i=pN+1}^{npN} \xi_i, \quad \zeta_2 = \tfrac{1}{\sqrt{N}} \sum_{i=pN+N+1}^{npN+(1-q)N} \xi_i.$$

Then, postponing the explanation of the last equality,

$$p_{0,n} = \mathbf{P}\Bigg(\eta + \zeta_1 + \max_{\Delta_0} \tfrac{1}{\sqrt{N}} \left(\sum_{i=k}^{pN} + \sum_{i=npN+1}^{(n+1)pN} + \sum_{i=(1-p)N+1}^{k+L-1} \right) \xi_i > u,$$

$$\eta + \zeta_2 + \max_{\Delta_n} \tfrac{1}{\sqrt{N}} \left(\sum_{i=k}^{(n+1)pN} + \sum_{i=(1-q)N+1}^{pN+N} + \sum_{i=npN+(1-q)N+1}^{k+L-1} \right) \xi_i > u \Bigg)$$

$$= \mathbf{P}\left(\eta + \zeta_1 + \max_{\Delta_0} \zeta'_L(k) > u, \ \ \eta + \zeta_2 + \max_{\Delta_0} \zeta''_L(k) > u \right),$$

where

$$\zeta_L^{'}(k) = \frac{1}{\sqrt{N}} \left(\sum_{i=k}^{k+L-(1-q-2p)N-1} \xi_i^{'} \right)$$

and

$$\zeta_L^{''}(k) = \frac{1}{\sqrt{N}} \left(\sum_{i=k}^{k+L-(1-2q-2p)N-1} \xi_i^{''} \right). \tag{2.4}$$

The main idea of this representation is that we consider $\zeta_L^{(N)}(k)$ for all possible pairs $(k, L) \in \Delta_0$ and "extract" the common summand $\eta + \zeta_1$. Analogously, for each $(k, L) \in \Delta_n$ we "extract" the summand $\eta + \zeta_2$. These summands are always present in $\zeta_L^{(N)}(k)$ when k, L are within the corresponding sets. It is easy to check that for $\varepsilon < 1/2$ and u large, the restriction on n ensures that the variables η, ζ_1, ζ_2 are independent. By construction they are also independent of the variables that remain inside the maximum signs. The latter are renumbered and called ξ_i' and ξ_i'' in such a way that (2.4) holds. Although ξ_i' and ξ_j'' may denote the same r.v. ξ_r, in our case the dependence between $\zeta_L'(k)$ and $\zeta_L''(k)$ does not matter. What is essential is that η, ζ_1, ζ_2, are independent of $\zeta_L'(k)$ and of $\zeta_L''(k)$. For the sake of brevity we omit the arguments for $\zeta_L'(k)$ and $\zeta_L''(k)$, as well as the set over which the maximum is taken.

From (2.4) it follows that

$$p_{0,n} \le \mathbf{P}\left(2\eta + \zeta_1 + \zeta_2 + \max\zeta' + \max\zeta'' > 2u\right)$$

$$= \tfrac{1}{\sqrt{2\pi\sigma^2}} \int\limits_{-\infty}^{\infty} \mathbf{P}\left(\tfrac{\zeta_1 + \ \zeta_2}{2} + \tfrac{\max\zeta' + \ \max\zeta''}{2} > u - v \right) e^{-\frac{v^2}{2\sigma^2}} dv,$$

where σ^2 now is equal to $\frac{[(1-q)N]-[(n+1)pN]}{N}$. After the change of variables $v = u - \sqrt{p}s$ we get that

$$p_{0,n} \le \tfrac{\sqrt{A}}{\sqrt{2\pi\sigma^2}u} e^{-\frac{u^2}{2\sigma^2}} \int\limits_{-\infty}^{\infty} \mathbf{P}\left(\tfrac{\zeta_1 + \ \zeta_2}{2\sqrt{p}} + \tfrac{\max\zeta' + \ \max\zeta''}{2\sqrt{p}} > s \right) e^{\frac{\sqrt{A}s}{\sigma^2}} ds$$

$$= \tfrac{\sigma}{\sqrt{2\pi}u} e^{-\frac{u^2}{2\sigma^2}} \mathbf{E} e^{\frac{\sqrt{A}}{\sigma^2}\left(\frac{\zeta_1 + \ \zeta_2}{2\sqrt{p}} + \frac{\max\zeta' + \ \max\zeta''}{2\sqrt{p}} \right)}$$

$$= \frac{\sigma}{\sqrt{2\pi u}} e^{-\frac{u^2}{2\sigma^2}} \mathbf{E} e^{\frac{\sqrt{A}}{\sigma^2}\left(\frac{\zeta_1 + \ \zeta_2}{2\sqrt{p}}\right)} \mathbf{E} e^{\frac{\sqrt{A}}{\sigma^2}\left(\frac{\max\zeta' + \ \max\zeta''}{2\sqrt{p}}\right)}. \qquad (2.5)$$

We now estimate the three factors that form the bound for $p_{0,n}$. Since $\sigma^2 = 1 - q - (n+1)p + o(u^{-2})$, for sufficiently large u the factor in front of the expectation is bounded by

$$\frac{\sigma}{\sqrt{2\pi u}} e^{-\frac{u^2}{2\sigma^2}} \leq \frac{2}{\sqrt{2\pi u}} e^{-\frac{u^2}{2}} e^{-\frac{A(n+1)+B}{2}}.$$

Next, since random variable $\frac{\zeta_1+\zeta_2}{2\sqrt{p}}$ is normally distributed, has zero mean and its variance is less than $(n-1)/2$ we have

$$\mathbf{E} e^{\frac{\sqrt{A}}{\sigma^2}\left(\frac{\zeta_1 + \ \zeta_2}{2\sqrt{p}}\right)} \leq e^{\frac{A(n-1)}{4\sigma^4}}.$$

In order to estimate the remaining expectation we will require an estimate of the probability

$$\mathbf{P}\left(\frac{\max\zeta' + \max\zeta''}{2\sqrt{p}} > s\right), \quad s > 0.$$

According to notation in (2.4) and denoting $S_k^{''} = \sum\limits_{i=1}^{k} \xi_i^{''}$ we see that the latter equals

$$\begin{aligned}
&\mathbf{P}\left(\max_{\Delta_0}\left(S^{'}_{k+L-(1-q-2p)N-1} - S^{'}_{k-1}\right) + \right. \\
&\qquad\qquad \left. + \max_{\Delta_0}\left(S^{''}_{k+L-(1-2q-2p)N-1} - S^{''}_{k-1}\right) > 2\sqrt{pN}s\right) \\
&\quad \leq \mathbf{P}\left(\max_{\Delta_0} S^{'}_{k+L+(q+2p)N-N-1} + \max_{0<k\leq pN} -S^{'}_{k-1} + \right. \\
&\qquad\qquad \left. + \max_{\Delta_0} S^{''}_{k+L+(2q+2p)N-N-1} + \max_{0<k\leq pN} -S^{''}_{k-1} > 2\sqrt{pN}s\right) \\
&\quad \leq 4\mathbf{P}\left(\max_{0<k\leq(2q+3p)N} S^{'}_k > \frac{\sqrt{pN}}{2}s\right) \leq 4e^{-\frac{1}{8}\left(\frac{A}{3A+2B}s^2\right)} < 4e^{-\frac{s^2}{24}},
\end{aligned} \qquad (2.6)$$

where we applied Lemma 1 Piterbarg (1991) in the second to the last step.

Thus, for any positive t we obtain the following estimate

$$\mathbf{E}e^{t\left(\frac{\max\zeta'+\ \max\zeta''}{2\sqrt{p}}\right)} = \int\limits_{-\infty}^{\infty} te^{ts}\mathbf{P}\left(\frac{\max\zeta'+\max\zeta''}{2\sqrt{p}} > s\right) ds$$
$$\leq 1+4t\int\limits_{0}^{\infty} e^{ts-\frac{s^2}{24}}\, ds \leq 1+4\sqrt{24\pi}te^{6t^2}. \qquad (2.7)$$

Then we put $t = \frac{\sqrt{A}}{\sigma^2}$ and when A is large, the estimate (2.7) gives

$$\mathbf{E}e^{\frac{\sqrt{A}}{\sigma^2}\left(\frac{\max\zeta'+\ \max\zeta''}{2\sqrt{p}}\right)} < \frac{8\sqrt{24\pi}}{\sigma^2}\sqrt{A}e^{\frac{6A}{\sigma^4}}.$$

We are now ready to estimate $p_{0,n}$. Gathering the estimates of the factors in (2.5) we get

$$p_{0,n} \leq \frac{\frac{16\sqrt{24\pi}}{\sigma^2}\sqrt{A}}{\sqrt{2\pi}u}e^{-\frac{u^2}{2}}e^{-An\left(\frac{1}{2}-\frac{1}{4\sigma^4}\right)+A\left(\frac{23}{4\sigma^4}-\frac{1}{2}\right)-\frac{B}{2}}.$$

Owing to the restriction $n_0 \leq n \leq \varepsilon p^{-1} - 1$ we have

$$\sigma^2 = 1 - q - (n+1)p + o(u^{-2}) > 1 - 2\varepsilon.$$

Choosing ε such that $4(1-2\varepsilon)^2 = 3$ we conclude that

$$p_{0,n} \leq \frac{C_1\sqrt{A}}{\sqrt{2\pi}u}e^{-\frac{u^2}{2}}e^{-A\frac{n-43}{6}-\frac{B}{2}}, \qquad (2.8)$$

where C_1 is some positive constant.

