Thesis for the Degree of Licentiate of Philosophy

Contribution to Extreme Value Theory with Applications in Bioinformatics

Dmitrii Zholud

Division of Mathematical Statistics
Department of Mathematical Sciences
Chalmers University of Technology
and University of Gothenburg
Göteborg, Sweden 2009

Contribution to Extreme Value Theory with Applications in Bioinformatics

Dmitrii Zholud

ISSN 1652-9715
Report 2009:8

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

Author e-mail: dmitrii@chalmers.se

Typeset with LaTeX.
Department of Mathematical Sciences
Printed in Göteborg, Sweden 2009

Contribution to Extreme Value Theory with Applications in Bioinformatics

Dmitrii Zholud

Department of Mathematical Sciences
Division of Mathematical Statistics
Chalmers University of Technology
and University of Gothenburg

Abstract

This thesis presents results in Extreme Value Theory with application to Bioinformatics. First, we obtain the asymptotic behavior of the probability of high level excursions for the maximum of the Wiener process increments, followed by the normalization sequence for the corresponding limiting Gumbel distribution. Next, we consider the Shepp statistics for Gaussian random walk and establish asymptotic formulas for the case of moderate and excessively large deviations. The latter is related to the problem of sequence comparing.

Further, we study extreme values of student's t-statistics under non-normality and various deviations from i.i.d assumption. The study is motivated by the analysis of systematic errors in a particular kind of biological experiments (BioScreen array experiments) which showed that the t-statistics distribution has a certain tail behavior regardless of what is the true model. We give a theoretical explanation of this phenomenon and a basis for new methods to correct theoretical p-values. The obtained asymptotic formulas are very accurate for small sample sizes and are of practical interest for quantile estimation in connection to High Throughput Screening.

Keywords: Extreme Value Statistics, Extreme Value Theory, Asymptotic Behavior, High Level Excursions, Quantile Estimation, Wiener Process, Gaussian Random Walk, Shepp Statistics, Student's t-statistics, Self Normalized Sums, Gumbel Law, Limit Theorems, Test Power, Small Sample Size, False Discovery Rate, High Throughput Screening.

Acknowledgments

First, I would like to thank my supervisor, Holger Rootzén. Not only for being so patient reading my manuscripts, but also because it has always been a great pleasure to work with him; the advice and help I got during my studies are very much appreciated! A separate credit goes to my co-supervisor Olle Nerman. It has never been easy to agree on a meeting with both supervisors, but Olle was a very good complement and this made my life easier in many aspects. It is a pleasure to thank the administrative staff as well. Next, I would like to acknowledge V.I. Piterbarg. The work on "Extremes of Shepp statistics of Gaussian random walk" was initiated during the last year of my studies at Lomonosov Moscow State University, and I am grateful for giving me a good start and constant support as a graduate student.

The friendly working environment gives its own contribution and is acknowledged. Among those who made my life more colorful are Daniel Ahlberg, Ottmar Cronie, Frank Eriksson, Sofia Tapani, Teresia Dahl, Anastassia Baxevani, Alexandra Jauhiainen and more.

Also I would like to express my sincere gratitude to Asja Grzibovska, Rauan Sagitov, Leonid Molokov and Mattias Bengtsson for their friendship. A lot of inspiration came from our non-work related conversations.

Finally, I can hardly imagine any of my academic achievements without my family, who stayed with me in good times and in bad times. A substantial part of what I am now is owing to you and your trust!

Dmitrii Zholud
Göteborg, March 6, 2009

This thesis consists of the following papers:

Paper I: Zholud, D.S. (2008). Extremes of Shepp statistics for the Wiener Process *Extremes,* v. 11, no 4, pp. 339-351.

Paper II: Zholud, D.S. (2009). Extremes of Shepp Statistics for Gaussian Random Walk, *Extremes,* v. 12, no 1, pp. 1-17.

Paper III: Zholud, D.S. (2009). Extremes of Student's t-statistics for non-normal and not necessarily i.i.d. random variables.

1 Introduction

In this section we give an overview of the three papers included in the thesis. For convenience we start with the third paper, where we give the motivation for the study and describe an important concept of Extreme Value Theory. Then follow the two other papers, which have a different motivation and are more theoretical.

Paper 3

The origin of the paper was the study of the experimental design for a particular kind of biological experiments, namely BioScreen array experiments, see [4], [5] and [6]. The aim of the experiments is to identify gene modified strains or conditions leading to differential growth behavior of Yeast colonies. Omitting the details, the phenotypic growth property of interest of the mutant cell (i.e. in which the gene is "knocked out") is compared to the same property of a wild type cell, and this is done for many different modified strains and/or growth conditions. Moreover, there is an experimental spatial layout of the single growth experiments in batches on micro-titer arrays. The "difference" is associated with the parameter called LSC, which is assumed to be measured with normally distributed error with mean zero. The hypothesis testing is based on the one-sample t-test for the two replicates of LSC. However, a histogram of the LSC values in a wild type data set (for which the null hypothesis is known to be true) showed that the distribution of the LSC deviates from normal. The latter results in the deviation of the distribution of the corresponding t-statistics from the theoretical t-distribution with one degree of freedom. Furthermore, the analysis of the experimental setup revealed spatial systematic biases (see [6]), making it practically impossible to model the true distribution of the LSC values. Even if the distribution was known, obtaining the exact theoretical distribution of the corresponding t-statistics would be questionable.

To overcome this difficulty we instead look at the resulting t-statistics. To be specific, we wish to correct the theoretical quantile by approximating the tail of the distribution of the t-statistics under the null hypothesis. The wild type data set consists of 1584 t-values, which limits the use of the non-extreme methods of quantile estimation to as far as, say, 0.99. The reason for going further out in the tail is, for example, that [5] uses the significance level of 0.001. Another reason is a multiple comparison problem (common to High-Throughput screenings), which arises when one considers a set, or family, of statistical inferences simultaneously.

We now introduce a Peaks over Threshold method. For a random variable X with distribution function F, the probability of exceedance over large thresholds is approximated using generalized Pareto distribution. Assume that the distribution tail has an approximation property

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

for u large and $\{x : x > 0 \text{ and } 1 + \xi x/\tilde{\sigma} > 0\}$, where $\tilde{\sigma} = \sigma + \xi(u - \mu)$ and μ, σ and ξ are some constants; ξ and $\tilde{\sigma}$ are called the parameters of the generalized Pareto distribution. For $\xi = 0$, we interpret the distribution as the limit as $\xi \to 0$, i.e.

$$H(x) = exp(-x/\tilde{\sigma}), \quad x > 0.$$

For an independent sample, given large fixed u one typically estimates the probability $\mathbf{P}\left(X > u\right)$ with a relative frequency and fit the parameters ξ and $\tilde{\sigma}$ using only observations larger than u. Once the parameters of the generalized Pareto distribution are "known", the p upper quantile, x_p, is estimated as

$$x_p = u + \frac{\tilde{\sigma}}{\xi}\left(\left[\frac{\mathbf{P}\left(X > u\right)}{1 - p}\right]^{\xi} - 1\right).$$

Indeed, for $x > u$,

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

therefore,

$$\mathbf{P}\left(X > x\right) \approx \mathbf{P}\left(X > u\right)\left(1 + \frac{\xi(x - u)}{\tilde{\sigma}}\right)^{-1/\xi}.$$

For modeling threshold excesses and the parameter estimation for generalized Pareto families we refer to [2] and [1]. The book by Beirlant et al. (2004) focuses on the Extreme Value Theory from a mathematical statistician point of view, while Coles (2001) puts more emphasis on applications.

Within the BioScreen problem, however, we obtain a stronger result. Inspired by the p-p plot of the t-values for the wild type data set, fig.1 of [9], the focus was shifted towards the behavior of the one-sample t-statistics under non-normality, dependency and non-stationarity. The main result of [9] follows.

Let $X = (X_1, X_2, .., X_n)$, $n \geq 2$ be a general random vector, where components need not be independent or identically distributed. Assume

X has a continuous density function $g(x_1, x_2, ..x_n)$ and denote T_n the corresponding t-statistics. We prove that under some (quite mild) regularity conditions on g,

$$\frac{\mathbf{P}\left(T_n > u\right)}{t_{n-1}(u)} = K_g + o(1) \quad \text{as} \quad u \to \infty, \tag{1.1}$$

where t_{n-1} stands for a t-distribution tail with $n-1$ degrees of freedom. The exact expression for the constant K_g is given in Theorem 2.1 of [9].

Note that in terms of the Peaks over Threshold method this means that the parameters of the generalized Pareto distribution for the t-statistics are invariant to the sample distribution. Indeed, assuming $K_g > 0$, which holds if $g(x, x, .., x)$ is positive for some $x > 0$,

$$\mathbf{P}\left(T_n > u + x | T_n > u\right) = \frac{\mathbf{P}\left(T_n > u + x\right)}{\mathbf{P}\left(T_n > u\right)} \approx \frac{t_{n-1}(u + x)}{t_{n-1}(u)},$$

and the latter expression does not depend on g.

Estimating quantile x_p is thus equivalent to estimating the probability $\mathbf{P}\left(T_n > u\right)$ for some large enough u. The simulation study of [9] shows that for $n = 2$, the threshold u can be as low as 0.95 quantile of the t_1 distribution.

The result is complemented by a second order approximation formula. Assuming g is twice-differentiable and satisfies some additional regularity conditions,

$$\frac{\mathbf{P}\left(T_n > u\right) - K_g t_{n-1}(u)}{t_{n+1}\left(\sqrt{\frac{n+1}{n-1}}u\right)} = M_g - L_g + o(1),$$

where constants M_g and L_g are defined in Theorem 2.2. This is compared to several other suggested tail approximations in the case of i.i.d. random variables. For small n the new method looks superior.

Paper 2

The starting point for this paper was the problem of detecting homology (similarity) between long DNA or protein molecules, [3]. Each of the protein molecules is represented as a sequence of letters that denote chemical groups in a poly-peptide chain. Element-by-element comparison is rarely used in practice. One of the reasons is the computational time it takes to compare two very long sequences (e.g. sequence against a huge

database) and another is the possible mismatches, say, due to sequencing errors. Moreover, some elements of the molecule are more similar in their chemical functions than others, and the mismatches might be because of the evolutionary mutations rather than that the molecules come from unrelated organisms or perform totally different functions. This argument led to invention of weight matrixes (such as BLOSUM62 or PAM250) and different kinds of scoring functions. For further reading we refer to such algorithms as BLAST or FASTA. Inspired by these practical bioinformatics methods we study a mathematically relevant matching problem and present a theoretical result on extreme values of the Shepp statistics for 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 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

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

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 corresponding normalization sequences.

Paper 1

This result is essential for the proof of [7].

We study

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

where $W(\cdot)$ is a standard Wiener process and determine an asymptotic expression for $\mathbf{P}(M_T > u)$ when $u \to \infty$. Further we establish the limiting Gumbel distribution of M_T as $T \to \infty$ and present the corresponding normalization sequence.

References

[1] Beirlant, J., Goegebeur, Y., Segers, J., Teugels, J., De Waal, D. and Ferro, C., Statistics of Extremes: Theory and Applications. Wiley, New York, 2004.

[2] Coles, S.G., An introduction to Statistical Modelling of Extreme Values. Springer, London, 2001.

[3] Leontovich, A.M., Nikolaev, V.K., On Power Threshold Values in Homology Searches. Manuscript.

[4] Warringer J. and Blomberg A. (2003). Automated screening in environmental arrays allows analysis of quantitative phenotypic profiles in Saccharomyces cerevisiae. Yeast 20, pp. 53-67

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

[6] Zholud, D., Rootzén, H., Nerman, O., and Blomberg, A., Positional effects in biological array experiments and their impact on the false discovery rate, 2007.

[7] Zholud, D., Extremes of Shepp statistics for Gaussian random walk, 2009, Extremes, Vol. 12, no 1, pp. 1-17.

[8] Zholud, D., Extremes of Shepp statistics for the Wiener process, 2008, Extremes, v. 11, no 4, pp. 339-351.

[9] Zholud, D.S., Extremes of Student's t-statistics for non-normal and not necessarily i.i.d. random variables, 2009.

PAPER I

Extremes DOI:10.1007/s10687-008-0061-7

Volume 11, Number 4 December, 2008 pp. 339-351

Extremes of Shepp Statistics for The Wiener Process

by

DMITRII ZHOLUD

Department of Mathematical Statistics, University of Gothenburg, Department of Mathematical Statistics, Chalmers University of Technology, Göteborg, Sweden. ✉ *dmitrii@math.chalmers.se*

Extremes of Shepp statistics for the Wiener process

DMITRII ZHOLUD

Abstract. Define $Y(t) = \max\limits_{0\le s\le 1} W(t+s) - W(t)$, where $W(\cdot)$ is a standard Wiener process. We study the maximum of Y up to time T: $M_T = \max\limits_{0\le t\le T} Y(t)$ 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.

Key words. Wiener process increments, Shepp statistics, high level excursions, analysis of extreme values, large deviations, asymptotic behavior, distribution tail, Gumbel law, limit theorems, weak theorems.

AMS 2000 Subject Classifications: $\frac{\text{Primary-60G70;}}{\text{Secondary-60G15, 60F05;}}$.

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 (1969).

In the non-stationary case, besides [1], [4] and related papers there are a number of results for Gaussian processes with a unique point of maximum variance. When $Z(t)$ is a Gaussian process with continuous paths, zero mean and nonconstant 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 [9].

Next, define $X(t,s) = W(t+s) - W(t)$ and $Y(t) = \max\limits_{0\le s\le 1} X(t,s)$, for $W(\cdot)$ a 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 $\mathbf{P}\left(M_T > u\right)$, 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 was obtained by combining standard techniques for 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, of length S, and short blocks B_i of length 1: $[0,T] = \bigcup_{i=0}^{n}(A_i \cup B_i)$, where

$$A_i = [i(S+1), i(S+1)+S)\,, \quad B_i = [i(S+1)+S, (i+1)(S+1))\,.$$

Then define a sequence of independent identically distributed random variables (i.i.d. r.v.) $Y_i = \max\limits_{A_i} Y(t),\ i = 1,2,..$ Letting S to infinity and following the proof of J. Pickands theorem [6] for $\max\{Y_1, Y_2, ...\}$, the only thing left is to show that random variables $\bar{Y}_i = \max\limits_{B_i} Y(t)$ give negligible contributions to the limiting distribution of $M_T = \max\{Y_1, \bar{Y}_1, Y_2, \bar{Y}_2, Y_3, \bar{Y}_3 ...\}$. However, this idea is extended to obtain a more general result (Lemma 3.1). It will be used when building limit theorems for Shepp statistics for a Gaussian random walk [10]. 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 standardized Wiener process increments is obtained in [5]. There are also a number of works about strong laws for increments of Wiener processes [2], [3].

One of the applications of the result derived in this paper is given in [10]. 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 $\zeta_L^{(N)}(k) = \frac{1}{\sqrt{N}} (S_{k+L-1} - S_{k-1})$. 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 a 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 [10] shows that this equation also holds when $u \to \infty$, $u/\sqrt{N} \to 0$.

2 Asymptotic behavior of the tail of distribution 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 a maximum of X over a thin layer $[0,T]\times[1-Bu^{-2},1]$. Within this area and assuming that u is large, $X(t,s)$ and $X(t,1)$ will behave in a similar way and it will be shown that it is possible to determine the asymptotic behavior using the standard technique for stationary processes.