Case 1.3: $\varepsilon p^{-1} \leq n \leq p^{-1} + 1$. In much the same way the representation (2.4) gives

$$p_{0,n} \leq \mathbf{P}\left(2\eta + \zeta_1 + \zeta_2 + \max_{\Delta_0}\zeta_L'(k) + \max_{\Delta_0}\zeta_L''(k) > 2u\right).$$

However, when $n \geq \varepsilon p^{-1}$, it may turn out that the sum in the expression for η is empty. In this case we set $\eta = 0$. We should also change the upper limit of summation in the definition of ζ_1 to $\min\{npN, (1-p)N\}$, and the lower limit of summation for ζ_2 to $\max\{2pN + (1-p)N + 1, (n+1)pN + 1\}$. Therefore, ζ' and ζ'', may consist of a smaller number of summands.

For any positive t, multiplying both parts of the inequality under the probability sign by t/2 and applying Chebyshev's inequality to the exponents, we obtain that

$$\begin{aligned} p_{0,n} &\leq e^{-tu}\mathbf{E}e^{t\left(\eta+\frac{\zeta_1+\zeta_2}{2}+\frac{\max\zeta'+\max\zeta''}{2}\right)} \\ &= e^{-tu}\mathbf{E}e^{t\left(\eta+\frac{\zeta_1+\zeta_2}{2}\right)}\mathbf{E}e^{t\left(\frac{\max\zeta'+\max\zeta''}{2}\right)}. \qquad (2.9)\end{aligned}$$

Although ζ' and ζ'' may contain smaller number of summands, it can be seen that this does not change the proof of (2.6) sufficiently. Thus the estimate (2.7) remains valid and

$$\mathbf{E}e^{t\left(\frac{\max\zeta'+\max\zeta'}{2}\right)} < 1+4\sqrt{24\pi}t\sqrt{p}e^{6t^2p}. \tag{2.10}$$

Next, according to the remark about limits of summation in ζ_1 and ζ_2, the variance of $\frac{\zeta_1+\zeta_2}{2}$ does not exceed $\frac{(n-1)p}{2}$. The variance of η does not exceed $\max\{0, 1-(n-1)p\}$. Applying Laplace transformation to the sum $\eta+\frac{\zeta_1+\zeta_2}{2}$, and since restrictions on n provide $(\varepsilon-p)/2 \le (n-1)p/2 \le 1/2$,

$$\mathbf{E}e^{t\left(\eta+\frac{\zeta_1+\zeta_2}{2}\right)} \le e^{\frac{t^2\max\left\{\frac{(n-1)p}{2},1-\frac{(n-1)p}{2}\right\}}{2}} < e^{\frac{t^2(1-\varepsilon/4)}{2}}. \tag{2.11}$$

So, gathering (2.11), (2.10) and (2.9),

$$p_{0,n} \le (1+4\sqrt{24\pi}t\sqrt{p}e^{6t^2p})e^{\frac{t^2(1-\varepsilon/4)}{2}}e^{-tu}.$$

Setting $t=\frac{u}{1-\varepsilon/4}$, we obtain the desired estimate:

$$p_{0,n} \le C_2\sqrt{A}e^{6A}e^{-\frac{u^2}{2(1-\frac{\varepsilon}{4})}}. \tag{2.12}$$

Case 1.4: $n > p^{-1}+1$. In this case the two events inside the probability $p_{0,n}$ are independent and Lemma 2.1 gives

$$p_{0,n} \le 2(H_A^B)^2\psi(u)^2. \tag{2.13}$$

The bounds obtained in cases 1.1-1.4 allow us to estimate $p_{0,n}$ for any value of n. Let $p_0(u)=\frac{1}{\sqrt{2\pi u}}e^{-\frac{1}{2}u^2}$. Estimates (2.3), (2.8), (2.12), (2.13) imply that

$$\begin{aligned}2(Tp^{-1}+1)\sum_{n=1}^{Tp^{-1}+1} p_{0,n} \le 2(Tp^{-1}+1)\times\\ \times\left\{\left(\sum_{n=1}^{n_0-1}\left(2H^B_{A(n+1)/2}-H^B_{A(n+1)}\right)(1+g_n(u,N))+\right.\right.\\ \left.+\sum_{n=n_0}^{\infty}C_1\sqrt{A}e^{-A\frac{n-43}{6}-\frac{B}{2}}\right)p_0(u)+\end{aligned}$$

$$+ p^{-1}C_2\sqrt{A}e^{6A}e^{-\frac{u^2}{2(1-\frac{\varepsilon}{4})}} + Tp^{-1}2(H_A^B)^2\psi(u)^2\Bigg\}.$$

Recall that $p^{-1} = u^2/A$. If $Tu^2 \to \infty$ and $Tu^2\psi(u) \to 0$, then using the estimate above and the Bonferroni inequality on page 121 we conclude that

$$\overline{\lim_{u,N}}\, \mathbf{P}\left(\max_{\substack{0<k\leq TN\\(1-q)N<L\leq N}} \zeta_L^{(N)}(k) > u\right)\Bigg/ Tu^2p_0(u) \leq A^{-1}H_A^B$$

and (2.14)

$$\underline{\lim_{u,N}}\, \mathbf{P}\left(\max_{\substack{0<k\leq TN\\(1-q)N<L\leq N}} \zeta_L^{(N)}(k) > u\right)\Bigg/ Tu^2p_0(u) \geq A^{-1}H_A^B - \\ -2A^{-1}\sum_{n=1}^{n_0-1}\left(2H^B_{A(n+1)/2} - H^B_{A(n+1)}\right) - 2\frac{C_1e^{-\frac{B}{2}}}{\sqrt{A}}\sum_{n=n_0}^{\infty} e^{-A\frac{n-43}{6}}.$$

It was proved in Zholud (2008) that the limit

$$H^B = \lim_{A\to\infty} A^{-1}H_A^B, \quad H^B > 0$$

exists. Thus $A^{-1}\left(2H^B_{A(n+1)/2} - H^B_{A(n+1)}\right) \to 0$, when $A \to \infty$. Choosing n_0 greater than 43 and letting A in (2.14) tend to infinity we obtain the asymptotic behavior of the probability of high level excursions for maximum of $\zeta_L^{(N)}(k)$ over the "upper" layer,

$$\mathbf{P}\left(\max_{\substack{0<k\leq TN\\(1-q)N<L\leq N}} \zeta_L^{(N)}(k) > u\right) = H^BTu^2p_0(u)(1+o(1)). \quad (2.15)$$

The second part of the proof is to show that the asymptotic behavior of the probability (1.1) is determined by the behavior of $\zeta_L^{(N)}(k)$ over the upper layer, which corresponds to the area of maximal variance of the field. Thus we need to estimate the probability of the high level excursion of the maximum of the random walk over the complementary set. Applying stationarity of $\zeta_L^{(N)}(k)$ with respect to k we obtain the following estimate

$$\mathbf{P}\left(\max_{\substack{0<k\le TN\\ 0<L\le(1-q)N}} \zeta_L^{(N)}(k)>u\right)\le (Tp^{-1}+1)\sum_{n=1}^{p^{-1}-1}\times$$
$$\times\,\mathbf{P}\left(\max_{\substack{0<k\le pN\\ (1-(n+1)q)N<L\le(1-nq)N}} \zeta_L^{(N)}(k)>u\right). \qquad (2.16)$$

Let p_n denote the probability under the sum sign. Bounds for p_n will be obtained in two steps.