Second, knowing the asymptotic behavior of 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 a 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, [9]. 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, [9], our proof does not follow the standard procedure. Instead of passing on to the family of conditional distributions as in [9], we 'extract' the common part of the increment $X(t,s)$ for all $(t,s) \in A_0(u)$ and use independence of 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,

$$\begin{aligned}
&\mathbf{P}\left(\max_{A_0(u)} W(t+s) - W(t) > u\right) \\
&\qquad = \mathbf{P}\left(W(1-q) - W(p) + \right. \\
&\qquad\qquad \left. + \max_{A_0(u)} W(t+s) - W(1-q) + W(p) - W(t) > u\right),
\end{aligned}$$

and by stationarity and independence of Wiener process increments $W(t+s)-W(1-q)$ and $W(p)-W(t)$, the probability above is equal to

$$\begin{aligned}
&\mathbf{P}\left(\xi + \max_{A_0(u)} W(t+s-(1-q)+p) - W(t) > u\right) \\
&\qquad = \mathbf{P}\left(\xi + \max_{\substack{0\le t\le p\\ 0\le s\le q}} W(t+s+p) - W(t) > u\right),
\end{aligned}$$

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,

$$\begin{aligned}
&\mathbf{P}\left(\max_{A_0(u)} W(t+s) - W(t) > u\right) \\
&\qquad = \tfrac{1}{\sqrt{2\pi}\sigma} \int\limits_{-\infty}^{\infty} e^{-\frac{v^2}{2\sigma^2}} \mathbf{P}\left(\max_{\substack{0\le t\le p\\ 0\le s\le q}} W(t+s+p) - W(t) > u - v\right) dv.
\end{aligned}$$

After the change of variables $v = u - \frac{w}{u}$, the last expression equals

$$\frac{\sigma^{-1}}{\sqrt{2\pi}u}\int\limits_{-\infty}^{\infty} e^{-\frac{(u-\frac{w}{u})^2}{2\sigma^2}}\mathbf{P}\left(\max_{\substack{0\le t\le p\\ 0\le s\le q}} u(W(t+s+p)-W(t))>w\right)dw$$

$$=\frac{\sigma^{-1}}{\sqrt{2\pi}u}e^{-\frac{u^2}{2\sigma^2}}\int\limits_{-\infty}^{\infty} e^{-\frac{w^2/u^2}{2\sigma^2}}e^{\frac{w}{\sigma^2}}\mathbf{P}\left(\max_{\substack{0\le t\le A\\ 0\le s\le B}} W(t+s+A)-W(t)>w\right)dw.$$

Next, by dominated convergence, which follows from the upper estimate of the probability under the integral sign (Borel Theorem [9], 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 Bonferroni inequality to obtain estimates of the probability of high level excursions of the maximum of X. Then we will 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

$$\begin{aligned}(\tfrac{Tu^2}{A}+1)\mathbf{P}\left(\max_{(t,s)\in A_0}X(t,s)>u\right)&\ge\mathbf{P}\left(\max_{\substack{0\le t\le T\\ 1-Bu^{-2}\le s\le 1}}X(t,s)>u\right)\ge\\ &\ge(\tfrac{Tu^2}{A}-1)\mathbf{P}\left(\max_{(t,s)\in A_0}X(t,s)>u\right)-\\ &-\sum_{\substack{0\le l,m\le\frac{Tu^2}{A}+1\\ l\ne m}}\mathbf{P}\left(\max_{(t,s)\in A_l}X(t,s)>u,\max_{(t,s)\in A_m}X(t,s)>u\right).\end{aligned}\tag{2.2}$$

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, [9]. The next lemma is a modification of Lemma 6.3, [9].

Lemma 2.2. *There exists an absolute constant C such that inequality*

$$\boldsymbol{P}\left(\max_{(t,s)\in A_0} X(t,s) > u, \max_{(t,s)\in A_r} X(t,s) > u\right) \le C(AB)^2\psi(u)e^{-\frac{1}{4}(r-1)A}$$

holds for any A, B any $1 < r \le 1 + \frac{u^2}{A}$, and for any u, $u \ge u_0$,

$$u_0 = \inf\left\{u : e^{-4Au^{-2}} \le 1 - 2Au^{-2}\ ,\quad Bu^{-2} \le \frac{1}{2}\right\}.$$

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 the notation $Y(\mathbf{v},\mathbf{w}) = X(\mathbf{v}) + X(\mathbf{w})$, where $\mathbf{v} = (t,s)$ and $\mathbf{w} = (t_1,s_1)$, we get:

$$\mathbf{P}\left(\max_{A_0} X(t,s) > u, \max_{A_r} X(t,s) > u\right) \le \mathbf{P}\left(\max_{A_0\times A_r} Y(\mathbf{v},\mathbf{w}) > 2u\right).$$

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, [9]. 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 makes up the proof.

Corollary 2.2.2. *When $r = 1$ and $u \geq u_0$, the following inequality holds*

$$\boldsymbol{P}\Big(\max_{(t,s)\in A_0} X(t,s) > u, \max_{(t,s)\in A_r} X(t,s) > u\Big) \leq \Big(C(AB)^2 e^{-\frac{1}{4}\sqrt{A}} + (\sqrt{A}+1)H_1^B\Big)\psi(u).$$

The proof follows the corresponding method on p.107 in Lemma 7.1, [9].

We are now ready to estimate (2.3) from above. Since

$$\sum_{n=1}^{\frac{Tu^2}{A}+1} p_{0,n} = p_{0,1} + \sum_{n=2}^{\frac{u^2}{A}+1} p_{0,n} + \sum_{n=\frac{u^2}{A}+2}^{\frac{Tu^2}{A}+1} p_{0,n} \ ,$$

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

$$(2.3) \leq 2\left(\frac{Tu^2}{A}+1\right)\psi(u)\left\{\left(C(AB)^2 e^{-\frac{1}{4}\sqrt{A}} + (\sqrt{A}+1)H_1^B\right) + \right.$$
$$\left. + C(AB)^2 \sum_{n=2}^{\infty} e^{-\frac{1}{4}(n-1)A} + \frac{Tu^2}{A} C(AB)^2 \psi(u)\right\}.$$

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

$$\overline{\lim_{u\to\infty}} \frac{\mathbf{P}\left(\max_{\substack{0\leq t\leq T \\ 1-Bu^{-2}\leq s\leq 1}} X(t,s) > u\right)}{Tu^2\psi(u)} \leq A^{-1}H_A^B$$

and (2.6)

$$\underline{\lim_{u\to\infty}} \frac{\mathbf{P}\left(\max_{\substack{0\leq t\leq T \\ 1-Bu^{-2}\leq s\leq 1}} X(t,s) > u\right)}{Tu^2\psi(u)} \geq (A')^{-1}H_{A'}^B -$$
$$- 2C(A')^{-1}\left\{\left((A'B)^2 e^{-\frac{1}{4}\sqrt{A'}} + \frac{\sqrt{A'}+1}{C}H_1^B\right) + (A'B)^2 \sum_{n=2}^{\infty} e^{-\frac{1}{4}(n-1)A'}\right\}.$$

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:

$$\infty > H_1^B \geq \underline{\lim_{A\to\infty}} A^{-1}H_A^B \geq \overline{\lim_{A'\to\infty}} (A')^{-1}H_{A'}^B.$$

Finally, we want to show that the limit

$$H^B = \lim_{A\to\infty} A^{-1}H_A^B, \quad \infty > H_1^B \geq H^B > 0, \tag{2.7}$$

that exists as a consequence of the just obtained estimate, is positive. This is done by considering the probability of high excursions over the subset $D = \bigcup_i A_{2i} \cap [0,T]\times[0,1]$ and following the proof of D.16 in [9].

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 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:*

$$\boldsymbol{P}\left(\max_{\substack{0\leq t\leq T\\ 1-Bu^{-2}\leq s\leq 1}} X(t,s) > u\right) = H^B T u^2 \psi(u)(1+o(1)).$$

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 $nBu^{-2} \leq \frac{1}{2}$, then*

$$\boldsymbol{P}\left(\max_{(t,s)\in B_n} X(t,s) > u\right) \leq 4H^{2B}e^{-\frac{1}{2}nB}Tu^2\psi(u)(1+c(u)),$$

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

$$\mathbf{P}\left(\max_{(t,s)\in B_n} X(t,s) > u\right) = \mathbf{P}\left(\max_{(t,s)\in B_n} \tfrac{X(t,s)}{\sqrt{1-nBu^{-2}}} > \tfrac{u}{\sqrt{1-nBu^{-2}}}\right)$$

$$= \mathbf{P}\left(\max_{\substack{0\leq t\leq T/(1-nBu^{-2})\\ 1-\frac{Bu^{-2}}{1-nBu^{-2}}\leq s\leq 1}} X(t,s) > \tfrac{u}{\sqrt{1-nBu^{-2}}}\right)$$

$$\leq \mathbf{P}\left(\max_{\substack{0\leq t\leq 2T\\ 1-2Bu^{-2}\leq s\leq 1}} X(t,s) > \tfrac{u}{\sqrt{1-nBu^{-2}}}\right).$$

The expression on the right-hand side satisfies 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. $\square$

Lemma 2.5. *If $nBu^{-2} > \frac{1}{2}$, then*

$$\boldsymbol{P}\left(\max_{(t,s)\in[0,T]\times[0,1]\setminus\bigcup_{i=0}^{n}B_i} X(t,s) > u\right) \le CTu^4\psi(\sqrt{2}u).$$

Proof: Expanding the set over which the maximum is taken, we get

$$\mathbf{P}\left(\max_{(t,s)\in[0,T]\times[0,1]\setminus\bigcup_{i=0}^{n}B_i} X(t,s) > u\right) \le \mathbf{P}\left(\max_{\substack{0\le t\le T\\ 0\le s\le \frac{1}{2}}} X(t,s) > u\right).$$

The maximum of the variance of $X(t,s)$ over the set $[0,T]\times[0,\frac{1}{2}]$ equals $\frac{1}{2}$. Applying Theorem 8.1, [9] we conclude the proof. $\square$

Below follows the proof of Theorem 1.1:

Lemmas 2.3, 2.4 and 2.5 imply that

$$\underline{\lim}_{u\to\infty} \frac{\mathbf{P}\left(\max_{[0,T]\times[0,1]} X(t,s)>u\right)}{Tu^2\psi(u)} \ge \underline{\lim}_{u\to\infty} \frac{\mathbf{P}\left(\max_{(t,s)\in B_0} X(t,s)>u\right)}{Tu^2\psi(u)} = H^B$$

and

$$\begin{aligned}\overline{\lim}_{u\to\infty} \frac{\mathbf{P}\left(\max_{[0,T]\times[0,1]} X(t,s)>u\right)}{Tu^2\psi(u)} &\le \overline{\lim}_{u\to\infty} \frac{1}{Tu^2\psi(u)}\left[\mathbf{P}\left(\max_{(t,s)\in B_0} X(t,s) > u\right) + \right.\\ &\left. + \sum_{n=1}^{\frac{u^2}{2B}} \mathbf{P}\left(\max_{(t,s)\in B_n} X(t,s) > u\right) + \mathbf{P}\left(\max_{(t,s)\in \hat{B}} X(t,s) > u\right)\right]\\ &\le H^{B'} + 4H^{2B'} \times \sum_{n=1}^{\infty} e^{-\frac{1}{2}nB'},\end{aligned}$$

where $\hat{B}$ denotes $[0,T]\times[0,1]\setminus\bigcup_{n=0}^{\frac{u^2}{2B}+1} B_n$. Constant $H^B = \lim_{A\to\infty} A^{-1}H_A^B$ is non-decreasing with respect to the parameter B, and the last inequalities show that it is bounded from above. Thus, $\lim_{B\to\infty} H^B = H$, say, exists, finite and positive, and $\lim_{B'\to\infty} H^{B'} + 4H^{2B'} \times \sum_{n=1}^{\infty} e^{-\frac{1}{2}nB'}$ also equals H. $\square$

3 Limit theorem for M_T.

In this section we consider the case where T goes to infinity, and we obtain limit distribution of $(M_T - a_T)/b_T$ for 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 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 a family of fields defined on the set $[0, \infty) \times [0, 1]$, which are stationary with respect to parameter t.
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 mathematical statement that involves variables u, N, S and such that if $\mathcal{D}(u, N, S)$ is fulfilled 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, defining $u_T = b_T + \frac{x}{a_T}$, $\lim\limits_{\substack{T\to\infty \\ (N\to\infty)}} TF(u_T, N) = e^{-x}$ for any fixed x . 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 $n = T/(S+1) \to \infty$ when $T \to \infty$.

Then, if $\mathcal{D}(u_T, N, 1)$ and $\mathcal{D}(u_T, N, S(T))$ hold,

$$\boldsymbol{P}\left(\max_{D_T} X^N(t, s) > u_T\right) \to 1 - e^{-e^{-x}}. \tag{3.2}$$

Proof: Let us introduce a partition $[0,T] = \bigcup_{i=0}^{n} (A_i \cup B_i)$, with

$$A_i = [i(S+1), i(S+1)+S] \quad \text{and} \quad B_i = [i(S+1)+S, (i+1)(S+1)]\,,$$

$$\text{so that} \quad |A_i| = S,\ |B_i| = 1, \quad i = 0, 1, .. \ .$$

For the expression on the left-hand side of (3.2) we have that

$$\mathbf{P}\left(\max_{D_T} X^N(t,s) \le u_T\right)$$
$$= 1 - \mathbf{P}\left(\bigcup_{i=0}^{n} \left\{ \max_{A_i \times [0,1]} X^N(t,s) > u_T \ \cup \max_{B_i \times [0,1]} X^N(t,s) > u_T \right\}\right).$$

Applying stationarity of $X^N(t,s)$ with respect to t we obtain the following estimate:

$$1 - n\mathbf{P}\left(\max_{[0,1]^2} X^N(t,s) > u_T\right) - \mathbf{P}\left(\bigcup_{i=0}^{n} \max_{A_i \times [0,1]} X^N(t,s) > u_T\right) \le$$

$$\le \mathbf{P}\left(\max_{D_T} X^N(t,s) \le u_T\right) \le 1 - \mathbf{P}\left(\bigcup_{i=0}^{n} \max_{A_i \times [0,1]} X^N(t,s) > u_T\right). \quad (3.3)$$

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)

$$n\mathbf{P}\left(\max_{[0,1]^2} X^N(t,s) > u_T\right) = nF(u_T,N)(1+o(1)) = \tfrac{TF(u_T,N)}{S+1}(1+o(1))$$
$$= \tfrac{e^{-x}(1+o(1))}{S+1} = o(1).$$

Using the fact that for $i \neq j$ r.v. $\max_{A_i \times [0,1]} X^N(t,s)$ and $\max_{A_j \times [0,1]} X^N(t,s)$ are independent, 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: 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 very 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

[1] Berman S.M., An asymptotic formula for the distribution of the maximum of a Gaussian process with stationary increments, 1985, J. Applied Probability, t. 22, pp. 454-460.

[2] Csörgő, M., Révész, P., How Big are the Increments of a Wiener Process?, 1979, The Annals of Probability, Vol. 7, no 4, pp. 731-737.

[3] Frolov, A. N., Unified Limit Theorems for Increments of Processes with Independent Increments, 2004, Theory of Probability and Its Applications C/C of Teoriia Veroiatnostei I Ee Primenenie, Vol. 49; p. 3, pp. 531-539

[4] Hüsler, J., Extreme values and high boundary crossings of locally stationary Gaussian processes, 1990, Ann. Probab. Vol. 18, no 3, pp. 1141-1158.

[5] Kabluchko.WP, Z., Extreme-Value Analysis of Standardized Gaussian Increments, 2007, Bernoulli, submitted

[6] Leadbetter, M.R., Lindgren, G., Rootzén, H., Extremes and Related Properties of Random Sequences and Processes, Springer-Verlag, 1983.

[7] Pickands, J., Upcrossings probabilities for stationary Gaussian processes, 1969, Trans. Amer. Math. Soc., Vol. 145, pp. 51-73.

[8] Pickands, J., Asymptotic Properties of the Maximum in a Stationary Gaussian Process, 1969, Trans. Amer. Math. Soc., Vol. 145, pp. 75-86

[9] Piterbarg, V.I. Asymptotic methods in Theory of Gaussian Processes and Fields. Translation of mathematical Monographs, AMS, Providence, Rhode island, 1996.

[10] Zholud, D.S., Extremes of Shepp Statistics for Gaussian Random Walk, 2009, Extremes, Vol. 12, no 1, pp. 1-17.

PAPER II

Extremes DOI:10.1007/s10687-008-0065-3

Volume 12, Number 1 March, 2009 pp. 1-17

Extremes of Shepp Statistics for Gaussian Random Walk

by

DMITRII ZHOLUD

Department of Mathematical Statistics, University of Gothenburg, Department of Mathematical Statistics, Chalmers University of Technology, Göteborg, Sweden. ✉ *dmitrii@math.chalmers.se*

Extremes of Shepp statistics for 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 $\mathbf{P}\left(M_T^{(N)} > u\right)$ 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.

Key words. Gaussian random walk increments, Shepp statistics, high excursions, extreme values, large deviations, moderate deviations, asymptotic behavior,
distribution tail, Gumbel law, limit theorems, weak theorems.

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

1 Introduction

Let $(\xi_i, i \geq 1)$ be a sequence of independent standard normal random variables, and let $S_k = \sum_{i=1}^{k} \xi_i$, with $S_0 = 0$, be the corresponding random walk. Introduce the Shepp and the closely related Erdös-Rényi statistics

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

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 statistics $W_{TN,N}$. Related problems were described in [2], [9], [6] and [10]. Paper [9] presents

the asymptotic behavior of the probability of moderate deviations for Erdös-Rényi statistics under the assumption of sub-gaussian distribution of random walk step and papers [6] and [10] study large deviations of Erdös-Rényi 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 [6]. 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

$$J_\theta = \lim_{l\to\infty} \frac{1}{\theta l} \mathbf{E} exp\left\{\theta \max_{0\le n<l} (\sqrt{2}S_n - \theta n)\right\}.$$

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*

$$\boldsymbol{P}\left(\max_{\substack{0<k\le TN\\ 0<L\le N}} \zeta_L^{(N)}(k) > u\right) \sim J_\theta T u^2 \psi(u).$$

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 [13] 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 a 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 a 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 indeed is 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* $TN\psi(u) \to 0$ *and* $TN \geq 1$, *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 a full description of the possible asymptotic behavior of (1.1) under various relations between u and N.

Finally, using the results of sections 2-3 we obtain a limit Gumbel distribution for $M_T^{(N)}$ when $T, N \to \infty$. If 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, 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 [4].

There is also extensive literature on a.s. convergence of related quantities, see e.g. [11], [2] and [3].

2 Moderate deviations of Shepp statistics.

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 [13] and we divide it into two main parts.

First, for any positive constant B we focus on the asymptotic behavior of a maximum of $\zeta_L^{(N)}(k)$ over a thin layer

$$\{(k,L): 0<k\le TN, (1-Bu^{-2})N<L\le N\}.$$

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 [9].

Second, knowing the asymptotic behavior of maximum of $\zeta_L^{(N)}(k)$ 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 [13]. Let A and B be positive constants and set $p=Au^{-2}$, $q=Bu^{-2}$. By $A_0(u)$ we will 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*

$$\boldsymbol{P}\left(\max_{A_0(u)}\zeta_L^{(N)}(k)>u\right)=H_A^B\frac{1}{\sqrt{2\pi}u}e^{-\frac{u^2}{2}}(1+o(1)), \tag{2.1}$$

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,

$$\mathbf{P}\left(\max_{A_0(u)}\zeta_L^{(N)}(k)>u\right)$$

$$= \mathbf{P}\left(\frac{1}{\sqrt{N}} \sum_{i=[pN]+1}^{[(1-q)N]} \xi_i + \max_{\substack{0<k\le pN \\ 0<L\le qN}} \frac{1}{\sqrt{N}} \sum_{i=k}^{k+L+[pN]-1} \xi_i' > u \right)$$

$$= \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} e^{-\frac{v^2}{2\sigma^2}} \mathbf{P}\left(\max_{\substack{0<k\le pN \\ 0<L\le qN}} \frac{1}{\sqrt{N}} (S'_{k+L+[pN]-1} - S'_{k-1}) > u - v \right) dv,$$

where $\sigma^2 = \frac{[(1-q)N]-[pN]}{N}$ and S'_k stands for the sum $\sum_{i=1}^{k} \xi_i'$ with $S'_0 = 0$.

For the sake of briefness introduce

$$M(k,L) = \max_{\substack{0<k\le pN \\ 0<L\le qN}} \frac{1}{\sqrt{pN}} (S'_{k+L+[pN]-1} - S'_{k-1}).$$

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

$$\frac{\sqrt{A}}{\sqrt{2\pi\sigma^2}u} \int_{-\infty}^{\infty} e^{-\frac{(u-\sqrt{A}w/u)^2}{2\sigma^2}} \mathbf{P}\left(M(k,L) > w\right) ds$$

$$= \frac{\sqrt{A}}{\sqrt{2\pi\sigma^2}u} e^{-\frac{u^2}{2\sigma^2}} \int_{-\infty}^{\infty} e^{-\frac{Aw^2/u^2}{2\sigma^2}} e^{\frac{\sqrt{A}w}{\sigma^2}} \mathbf{P}\left(M(k,L) > w\right) dw. \qquad (2.2)$$

By weak convergence of a random walk to a Wiener process, for any w,

$$\lim_{pN\to\infty} \mathbf{P}\left(M(k,L) > w\right) = \mathbf{P}\left(\max_{\substack{0\le t\le 1 \\ 0\le s\le B/A}} W(t+s+1) - W(t) > w \right),$$

where $W(\cdot)$ is a standard Wiener process; using Lemma 1 [9] it is straightforward to show that

$$\mathbf{P}\left(M(k,L) > w\right) \le 2e^{-\frac{w^2}{24}}.$$

Thus, by dominated convergence

$$\int_{-\infty}^{\infty} e^{-\frac{Aw^2/u^2}{2\sigma^2}} e^{\frac{\sqrt{A}w}{\sigma^2}} \mathbf{P}\left(M(k,L) > w\right) dw$$

$$= \int_{-\infty}^{\infty} e^{\sqrt{A}w} \mathbf{P}\left(\max_{\substack{0\le t\le 1 \\ 0\le s\le B/A}} (W(t+s+1) - W(t)) > w \right) dw + o(1)$$

$$= \frac{1}{\sqrt{A}} \mathbf{E} \exp\left(\max_{\substack{0\le t\le A \\ 0\le s\le B}} W(t+s+A) - W(t) \right) + o(1).$$

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 move on from the rectangle $[0, pN] \times ((1-q)N, N]$ to the layer $[0, TN] \times ((1-q)N, N]$. We use Lemma 2.1 and 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

$$(Tp^{-1}+1)\mathbf{P}\left(\max_{\Delta_0}\zeta_L^{(N)}(k) > u\right) \geq \mathbf{P}\left(\max_{\substack{0<k\leq TN\\(1-q)N<L\leq N}}\zeta_L^{(N)}(k) > u\right)$$

$$\geq (Tp^{-1}-1)\mathbf{P}\left(\max_{\Delta_0}\zeta_L^{(N)}(k) > u\right)$$

$$- \sum_{\substack{0\leq l,m\leq Tp^{-1}+1\\ l\neq m}} \mathbf{P}\left(\max_{\Delta_l}\zeta_L^{(N)}(k) > u, \max_{\Delta_m}\zeta_L^{(N)}(k) > u\right).$$

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 Bonferroni inequality will be asymptotically equivalent. The estimates of $p_{0,n}$ are obtained in the same manner as in [9]. The proof will be divided into four parts.

Case 1.1: $1 \leq n < n_0$. The value of n_0 will be chosen later. We have:

$$p_{0,n} \leq \mathbf{P}\left(\max_{\substack{0<k\leq pN(n+1)/2\\(1-q)N<L\leq N}}\zeta_L^{(N)}(k) > u, \max_{\substack{pN(n+1)/2<k\leq pN(n+1)\\(1-q)N<L\leq N}}\zeta_L^{(N)}(k) > u\right)$$

$$= 2\mathbf{P}\left(\max_{\substack{0<k\leq pN(n+1)/2\\(1-q)N<L\leq N}}\zeta_L^{(N)}(k) > u\right) - \mathbf{P}\left(\max_{\substack{0<k\leq pN(n+1)\\(1-q)N<L\leq N}}\zeta_L^{(N)}(k) > u\right).$$

Applying Lemma 2.1 we obtain that

$$p_{0,n} \le \frac{1}{\sqrt{2\pi}u}(2H^B_{A(n+1)/2} - H^B_{A(n+1)})e^{-\frac{u^2}{2}}(1+g_n(u,N)), \qquad (2.3)$$

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

$$\eta = \frac{1}{\sqrt{N}} \sum_{i=(n+1)pN+1}^{(1-q)N} \xi_i, \quad \zeta_1 = \frac{1}{\sqrt{N}} \sum_{i=pN+1}^{npN} \xi_i, \quad \zeta_2 = \frac{1}{\sqrt{N}} \sum_{i=pN+N+1}^{npN+(1-q)N} \xi_i.$$

Then, postponing the explanation of the last equality,

$$p_{0,n} = \mathbf{P}\left(\eta + \zeta_1 + \max_{\Delta_0} \frac{1}{\sqrt{N}} \left(\sum_{i=k}^{pN} + \sum_{i=npN+1}^{(n+1)pN} + \sum_{i=(1-p)N+1}^{k+L-1}\right) \xi_i > u,\right.$$

$$\left. \eta + \zeta_2 + \max_{\Delta_n} \frac{1}{\sqrt{N}} \left(\sum_{i=k}^{(n+1)pN} + \sum_{i=(1-q)N+1}^{pN+N} + \sum_{i=npN+(1-q)N+1}^{k+L-1}\right) \xi_i > u\right)$$

$$= \mathbf{P}\left(\eta + \zeta_1 + \max_{\Delta_0} \zeta_L^{'}(k) > u, \quad \eta + \zeta_2 + \max_{\Delta_0} \zeta_L^{''}(k) > u\right),$$

where

$$\zeta_L^{'}(k) = \frac{1}{\sqrt{N}} \left(\sum_{i=k}^{k+L-(1-q-2p)N-1} \xi_i^{'}\right)$$

and (2.4)

$$\zeta_L^{''}(k) = \frac{1}{\sqrt{N}} \left(\sum_{i=k}^{k+L-(1-2q-2p)N-1} \xi_i^{''}\right).$$

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 $\eta,\ \zeta_1,\ \zeta_2$ 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 will be essential is that $\eta,\ \zeta_1,\ \zeta_2$, 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

$$\begin{aligned} p_{0,n} &\le \mathbf{P}\left(2\eta+\zeta_1+\zeta_2+\max\zeta'+\max\zeta''>2u\right) \\ &= \tfrac{1}{\sqrt{2\pi\sigma^2}}\int\limits_{-\infty}^{\infty}\mathbf{P}\left(\tfrac{\zeta_1+\ \zeta_2}{2}+\tfrac{\max\zeta'+\ \max\zeta''}{2}>u-v\right)e^{-\frac{v^2}{2\sigma^2}}\,dv, \end{aligned}$$

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

$$\begin{aligned} p_{0,n} &\le \tfrac{\sqrt{A}}{\sqrt{2\pi\sigma^2}u}e^{-\frac{u^2}{2\sigma^2}}\int\limits_{-\infty}^{\infty}\mathbf{P}\left(\tfrac{\zeta_1+\ \zeta_2}{2\sqrt{p}}+\tfrac{\max\zeta'+\ \max\zeta''}{2\sqrt{p}}>s\right)e^{\frac{\sqrt{A}s}{\sigma^2}}\,ds \\ &= \tfrac{\sigma}{\sqrt{2\pi}u}e^{-\frac{u^2}{2\sigma^2}}\mathbf{E}e^{\frac{\sqrt{A}}{\sigma^2}\left(\frac{\zeta_1+\ \zeta_2}{2\sqrt{p}}+\frac{\max\zeta'+\ \max\zeta''}{2\sqrt{p}}\right)} \\ &= \tfrac{\sigma}{\sqrt{2\pi}u}e^{-\frac{u^2}{2\sigma^2}}\mathbf{E}e^{\frac{\sqrt{A}}{\sigma^2}\left(\frac{\zeta_1+\ \zeta_2}{2\sqrt{p}}\right)}\mathbf{E}e^{\frac{\sqrt{A}}{\sigma^2}\left(\frac{\max\zeta'+\ \max\zeta''}{2\sqrt{p}}\right)}. \quad (2.5) \end{aligned}$$

We will 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

$$\frac{\sigma}{\sqrt{2\pi}u}e^{-\frac{u^2}{2\sigma^2}}\le\frac{2}{\sqrt{2\pi}u}e^{-\frac{u^2}{2}}e^{-\frac{A(n+1)+B}{2}}.$$

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

$$\mathbf{E}e^{\frac{\sqrt{A}}{\sigma^2}\left(\frac{\zeta_1+\ \zeta_2}{2\sqrt{p}}\right)}\le e^{\frac{A(n-1)}{4\sigma^4}}.$$

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 $S'_k=\sum\limits_{i=1}^{k}\xi'_i$ and $S''_k=\sum\limits_{i=1}^{k}\xi''_i$ we see that the latter equals

$$\mathbf{P}\left(\max_{\Delta_0}\left(S'_{k+L-(1-q-2p)N-1}-S'_{k-1}\right)+\right.$$

$$+\max_{\Delta_0}\left(S^{''}_{k+L-(1-2q-2p)N-1}-S^{''}_{k-1}\right)>2\sqrt{pN}s\Bigg)$$

$$\le \mathbf{P}\left(\max_{\Delta_0}S^{'}_{k+L+(q+2p)N-N-1}+\max_{0<k\le pN}-S^{'}_{k-1}+\right.$$

$$\left.+\max_{\Delta_0}S^{''}_{k+L+(2q+2p)N-N-1}+\max_{0<k\le pN}-S^{''}_{k-1}>2\sqrt{pN}s\right)$$

$$\le 4\mathbf{P}\left(\max_{0<k\le(2q+3p)N}S^{'}_{k}>\frac{\sqrt{pN}}{2}s\right)\le 4e^{-\frac{1}{8}\left(\frac{A}{3A+2B}s^2\right)}<4e^{-\frac{s^2}{24}}, \quad (2.6)$$

where we applied Lemma 1 [9] in the second to the last step.
Thus, for any positive t we obtain the following estimate

$$\mathbf{E}e^{t\left(\frac{\max\zeta'+\max\zeta''}{2\sqrt{p}}\right)}=\int\limits_{-\infty}^{\infty}te^{ts}\mathbf{P}\left(\frac{\max\zeta^{'}+\max\zeta^{''}}{2\sqrt{p}}>s\right)ds\le 1+4t\int\limits_{0}^{\infty}e^{ts-\frac{s^2}{24}}ds$$

$$=1+4te^{6t^2}\int\limits_{0}^{\infty}e^{-\frac{(s-12t)^2}{24}}ds\le 1+4\sqrt{24\pi}te^{6t^2}. \quad (2.7)$$

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

$$p_{0,n}\le\frac{\frac{16\sqrt{24\pi}}{\sigma^2}\sqrt{A}}{\sqrt{2\pi}u}e^{-\frac{u^2}{2}}e^{-An\left(\frac{1}{2}-\frac{1}{4\sigma^4}\right)+A\left(\frac{23}{4\sigma^4}-\frac{1}{2}\right)-\frac{B}{2}}.$$

Owing to the restriction $n_0\le n\le\varepsilon p^{-1}-1$ 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

$$p_{0,n}\le\frac{C_1\sqrt{A}}{\sqrt{2\pi}u}e^{-\frac{u^2}{2}}e^{-A\frac{n-43}{6}-\frac{B}{2}}, \quad (2.8)$$

where C_1 is some positive constant.

Case 1.3: $\varepsilon p^{-1}\le n\le p^{-1}+1$. In much the same way the representation (2.4) gives

$$p_{0,n}\le\mathbf{P}\left(2\eta+\zeta_1+\zeta_2+\max_{\Delta_0}\zeta'_L(k)+\max_{\Delta_0}\zeta''_L(k)>2u\right).$$

However, when $n\ge\varepsilon p^{-1}$, 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

$$p_{0,n} \le e^{-tu}\mathbf{E}e^{t\left(\eta+\frac{\zeta_1+\zeta_2}{2}+\frac{\max\zeta'+\max\zeta''}{2}\right)}$$

$$= e^{-tu}\mathbf{E}e^{t\left(\eta+\frac{\zeta_1+\zeta_2}{2}\right)}\mathbf{E}e^{t\left(\frac{\max\zeta'+\max\zeta''}{2}\right)}. \quad (2.9)$$

Although ζ' and ζ'' may contain smaller number of summands, it can be seen that this does sufficiently change the proof of (2.6). Thus the estimate (2.7) remains valid and

$$\mathbf{E}e^{t\left(\frac{\max\zeta'+\max\zeta'}{2}\right)} < 1 + 4\sqrt{24\pi}t\sqrt{p}e^{6t^2p}. \quad (2.10)$$

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 $(\varepsilon-p)/2 \le (n-1)p/2 \le 1/2$,

$$\mathbf{E}e^{t\left(\eta+\frac{\zeta_1+\zeta_2}{2}\right)} \le e^{\frac{t^2\max\left\{\frac{(n-1)p}{2}, 1-\frac{(n-1)p}{2}\right\}}{2}} < e^{\frac{t^2(1-\varepsilon/4)}{2}}. \quad (2.11)$$

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:

$$p_{0,n} \le C_2\sqrt{A}e^{6A}e^{-\frac{u^2}{2\left(1-\frac{\varepsilon}{4}\right)}}. \quad (2.12)$$

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

$$p_{0,n} \le 2(H_A^B)^2\psi(u)^2. \quad (2.13)$$

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

$$2(Tp^{-1}+1)\sum_{n=1}^{Tp^{-1}+1} p_{0,n} \le 2(Tp^{-1}+1) \times$$

$$\times \left\{ \left(\sum_{n=1}^{n_0-1} \left(2H^B_{A(n+1)/2} - H^B_{A(n+1)} \right) (1 + g_n(u, N)) + \right.\right.$$
$$\left. + \sum_{n=n_0}^{\infty} C_1 \sqrt{A} e^{-A\frac{n-43}{6} - \frac{B}{2}} \right) p_0(u) +$$
$$\left.\left. + p^{-1} C_2 \sqrt{A} e^{6A} e^{-\frac{u^2}{2\left(1-\frac{\varepsilon}{4}\right)}} + Tp^{-1} 2 (H^B_A)^2 \psi(u)^2 \right\}.$$