Case 2.1: $n<\frac{13}{16}p^{-1}-1$. The restriction on n ensures that the sum extracted from $\zeta_L^{(N)}(k)$ in the equality below is not empty:

$$\max_{\substack{0<k\le pN\\ (1-(n+1)q)N<L\le(1-nq)N}} \zeta_L^{(N)}(k)=\tfrac{1}{\sqrt{N}}\sum_{i=[pN]+1}^{[(1-(n+1)q)N]}\xi_i+$$
$$+\max_{\substack{0<k\le pN\\ (1-(n+1)q)N<L\le(1-nq)N}}\tfrac{1}{\sqrt{N}}\left(\sum_{i=k}^{[pN]}\xi_i+\sum_{i=[(1-(n+1)q)N]+1}^{k+L-1}\xi_i\right).$$

Repeating the proof of Lemma 2.1 we obtain the following analog of the equality (2.2),

$$p_n=\frac{\sqrt{A}}{\sqrt{2\pi\sigma'^2}u}e^{-\frac{u^2}{2\sigma'^2}}\int\limits_{-\infty}^{\infty}e^{-\frac{Aw^2/u^2}{2\sigma'^2}}e^{\frac{\sqrt{A}w}{\sigma'^2}}\mathbf{P}\left(M(k,L)>w\right)dw, \qquad (2.17)$$

where σ'^2 is equal to $\frac{[(1-(n+1)q)N]-[pN]}{N}$.

The expression (2.2) for the probability in Lemma 2.1 differs from (2.17) only in the variance σ'^2 of the extracted summand. Recall that σ^2 in Lemma 2.1 is equal to $\frac{[(1-q)N]-[pN]}{N}$. It is straightforward to show that

$$\frac{\sigma^2}{\sigma'^2}=1+\frac{nq}{1-(n+1)q-p}+o(u^{-2})=1+z.$$

With this notation the right-hand side of (2.17) takes form

$$\frac{\sqrt{A}e^{-\frac{u^2}{2\sigma^2}(1+z)}}{\sqrt{2\pi\sigma'^2}u}\int\limits_{-\infty}^{\infty}e^{-\frac{Aw^2/u^2}{2\sigma^2}(1+z)+\frac{\sqrt{A}w}{\sigma^2}z}e^{\frac{\sqrt{A}w}{\sigma^2}}\mathbf{P}\left(M(k,L)>w\right)dw.$$

The first exponent under the integral sign is a parabola with respect to w and attains its maximum at the point $w = \frac{z}{z+1}\frac{u^2}{\sqrt{A}}$. Straightforward calculation then show that

$$p_n \leq \frac{\sqrt{A}}{\sqrt{2\pi\sigma'^2}u} e^{-\frac{u^2}{2\sigma^2}K} \int\limits_{-\infty}^{\infty} e^{\frac{\sqrt{A}w}{\sigma^2}} \mathbf{P}\left(M(k,L) > w\right) dw,$$

where

$$K = 1 + \frac{z}{1+z} = 1 + \frac{nq}{1-q-p} \geq 1 + nq.$$

Finally, owing to Lemma 2.1 there exists a constant C such that

$$p_n \leq \frac{\sigma}{\sigma'} e^{-\frac{nB}{2}} H_A^B \frac{1}{\sqrt{2\pi}u} e^{-\frac{u^2}{2}}(1+o(1)) \leq Ce^{-\frac{nB}{2}} H_A^B p_0(u),$$

where $o(1) \to 0$ uniformly in n when $u, N \to \infty$.

Case 2.2: $np \geq \frac{13}{16}$. Now σ'^2 can be arbitrary small and we estimate p_n using Lemma 1 of Piterbarg (1991):

$$\begin{aligned} p_n &\leq \mathbf{P}\left(\max_{\substack{0<k\leq pN \\ 0<L\leq \frac{3}{16}N}} \zeta_L^{(N)}(k) > u \right) \leq 2\mathbf{P}(\max_{0<k\leq \frac{3}{16}N+pN} S_k > \tfrac{1}{2}u\sqrt{N}) \\ &\leq 2e^{-\frac{u^2}{4(\frac{3}{16}+p)}} \leq 2e^{-u^2}. \end{aligned}$$

Thus, combining the estimates for p_n obtained in cases 2.1 and 2.2 with (2.16) and (2.15) we have

$$\overline{\lim_{u,N}} \mathbf{P}\left(\max_{\substack{0<k\leq TN \\ 0<L\leq N}} \zeta_L^{(N)}(k) > u \right) \Big/ (Tu^2 p_0(u)) \leq H^B + \frac{H_A^B C}{A} \sum_{n=1}^{\infty} e^{-\frac{nB}{2}},$$

$$\underline{\lim_{u,N}} \mathbf{P}\left(\max_{\substack{0<k\leq TN \\ 0<L\leq N}} \zeta_L^{(N)}(k) > u \right) \Big/ (Tu^2 p_0(u)) \geq H^B.$$

It was proved in Zholud (2008) that the limit $H = \lim\limits_{B\to\infty} H^B$ exists and is positive. Letting first A, and then B tend to infinity, we conclude that the upper and lower limits coincide and equal H. This finishes the proof of Theorem 1.3.

3 Very large deviations of the Shepp statistic

Here we prove Theorem 1.4. The asymptotic behavior of the probability (1.1) under assumption that $u/\sqrt{N} \to \infty$ is considered. First, we find the asymptotic behavior of the probability

$$\mathbf{P}\left(\max_{0<k\le TN} \zeta_N^{(N)}(k) > u\right). \tag{3.1}$$

As in the previous section, we then show that the maximum of the field $\zeta_L^{(N)}(k)$ over the complementary set $\{(k,L) : 0 < k \le TN,\ 0 < L \le N-1\}$ gives neglible contribution to the probability (1.1).

Now a key lemma that plays an essential role in establishing the asymptotic formula for (3.1).

Lemma 3.1. *Let (ξ_1, ξ_2) be a Gaussian random vector such that ξ_1 and ξ_2 are standard normal variables with correlation coefficient $\alpha < 1$. Then,*

$$\boldsymbol{P}(\xi_1 > u, \xi_2 > u) < \frac{1}{\sqrt{2\pi}u} e^{-\frac{1}{2}u^2} \frac{1}{\sqrt{2\pi}u}(1+\alpha)\frac{\sqrt{1+\alpha}}{\sqrt{1-\alpha}} e^{-\frac{1}{2}u^2\frac{1-\alpha}{1+\alpha}}.$$

Proof. The variable ξ_2 can be expressed as the sum of two independent variables $\alpha\xi_1$ and ζ, where $\zeta \sim N(0, 1-\alpha^2)$. By $\varphi_\zeta(\cdot)$ we will refer to the density function of ζ. Denoting the probability in the statement of the lemma by $I(u)$ we have

$$\begin{aligned} I(u) &= \mathbf{P}(\xi_1 > u, \alpha\xi_1 + \zeta > u) \\ &= \tfrac{1}{\sqrt{2\pi}} \int_u^\infty e^{-\frac{v^2}{2}} \mathbf{P}(\zeta > u - \alpha v)dv = -\tfrac{1}{\sqrt{2\pi}} \int_u^\infty \tfrac{\mathbf{P}(\zeta > u-\alpha v)}{v} de^{-\frac{v^2}{2}} \\ &= -\tfrac{\mathbf{P}(\zeta>u-\alpha v)}{\sqrt{2\pi}v} e^{-\frac{v^2}{2}}\Big|_u^\infty + \tfrac{1}{\sqrt{2\pi}} \int_u^\infty e^{-\frac{v^2}{2}} d\tfrac{\mathbf{P}(\zeta>u-\alpha v)}{v} \\ &= \tfrac{e^{-\frac{u^2}{2}}}{\sqrt{2\pi}u} \mathbf{P}(\zeta > u(1-\alpha)) + \int_u^\infty \tfrac{e^{-\frac{v^2}{2}}}{\sqrt{2\pi}} \Big(\alpha \tfrac{\varphi_\zeta(u-\alpha v)}{v} - \tfrac{\mathbf{P}(\zeta>u-\alpha v)}{v^2}\Big) dv. \end{aligned}$$

Write $K(u)$ for the first summand in the last expression. The second summand is less than

$$\frac{\alpha}{\sqrt{2\pi}u}\int\limits_{u}^{\infty} e^{-\frac{v^2}{2}}\varphi_\zeta(u-\alpha v)dv$$

and thus $I(u)$ is bounded by

$$\begin{aligned}
&K(u)+\tfrac{\alpha}{\sqrt{2\pi}u}\int\limits_{u}^{\infty}\tfrac{1}{\sqrt{2\pi(1-\alpha^2)}}e^{-\frac{1}{2}\left(v^2+\frac{(u-\alpha v)^2}{1-\alpha^2}\right)}dv\\
&\quad=K(u)+\alpha\tfrac{e^{-\frac{u^2}{2}}}{\sqrt{2\pi}u}\int\limits_{u}^{\infty}\tfrac{1}{\sqrt{2\pi(1-\alpha^2)}}e^{-\frac{1}{2}\frac{(v-\alpha u)^2}{1-\alpha^2}}dv\\
&\qquad=K(u)+\alpha K(u)=\tfrac{1}{\sqrt{2\pi}u}e^{-\frac{1}{2}u^2}(1+\alpha)\mathbf{P}\left(\tfrac{\zeta}{\sqrt{1-\alpha^2}}>u\tfrac{\sqrt{1-\alpha}}{\sqrt{1+\alpha}}\right).
\end{aligned}$$

The lemma now follows from the standard upper bound of the standard normal distribution tail. □

Next, we estimate (3.1) using the Bonferroni inequality:

$$\begin{aligned}
&[TN]\mathbf{P}\left(\zeta_N^{(N)}(1)>u\right)\geq\mathbf{P}\left(\max_{0<k\leq TN}\zeta_N^{(N)}(k)>u\right)\\
&\geq[TN]\mathbf{P}\left(\zeta_N^{(N)}(1)>u\right)-\sum_{\substack{1\leq l,m\leq TN\\ l\neq m}}\mathbf{P}\left(\zeta_N^{(N)}(l)>u,\zeta_N^{(N)}(m)>u\right).
\end{aligned}$$

By stationarity, and applying Lemma 3.1 with

$$\alpha=\alpha_n=\mathbf{E}\zeta_N^{(N)}(1)\zeta_N^{(N)}(n)=\max\{0,\frac{N-(n-1)}{N}\},$$

we get that the double sum is bounded by

$$\begin{aligned}
&2TN\sum_{n=2}^{TN}\mathbf{P}\left(\zeta_N^{(N)}(1)>u,\zeta_N^{(N)}(n)>u\right)\\
&\quad<2TN\sum_{n=N+1}^{TN}\mathbf{P}\left(\zeta_N^{(N)}(1)>u\right)^2\\
&\qquad\qquad+2TN\sum_{n=2}^{N}\tfrac{1}{\sqrt{2\pi}u}e^{-\frac{1}{2}u^2}\tfrac{1}{\sqrt{2\pi}u}(1+\alpha_n)\tfrac{\sqrt{1+\alpha_n}}{\sqrt{1-\alpha_n}}e^{-\frac{1}{2}u^2\frac{1-\alpha_n}{1+\alpha_n}}.
\end{aligned}$$

As before let $p_0(u)$ denote $\frac{1}{\sqrt{2\pi}u}e^{-\frac{1}{2}u^2}$, the asymptotic bound for the standard normal distribution tail. The first summand is then less than

$$2(TN)^2\mathbf{P}\left(\zeta_N^{(N)}(1) > u\right)^2 = 2(TN)^2 p_0(u)^2(1+o(1))$$

and the second is estimated from above by

$$2TNp_0(u)\frac{2\sqrt{2N}}{\sqrt{2\pi}u}\sum_{n=2}^{N}\left(e^{-\frac{u^2/N}{4}}\right)^{n-1} = o(TNp_0(u)),$$

where we took into account that $u/\sqrt{N} \to \infty$.

Replacing the double sum by its upper estimate and dividing both sides of the Bonferroni inequality by $[TN]p_0(u)$, and assuming $TN \geq 1$, we get that

$$1+o(1) \geq \frac{\mathbf{P}\left(\max\limits_{0<k\leq TN}\zeta_N^{(N)}(k) > u\right)}{[TN]p_0(u)} \geq 1-4TNp_0(u)(1+o(1))+o(1).$$

Finally, for $TNp_0(u) \to 0$ we obtain the following asymptotic formula for the probability (3.1),

$$\mathbf{P}\left(\max_{0<k\leq TN}\zeta_N^{(N)}(k) > u\right) = [TN]p_0(u)(1+o(1)). \tag{3.2}$$

The remaining step is to note that the probability for the maximum over the complementary set is neglible. Since

$$\begin{aligned}\mathbf{P}\left(\max_{\substack{0<k\leq TN\\0<L\leq N-1}}\zeta_L^{(N)}(k) > u\right) &\leq TN\sum_{L=1}^{N-1}\mathbf{P}\left(\zeta_L^{(N)}(1) > u\right)\\ &\leq TN\sum_{L=1}^{N-1}p_0\left(u\sqrt{\tfrac{N}{L}}\right) = TNp_0(u)\sum_{L=1}^{N-1}e^{-\frac{u^2(N-L)}{2L}}\\ &\leq TNp_0(u)\sum_{L=1}^{N-1}\left(e^{-\frac{u^2/N}{2}}\right)^{N-L} = o(TNp_0(u)),\end{aligned}$$

the latter estimate and (3.2) conclude the proof of Theorem 1.4.

4 Limit theorems for $M_T^{(N)}$

In this section we consider the case when T, N go to infinity. It can be shown that for appropriate normalization constants a_T and

b_T the limit distribution of $\left(M_T^{(N)} - a_T\right) \Big/ b_T$ is Gumbel. Theorem 4.1 exhibits the normalizing constants for three different limit relations between T and N.

Theorem 4.1. *Assume that one of the following relations hold:*

$$1)\ \frac{2\ln T}{N} \to 0. \quad 2)\ \frac{2\ln T}{N} \to \theta^2 > 0. \quad 3)\ \frac{2\ln T}{N} \to \infty.$$

Then, for any fixed x,

$$\boldsymbol{P}\left(\max_{\substack{0<k\le TN \\ 0<L\le N}} a_T(\zeta_L^{(N)}(k) - b_T) \le x \right) = e^{-e^{-x}} + o(1),$$

where

$$a_T = \sqrt{2\ln T}, \qquad b_T = \sqrt{2\ln T} + \frac{F(T,N) + \frac{1}{2}\left(\ln\ln T - \ln\pi\right)}{\sqrt{2\ln T}}$$

and the function $F(T,N)$ is given by

$$1)\ F(T,N) = \ln H \quad 2)\ F(T,N) = \ln\frac{J_\theta}{\theta} \quad 3)\ F(T,N) = -\ln\frac{2\ln T}{N}.$$

The proof follows from Lemma 3.1 of Zholud (2008) closely, and is hence omitted.

The limit distribution for the case $\frac{2\ln T}{N} = \theta^2,\ 0 < \theta < \infty$ was obtained by A.M. Kozlov in Kozlov (2004) and was reformulated in Theorem 4.1 for comparison purpose.

Acknowledgements The author gives his deep thanks to Professor Piterbarg V.I. for stating the problem and for constant assistance, and wishes to thank Holger Rootzén for pointing out numerous typos and unclear points.

References

P. Erdös and A. Rényi. On a new law of large numbers. *Journal d'Analyse Mathématique*, 23(1):103–111, 1970. 116, 118

A.N. Frolov. Limit theorems for increments of sums of independent random variables. *Theory of Probability and Its Applications*, 48 (1):93–107, 2004. 118

Z. Kabluchko. Extreme-value analysis of standardized Gaussian increments. Technical Report arXiv:0706.1849, The University of Ulm, 2007. 118

A.M. Kozlov. On large deviations for the Shepp statistic. *Discrete Mathematics and Applications*, 14(2):211–216, 2004. 116, 133

V.I. Piterbarg. On large jumps of a random walk. *Theory of probability and its applications*, 36(1):50–62, 1991. 116, 119, 120, 122, 124, 129

V.I. Piterbarg and A.M. Kozlov. On large jumps of a Cramer random walk. *Theory of probability and its applications*, 47(4): 719–729, 2002. 116

L.A. Shepp. A limit law concerning moving averages. *Ann. Math. Statist.*, 35(1):424–428, 1964. 118

D.S. Zholud. Extremes of the Shepp statistic for the Wiener process. *Extremes*, 11(4):339–351, 2008. 117, 118, 119, 127, 129, 133

PAPER IV

Extremes
2008, Vol. 11, No. 4, pp. 339-351.

Extremes of the Shepp statistic for the Wiener process

Dmitrii Zholud [*]

Abstract

Define $Y(t) = \max_{0 \leq s \leq 1} W(t+s) - W(t)$, where $W(\cdot)$ is the standard Wiener process. We study the maximum of Y up to time T: $M_T = \max_{0 \leq t \leq T} Y(t)$ and determine an asymptotic expression for $\mathbf{P}\left(M_T > u\right)$ when $u \to \infty$. Further we establish the limiting Gumbel distribution of M_T when $T \to \infty$ and present the corresponding normalization sequence.