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 2 we conclude that

$$\overline{\lim_{u,N}} \mathbf{P} \left(\max_{\substack{0<k\le TN \\ (1-q)N<L\le N}} \zeta_L^{(N)}(k) > u \right) \Bigg/ Tu^2 p_0(u) \le A^{-1} H^B_A$$

and (2.14)

$$\underline{\lim_{u,N}} \mathbf{P} \left(\max_{\substack{0<k\le TN \\ (1-q)N<L\le N}} \zeta_L^{(N)}(k) > u \right) \Bigg/ Tu^2 p_0(u) \ge A^{-1} H^B_A -$$
$$- 2A^{-1} \sum_{n=1}^{n_0-1} \left(2H^B_{A(n+1)/2} - H^B_{A(n+1)} \right) - 2\frac{C_1 e^{-\frac{B}{2}}}{\sqrt{A}} \sum_{n=n_0}^{\infty} e^{-A\frac{n-43}{6}}.$$

It was proved in [13] that the limit

$$H^B = \lim_{A\to\infty} A^{-1} H^B_A, \quad H^B > 0$$

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 to be 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,

$$\mathbf{P} \left(\max_{\substack{0<k\le TN \\ (1-q)N<L\le N}} \zeta_L^{(N)}(k) > u \right) = H^B Tu^2 p_0(u)(1 + o(1)). \tag{2.15}$$

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 the maximum 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

$$\mathbf{P}\left(\max_{\substack{0<k\le TN\\ 0<L\le(1-q)N}}\zeta_L^{(N)}(k)>u\right)$$
$$\le (Tp^{-1}+1)\sum_{n=1}^{p^{-1}-1}\mathbf{P}\left(\max_{\substack{0<k\le pN\\ (1-(n+1)q)N<L\le(1-nq)N}}\zeta_L^{(N)}(k)>u\right). \quad (2.16)$$

Let p_n denote 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),

$$p_n=\frac{\sqrt{A}}{\sqrt{2\pi\sigma'^2}u}e^{-\frac{u^2}{2\sigma'^2}}\int_{-\infty}^{\infty}e^{-\frac{Aw^2/u^2}{2\sigma'^2}}e^{\frac{\sqrt{A}w}{\sigma'^2}}\mathbf{P}\left(M(k,L)>w\right)dw, \quad (2.17)$$

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) can be rewritten as

$$\frac{\sqrt{A}}{\sqrt{2\pi\sigma'^2}u}e^{-\frac{u^2}{2\sigma^2}(1+z)}\int_{-\infty}^{\infty}e^{-\frac{Aw^2/u^2}{2\sigma^2}(1+z)+\frac{\sqrt{A}w}{\sigma^2}z}e^{\frac{\sqrt{A}w}{\sigma^2}}\mathbf{P}\left(M(k,L)>w\right)dw.$$

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

$$p_n \le \frac{\sqrt{A}}{\sqrt{2\pi\sigma'^2}u} e^{-\frac{u^2}{2\sigma^2}K} \int\limits_{-\infty}^{\infty} e^{\frac{\sqrt{A}w}{\sigma^2}} \mathbf{P}\left(M(k,L) > w\right) dw,$$

where

$$K = 1 + \frac{z}{1+z} = 1 + \frac{nq}{1-q-p} \ge 1 + nq.$$

Finally, owing to Lemma 2.1 there exists a constant C such that

$$p_n \le \frac{\sigma}{\sigma'} e^{-\frac{nB}{2}} H_A^B \frac{1}{\sqrt{2\pi}u} e^{-\frac{u^2}{2}}(1+o(1)) \le Ce^{-\frac{nB}{2}} H_A^B p_0(u),$$

where $o(1) \to 0$ uniformly in n when $u, N \to \infty$.

Case 2.2: $np \ge \frac{13}{16}$. Now σ'^2 can be arbitrary small and we estimate p_n using Lemma 1 of [9]:

$$p_n \le \mathbf{P}\left(\max_{\substack{0<k\le pN \\ 0<L\le \frac{3}{16}N}} \zeta_L^{(N)}(k) > u\right) \le 2\mathbf{P}\Big(\max_{0<k\le \frac{3}{16}N+pN} S_k > \tfrac{1}{2}u\sqrt{N}\Big)$$
$$\le 2e^{-\frac{u^2}{4(\frac{3}{16}+p)}} \le 2e^{-u^2}.$$

Thus, combining the estimates for p_n obtained in cases 2.1 and 2.2, with (2.16) and (2.15) we have

$$\overline{\lim_{u,N}} \mathbf{P}\left(\max_{\substack{0<k\le TN \\ 0<L\le N}} \zeta_L^{(N)}(k) > u\right) \Big/ \left(Tu^2 p_0(u)\right) \le H^B + A^{-1} H_A^B C \sum_{n=1}^{\infty} e^{-\frac{nB}{2}}$$

and

$$\underline{\lim_{u,N}} \mathbf{P}\left(\max_{\substack{0<k\le TN \\ 0<L\le N}} \zeta_L^{(N)}(k) > u\right) \Big/ \left(Tu^2 p_0(u)\right) \ge H^B.$$

It was proved in [13] 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 upper and lower limits coincide and equal H. This finishes the proof of Theorem 1.3.

3 Very large deviations of Shepp statistics.

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 a 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 and ξ_2 be 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

$$\begin{aligned}
I(u)&=\mathbf{P}\left(\xi_1>u,\alpha\xi_1+\zeta>u\right)\\
&=\tfrac{1}{\sqrt{2\pi}}\int_u^\infty e^{-\frac{v^2}{2}}\mathbf{P}(\zeta>u-\alpha v)\,dv=-\tfrac{1}{\sqrt{2\pi}}\int_u^\infty\tfrac{\mathbf{P}(\zeta>u-\alpha v)}{v}\,de^{-\frac{v^2}{2}}\\
&=-\tfrac{\mathbf{P}(\zeta>u-\alpha v)}{\sqrt{2\pi}v}e^{-\frac{v^2}{2}}\Big|_u^\infty+\tfrac{1}{\sqrt{2\pi}}\int_u^\infty e^{-\frac{v^2}{2}}\,d\tfrac{\mathbf{P}(\zeta>u-\alpha v)}{v}\\
&=\tfrac{e^{-\frac{u^2}{2}}}{\sqrt{2\pi}u}\mathbf{P}\left(\zeta>u(1-\alpha)\right)+\tfrac{1}{\sqrt{2\pi}}\int_u^\infty e^{-\frac{v^2}{2}}\left(\alpha\tfrac{\varphi_\zeta(u-\alpha v)}{v}-\tfrac{\mathbf{P}(\zeta>u-\alpha v)}{v^2}\right)dv.
\end{aligned}$$

Write $K(u)$ for the first summand in the last expression. The second summand is less than

$$\frac{\alpha}{\sqrt{2\pi}u}\int_u^\infty e^{-\frac{v^2}{2}}\varphi_\zeta(u-\alpha v)\,dv$$

and thus $I(u)$ is bounded by

$$\begin{aligned}
&K(u)+\tfrac{\alpha}{\sqrt{2\pi}u}\int_u^\infty\tfrac{1}{\sqrt{2\pi(1-\alpha^2)}}e^{-\frac{1}{2}\left(v^2+\frac{(u-\alpha v)^2}{1-\alpha^2}\right)}\,dv\\
&\quad=K(u)+\alpha\tfrac{e^{-\frac{u^2}{2}}}{\sqrt{2\pi}u}\int_u^\infty\tfrac{1}{\sqrt{2\pi(1-\alpha^2)}}e^{-\frac{1}{2}\frac{(v-\alpha u)^2}{1-\alpha^2}}\,dv\\
&\quad\quad=K(u)+\alpha K(u)=\tfrac{1}{\sqrt{2\pi}u}e^{-\frac{1}{2}u^2}(1+\alpha)\mathbf{P}\left(\tfrac{\zeta}{\sqrt{1-\alpha^2}}>u\tfrac{\sqrt{1-\alpha}}{\sqrt{1+\alpha}}\right).
\end{aligned}$$

The lemma now follows from the standard upper estimate of standard normal distribution tail. □

Next, we estimate (3.1) using Bonferroni inequality:

$$[TN]\mathbf{P}\left(\zeta_N^{(N)}(1) > u\right) \geq \mathbf{P}\left(\max_{0<k\leq TN} \zeta_N^{(N)}(k) > u\right)$$
$$\geq [TN]\mathbf{P}\left(\zeta_N^{(N)}(1) > u\right) - \sum_{\substack{1\leq l,m\leq TN \\ l\neq m}} \mathbf{P}\left(\zeta_N^{(N)}(l) > u, \zeta_N^{(N)}(m) > u\right).$$

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

$$2TN \sum_{n=2}^{TN} \mathbf{P}\left(\zeta_N^{(N)}(1) > u, \zeta_N^{(N)}(n) > u\right) < 2TN \sum_{n=N+1}^{TN} \mathbf{P}\left(\zeta_N^{(N)}(1) > u\right)^2$$
$$+ 2TN \sum_{n=2}^{N} \tfrac{1}{\sqrt{2\pi}u} e^{-\frac{1}{2}u^2} \tfrac{1}{\sqrt{2\pi}u}(1+\alpha_n)\tfrac{\sqrt{1+\alpha_n}}{\sqrt{1-\alpha_n}} e^{-\frac{1}{2}u^2\frac{1-\alpha_n}{1+\alpha_n}}.$$

As before let $p_0(u)$ denote $\frac{1}{\sqrt{2\pi}u}e^{-\frac{1}{2}u^2}$, the asymptotic bound for 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

$$1 + o(1) \geq \frac{\mathbf{P}\left(\max_{0<k\leq TN} \zeta_N^{(N)}(k) > u\right)}{[TN]p_0(u)} \geq 1 - 4TNp_0(u)(1+o(1)) + o(1).$$

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

$$\mathbf{P}\left(\max_{\substack{0<k\le TN\\ 0<L\le N-1}} \zeta_L^{(N)}(k) > u\right) \le TN \sum_{L=1}^{N-1} \mathbf{P}\left(\zeta_L^{(N)}(1) > u\right)$$
$$\le TN \sum_{L=1}^{N-1} p_0\left(u\sqrt{\tfrac{N}{L}}\right) = TNp_0(u) \sum_{L=1}^{N-1} e^{-\frac{u^2(N-L)}{2L}}$$
$$\le TNp_0(u) \sum_{L=1}^{N-1} \left(e^{-\frac{u^2/N}{2}}\right)^{N-L} = o(TNp_0(u)).$$

Combining this estimate and (3.2) we get 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 [13] 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 [6] and was reformulated in Theorem 4.1 for comparison purpose.

Acknowledgement:
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

[1] Dembo, A., Karlin, S. and Zeitouni, O. Limit Distribution of Maximal Non-Aligned Two-Sequence Segmental Score The Annals of Probability, 1994, v. 22, no. 4, pp. 2022-2039.

[2] Erdös P., Rényi A. On a new law of large numbers.-Annal.Math., 1970, v. 23, p.103-111.

[3] Frolov, A.N., Limit Theorems for Increments of Sums of Independent Random Variables, Theory of Probability and its Applications, 2004, v. 48, no 1, pp. 93-107.

[4] Kabluchko, Z., Extreme-Value Analysis of Standardized Gaussian Increments, 2008, Bernoulli. http://arxiv.org/abs/0706.1849

[5] Kozlov, V.M., On the Erdos-Renyi Partial Sums: Large Deviations, Conditional Behavior, Theory of Probability and its Applications, 2002, v. 46, no 4, pp. 636-651.

[6] Kozlov A.M. On large deviations for the Shepp statistic. Discrete Mathematics and Applications, 2004, v. 14, no 2, pp. 211-216

[7] Leadbetter, M. R., Lindgren, G., Rootzén, H., Extremes and Related Properties of Random Sequences andProcesses, Springer-Verlag, 1983.

[8] Piterbarg V.I. Asymptotic methods in Theory of Gaussian Processes and Fields. Translation of mathematical Monographs, AMS, Providence, Rhode island,1996.

[9] Piterbarg, V.I., On large jumps of a random walk, 1991, Theory of probability and its applications, v. 36, no 1, pp. 50-62.

[10] Piterbarg, V.I., Kozlov A.M., On large jumps of a Cramer random walk, 2002, Theory of Probability and its Applications, v. 47, no 4, pp. 719-729.

[11] Shepp L.A. A limit law concerning moving averages.-Ann. Math. Statist., 1964, v.35, pp. 424-428.

[12] Waterman, M.S. Introduction to Computational Biology. Chapman&Hall, 1995.

[13] Zholud, D.S., Extremes of Shepp statistics for the Wiener Process, 2008, Extremes, v. 11, no 4, pp. 339-351.

PAPER III

Extremes of Student's t-statistics for non-normal and not necessarily i.i.d. random variables.

DMITRII ZHOLUD
Department of Mathematical Statistics, University of Gothenburg,
Department of Mathematical Statistics, Chalmers University of Technology,
Göteborg, Sweden. *dmitrii@math.chalmers.se*

Abstract. Let $X = (X_1, X_2, .., X_n),\ n \geq 2$, be a random vector with continuous joint density g. Consider Student's t-statistic $T_n = \sqrt{n}\frac{\overline{X}-\mu_0}{\sqrt{S^2}}$, where $\overline{X}$ and S^2 stand for the sample mean and the sample variance respectively. We determine asymptotic expressions for $\mathbf{P}(T_n > u)$ when $u \to \infty$ and show they are accurate for small n. This gives a basis for new methods to correct theoretical p-values in high-throughput screenings, where sample size can be as low as two to five. The results are complemented by the examples and a simulation study.

Key words. Student's t-statistic, self normalized sums, asymptotic behavior, extreme values, high-throughput screening, false positive/discovery rate, test power, small sample size, non-normal, skewed, heavy tailed, dependent, non-stationary, non-i.i.d.

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

1 Introduction

The origin of this article was the paper [18] which studies systematic errors in a particular kind of biological experiments (so called BioScreen array experiments, see [2] and [3]) and their impact on false positive and false discovery rates. Omitting details, the parameter of interest, called LSC, was assumed to be normally distributed if the null hypothesis was true. However, a histogram of the LSC values in a wild type data set (for which the null hypothesis is known to be true) in fact is somewhat skewed. We therefore made a p-p plot of all the (one sample) t-statistics computed in this experiment, see Fig. 1. Each test was based on two LSC replicates and the p-p plot showed clear deviations from the theoretical t_1 distribution. Interestingly, both the lower and upper tails of the plot approached straight lines, as indicated by the two arrows in Fig. 1. However the slopes of these lines were different from the theoretical 45° slope. Similar behavior was observed in a number of related experiments.

The aim of the present paper is to give a theoretical explanation of this phenomenon, and to give a basis for new methods to correct theoretical

p-values - the latter will be pursued further in a subsequent paper.

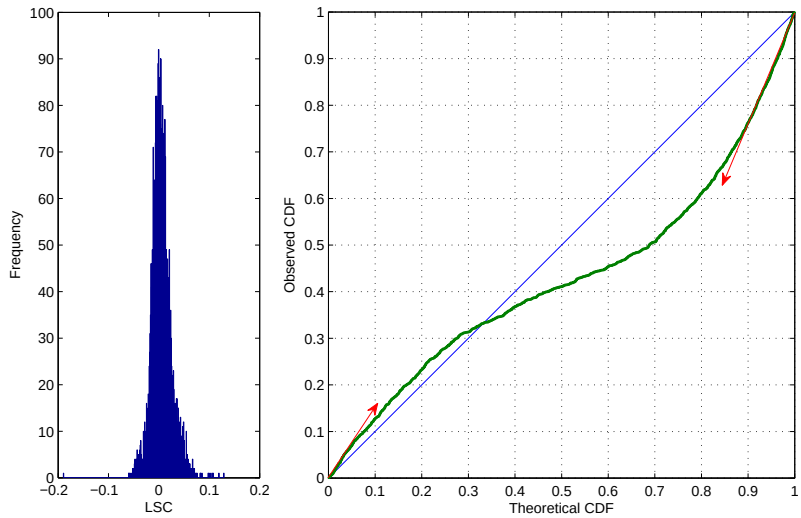

Figure 1: *A histogram of* 3456 *LSC values from the wild type dataset (left). The p-p plot for the corresponding t-tests (right). The diagonal line corresponds to* t_1 *distribution.*

It is worth mentioning that the data in [18] showed similar behavior for a two-sample t-test as well. The corresponding theoretical problem will also be studied in a later paper.

Now follows the mathematical formulation of the problem. Let $X = (X_1, X_2, ..X_n)$, $n \geq 2$, be a random vector with independent and normally distributed components with mean μ and variance σ^2. The t-test of the hypothesis $H_0 : \mu = \mu_0$ is based on

$$T_n = \sqrt{n}\frac{\overline{X} - \mu_0}{\sqrt{S^2}},$$

where $\overline{X} = 1/n \sum_{i=1}^{n} X_i$ and $S^2 = 1/(n-1) \sum_{i=1}^{n} (X_i - \overline{X})^2$.

Assuming the null hypothesis is true, the statistics T_n has a Student's t-distribution with $n-1$ degrees of freedom. The t-distribution function is given by

$$1 - t_{n-1}(x) = \mathbf{P}\left(T_n \leq x\right) = K \int_{-\infty}^{x} (1 + t^2/(n-1))^{-n/2} dt,$$

where $K = \frac{\Gamma(\frac{n}{2})}{\sqrt{(n-1)\pi}\Gamma(\frac{n-1}{2})}$ and Γ stands for the Gamma function.

In hypothesis testing, tail probabilities of the random variable T_n and accurate approximations of those are important. New high-throughput screening methods used in e.g drug discovery may lead to millions of biochemical, genetic or pharmacological tests and the sample size used in each test may then have to be as small as 2-5. This often makes a "standard" Normal approximation inapplicable. Furthermore, the significance level in a multiple testing situations then often is smaller than say 0.001. This motivates study of the *asymptotic* behavior of the tail t-statistic distribution for *small sample size* n.

When the sample size is small, the Normal approximation and the somewhat more accurate Edgeworth expansion (see e.x. [6], [7]) perform poorly, especially in the tail area. The so-called saddlepoint approximation is more accurate. We cite [14]: *" To overcome the difficulties encountered by the Normal approximation and the Edgeworth expansion, one can consider using a saddlepoint approximation, which provides a very good approximation to the tail, as well as in the center of the distribution. By a 'good' approximation, here, we imply one with a small relative error. By comparison, the Edgeworth expansion gives only absolute errors. "* Except for numerical results, however, the saddlepoint approximation [14], [15] does not provide any estimates for the relative error. It provides good but complicated approximations and does not tell anything explicit about the behavior of the t-statistics as one goes further out in the tail.

Assume now $X = (X_1, X_2, .., X_n),\ n \geq 2$ is a general random vector, where components need not be independent or identically distributed. Let $g(x_1, x_2, ..x_n)$ be its density function, which we assume to be continuous. We will show that under some (quite mild) regularity conditions on g,

$$\frac{\mathbf{P}\left(T_n > u\right)}{t_{n-1}(u)} = K_g + o(1) \ \text{ as } \ u \to \infty, \tag{1.1}$$

with an exact expression for the constant K_g given in Theorem 2.1.

Note that neither independence nor stationarity of the r.v.-s that constitute the vector X is required. The regularity assumptions for g make the result applicable to a wide class of densities. In particular this includes all the examples in numerical section of [14]. Furthermore, the assumption of continuity of g can be weakened, see Corollary 2.1.3.

For $K_g > 0$, equation (1.1) implies that the relative error of approximation of the probability $\mathbf{P}\left(T_n > u\right)$ by $K_g t_{n-1}(u)$ tends to zero as $u \to \infty$.

Under some extra assumptions on g, we obtain the limit behavior of the absolute error in Theorem 2.2. Together with (1.1) this gives that the relative error converges to zero as $O(1/u^2)$ when $u \to \infty$. Theorem 2.2 can be viewed as a more accurate approximation formula. In contrast to other approximation methods, the speed of convergence in (1.1) is inversely proportional to n. This is important for high throughput screening methods.

Consider a right-tailed t-test and *small enough* significance level α. Assume normality and independence under H_0 and let g be the distribution under H_1. Equation (1.1) then gives that the test power and false discovery rate (FDR) are approximately

$$K_g\alpha \quad \text{and} \quad \frac{m_0}{m_0 + K_g(m - m_0)},$$

where m_0 is the number of true null hypotheses and $m-m_0$ is the number of false null hypotheses. Note that the power of the test is proportional to the significance level, and that, perhaps in contradiction to what could be expected, FDR does not improve as one goes further out in the tail. This observation can be extended to the case when the tests have not necessarily the same distribution under H_1.

The other way around, let g be the distribution under the null hypothesis. Using $1 - \alpha$ quantile of t-distribution, as if g was normal, is thus misleading if g in fact is not normal. By Theorem 2.1 the true false positive rate is instead approximately $K_g\alpha$. Given the proportionality constant K_g we can thus adjust the quantile. For the case when g is known, K_g may be calculated directly. However, a key point now is that in case of "unknown" g, the constant can be estimated from data under the null hypothesis. But this is a subject of the later paper. We just mention the advantage compared to other methods: the saddlepoint approximation, for example, requires an exact analytical expression for the g.

There is an extensive literature on approximations of t-statistics. Papers [12], [13] and [8] describe the saddlepoint approximation method in general, while [4], [14], [17] and [15] introduce saddlepoint approximation for t-statistics. Related approximation methods, such as normal approximation and Edgeworth expansion can be found in [6], [7], [5] and [10].

The structure of the paper is as follows. Section 2 contains the proof of the main result (1.1) and the limit expression for the relative error (see Theorem 2.1 and Theorem 2.2). The proofs are given in terms of self normalized sums and the corresponding formulas for t-statistics are presented in Corollary 2.1.3. Section 3 includes as examples the approximation formulas for normal, Cauchy, t_2 and centered exponential distributions and a simulation study of the relative errors.

2 Asymptotic expressions for $\mathbf{P}\left(T_n > u\right)$.

The aim of the current section is to establish an asymptotic formula that provides accurate tail approximation for the case when population distribution is not necessarily normal or i.i.d. Without loss of generality we assume that μ_0 in the definition of statistics T_n is zero. Otherwise we could just subtract μ_0 from each component of X and use the density $\tilde{g} = g(\mathbf{x} - \mu_0)$ instead. Variables in bold letters, if not defined otherwise, denote the corresponding vectors, i.e $\mathbf{x} = (x_1, x_2, .., x_n)$. We will do the computations in terms of the so-called self-normalized sum S_n/V_n where S_n and V_n^2 are defined to be $\sum X_i$ and $\sum X_i^2$ respectively. The standard identity

$$\{T_n \geq u\} = \left\{ \frac{S_n}{V_n} \geq u \left(\frac{n}{u^2 + n - 1} \right)^{1/2} \right\}$$

shows that the asymptotic behavior of $\mathbf{P}\left(T_n > u\right)$ as $u \to \infty$ can be derived from the asymptotic behavior of

$$p(\varepsilon) = \mathbf{P}\left(S_n/V_n \geq \sqrt{n}(1-\varepsilon)\right)$$

for $\varepsilon \to 0+$.

Theorem 2.1. *If g is continuous and there exists $c < \sqrt{n}$ such that*

$$\int_0^\infty h^{n-1} \max_{\substack{\sum x_i^2 = h^2 \\ \sum x_i > ch}} g(x_1, x_2, .., x_n) dh < \infty \tag{2.1}$$

then

$$\frac{p(\varepsilon)}{t_{n-1}(u(\varepsilon))} = K_g + o(1) \quad as \quad \varepsilon \to 0+, \tag{2.2}$$

where the constant K_g is given by

$$K_g = \frac{2(\pi n)^{n/2}}{\Gamma(\frac{n}{2})} \int_0^\infty h^{n-1} g(h, h, .., h)\ dh \tag{2.3}$$

and

$$u(\varepsilon) = \frac{\sqrt{n-1}(1-\varepsilon)}{\sqrt{1-(1-\varepsilon)^2}}. \tag{2.4}$$

Proof: The starting point of the proof is equality

$$p(\varepsilon) = \int_{D_1} g(\mathbf{x}) d\mathbf{x},$$

where $D_1 = \{\mathbf{x} :\ S_n/V_n \geq \sqrt{n}(1-\varepsilon)\}$ and $d\mathbf{x}$ is the standard notation for $dx_1 dx_2..dx_n$. Let $\mathbf{e_1}, \mathbf{e_2}, .., \mathbf{e_n}$ be the standard basis in $\mathbb{R}^n$, define a unit vector $\mathbf{I_n} = (1/\sqrt{n}, 1/\sqrt{n}, .., 1/\sqrt{n})$ and let A be some orthogonal linear operator such that

$$A(\mathbf{e_n}) = \mathbf{I_n}. \tag{2.5}$$

We now conduct a series of variable changes. First, changing coordinate system $\mathbf{x} = A\mathbf{y}$ we have $S_n = \sqrt{n}y_n \quad \text{and} \quad V_n^2 = \sum\limits_{i=1}^{n} y_i^2$, and therefore

$$p(\varepsilon) = \int\limits_{D_2} g(A\mathbf{y})d\mathbf{y},$$

where $D_2 = \left\{\mathbf{y} :\ \sum\limits_{i=1}^{n} y_i^2 < \frac{y_n^2}{(1-\varepsilon)^2},\ y_n > 0\right\}$.

Next, the variable change $y_i = t_i t_n \quad \text{for} \quad i = 1, 2, ..n-1 \quad \text{and} \quad y_n = t_n$ gives

$$p(\varepsilon) = \int\limits_{D_3} t_n^{n-1} g\left(t_n A(t_1, t_2, .., t_{n-1}, 1)\right) d\mathbf{t},$$

where $D_3 = \left\{\mathbf{t} :\ \sum\limits_{i=1}^{n-1} t_i^2 < \frac{1}{(1-\varepsilon)^2} - 1,\ t_n > 0\right\}$. From Fubini's theorem we obtain that

$$p(\varepsilon) = \int\limits_{B} \cdots \int\limits_{0}^{\infty} t_n^{n-1} g\left(t_n A(t_1, .., t_{n-1}, 1)\right) dt_n dt_1..dt_{n-1}, \tag{2.6}$$

where $B = \left\{(t_1, ..t_{n-1}) : \sum\limits_{i=1}^{n-1} t_i^2 < \frac{1}{(1-\varepsilon)^2} - 1\right\}$.

We now split B up into the disjoint union of two half-balls

$$B_1 = \{(t_1, ..t_{n-1}) \in B, t_1 \geq 0\} \quad \text{and} \quad B_2 = \{(t_1, ..t_{n-1}) \in B, t_1 < 0\}$$

and study the integrals over B_1 and B_2 separately. It will be seen that the integrals have the same asymptotic behavior when $\varepsilon \to 0+$. We start with the "upper" half-ball B_1. Introduce new variables $r, k_1, k_2, ..k_{n-2}, h$ by

$$t_1 = (n-1)^{1/2} r^{-1} \sqrt{1 - (k_1^2 + k_2^2 + ... + k_{n-2}^2)},$$

$$t_i = (n-1)^{1/2} r^{-1} k_{i-1}, \qquad \text{for} \quad 2 \leq i \leq n-1,$$

and

$$t_n = (n-1)^{-1/2} \frac{hr}{\sqrt{1 + r^2/(n-1)}}, \qquad h, r > 0.$$

By construction

$$\sum_{i=1}^{n-1} t_i^2 = (n-1)r^{-2}$$

and recalling the definition of $u(\varepsilon)$ in (2.4), it is straightforward to show that the equations above define a bijection between the sets

$$\left\{\sum k_i^2 < 1,\ r > u(\varepsilon), h > 0\right\} \quad \text{and} \quad B_1 \times \{t_n > 0\}.$$

The Jacobian determinant of the transformation is

$$(n-1)^{n/2-1} \frac{\left(1-\sum k_i^2\right)^{-1/2}}{r^{n-1}\sqrt{1+r^2/(n-1)}}$$

and according to the variable change above the integral (2.6) over B_1 equals

$$\int_{u(\varepsilon)}^{\infty} \int_{\sum k_i^2<1} \cdots \int \int_0^{\infty} \left(1+r^2/(n-1)\right)^{-n/2} H(r,k_1,..k_{n-2},h)\ dhdk_1...dk_{n-2}dr,$$

where

$$H(r,k_1,..k_{n-2},h) = \frac{\left(1-\sum k_i^2\right)^{-1/2}}{\sqrt{n-1}} \times$$

$$h^{n-1} g\left(\frac{h}{\sqrt{1+r^2/(n-1)}} A\left(\sqrt{1-\sum k_i^2}, k_1, ..k_{n-2}, r/\sqrt{n-1}\right)\right).$$

Letting

$$F(r) = \int_{\sum k_i^2<1} \cdots \int \int_0^{\infty} H(r,k_1,..k_{n-2},h)\ dhdk_1..dk_{n-2}, \tag{2.7}$$

we rewrite this as

$$\int_{u(\varepsilon)}^{\infty} \left(1+r^2/(n-1)\right)^{-n/2} F(r)dr. \tag{2.8}$$

Note that $u(\varepsilon) \to \infty$ as $\varepsilon \to 0+$ and thus our next aim is to study the asymptotic behavior of $F(r)$ when $r \to \infty$.

First we find the point-wise limit of $H(r, k_1, .., k_{n-2}, h)$ as $r \to \infty$. Introduce a unit vector

$$\mathbf{w}(k_1, k_2, .., k_{n-2}) = \left(\sqrt{1 - \sum k_i^2}, k_1, ..k_{n-2}, 0\right) \tag{2.9}$$

and set

$$x(r) = \frac{1}{\sqrt{1 + r^2/(n-1)}}. \tag{2.10}$$

Suppressing the argument of $\mathbf{w}$, we rewrite $H(r, k_1, ..k_{n-2}, h)$ as

$$\frac{\left(1 - \sum k_i^2\right)^{-1/2}}{\sqrt{n-1}} h^{n-1} g\left(hA\left(x(r)\mathbf{w} + \sqrt{1 - x(r)^2}\mathbf{e_n}\right)\right). \tag{2.11}$$

Now recall that the linear operator A in (2.11) is orthogonal, which gives $||A(\mathbf{w})|| = 1$. Thus noting that $x(r) \to 0$ as $r \to \infty$ and using (2.5) we get that for fixed h,

$$\lim_{r\to\infty} hA\left(x(r)\mathbf{w} + \sqrt{1 - x(r)^2}\mathbf{e_n}\right)$$
$$= \lim_{r\to\infty}\left[hx(r)A\left(\mathbf{w}\right) + h\sqrt{1 - x(r)^2}\mathbf{I_n}\right] = h\mathbf{I_n}.$$

By assumption, the density g is continuous and from (2.11) we conclude that

$$\lim_{r\to\infty} H(r, k_1, ..k_{n-2}, h) = \frac{\left(1 - \sum k_i^2\right)^{-1/2}}{\sqrt{n-1}} h^{n-1} g\left(h\mathbf{I_n}\right). \tag{2.12}$$

Next, we construct an integrable bound for H. The bound is obtained by replacing the factor

$$h^{n-1} g\left(hA\left(x(r)\mathbf{w} + \sqrt{1 - x(r)^2}\mathbf{e_n}\right)\right) \tag{2.13}$$

in (2.11) by

$$h^{n-1} \max_{\substack{\sum x_i^2 = h^2 \\ \sum x_i > ch}} g(x_1, x_2, .., x_n), \tag{2.14}$$

where constant c is defined by (2.1).

Indeed, using orthogonality of A, the argument of g in (2.13) satisfies

$$\left|\left|hA\left(x(r)\mathbf{w} + \sqrt{1 - x(r)^2}\mathbf{e_n}\right)\right|\right| = h$$

and the sum of its coordinates, owing to (2.5), is equal to

$$\left\langle hA\left(x(r)\mathbf{w} + \sqrt{1 - x(r)^2}\mathbf{e_n}\right), \sqrt{n}\mathbf{I_n}\right\rangle = h\sqrt{n}\sqrt{1 - x(r)^2} > ch,$$

where $\langle \cdot, \cdot \rangle$ stands for scalar product and the last inequality holds for u large enough. Thus (2.13) is bounded by (2.14).

We can now conclude that $F(r)$ is bounded by a constant as follows. Expression in (2.14) does not depend on $k_1, k_2, .., k_{n-2}$ and hence, the integral for the corresponding bound of H is a product of two finite integrals: the integral of (2.14) is finite according to (2.1) and the integral of the remaining factor (which does not depend on h) is finite and equals $I_{n-2}/\sqrt{n-1}$, where constant I_n is defined in Lemma 2.1.