Keywords Wiener process, increments, maximum, extreme values, high level excursions, large deviations, asymptotic behavior, Shepp statistic, distribution tail, Gumbel law, limit theorems, weak theorems, Gaussian random walk.

AMS 2000 Subject Classifications: $\frac{\text{Primary-60G70;}}{\text{Secondary-60G15, 60F05;}}$.

[*] *Department of Mathematical Statistics*
Chalmers University of Technology and University of Göteborg, Sweden.
E-mail: dmitrii@zholud.com

1 Introduction

First, we introduce two different techniques used in the asymptotic theory of Gaussian processes and fields. For a Gaussian process $Z(t)$, consider asymptotic behavior of the probability

$$\mathbf{P}\left(\max_{[0,T]} Z(t) > u\right),\ u \to \infty. \tag{1.1}$$

In the case when $Z(t)$ is a stationary Gaussian process with a covariance function $r(t)$ such that $r(t) - r(0)$ is a regularly varying function of index α for $t \to 0$, the exact asymptotic forms of (1.1) were given by Pickands, see Pickands (1969a,b).

In the non-stationary case there are a number of results for Gaussian processes with a unique point of maximum variance, see e.g. Berman (1985), Hüsler (1990) and related papers. When $Z(t)$ is a Gaussian process with continuous paths, zero mean and non-constant variance, and there is a unique fixed point of maximum variance t_0 in the interval $[0, T]$, the asymptotic behavior of probability in (1.1) is known. The theory sketched out above is described in detail in Piterbarg (1996).

Next, define $X(t, s) = W(t+s)-W(t)$ and $Y(t) = \max\limits_{0\le s\le 1} X(t, s)$, for $W(\cdot)$ the standard Wiener process. Let $M_T = \max\limits_{[0,T]} Y(t)$ be the maximum up to time T of $Y(t)$. The aim of this paper is to find the asymptotic behavior of $\mathbf{P}(M_T > u)$, the probability of high level excursions of $Y(t)$ as $u \to \infty$ and to obtain the limiting distribution of M_T when $T \to \infty$.

For the first task it is crucial to use a representation of M_T as a maximum of the Gaussian field $X(t, s)$ over rectangle $[0, T] \times [0, 1]$:

$$M_T = \max_{[0,T]\times[0,1]} X(t, s).$$

Since for fixed s, $X(\cdot, s)$ is a stationary process, and for fixed t, $X(t, \cdot)$ is a process with a unique point of maximum variance, the asymptotic behavior is obtained by combining standard techniques for the corresponding cases. Let $\psi(u)$ be the tail of the standard normal distribution function. The following result and its proof, as well as the expression for the constant H are given in Section 2.

Theorem 1.1. *If $Tu^2 \to \infty$ and $Tu^2\psi(u) \to 0$ when $u \to \infty$, then*

$$\boldsymbol{P}(M_T > u) = HTu^2\psi(u)(1 + o(1)).$$

When the asymptotic behavior of the tail of distribution of M_T is known, we find a limiting distribution of M_T when $T \to \infty$. In this case it is essential to use the representation of M_T as a maximum up to time T of stationary process $Y(t)$. When $|t_1 - t_2| > 1$, the random variables $Y(t_1)$ and $Y(t_2)$ are independent . The method of establishing the limit theorem is common. Introduce a partition of $[0, T]$ into long blocks $A_i = [i(S + 1), i(S + 1) + S)$ of length S, and short blocks $B_i = [i(S + 1) + S, (i + 1)(S + 1))$ of length 1 such that $[0, T] = \bigcup_{i=0}^{n} (A_i \cup B_i)$. Then define a sequence of independent identically distributed random variables (i.i.d. r.v.) $Y_i = \max_{A_i} Y(t)$, $i = 1, 2, ..$ Letting S to infinity and following the proof of J. Pickands theorem for $\max\{Y_1, Y_2, ...\}$, see Leadbetter et al. (1983), the only thing left is to show that random variables $\bar{Y}_i = \max Y(t)$ over B_i give negligible contributions to the limiting distribution of the maximum $M_T = \max\{Y_1, \bar{Y}_1, Y_2, \bar{Y}_2, Y_3, \bar{Y}_3...\}$. However, this idea is extended to obtain a more general result, see Lemma 3.1. It will be used when building limit theorems for the Shepp statistic for a Gaussian random walk, see Zholud (2009). As a corollary of the lemma stated in Section 3 we obtain the limiting Gumbel distribution for M_T, when $T \to \infty$.

Theorem 1.2. *For any fixed x and $T \to \infty$, the following relation holds:*

$$\boldsymbol{P}\left(\max_{(t,s)\in[0,T]\times[0,1]} a_T(W(t + s) - W(t) - b_T) \le x\right) = e^{-e^{-x}} + o(1),$$

where

$$a_T = \sqrt{2\ln T}, \quad b_T = \sqrt{2\ln T} + \frac{\ln H + \frac{1}{2}(\ln\ln T - \ln \pi)}{\sqrt{2\ln T}}.$$

A similar result for the standardized Wiener process increments is obtained in Kabluchko (2007). There are also a number of works about strong laws for the increments of the Wiener process, see e.g. Csörgő and Révész (1979), Frolov (2005).

One of the applications of the result derived in this paper is given in Zholud (2009). Let $(\xi_i, i \geq 1)$ be standard normal random variables, and S_k be the corresponding random walk, $S_k = \sum\limits_{i=1}^{k} \xi_i$, $S_0 = 0$. Define a random variable $\zeta_L^{(N)}(k) = \frac{1}{\sqrt{N}} \left(S_{k+L-1} - S_{k-1}\right)$. Asymptotic behavior of the probability

$$\mathbf{P}\left(\max_{\substack{0<k\leq TN\\0<L\leq N}} \zeta_L^{(N)}(k) > u\right),$$

when $u \to \infty$, $N \to \infty$ in some synchronized way is then examined. For fixed u, owing to the weak convergence of a random walk to the Wiener process,

$$\mathbf{P}\left(\max_{\substack{0<k\leq TN\\0<L\leq N}} \zeta_L^{(N)}(k) > u\right) = \mathbf{P}\left(M_T > u\right)\left(1 + o(1)\right),\ N \to \infty.$$

Paper Zholud (2009) shows that this equation also holds when $u \to \infty$ and $u/\sqrt{N} \to 0$.

2 Asymptotic behavior of the distribution tail of M_T

In this section we find the asymptotic behavior of the probability

$$\mathbf{P}\left(M_T > u\right) = \mathbf{P}\left(\max_{\substack{0\leq t\leq T\\0\leq s\leq 1}} W(t+s) - W(t) > u\right), \tag{2.1}$$

when $u \to \infty$ and $T \to \infty$ in an appropriate way. As before, we denote $X(t,s) = W(t+s) - W(t)$. The proof is divided into two steps:

First, for any positive constant B we focus on the asymptotic behavior of maximum of X over a thin layer $[0,T] \times [1 - Bu^{-2}, 1]$. It will be shown that within this area and assuming that u is large, $X(t,s)$ and $X(t,1)$ behave in a similar way, and it is possible to determine the asymptotic behavior using the standard technique for stationary processes.

Second, knowing the asymptotic behavior of the maximum of X over the area of its maximum variance, we will show that the maximum over the complementary set $[0,T] \times [0, 1 - Bu^{-2}]$ gives neglible contribution to the probability in (2.1).

The proof of the first part is based on the Double Sum Method: the lemma below is the analog of Lemma 6.1, Piterbarg (1996). To proceed, let A and B be any positive constants and denote $p = Au^{-2}$, $q = Bu^{-2}$ and $A_0(u) = [0,p] \times [1-q, 1]$. Although it is possible to obtain a representation similar to what we get in Lemma 2.1 by repeating the proof of Lemma 6.1, Piterbarg (1996), our proof does not follow the standard procedure. Instead of passing on to the family of conditional distributions as in Piterbarg (1996), we extract the common part of the increment $X(t,s)$ for all $(t,s) \in A_0(u)$ and use independence of the Wiener process increments.