The dominated convergence theorem, (2.12) and Lemma 2.1 below hence imply that

$$\lim_{r\to\infty} F(r) = \int_{\sum k_i^2<1} \cdots \int \int_0^\infty \frac{\left(1-\sum k_i^2\right)^{-1/2}}{\sqrt{n-1}} h^{n-1} g(h\mathbf{I_n})\ dh dk_1..dk_{n-2}$$

$$= \frac{I_{n-2}}{\sqrt{n-1}} \int_0^\infty h^{n-1} g(h\mathbf{I_n})\ dh = \frac{(\pi n)^{n/2}}{\sqrt{(n-1)\pi}\Gamma(\frac{n-1}{2})} \int_0^\infty h^{n-1} g(h,h,..,h)\ dh.$$

We now go back to (2.6) and estimate the integral over the set B_2. The calculations are similar, except that the variable change on page 6 is modified by taking $t_1 = -(n-1)^{1/2} r^{-1} \sqrt{1-(k_1^2+k_2^2+...+k_{n-2}^2)}$. It can be seen that further calculations are equivalent to estimating the integral over B_1 but with $\tilde{A}(x) = A((-x_1, x_2, .., x_n))$ instead of A. The linear operator $\tilde{A}$ is orthogonal and satisfies (2.5). Redefine $F(r)$ in (2.7) accordingly and call it $\tilde{F}(r)$. Since the limit of $F(r)$ does not depend on A, the limits for $\tilde{F}(r)$ and $F(r)$ coincide. Finally, $p(\varepsilon)$ is the sum of (2.8) and (2.8) with $F(r)$ replaced by $\tilde{F}(r)$. Noting that $u(\varepsilon) \to \infty$ as $\varepsilon \to 0+$ finishes the proof. □

To get a second order version of this result we next assume that $g \in \mathcal{C}^2(\mathbb{R}^n)$. Let ∇g and $Hess(g)$ denote the gradient and the Hessian matrix of g, respectively

$$\nabla g = \left(\frac{\partial g}{\partial x_1}, \frac{\partial g}{\partial x_2}, .., \frac{\partial g}{\partial x_n} \right)$$

and

$$Hess(g) = \begin{pmatrix} \frac{\partial^2 g}{\partial x_1 \partial x_1} & \frac{\partial^2 g}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 g}{\partial x_1 \partial x_n} \\ \frac{\partial^2 g}{\partial x_2 \partial x_1} & \frac{\partial^2 g}{\partial x_2 \partial x_2} & \cdots & \frac{\partial^2 g}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 g}{\partial x_n \partial x_1} & \frac{\partial^2 g}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 g}{\partial x_n \partial x_n} \end{pmatrix}.$$

Assume also that

$$\int_0^\infty h^n ||\nabla g|| dh < \infty \quad \text{and} \quad \int_0^\infty h^{n+1} \max_{\substack{\sum x_i^2=1 \\ \sum x_i=0}} \mathbf{x}\, Hess(g)\, \mathbf{x}^T dh < \infty, \quad (2.15)$$

where gradient and Hessian are computed at the point $(h, h, ..h)$. With the notation of Theorem 2.1 we then have the following result

Theorem 2.2. *If $g \in \mathcal{C}^2(\mathbb{R}^n)$ and satisfies (2.1) and (2.15), then*

$$\frac{p(\varepsilon) - K_g t_{n-1}(u(\varepsilon))}{t_{n+1}\left(\sqrt{\frac{n+1}{n-1}} u(\varepsilon)\right)} = M_g - L_g + o(1),$$

where

$$L_g = \frac{(n-1)(\pi n)^{n/2}}{\Gamma(\frac{n}{2})} \int_0^\infty h^n \frac{1}{n} \sum_i \frac{\partial g}{\partial x_i}\, dh$$

and

$$M_g = \frac{(n-1)(\pi n)^{n/2}}{\Gamma(\frac{n}{2})} \int_0^\infty h^{n+1} \left[\frac{1}{n} \sum_i \frac{\partial^2 g}{\partial^2 x_i} - \frac{1}{n(n-1)} \sum_{i \neq j} \frac{\partial^2 g}{\partial x_i \partial x_j} \right] dh.$$

All the partial derivatives are computed at the point $(h, h, ..h)$.

Proof: We start from equation (2.11). Our aim is to replace it by an asymptotic expression for $H(r, k_1, .., k_{n-2}, h)$ as $r \to \infty$. Next, we will obtain asymptotic expressions for $F(r)$ and $\tilde{F}(r)$ and then substitute them into (2.8). Recall that by $\tilde{F}(r)$ we denote the analog of (2.7) for the case when integral (2.6) is calculated over the set B_2, see Theorem 2.1. Define

$$f(x) = g\left(hA\left(x\mathbf{w} + \sqrt{1-x^2}\mathbf{e_n}\right)\right).$$

The derivatives $f'(x)$ and $f''(x)$ are then equal to

$$\left\langle \nabla g, hA(\mathbf{w}) - h\frac{x}{\sqrt{1-x^2}}\mathbf{I_n} \right\rangle$$

and

$$\left\langle \left(hA(\mathbf{w}) - h\frac{x}{\sqrt{1-x^2}}\mathbf{I_n}\right) Hess(g), hA(\mathbf{w}) - h\frac{x}{\sqrt{1-x^2}}\mathbf{I_n} \right\rangle -$$

$$- \left\langle \nabla(g), h\frac{1}{(1-x^2)^{3/2}}\mathbf{I_n} \right\rangle,$$

where $\nabla(g)$ and $Hess(g)$ are taken at the point $hA\left(x\mathbf{w}+\sqrt{1-x^2}\mathbf{e_n}\right)$. Taylor expansion for $f(x)$ at 0 gives

$$f(x)=g\left(h\mathbf{I_n}\right)+h\left\langle\nabla g,A(\mathbf{w})\right\rangle x+$$
$$+\frac{1}{2}\left[h^2\ \left\langle A(\mathbf{w})Hess(g),A(\mathbf{w})\right\rangle-h\langle\nabla g,\mathbf{I_n}\rangle\right]x^2+o(x^2), \tag{2.16}$$

where ∇g and $Hess(g)$ are computed at the point $h\mathbf{I_n}$. Setting $x=x(r)$ as defined by (2.10) and substituting it into (2.16) we get the asymptotic expression for $f(x(r))$ as $r\to\infty$. Substituting the latter into (2.11) we get the asymptotic expression for $H(r,k_1,..,k_{n-2},h)$. The regularity conditions (2.15) ensure that the dominated convergence theorem holds, and the asymptotic expression for $F(r)$ follows from (2.7). Replacing A by $\tilde{A}(x)=A((-x_1,x_2,..,x_n))$ we get the asymptotic expression for $\tilde{F}(r)$. Note that the second summand in (2.16) gives no contribution to the sum $F(r)+\tilde{F}(r)$ since the integral

$$\int\limits_{\sum k_i^2<1}\cdots\int\int\limits_0^\infty\frac{\left(1-\sum k_i^2\right)^{-1/2}}{\sqrt{n-1}}h^n\left\langle\nabla g,A(\mathbf{w})\right\rangle\ dhdk_1..dk_{n-2}$$

changes sign, but not absolute value, when A is replaced by $\tilde{A}$. The latter follows from the identity

$$\tilde{A}(\mathbf{w}(k_1,k_2,..,k_{n-2}))=-A(\mathbf{w}(-k_1,-k_2,..,-k_{n-2}))$$

and the symmetry of the ball $\sum k_i^2<1$. Analogously, the integrals of the third summand in (2.16) for A and $\tilde{A}$ coincide. This gives

$$F(r)+\tilde{F}(r)=\int\limits_{\sum k_i^2<1}\cdots\int\int\limits_0^\infty\frac{\left(1-\sum k_i^2\right)^{-1/2}}{\sqrt{n-1}}h^{n-1}\times$$
$$\times\left[2g\left(h\mathbf{I_n}\right)\ +h^2\left\langle A(\mathbf{w})Hess(g),A(\mathbf{w})\right\rangle x(r)^2-\right. \tag{2.17}$$
$$\left.-h\left\langle\nabla g,\mathbf{I_n}\right\rangle x(r)^2\right]\ dhdk_1..dk_{n-2}+x(r)^2o(1).$$

Substituting $F(r)+\tilde{F}(r)$ into (2.8) and using Lemma 2.1, straightforward calculations show that the first and the last summands in square brackets in (2.17) contribute to $p(\varepsilon)$ as

$$K_g t_{n-1}(u(\varepsilon))\quad\text{and}\quad-L_g t_{n+1}\left(\sqrt{\frac{n+1}{n-1}}u(\varepsilon)\right).$$

We next turn to the remaining summand and compute

$$\int\limits_{\sum k_i^2<1}\cdots\int\int\limits_0^\infty h^{n+1}\frac{\left\langle A(\mathbf{w})Hess(g),A(\mathbf{w})\right\rangle}{\sqrt{1-\sum k_i^2}}\ dhdk_1..dk_{n-2}. \tag{2.18}$$

The expression in the numerator is a quadratic form and can be written as

$$\sum_{1\le i\le j\le n} \alpha_{ij}(h) w_i w_j,$$

where w_i stands for i-th coordinate of the vector $\mathbf{w}$ and $\alpha_{ij}(h)$ are the coefficients computed at the point h. Note that since

$$w_j(k_1,..,-k_{j-1},..,k_{n-2}) = -w_j(k_1,..,k_{j-1},..,k_{n-2}) \text{ for } j>1$$

and

$$w_i(k_1,..,-k_{j-1},..,k_{n-2}) = w_i(k_1,..,k_{j-1},..,k_{n-2}) \text{ for } i\neq j,$$

then

$$\int\limits_{\sum k_i^2<1} \cdots \int\limits_{0}^{\infty} h^{n+1} \frac{\alpha_{ij}(h) w_i w_j}{\sqrt{1-\sum k_i^2}}\, dh dk_1..dk_{n-2} = 0 \text{ for } i\neq j.$$

Together with $w_n = 0$ the integral (2.18) then equals

$$\sum_{i=1}^{n-1} \int\limits_{\sum k_i^2<1} \cdots \int\limits_{0}^{\infty} h^{n+1} \frac{\alpha_{ii}(h) w_i^2}{\sqrt{1-\sum k_i^2}}\, dh dk_1..dk_{n-2},$$

and substituting $w_1 = \sqrt{1-\sum k_i^2}$, $w_i = k_{i-1}$ for $1<i<n$ and using Lemma 2.1 we get

$$I_{n-2} \int\limits_0^\infty h^{n+1} \alpha_{11}(h)\, dh + \sum_{i=2}^{n-1} \frac{I_{n-2}}{n-1} \int\limits_0^\infty h^{n+1} \left(\alpha_{ii}(h) - \alpha_{11}(h)\right)\, dh$$

$$= \frac{I_{n-2}}{n-1} \int\limits_0^\infty h^{n+1} \sum_{i=1}^{n-1} \alpha_{ii}(h)\, dh,$$

where I_n is defined in Lemma 2.1. Writing $\mathbf{A}$ for the matrix of the operator A, we have

$$\sum_{i=1}^{n} \alpha_{ii}(h) = tr(\mathbf{A}\ Hess(g)\mathbf{A}^T) = tr(Hess(g)) = \sum_i \frac{\partial^2 g}{\partial^2 x_i}$$

and

$$\alpha_{nn}(h) = \langle A(\mathbf{e_n}) Hess(g), A(\mathbf{e_n})\rangle = \frac{1}{n} \sum_{i,j} \frac{\partial^2 g}{\partial x_i \partial x_j}.$$

The integral (2.18) thus equals

$$I_{n-2} \int_0^\infty h^{n+1} \left[\frac{1}{n} \sum_i \frac{\partial^2 g}{\partial^2 x_i} - \frac{1}{n(n-1)} \sum_{i \neq j} \frac{\partial^2 g}{\partial x_i \partial x_j} \right] \, dh,$$

where all partial derivatives are computed at $h\mathbf{I_n}$. Substituting the expression above into (2.17) and then (2.17) into (2.8), simple algebra gives the remaining terms $M_g t_{n+1} \left(\sqrt{\frac{n+1}{n-1}} u(\varepsilon) \right)$ and $o(t_{n+1}(u(\varepsilon)))$. □

In the proof we have used the following lemma.

Lemma 2.1. *We have that*

$$I_n = \int_{\sum x_i^2 < 1} \frac{1}{\sqrt{1 - \sum x_i^2}} d\boldsymbol{x} = \int_{\sum x_i^2 < 1} \frac{(n+1)x_1^2}{\sqrt{1 - \sum x_i^2}} d\boldsymbol{x} = \frac{\pi^{(n+1)/2}}{\Gamma(\frac{n+1}{2})}.$$

Proof: Passing to spherical coordinates it can be seen that

$$I_n = \frac{\int_0^1 r^{n-1} \left(1 - r^2\right)^{-1/2} dr}{\int_0^1 r^{n-1} dr} \int_{\sum x_i^2 < 1} 1 \, d\mathbf{x}.$$

The integral on the right-hand side of the last equation is a volume of a unit n-ball and equals $\pi^{\frac{n}{2}} / \Gamma(\frac{n}{2} + 1)$. Combining this with

$$\int_0^1 r^{n-1} \left(1 - r^2\right)^{-1/2} dr = \frac{\sqrt{\pi} \Gamma(\frac{n}{2})}{2\Gamma(\frac{n+1}{2})}$$

we get the first equality. Similarly,

$$J_n = \frac{\int_0^1 r^{n+1} \left(1 - r^2\right)^{-1/2} dr}{\int_0^1 r^{n-1} dr} \frac{\int_0^\pi \cos^2(\varphi_1) \sin(\varphi_1)^{n-2} d\varphi_1}{\int_0^\pi \sin(\varphi_1)^{n-2} d\varphi_1} \int_{\sum x_i^2 < 1} 1 \, d\mathbf{x}$$

and direct calculation gives the second equality. □

Note that the dominated convergence theorem used in Theorems 2.1 and 2.2 can be relaxed to hold only almost everywhere and the assumption of continuity of g can then be weakened. We now develop this idea. Let C_0 be the set of such positive x that g is continuous at the point $x\mathbf{I_n}$. Similarly, let C_2 be the set of positive x for which g is twice differentiable at the point $x\mathbf{I_n}$. Below follows a summary of the above theorems in terms of tail probabilities of t-statistics.

Corollary 2.1.3. *If $\mu(C_0) = 1$ and g satisfies condition (2.1) of Theorem 2.1, then*

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

where the constant K_g is defined in Theorem 2.1 and positive if there exists a point $x \in C_1$ such that $g(x\mathbf{I_n}) > 0$. If, in addition, $\mu(C_2) = 1$ and g satisfies (2.15), then

$$\frac{\boldsymbol{P}(T_n > u) - K_g t_{n-1}(u)}{t_{n+1}\left(\sqrt{\frac{n+1}{n-1}}u\right)} = M_g - L_g + o(1),$$

where constants M_g and L_g are defined in Theorem 2.2.

An interesting consequence of the corollary is that *the tail of non-central t-distribution is asymptotically constant times the tail of the corresponding t-distribution.*

3 Simulation study

In this section we study the relative error of approximations of $\mathbf{P}(T_n > u)$ by $K_g t_{n-1}(u)$ and by $K_g t_{n-1}(u) + (M_g - L_g)\, t_{n+1}\left(\sqrt{\frac{n+1}{n-1}}u\right)$. For simplicity we limit our simulations to the i.i.d case so that $g(x_1, x_2, .., x_n) = g(x_1)g(x_2)..g(x_n)$ for some density $g(x)$. The constants K_g, L_g and M_g then take form

$$K_g = 2c_n \int_0^\infty h^{n-1} g_1(h)^n \, dh \quad , \quad L_g = (n-1)c_n \int_0^\infty h^n g_1'(h) g_1(h)^{n-1} \, dh$$

and

$$M_g = (n-1)c_n \int_0^\infty h^{n+1} \left[g_1''(h) g_1(h)^{n-1} - (g_1'(h))^2 g(h)^{n-2} \right] \, dh,$$

where $c_n = \frac{(\pi n)^{n/2}}{\Gamma(\frac{n}{2})}$. As in the simulation study of [15] we choose g to be the normal, Cauchy, t_2 or centered exponential density. It can be checked that these densities satisfy the conditions of Corollary 2.1.3 and that the constants K_g and $M_g - L_g$ are as given in Table 1. The constants for normal distribution with non-zero mean are calculated numerically, see Appendix A. In connection with the Cauchy and t_2 distributions it

is worth mentioning a more general formula. That is, when g is a t-distribution with ν degrees of freedom,

$$K_g = \frac{(\pi n)^{n/2}\,\mathrm{B}\left(\frac{n\nu}{2},\frac{1}{2}\right)}{\sqrt{\pi}\mathrm{B}\left(\frac{\nu}{2},\frac{1}{2}\right)^n} \quad \text{and} \quad M_g - L_g = \frac{n^2+n-2}{n(\nu+1)+2}K_g,$$

where $B(\cdot,\cdot)$ stands for the Beta function.

Distribution	Density g	K_g	$M_g - L_g$
$N(0,\sigma^2)$	$\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{1}{2\sigma^2}x^2}$	1	0
Cauchy	$\frac{1}{\pi(1+x^2)}$	$\frac{n^{n/2}}{(4\pi)^{\frac{n-1}{2}}\Gamma\left(\frac{n+1}{2}\right)}$	$\frac{n^2+n-2}{2n+2}K_g$
t_2	$\frac{1}{(2+x^2)^{3/2}}$	$\frac{(\pi n)^{n/2}\Gamma(n)}{2^n\Gamma\left(\frac{3n}{2}\right)}$	$\frac{n^2+n-2}{3n+2}K_g$
Centered Exponential	$e^{-(x+1)},\ x \geq -1$	$\frac{2\pi^{n/2}\Gamma(n)}{e^n n^{n/2}\Gamma\left(\frac{n}{2}\right)}$	$\frac{(n-1)}{2}K_g$

Table 1: *The constants K_g and $M_g - L_g$ for the normal, Cauchy, t_2 or centered exponential densities.*

We start with the analysis of the relative error for normal distribution. If $g \sim N(\mu,\sigma^2)$ then T_n has a non-central t-distribution with $n-1$ degrees of freedom and non-centrality parameter $\mu\sqrt{n}/\sigma$. Setting $\sigma^2 = 1$, the

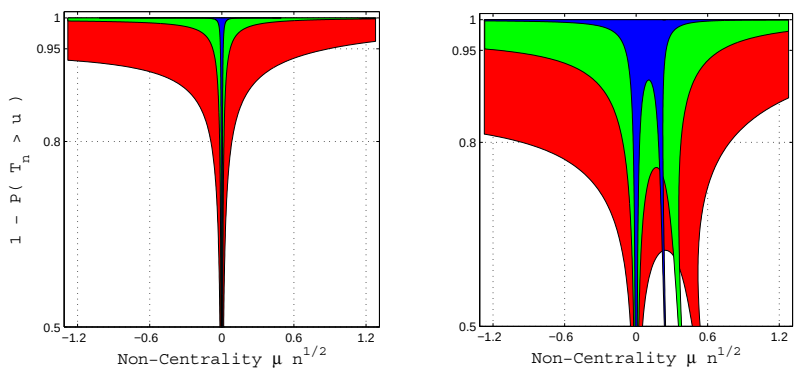

Figure 2: *The region of small (less than 0.01) absolute relative error for the first order (left plot) and second order (right plot) approximations of $\boldsymbol{P}(T_n > u)$. The red, green and blue areas correspond to $n = 2, 3$ and 5 accordingly. T_n has a non-central t-distribution.*

plots in fig. 2 show the behavior of the relative error for different combinations of threshold u and non-centrality parameter $\mu\sqrt{n}$.

First, we see that the more the initial distribution deviates from standard normal (the non-centrality parameter is proportional to μ), the slower the relative error converges to zero. Second, the 'nested' structure of the sets shows that given a fixed value of the non-centrality parameter ($\mu\sqrt{n} = const$), the relative error converges slower for larger n.

The behavior of the relative error for a fixed threshold u is illustrated by fig. 3.

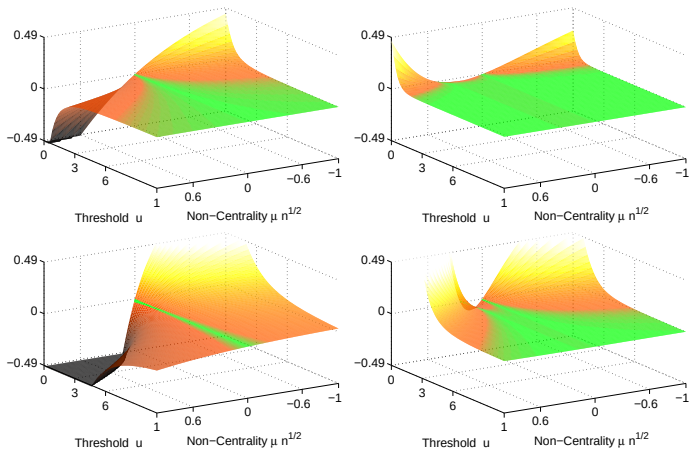

Figure 3: *The relative error versus non-centrality parameter $\mu\sqrt{n}$ and threshold u. The left and right plots correspond to the first and second order approximation formulas respectively; the upper plots correspond to $n = 2$ and the lower plots to $n = 5$. Small (less than 0.01) values of the absolute relative error are shown in green.*

It can be seen that the relative error tends to infinity when $\mu\sqrt{n} \to \infty$ (n is fixed) and also when $n \to \infty$ and the non-centrality parameter $\mu\sqrt{n}$ is fixed and non-zero. The first observation is purely empirical and the second can be derived from Theorem 2.4 of [1] and expressions for the constants K_g and $M_g - L_g$ in Appendix A.

We now proceed with the examples of [15]. Since the exact value of the probability $\mathbf{P}(T_n > u)$ is hard to calculate, we only consider sample sizes 2 and 5. When using Monte-Carlo simulations, it is important to take into account that the relative error of the method itself depends not only on the number of simulations, but also on the value of the estimated probability. We thus set the number of simulations varying from

$1,000,000$ to $5,000,000,000$ inversely proportionally to $\mathbf{P}(T_n > u)$. The range of values for u is such that $t_n(u)$ varies from 0.5 to 0.999 for $n = 2$ and from 0.5 to 0.9999 for $n = 5$. The simulation for standard normal distribution showed that the relative error of the Monte-Carlo method itself is less than 0.001 uniformly in u.

As we see from fig. 4, the approximations for $n = 2$ are extremely accurate starting from as low as $q_{0.95}$. The relative errors for the first and second order approximations are less than 0.01 and 0.0005 respectively for all three distributions.

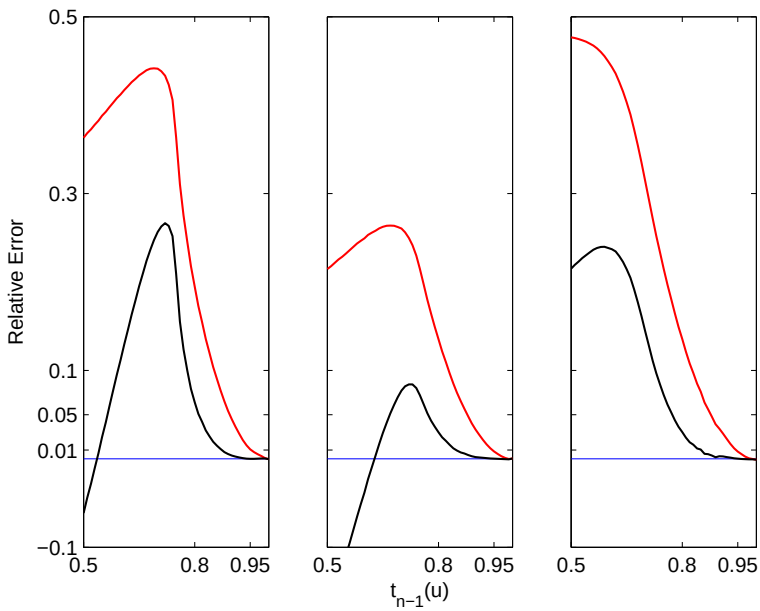

Figure 4: $n = 2$. *The plots show the relative error for the approximation of* $\boldsymbol{P}(T_n > u)$ *by* $K_g t_{n-1}(u)$ *(red line) and* $K_g t_{n-1}(u) + (M_g - L_g)\, t_{n+1}\left(\sqrt{\frac{n+1}{n-1}}u\right)$ *(black line). Left, middle and right plots correspond to Cauchy,* t_2 *and centered exponential distributions accordingly. The horizontal blue line is a zero mark.*

According to the simulation section of [14], this is much more accurate than the precision of the saddlepoint approximation for $n = 5$. For the case $n = 5$ similar precision is reached further out in the tail (from 0.0005 to 0.0001 quantile), see fig. 5.

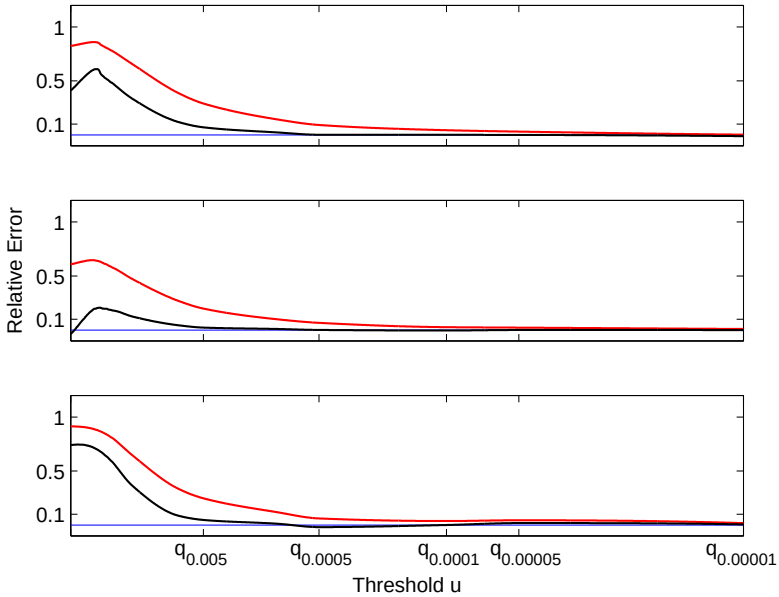

Figure 5: $n = 5$. *The same as in fig. 4 except that the order of the plots is from the upper to the lower. We use original scale for x-axis and q_r stand for r-quantile of the t_5 distribution.*

ACKNOWLEDGEMENTS:
The author wishes to thank Professor Holger Rootzén for careful reading and fruitful discussions, and Professor Olle Nerman for useful comments.

References

[1] Bentkus V., Jing B.-Y., Shao Q.M., and Zhou W. (2007). Limiting distributions of the non-central t-statistic and their applications to the power of t-tests under non-normality. Bernoulli v. 13, n. 2, pp. 346-364.

[2] Warringer J. and Blomberg A. (2003). Automated screening in environmental arrays allows analysis of quantitative phenotypic profiles in Saccharomyces cerevisiae. Yeast 20, pp. 53-67

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

[4] Daniels, H. E. and Young, G. A. (1991). Saddlepoint approximation for the Studentized mean, with an application to the bootstrap. Biometrika 78 169-179.

[5] Field, C. and Ronchetti, E. (1990). Small Sample Asymptotics. IMS, Hayward, CA.

[6] Giné, E., Götze, F. and Mason, D. M. (1997). When is the Student t -statistic asymptotically standard normal? Ann. Probab. 25 1514-1531.

[7] Hall, P. (1987). Edgeworth expansion for Student's t statistic under minimal moment conditions. Ann. Probab. 15 920-931.

[8] Jensen, J. L. (1995). Saddlepoint Approximations. Oxford Univ. Press.

[9] Kendall, M., and Stuart, A., (1977) The Advanced Theory of Statistics, Volume 1, MacMillan, New York.

[10] Kolassa, J. E. (1997). Series Approximation Methods in Statistics, 2nd ed. Lecture Notes in Statist. 88. Springer, New York.

[11] Logan, B. F., Mallows, C. L., Rice, S. O. and Shepp, L. A. (1973). Limit distributions of self-normalized sums. Ann. Probab. 1 788-809.

[12] Lugannani, R. and Rice, S. (1980). Saddlepoint approximation for the distribution of the sum of independent random variables. Adv. in Appl. Probab. 12 475-490.

[13] Reid, N. (1988). Saddlepoint methods and statistical inference (with discussion). Statist. Sci. 3 213-238.

[14] Jing, B.-Y., Shao, Q.M. and Zhou, W. (2004). Saddlepoint approximation for Student's -statistic with no moment conditions. Ann. Statist. 32, 2679-2711. MR 2153999 (2006a:62030)

[15] Zhou W., Jing B.-Y.. (2006). Tail probability approximations for Student's t-statistics. Probability Theory and Related Fields, 136(4), 541-559.

[16] Shao, Q.-M. (1997) Self-normalized large deviations. Ann. Probab. 25 285-328.

[17] Shao, Q.M., Recent progress on self-normalized limit theorems. Probability, Finance and Insurance. Edited by T.L. Lai, H. Yang, and S.P. Yung, World Scientific, 2004. MR 2189198 (2006i:60056)

[18] Zholud, D., Rootzén, H., Nerman, O., and Blomberg, A., Positional Effects in Biological Array Experiments and Their Impact on the False Discovery Rate. Manuscript.

Appendix A: Constant estimation.

Let $g(x) = N(\mu, 1)$. The computation of the constants K_g, L_g and M_g is done separately for $\mu > 0$ and $\mu < 0$. The Mathematica 6.0 code for $\mu > 0$ follows:

$g[\text{x\_}]:=PDF[NormalDistribution[\mu, 1], x];$

$c_n:=\frac{(\pi n)^{n/2}}{Gamma\left[\frac{n}{2}\right]};$

$Assuming\left[\mu > 0 \;\&\&\; n \in Integers \;\&\&\; n > 0, \{\right.$

$k = FullSimplify\left[2c_n Integrate\left[x^{n-1} g[x]^n, \{x, 0, \infty\}\right]\right],$

$l = FullSimplify\left[(n-1)c_n Integrate\left[x^n \partial_x g[x] g[x]^{n-1}, \{x, 0, \infty\}\right]\right],$

$m = FullSimplify\left[(n-1)c_n\right.$

$Integrate\left[x^{n+1}\left(\partial_{x,x} g[x] g[x]^{n-1} - \left(\partial_x g[x]\right)^2 g[x]^{n-2}\right), \{x, 0, \infty\}\right]$

$\left.\left.\}\right]\right]$

Executing the code above gives

$$K_g = \text{Hypergeometric1F1}\left(\frac{1-n}{2}, \frac{1}{2}, -\frac{n\mu^2}{2}\right) + \frac{\sqrt{2n}\mu\Gamma\left(\frac{n+1}{2}\right)\text{Hypergeometric1F1}\left(1-\frac{n}{2}, \frac{3}{2}, -\frac{n\mu^2}{2}\right)}{\Gamma\left(\frac{n}{2}\right)}, \tag{3.1}$$

and by $FullSimplify\,[m - l]$ we obtain that

$$M_g - L_g = -\frac{(n-1)\mu}{2\sqrt{n}\Gamma\left(\frac{n}{2}\right)} e^{-\frac{n\mu^2}{2}} \times \tag{3.2}$$

$$\times \left(n^{3/2}\mu\Gamma\left(\frac{n}{2}\right)\text{Hypergeometric1F1}\left(\frac{n}{2}+1, \frac{3}{2}, \frac{n\mu^2}{2}\right)\right.$$

$$+ \sqrt{2}\Gamma\left(\frac{n+1}{2}\right)\left((n+1)\text{Hypergeometric1F1}\left(\frac{n+3}{2}, \frac{3}{2}, \frac{n\mu^2}{2}\right)\right.$$

$$\left.\left.-n\text{Hypergeometric1F1}\left(\frac{n+1}{2}, \frac{3}{2}, \frac{n\mu^2}{2}\right)\right)\right),$$

where Γ stands for the Gamma function and Hypergeometric1F1 (further simply ${}_1F_1$) is the Kummer's confluent hypergeometric function of

the first kind.

The code for $\mu < 0$ is identical, except that $\mu > 0$ in the "Assuming" statement is replaced by $\mu < 0$. In this case

$$K_g = \frac{e^{-\frac{n\mu^2}{2}}\Gamma\left(\frac{n+1}{2}\right)\text{HypergeometricU}\left(\frac{n}{2},\frac{1}{2},\frac{n\mu^2}{2}\right)}{\sqrt{\pi}},$$

where HypergeometricU (further simply U) is the Kummer's confluent hypergeometric function of the second kind. Assuming $\mu < 0$ and using the identity

$$U(a,b,z) = \frac{\pi}{\sin\pi b}\left[\frac{{}_1F_1(a,b,z)}{\Gamma(a-b+1)\Gamma(b)} - \frac{z^{1-b}{}_1F_1(a-b+1,2-b,z)}{\Gamma(a)\Gamma(2-b)}\right]$$

and Kummer's transformation ${}_1F_1(a,b,z) = e^z{}_1F_1(b-a,b,-z)$ it is easy to show that the expression above is equivalent to (3.1). The expression for $M_g - L_g$ (after some re-arrangement) gives (3.2) and formulas (3.1) and (3.2) are thus valid for any μ. Interestingly, Mathematica 6.0 would not be able to compute the constants without considering the two separate cases.

The constants for the Cauchy, t_2 and centered exponential densities are obtained by replacing $g[x\_]:=\text{PDF}[\text{NormalDistribution}[\mu,1],x]$ by the corresponding distribution. The "Assuming" statement is modified accordingly.

References

[1] S.I. Hayek, Advanced Mathematical Methods in Science and Engineering, (Marcel Dekker, New York , 2001)

[2] http://mathworld.wolfram.com/topics/ConfluentHypergeometricFunctions.html

Appendix B: Contour plots.