Lemma 2.1. *Let $u \to \infty$. Then*

$$\boldsymbol{P}\left(\max_{A_0(u)} W(t+s) - W(t) > u\right) = H_A^B \frac{1}{\sqrt{2\pi}u} e^{-\frac{u^2}{2}}(1 + o(1)),$$

where

$$H_A^B = e^{-\frac{A+B}{2}} \mathbf{E} \exp\left(\max_{\substack{0 \le t \le A \\ 0 \le s \le B}} W(t+s+A) - W(t)\right).$$

Proof. We have that since $1 - q > p$ for large u,

$$\mathbf{P}\left(\max_{A_0(u)} W(t+s) - W(t) > u\right)$$
$$= \mathbf{P}\Bigg(W(1-q) - W(p) +$$
$$+ \max_{A_0(u)} W(t+s) - W(1-q) + W(p) - W(t) > u\Bigg),$$

and by stationarity and independence of the Wiener process increments
$W(t+s) - W(1-q)$ and $W(p) - W(t)$, the probability above is equal to

$$\mathbf{P}\left(\xi + \max_{A_0(u)} W(t+s-(1-q)+p) - W(t) > u\right)$$

$$= \mathbf{P}\left(\xi + \max_{\substack{0\le t\le p\\ 0\le s\le q}} W(t+s+p) - W(t) > u\right),$$

where random variable ξ is normally distributed with zero mean, variance $\sigma^2 = 1-p-q$ and is independent of the expression inside the maximum sign. Thus,

$$\mathbf{P}\left(\max_{A_0(u)} W(t+s) - W(t) > u\right)$$
$$= \frac{1}{\sqrt{2\pi}\sigma} \int\limits_{-\infty}^{\infty} e^{-\frac{v^2}{2\sigma^2}} \mathbf{P}\left(\max_{\substack{0\le t\le p\\ 0\le s\le q}} W(t+s+p) - W(t) > u - v\right) dv.$$

After the change of variables $v = u - \frac{w}{u}$, the last expression equals

$$\frac{\sigma^{-1}}{\sqrt{2\pi}u} \int\limits_{-\infty}^{\infty} e^{-\frac{(u-\frac{w}{u})^2}{2\sigma^2}} \mathbf{P}\left(\max_{\substack{0\le t\le p\\ 0\le s\le q}} u(W(t+s+p) - W(t)) > w\right) dw$$

$$= \frac{e^{-\frac{u^2}{2\sigma^2}}}{\sqrt{2\pi}\sigma u} \int\limits_{-\infty}^{\infty} e^{-\frac{w^2/u^2}{2\sigma^2}} e^{\frac{w}{\sigma^2}} \mathbf{P}\left(\max_{\substack{0\le t\le A\\ 0\le s\le B}} W(t+s+A) - W(t) > w\right) dw.$$

Next, by the dominated convergence theorem, which follows from the upper estimate of the probability under the integral sign (see Borel's theorem, Piterbarg (1996), p.13), and relations $\sigma^2 \to 1$ and

$$e^{-\frac{u^2}{2\sigma^2}} = e^{-\frac{u^2}{2}(1+p+q+o(u^{-2}))}(1+o(1)) = e^{-\frac{u^2}{2}} e^{-\frac{A+B}{2}}(1+o(1)),$$

when $u \to \infty$, we obtain the desired representation. □

Corollary 2.1.1.
1) H_A^B is nondecreasing with respect to the parameters A and B.
2) $H_{A_1+A_2}^B \le H_{A_1}^B + H_{A_2}^B$.
3) $H_A^B \le AH_1^B$, for any integer A.

Our next aim is to move on from the rectangle $[0, Au^{-2}] \times [1 - Bu^{-2}, 1]$ to the layer $[0, T] \times [1 - Bu^{-2}, 1]$. We use Lemma 2.1 and the Bonferroni inequality to obtain estimates of the probability of high level excursions of the maximum of X. Then we show that estimates from below and from above are asymptotically equivalent.

Let $A_r(u) = [rAu^{-2}, (r+1)Au^{-2}] \times [1 - Bu^{-2}, 1]$. For ease of notation we suppress dependence on u. Using stationarity of $X(t,s)$ with respect to t, we obtain that

$$(\tfrac{Tu^2}{A}+1)\mathbf{P}\left(\max_{(t,s)\in A_0} X(t,s) > u\right) \geq \mathbf{P}\left(\max_{\substack{0\leq t\leq T\\ 1-Bu^{-2}\leq s\leq 1}} X(t,s) > u\right) \geq$$
$$\geq (\tfrac{Tu^2}{A} - 1)\mathbf{P}\left(\max_{(t,s)\in A_0} X(t,s) > u\right) -$$
$$- \sum_{\substack{0\leq l,m\leq \frac{Tu^2}{A}+1\\ l\neq m}} \mathbf{P}\left(\max_{(t,s)\in A_l} X(t,s) > u, \max_{(t,s)\in A_m} X(t,s) > u\right). \tag{2.2}$$

Let $p_{l,m}$ denote the summands in the last sum in (2.2). The sum, owing to stationarity, does not exceed

$$2\left(\frac{Tu^2}{A} + 1\right)\sum_{n=1}^{\frac{Tu^2}{A}+1} p_{0,n}. \tag{2.3}$$

Estimating the probabilities $p_{0,n}$ from above, we will show that the sum (2.3) is negligible, and thus the upper and lower estimates in (2.2) are asymptotically equivalent.

The estimates are obtained in slightly different ways, in the same manner as in Lemma 7.1, Piterbarg (1996). The next lemma is a modification of Lemma 6.3, Piterbarg (1996).

Lemma 2.2. *There exists an absolute constant C such that inequality*

$$\boldsymbol{P}\left(\max_{(t,s)\in A_0} X(t,s) > u, \max_{(t,s)\in A_r} X(t,s) > u\right) \leq C(AB)^2\psi(u)e^{-\frac{(r-1)A}{4}}$$

holds for any A, B any $1 < r \leq 1 + \frac{u^2}{A}$, and for any u, $u \geq u_0$,

$$u_0 = \inf\left\{u : e^{-4Au^{-2}} \leq 1 - 2Au^{-2}\ ,\ \ Bu^{-2} \leq \frac{1}{2}\right\}.$$

Proof. The Gaussian field $X(t,s)$ has zero mean, is stationary in t, and its covariance function is

$$K(t,s;t_1,s_1) = mes\left([t,t+s]\bigcap[t_1,t_1+s_1]\right). \tag{2.4}$$

Consequently, a global Hölder condition holds:

$$\mathbf{E}\left(X(t,s)-X(t_1,s_1)\right)^2 \le 2(|s-s_1|+|t-t_1|). \tag{2.5}$$

Introducing $Y(\mathbf{v},\mathbf{w}) = X(\mathbf{v}) + X(\mathbf{w})$, where $\mathbf{v} = (t,s)$ and $\mathbf{w} = (t_1, s_1)$, we get:

$$\mathbf{P}\left(\max_{(t,s)\in A_0} X(t,s) > u, \max_{(t,s)\in A_r} X(t,s) > u\right)$$
$$\le \mathbf{P}\left(\max_{A_0\times A_r} Y(\mathbf{v},\mathbf{w}) > 2u\right).$$

Using (2.4), (2.5) and restrictions on r and u it is straightforward to estimate the minimum and maximum values of the variance of $Y(\mathbf{v},\mathbf{w})$ and then to obtain an estimate from below of the covariance function of normalized field $Y^*(\mathbf{v},\mathbf{w})$, see Lemma 6.3, Piterbarg (1996). Further steps repeat the proof of the lemma. □

Corollary 2.2.1. *When* $r > 1 + \frac{u^2}{A}$ *and* $u \ge u_0$ *the following inequality holds*

$$\boldsymbol{P}\left(\max_{(t,s)\in A_0} X(t,s) > u, \max_{(t,s)\in A_r} X(t,s) > u\right) \le C(AB)^2\psi(u)^2.$$

Condition $r > 1 + \frac{u^2}{A}$ implies that the events inside the probability are independent and finishes the proof.

Corollary 2.2.2. *When* $r = 1$ *and* $u \ge u_0$*, the following inequality holds*

$$\boldsymbol{P}\left(\max_{(t,s)\in A_0} X(t,s) > u, \max_{(t,s)\in A_r} X(t,s) > u\right)$$
$$\le \left(C(AB)^2 e^{-\frac{1}{4}\sqrt{A}} + (\sqrt{A}+1)H_1^B\right)\psi(u).$$

The proof follows Lemma 7.1 on p.107 in Piterbarg (1996). We are now ready to estimate (2.3) from above. Since

$$\sum_{n=1}^{\frac{Tu^2}{A}+1} p_{0,n} = p_{0,1} + \sum_{n=2}^{\frac{u^2}{A}+1} p_{0,n} + \sum_{n=\frac{u^2}{A}+2}^{\frac{Tu^2}{A}+1} p_{0,n}$$

and estimating the first summand by using Corollary 2.2.2, the second using Lemma 2.2 and the last using Corollary 2.2.1, we get that

$$(2.3) \leq 2\left(\tfrac{Tu^2}{A}+1\right)\psi(u)\left\{\left(C(AB)^2 e^{-\frac{1}{4}\sqrt{A}} + (\sqrt{A}+1)H_1^B\right) + \right. \\ \left. + C(AB)^2 \sum_{n=2}^{\infty} e^{-\frac{1}{4}(n-1)A} + \tfrac{Tu^2}{A} C(AB)^2 \psi(u)\right\}.$$

Assuming that $Tu^2 \to \infty$ and $Tu^2\psi(u) \to 0$ it follows from (2.2), (2.3), Lemma 2.1 and the estimate of (2.3) above that

$$\varlimsup_{u\to\infty} \frac{\mathbf{P}\left(\max\limits_{\substack{0\leq t\leq T \\ 1-Bu^{-2}\leq s\leq 1}} X(t,s)>u\right)}{Tu^2\psi(u)} \leq A^{-1}H_A^B$$

and (2.6)

$$\varliminf_{u\to\infty} \frac{\mathbf{P}\left(\max\limits_{\substack{0\leq t\leq T \\ 1-Bu^{-2}\leq s\leq 1}} X(t,s)>u\right)}{Tu^2\psi(u)} \geq (A')^{-1}H_{A'}^B - \\ - 2\tfrac{C}{A'}\left\{\left((A'B)^2 e^{-\frac{\sqrt{A'}}{4}} + \tfrac{\sqrt{A'}+1}{C}H_1^B\right) + (A'B)^2 \sum_{n=2}^{\infty} e^{-\frac{(n-1)A'}{4}}\right\}.$$

Thus, noticing that the expression in the last line tends to zero when $A' \to \infty$, and applying Corollary 2.1.1 3), we see that:

$$\varliminf_{A\to\infty} A^{-1}H_A^B \leq \varlimsup_{A'\to\infty} (A')^{-1}H_{A'}^B \leq H_1^B < \infty.$$

Finally, we want to show that the limit

$$H^B = \lim_{A\to\infty} A^{-1}H_A^B, \quad 0 < H^B \leq H_1^B < \infty, \tag{2.7}$$

that exists as a consequence of the estimate above, is positive. This is done by considering the probability of high level excursion over the subset $D = \bigcup\limits_i A_{2i} \cap [0,T]\times[0,1]$ and following the proof of D.16 in Piterbarg (1996).

Thus, assuming A and A' in (2.6) tend to infinity and applying (2.7), we obtain the asymptotic behavior of the probability of high level excursion of the maximum of $X(t,s)$ over the upper layer $[0,T]\times[1-Bu^{-2},1]$:

Lemma 2.3. *Assuming* $Tu^2 \to \infty$ *and* $Tu^2\psi(u) \to 0$*, the following equality holds:*

$$\boldsymbol{P}\left(\max_{\substack{0\le t\le T\\ 1-Bu^{-2}\le s\le 1}} X(t,s) > u\right) = H^B Tu^2\psi(u)(1+o(1)).$$

Below we give the second part of the proof. It shows that the asymptotic behavior of the probability of the high level excursion of the maximum of $X(t,s)$ over the upper layer, which corresponds to the area of the maximum variance of the field, gives the main contribution to (2.1).

Let $B_n(u) = [0,T]\times[1-(n+1)Bu^{-2}, 1-nBu^{-2}]$ and assume that the conditions $Tu^2 \to \infty$ and $Tu^2\psi(u) \to 0$ are satisfied. As before, for notational convenience we suppress the dependence of B_n on u.

Lemma 2.4. *Starting from large enough values of* u*, if* $nBu^{-2} \le \frac{1}{2}$*, then*

$$\boldsymbol{P}\left(\max_{(t,s)\in B_n} X(t,s) > u\right) \le 4H^{2B}e^{-\frac{1}{2}nB}Tu^2\psi(u)(1+c(u)),$$

where $c(u) \to 0$*, when* $u \to \infty$*.*

Proof. Normalizing by the maximum standard deviation of $X(t,s)$ over B_n we get

$$\mathbf{P}\left(\max_{(t,s)\in B_n} X(t,s) > u\right) = \mathbf{P}\left(\max_{(t,s)\in B_n} \tfrac{X(t,s)}{\sqrt{1-nBu^{-2}}} > \tfrac{u}{\sqrt{1-nBu^{-2}}}\right)$$

$$= \mathbf{P}\left(\max_{\substack{0\le t\le T/(1-nBu^{-2})\\ 1-\frac{Bu^{-2}}{1-nBu^{-2}}\le s\le 1}} X(t,s) > \tfrac{u}{\sqrt{1-nBu^{-2}}}\right)$$

$$\le \mathbf{P}\left(\max_{\substack{0\le t\le 2T\\ 1-2Bu^{-2}\le s\le 1}} X(t,s) > \tfrac{u}{\sqrt{1-nBu^{-2}}}\right).$$

The expression on the right-hand side satisfies all the conditions of Lemma 2.3, and for large enough u inequality $\psi(\frac{u}{\sqrt{1-nBu^{-2}}}) \le 2\psi(u)e^{-\frac{1}{2}nB}$ holds uniformly in n. □

Lemma 2.5. *If* $nBu^{-2} > \frac{1}{2}$, *then*

$$\boldsymbol{P}\left(\max_{(t,s)\in[0,T]\times[0,1]\setminus\bigcup\limits_{i=0}^{n} B_i} X(t,s) > u\right) \le CTu^4\psi(\sqrt{2}u).$$

Proof. Expanding the set under the maximum sign, we get

$$\mathbf{P}\left(\max_{(t,s)\in[0,T]\times[0,1]\setminus\bigcup\limits_{i=0}^{n} B_i} X(t,s) > u\right) \le \mathbf{P}\left(\max_{\substack{0\le t\le T\\ 0\le s\le \frac{1}{2}}} X(t,s) > u\right).$$

The maximum of the variance of $X(t,s)$ over the set $[0,T]\times[0,\frac{1}{2}]$ equals $\frac{1}{2}$. Theorem 8.1, Piterbarg (1996) finishes the proof. □

Now follows the proof of Theorem 1.1: Lemmas 2.3, 2.4 and 2.5 imply that

$$\underline{\lim}_{u\to\infty} \frac{\mathbf{P}\left(\max\limits_{[0,T]\times[0,1]} X(t,s)>u\right)}{Tu^2\psi(u)} \ge \underline{\lim}_{u\to\infty} \frac{\mathbf{P}\left(\max\limits_{(t,s)\in B_0} X(t,s)>u\right)}{Tu^2\psi(u)} = H^B \quad \text{and}$$

$$\begin{aligned}\overline{\lim}_{u\to\infty} \frac{\mathbf{P}\left(\max\limits_{[0,T]\times[0,1]} X(t,s)>u\right)}{Tu^2\psi(u)} &\le \overline{\lim}_{u\to\infty} \frac{1}{Tu^2\psi(u)} \left[\mathbf{P}\left(\max_{(t,s)\in B_0} X(t,s) > u\right)\right.\\ &\left. + \sum_{n=1}^{\frac{u^2}{2B}} \mathbf{P}\left(\max_{(t,s)\in B_n} X(t,s) > u\right) + \mathbf{P}\left(\max_{(t,s)\in \hat{B}} X(t,s) > u\right)\right]\\ &\le H^{B'} + 4H^{2B'} \times \sum_{n=1}^{\infty} e^{-\frac{1}{2}nB'},\end{aligned}$$

where $\hat{B}$ denotes $[0,T]\times[0,1]\setminus\bigcup\limits_{n=0}^{\frac{u^2}{2B}+1} B_n$. Now note that the constant $H^B = \lim\limits_{A\to\infty} A^{-1}H_A^B$ is non-decreasing with respect to the parameter B, and the last inequalities show that it is bounded from above. Thus, $\lim\limits_{B\to\infty} H^B = H$, say, exists, finite and positive, and $\lim\limits_{B'\to\infty} H^{B'} + 4H^{2B'} \times \sum\limits_{n=1}^{\infty} e^{-\frac{1}{2}nB'}$ also equals H. □

3 Limit theorem for M_T

In this section we consider the case where T goes to infinity, and we obtain the limit distribution of $(M_T - a_T)/b_T$ for the appropriate normalization functions a_T and b_T. First we prove a general lemma, which can serve as a template for obtaining limiting theorems not only for random fields, but for a family of fields as well. We assume that the specific asymptotic behavior of the tail of the distribution of the maximum of some field takes place and that this asymptotic behavior is defined by an asymptotic relation between threshold u, parameter S that defines the set over which the maximum is taken, and parameter N discussed below. The condition that defines the asymptotic behavior will be denoted by, say, $\mathcal{D}(u, N, S)$. The following lemma shows that knowing asymptotic behavior under $\mathcal{D}(u, N, S)$ we can derive a new condition involving T and N such that if it holds when T goes to infinity, M_T has limiting Gumbel distribution.