The construction of fig. 2 and fig. 3 in the simulation section involves non straight-forward computations. First, consider the Mathematica 6.0 code that would have produced produce fig.2 and fig.3. Let $k[\mu\_, \mathrm{n\_}]$ and $k[\mu\_, \mathrm{n\_}]$ be the constants K_g and $M_g - L_g$ as defined in Appendix A. The probability $\mathbf{P}\left(T_n > u\right)$ is computed by

$$pt[\mu\_, n\_, u\_] := 1 - CDF\left[NoncentralStudentTDistribution\left[n-1, \mu\sqrt{n}\right], u\right]$$

and the (first and second order) approximation formulas take form

$$f1[\mu\_, n\_, u\_] := k[\mu, n](1 - CDF[StudentTDistribution[n-1], u])$$

and

$$f2[\mu\_, n\_, u\_] := k[\mu, n](1 - CDF[StudentTDistribution[n-1], u]) + ml[\mu, n]\left(1 - CDF\left[StudentTDistribution[n+1], \sqrt{\tfrac{n+1}{n-1}}u\right]\right).$$

The relative errors $\mathrm{re1}[\mu\_, \mathrm{n\_}, \mathrm{u\_}]$ and $\mathrm{re2}[\mu\_, \mathrm{n\_}, \mathrm{u\_}]$ equal

$$1 - \mathrm{f1}[\mu, n, u]/\mathrm{pt}[\mu, n, u] \quad \text{and} \quad 1 - \mathrm{f2}[\mu, n, u]/\mathrm{pt}[\mu, n, u].$$

The following fragment illustrates that Mathematica 6.0 does not provide flexible enough numerical computation tools.

RegionPlot $\left[\left\{\mathrm{Abs}\left[\mathrm{re1}\left[\mu/\sqrt{2}, 2, u\right]\right] < 0.01,\right.\right.$

$\mathrm{Abs}\left[\mathrm{re1}\left[\mu/\sqrt{3}, 3, u\right]\right] < 0.01,$

$\left.\left.\mathrm{Abs}\left[\mathrm{re1}\left[\mu/\sqrt{5}, 5, u\right]\right] < 0.01\right\}, \{\mu, -2, 2\}, \{u, 0, 11\}\right.$,

PerformanceGoal $\rightarrow$ "Quality", PlotPoints $\rightarrow$ 20,

MaxRecursion $\rightarrow$ 1]

For simplicity we use the scale which is different from the scale in fig.2. Constructing the plots in the same scale as in fig. 2 would result in much more severe problems. The output of the above code is displayed in fig. 6. Comparing with the plots in fig. 2 we notice that the behavior of the relative errors in the lower part of the graph differs (the thin regions are cut off). Increasing the PlotPoints or MaxRecursion parameters would result in sufficiently longer computational time (presumably, due to the way Mathematica 6.0 computes the non-central t-distribution) and huge image size, while down-sampling to bitmap format causes blurriness.

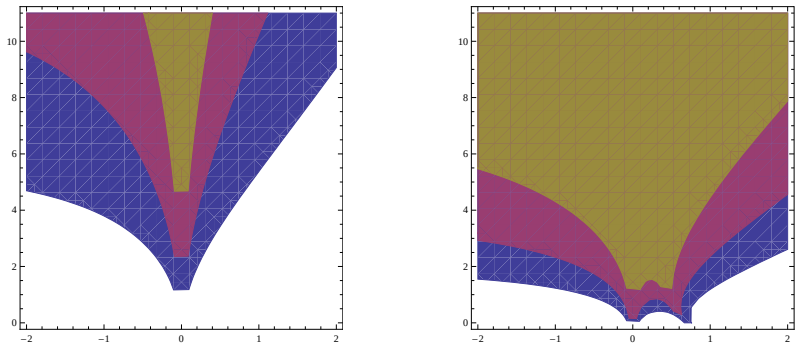

Figure 6: *The contour plots produced by Mathematica 6.0.*

To overcome this problem we considered using MathLab 2008. Note that in order to estimate the contour lines both MatLab and Mathematica evaluate the function on the pre-defined grid. The disadvantage of such approach in Mathematica, for example, is that the grid is equally-spaced. Thus setting the PlotPoints parameter to an even number one would never get the whole peak that corresponds to zero non-centrality parameter (assuming the interval is symmetric). Setting PlotPoints to 21, on the other hand, would give much better quality for the first order approximation formula. The thin regions for the second plot will still be missing, since to capture all the "locations" of the thin regions one would need to choose too small grid step. The latter results in days of computation and megabytes of space for storing the figure with such resolution in a vector format. To overcome the problem we consider using our own algorithm for contour line estimation. Let $F(x, y) = 0$ be the twice differentiable contour line with bounded curvature and assume that the initial point (x_0, y_0) with $dist((x_0, y_0), F(x, y) = 0) < \varepsilon$ for some small enough positive ε is given. The contour line is estimated point-by-point, starting at (x_0, y_0) and following those of the two possible directions which is "closer" to the initial vector. The initial vector must not coincide with $\nabla F(x_0, y_0)$. Unlike in the "grid approach", the resulting contour line is a locally simply connected component. If the contour line has more than one path connected components then the step r should be set smaller than minimum of the distances between the current component and all the other components. The latter ensures that the algorithm does not "jump" from one component to another.

```
function [x y]=SmoothContour(x0,y0,Fun,r,boundx,boundy)
% alpha - the angel between the direction of the contour
% line and OX axis.  boundx and boundy are boundary
% intervals for x and y. MaxIterations - maximum allowed
% iterations of the algorithm.
```

```
MaxIterations=1000;
MinCycleLength=8;
eps=0.0000001;
a_eps=0.001;

% Setting initial guess for the countour line direction.
% In this example - towards the region center. Choose other
% direction if needed.

alpha=angle((mean(boundx)-x0)+sqrt(-1)*(mean(boundy)-y0));

% Check if (x0,y0) belongs to the contour line.

if abs(Fun(x0,y0))<eps
    SolutionFound=1;
    x=zeros(1,MaxIterations);
    y=zeros(1,MaxIterations);
    x(1)=x0;
    y(1)=y0;
else
    x=[];
    y=[];
    return
end

Iter=1;
Cycled=0;

% Iteration cycle is repeated while solution F(x,y)=0 exists
% and within bounds, maximum number of iterations is not
% reached and contour plot does not intersect with itself.

while (SolutionFound)&&(Iter<MaxIterations)&&(~Cycled)...
                     &&(boundx(1)<=x(Iter))...
                     &&(x(Iter)<boundx(2)) ...
                     &&(boundy(1)<=y(Iter))...
                     &&(y(Iter)<boundy(2)) ...


% Check if Fun(x,y) has different signs alone the contour
% line. The estimated angel alpha of the contour line
% direction must not be orthogonal to the contour line
% direction.
```

```
   if  Fun(x(Iter)+r*cos(alpha)+r*cos(alpha-pi+a_eps)...
           ,y(Iter)+r*sin(alpha)+r*sin(alpha-pi+a_eps))...
       *Fun(x(Iter)+r*cos(alpha)+r*cos(alpha+pi-a_eps)...
            ,y(Iter)+r*sin(alpha)+r*sin(alpha+pi-a_eps))>0
       break;
   end

% Find the next contour point.

   [alpha_new, fval, SolutionFound]=fzero(...
             @(a) Fun(x(Iter)+r*cos(alpha)+r*cos(a),...
             y(Iter)+r*sin(alpha)+r*sin(a))...
             ,[alpha-pi+a_eps alpha+pi-a_eps]);

% If no solution found - terminate.

   if (~SolutionFound)
       break;
   end

% Else - compute the coordinates of the contour point.

   x(Iter+1)=x(Iter)+r*cos(alpha)+r*cos(alpha_new);
   y(Iter+1)=y(Iter)+r*sin(alpha)+r*sin(alpha_new);

% If contour line intersects with itself - terminate.

   if (Iter>MinCycleLength)&&...
      (...
      min((x(1:Iter-MinCycleLength)-x(Iter+1)).^2 ...
           +(y(1:Iter-MinCycleLength)-y(Iter+1)).^2)...
       <=4*r^2 ...
      )
       disp('Self-intersection found.');
       break;
   end

% Re-estimate the direction of the contour line and continue
% with the next iteration

   alpha=angle((x(Iter+1)-x(Iter))...
               +sqrt(-1)*(y(Iter+1)-y(Iter)));
   Iter=Iter+1;
```

```
end

x(Iter:end)=[];
y(Iter:end)=[];
```

Next follows the implementation of the Kummer's confluent hypergeometric function of the first kind (see constant estimation, Appendix A).

```
function y=Hypergeometric1F1(a,b,z)

MaxIterations=100;
Prescision=0.0000000000001;
x=zeros(length(z),MaxIterations);
n=0;

while True
    n=n+1;
    x(:,n)=prod((a:a+n-1)./(b:b+n-1)./(n:-1:1))*(z.^n)';
    if ((n>MaxIterations)||(sum(abs(x(:,n))>Prescision))==0)
        break;
    end
end

y=sum(x,2)'+1;
```

And the relative errors for the first and second order approximation formulas are defined as follows:

```
function y=RelativeError1(mu,n,u)
% Computing relative error of approximation formula (1)

k=Hypergeometric1F1((1-n)/2, 1/2, -n*mu.^2/2)...
      + sqrt(2*n)*gamma((n + 1)/2)/gamma(n/2)*mu...
      .*Hypergeometric1F1(1 - n/2, 3/2, -n*mu.^2/2);

f= k*(1 - tcdf(u,n-1));
ptrue= 1 - nctcdf(u,n-1,mu*sqrt(n));
y = 1 - f./ptrue;



function y=RelativeError2(mu,n,u)
% Computing relative error of approximation formula (2)

k=Hypergeometric1F1((1-n)/2, 1/2, -n*mu.^2/2)...
      + sqrt(2*n)*gamma((n + 1)/2)/gamma(n/2)*mu...
      .*Hypergeometric1F1(1 - n/2, 3/2, -n*mu.^2/2);
```

```
ml= -1/(2*sqrt(n)*gamma(n/2))*(n - 1)*mu...
    .*exp(-n*mu.^2/2)...
    .*(...
      n^(3/2)*gamma(n/2)*mu...
      .*Hypergeometric1F1(1 + n/2, 3/2, n*mu.^2/2)...
      + sqrt(2)*gamma((n + 1)/2)...
      *(...
        -n*Hypergeometric1F1((n + 1)/2, 3/2, n*mu.^2/2)...
        +(n + 1)...
        *Hypergeometric1F1((n + 3)/2, 3/2, n*mu.^2/2)...
       )...
      );

f = k*(1 - tcdf(u,n-1)) + ...
   ml*(1 - tcdf(sqrt((n + 1)/(n - 1))*u,(n + 1)));
ptrue= 1 - nctcdf(u,n-1,mu*sqrt(n));
y = 1 - f./ptrue;
```

Finally, we build the plots.

```
LE=norminv(0.1,0,1);
RE=norminv(0.9,0,1);
Step=0.01;
MaxP=0.99999;
MaxU=13;
Scale='No P-Value';
Color=['r' 'g' 'b'];

figure()     %First plot

i=1;
for n=[2 3 5]
    r1=fzero(@(mu) RelativeError1(mu,n,0)-0.01,[ -1  0]);
    r2=fzero(@(mu) RelativeError1(mu,n,0)+0.01,[  0  1]);
    if strcmp(Scale,'P-Value')
        [x1 y1]=SmoothContour(r1,0,...
        @(mu,u) RelativeError1(mu,n,u)-0.01,Step,...
        [LE RE],[0 tinv(MaxP,n-1)]);
        [x2 y2]=SmoothContour(r2,0,...
        @(mu,u) RelativeError1(mu,n,u)+0.01,Step,...
        [LE RE],[0 tinv(MaxP,n-1)]);
        y=tcdf([y1(length(y1):-1:1) y2],n-1);
    else
        [x1 y1]=SmoothContour(r1,0,...
        @(mu,u) RelativeError1(mu,n,u)-0.01,Step,...
```

```
        [LE RE],[0 MaxU]);
        [x2 y2]=SmoothContour(r2,0,...
        @(mu,u) RelativeError1(mu,n,u)+0.01,Step,...
        [LE RE],[0 MaxU]);
        y=[y1(length(y1):-1:1) y2];
    end

    x=[x1(length(x1):-1:1) x2];

    area(x,y,1+(MaxU-1)*(1-strcmp(Scale,'P-Value')),...
        'FaceColor',Color(i));
    hold on;
    i=i+1;
end

figure()     %Second plot

i=1;
for n=[2 3 5]
    r1=fzero(@(mu) RelativeError2(mu,n,0)-0.01,[ -1  0  ]);
    r2=fzero(@(mu) RelativeError2(mu,n,0)+0.01,[  0  0.1]);
    r3=fzero(@(mu) RelativeError2(mu,n,0)+0.01,[  0.1  1]);
    r4=fzero(@(mu) RelativeError2(mu,n,0)-0.01,[ 0.1  1 ]);
    if strcmp(Scale,'P-Value')
        [x1 y1]=SmoothContour(r1,0,...
        @(mu,u) RelativeError2(mu,n,u)-0.01,Step,...
        [LE RE],[0 tinv(MaxP,n-1)]);
        [x2 y2]=SmoothContour(r2,0,...
        @(mu,u) RelativeError2(mu,n,u)+0.01,Step,...
        [LE RE],[0 tinv(MaxP,n-1)]);
        [x3 y3]=SmoothContour(r4,0,...
        @(mu,u) RelativeError2(mu,n,u)-0.01,Step,...
        [LE RE],[0 tinv(MaxP,n-1)]);
        y=tcdf([y1(length(y1):-1:1) y2 0 y3],n-1);
    else
        [x1 y1]=SmoothContour(r1,0,...
        @(mu,u) RelativeError2(mu,n,u)-0.01,Step,...
        [LE RE],[0 MaxU]);
        [x2 y2]=SmoothContour(r2,0,...
        @(mu,u) RelativeError2(mu,n,u)+0.01,Step,...
        [LE RE],[0 MaxU]);
        [x3 y3]=SmoothContour(r4,0,...
        @(mu,u) RelativeError2(mu,n,u)-0.01,Step,...
        [LE RE],[0 MaxU]);
```

```
        y=[y1(length(y1):-1:1) y2 0 y3];
    end

    x=[x1(length(x1):-1:1) x2 r3 x3];

    area(x,y,1+(MaxU-1)*(1-strcmp(Scale,'P-Value')),...
        'FaceColor',Color(i));
    hold on;
    i=i+1;
end
```

The appearance of the plots is adjusted by the following code

```
if strcmp(Scale,'P-Value')
    set(gca,'YLim',[0.5 MaxP+0.01]);
    set(gca,'YTick',[0.5 0.8 0.95 0.99999]);
    ylabel('1 - P( T_n > u )');
else
    set(gca,'YLim',[0   MaxU+0.21]);
    set(gca,'YTick',[0 3 6 MaxU]);
    ylabel('Threshold  u');
end

axis square;
set(gca,'XLim',[LE-0.03  RE+0.03]);
set(gca,'XTick',[-1.2 -0.6 0 0.6 1.2 ]);
xlabel('Non-Centrality \mu n^{1/2}');
grid on;
box on;
```

Appendix C: Monte-Carlo simulations.

The true probabilities for the case of the Cauchy, t_2 and centered exponential densities are estimated by means of Monte-Carlo simulations. Consider the case $n = 2$. The number of Monte-Carlo simulations varies from $1,000,000$ to $1000,000,000$ inversely proportionally to the value of $tcdf(u, n-1)$. The MatLab code below corresponds to the Cauchy density. The other densities are handled accordingly.

```
function [P_Montecarlo]=PEstimate(n)

alpha=[0.5:-0.02:0.02 0.05:-0.01:0.01 0.005 0.001]';
u=tinv(1-alpha,n-1);

P_Montecarlo=zeros(1,length(u));

for i=1:l
    k=0;
    j=0;
    while j<1/alpha(i)
        A=trnd(1,n,1000000);
        Xbar=mean(A);
        Xstd=std(A);
        T=sqrt(n)*Xbar./Xstd;
        k=k+sum(T>u(i));
        j=j+1;
    end

    P_Montecarlo(i)=k/(j*1000000);
    disp(['Cauchy ' num2str(alpha(i)) ' done']);
end
```

The constants K_g and $M_g - L_g$ are given by

$$K\_Cauchy = n\hat{}(n/2)/((4 * pi)\hat{}(n/2 - 0.5) * gamma((n+1)/2))$$

and

$$ML\_Cauchy = (n\hat{}2 + n - 2)/(2 * n + 2) * K\_Cauchy,$$

and the relative error estimation becomes straightforward.