Lemma 3.1. *Assume that:*
1) $X^N(t, s)$ $N = 1, 2...$ is a family of fields stationary with respect to the parameter t, and defined on the set $[0, \infty) \times [0, 1]$.
2) For any N, any t, t_1 such that $|t - t_1| > 1$ and any $s, s_1 \in [0, 1]$, the random variables $X^N(t, s)$ and $X^N(t_1, s_1)$ are independent.
3) By $\mathcal{D}(u, N, S)$ we refer to some logical statement that involves variables u, N, S and such that if $\mathcal{D}(u, N, S)$ holds then the following asymptotic behavior of the tail of the distribution of a maximum of $X^N(t, s)$ over the set $D_S = [0, S] \times [0, 1]$ takes place:

$$\boldsymbol{P}\left(\max_{D_S} X^N(t, s) > u\right) \sim SF(u, N) \tag{3.1}$$

for some function $F(u, N)$. We also demand that if $\mathcal{D}(u, N, 1)$, then (3.1) holds for $S \equiv 1$.

4)Let $T \to \infty$ and suppose there exist appropriate normalizing functions a_T and b_T such that

$$\lim_{\substack{T\to\infty \\ (N\to\infty)}} TF(u_T, N) = e^{-x}$$

for any fixed x, where $u_T = b_T + \frac{x}{a_T}$. Functions a_T and b_T may also depend on N.

5)Let $S = S(T)$ be such a function that $S \to \infty$ and $n = \frac{T}{S+1} \to \infty$ when $T \to \infty$.

Then, if $\mathcal{D}(u_T, N, 1)$ and $\mathcal{D}(u_T, N, S(T))$ hold,

$$\boldsymbol{P}\left(\max_{D_T} X^N(t,s) > u_T\right) \to 1 - e^{-e^{-x}}. \qquad (3.2)$$

Proof. Let us introduce a partition $[0,T] = \bigcup_{i=0}^{n} (A_i \cup B_i)$, with

$$\begin{aligned} A_i &= [i(S+1), i(S+1)+S] \\ B_i &= [i(S+1)+S, (i+1)(S+1)], \end{aligned}$$

so that $|A_i| = S$ and $|B_i| = 1$ for all i. For the expression on the left-hand side of (3.2) we have that

$$\begin{aligned} &\mathbf{P}\left(\max_{D_T} X^N(t,s) \le u_T\right) \\ &= 1 - \mathbf{P}\left(\bigcup_{i=0}^{n}\left\{\max_{A_i\times[0,1]} X^N(t,s) > u_T \ \cup \max_{B_i\times[0,1]} X^N(t,s) > u_T\right\}\right). \end{aligned}$$

Applying stationarity of $X^N(t,s)$ with respect to t we obtain the following estimate:

$$\begin{aligned} &1 - n\mathbf{P}\left(\max_{[0,1]^2} X^N(t,s) > u_T\right) - \mathbf{P}\left(\bigcup_{i=0}^{n} \max_{A_i\times[0,1]} X^N(t,s) > u_T\right) \le \\ &\le \mathbf{P}\left(\max_{D_T} X^N(t,s) \le u_T\right) \le 1 - \mathbf{P}\left(\bigcup_{i=0}^{n} \max_{A_i\times[0,1]} X^N(t,s) > u_T\right). \end{aligned} \qquad (3.3)$$

The term $n\mathbf{P}\left(\max_{[0,1]^2} X^N(t,s) > u_T\right)$ is estimated using $\mathcal{D}(u_T, N, 1)$ and 3) and, for penultimate equality, 4)

$$\begin{aligned} n\mathbf{P}\left(\max_{[0,1]^2} X^N(t,s) > u_T\right) &= nF(u_T, N)(1+o(1)) \\ &= \tfrac{TF(u_T,N)}{S+1}(1+o(1)) = \tfrac{e^{-x}(1+o(1))}{S+1} = o(1). \end{aligned}$$

Using the fact that $\max_{A_i\times[0,1]} X^N(t,s)$ and $\max_{A_j\times[0,1]} X^N(t,s)$ are independent for $i \ne j$, see 2), and, again, stationarity, we estimate the expression on the right-hand side of (3.3) using $\mathcal{D}(u_T, N, S(T))$ and 3) in the third step, and 4) and 5) in the fifth

$$1 - \mathbf{P}\left(\bigcup_{i=0}^{n} \max_{A_i\times[0,1]} X^N(t,s) > u_T\right) = \prod_{i=0}^{n} \mathbf{P}\left(\max_{A_i\times[0,1]} X^N(t,s) \le u_T\right)$$

$$= \left(1 - \mathbf{P}\left(\max_{A_0\times[0,1]} X^N(t,s) > u_T\right)\right)^n = (1 - SF(u_T, N))^n$$

$$= e^{n\ln(1-SF(u_T,N))} = e^{-nSF(u_T,N)(1+o(SF(u_T,N)))}$$

$$= e^{-TF(u_T,N)(1+o(1))} = e^{-e^{-x}}(1+o(1)).$$

It therefore follows from (3.3) that

$$e^{-e^{-x}}(1+o(1))+o(1) \le \mathbf{P}\left(\max_{D_T} X^N(t,s) \le u_T\right) \le e^{-e^{-x}}(1+o(1)),$$

and this finishes the proof. □

Corollary 3.1.1 (The Wiener process)**.**

Put $X^N(t,s) \equiv W(t+s) - W(t)$. We say that $\mathcal{D}(u,N,S)$ holds if and only if $Su^2 \to \infty$ and $Su^2\psi(u) \to 0$, $u \to \infty$. Thus, conditions 1), 2) and 3) of the lemma are satisfied by Theorem 1.1.

It is easy to verify that Condition 4) is satisfied for

$$u_T = \frac{x}{\sqrt{2\ln T}} + \sqrt{2\ln T} + \frac{\ln H + \frac{1}{2}(\ln\ln T - \ln\pi)}{\sqrt{2\ln T}}.$$

In 5) we set $S(T) = \sqrt{T}$. Condition $\mathcal{D}(u_T, N, 1)$ becomes equivalent to $u_T \to \infty$ that is equivalent to $T \to \infty$ owing to our choice of u_T. Finally, using 3) it is easy to show that

$$S(T)u_T^2\psi(u_T) = S(T)/T \times TF(u_T,N) = e^{e^{-x}}(1+o(1))/\sqrt{T} = o(1).$$

Thus $\mathcal{D}(u_T, N, S)$ is equivalent to $T \to \infty$ and Theorem 1.2 holds.

Acknowledgements The author is grateful to Professor Piterbarg V.I. for constant assistance and fruitful discussions on the Double Sum method, and wishes to thank Professor Holger Rootzén for careful reading and pointing out numerous typographical errors and unclear points.

References

S.M. Berman. An asymptotic formula for the distribution of the maximum of a Gaussian process with stationary increments. *Journal of Applied Probability*, 22(2):454–460, 1985. 138

M. Csörgő and P. Révész. How big are the increments of a Wiener process? *The Annals of Probability*, 7(4):731–737, 1979. 139

A.N. Frolov. Unified limit theorems for increments of processes with independent increments. *Theory of Probability and Its Applications*, 49(3):531–539, 2005. 139

J. Hüsler. Extreme values and high boundary crossings of locally stationary Gaussian processes. *Annals of Probability*, 18(3):1141–1158, 1990. 138

Z. Kabluchko. Extreme-value analysis of standardized Gaussian increments. Technical Report arXiv:0706.1849, The University of Ulm, 2007. 139

M.R. Leadbetter, G. Lindgren, and H. Rootzén. *Extremes and related properties of random sequences and processes.* Springer series in statistics. Springer-Verlag, 1983. 139

J. Pickands. Upcrossings probabilities for stationary Gaussian processes. *Trans. Amer. Math. Soc.*, 145:51–73, 1969a. 138

J. Pickands. Asymptotic properties of the maximum in a stationary Gaussian process. *Trans. Amer. Math. Soc.*, 145:75–86, 1969b. 138

V.I. Piterbarg. *Asymptotic Methods in the Theory of Gaussian Processes and Fields.* Translation of mathematical monographs. AMS, 1996. 138, 141, 142, 143, 144, 145, 147

D.S. Zholud. Extremes of the Shepp statistic for a Gaussian random walk. *Extremes*, 12(1):1–17, 2009. 139, 140

Notes

Notes

For further information please visit www.zholud.com

ISBN 978-91-628-8354-6