Self-Normalized Sums and Directional Conclusions

Fredrik Jonsson

Department of Mathematics
Uppsala University
UPPSALA 2012
Dissertation presented at Uppsala University to be publicly examined in Polhemssalen, Ångström Laboratory, Lägerhyddsvägen 1, Uppsala. Friday, January 20, 2012 at 13:15 for the degree of Doctor of Philosophy. The examination will be conducted in English.

Abstract

This thesis consists of a summary and five papers, dealing with self-normalized sums of independent, identically distributed random variables, and three-decision procedures for directional conclusions.

In Paper I, we investigate a general set-up for Student's t-statistic. Finiteness of absolute moments is related to the corresponding degree of freedom, and relevant properties of the underlying distribution, assuming independent, identically distributed random variables.

In Paper II, we investigate a certain kind of self-normalized sums. We show that the corresponding quadratic moments are greater than or equal to one, with equality if and only if the underlying distribution is symmetrically distributed around the origin.

In Paper III, we study linear combinations of independent Rademacher random variables. A family of universal bounds on the corresponding tail probabilities is derived through the technique known as exponential tilting. Connections to self-normalized sums of symmetrically distributed random variables are given.

In Paper IV, we consider a general formulation of three-decision procedures for directional conclusions. We introduce three kinds of optimality characterizations, and formulate corresponding sufficiency conditions. These conditions are applied to exponential families of distributions.

In Paper V, we investigate the Benjamini-Hochberg procedure as a means of confirming a selection of statistical decisions on the basis of a corresponding set of generalized p-values. Assuming independence, we show that control is imposed on the expected average loss among confirmed decisions. Connections to directional conclusions are given.

Keywords: Self-normalized sums, heavy-tailedness, Student's t-statistic, distributional symmetry, exponential tilting, directional conclusions, reversal rates, multiple statistical inference, the Benjamini-Hochberg procedure

Fredrik Jonsson, Uppsala University, Department of Mathematics, Box 480, SE-751 06 Uppsala, Sweden.

© Fredrik Jonsson 2012

ISSN 1401-2049
urn:nbn:se:uu:diva-162168 (http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-162168)
All we know about the world teaches us that the effects of A and B are always different—in some decimal place—for any A and B. Thus asking “Are the effects different?” is foolish.

List of Papers

This thesis is based on the following papers, which are referred to in the text by their Roman numerals.


Reprints were made with permission from the publishers.
Contents

1 Introduction ..................................................... 1
  1.1 Self-normalized sums ....................................... 2
    1.1.1 Normalized sums in probability theory ............. 2
    1.1.2 Self-normalized sums in statistics ................. 4
    1.1.3 Gosset and “The probable error of a mean” ........ 5
    1.1.4 The assumption of normality ......................... 6
    1.1.5 Contributions ........................................ 6
  1.2 Directional conclusions .................................... 8
    1.2.1 Statistical significance .............................. 8
    1.2.2 Investigating public opinions ....................... 9
    1.2.3 Type III errors ...................................... 10
    1.2.4 Contributions ........................................ 11
  1.3 The problem of multiplicity ................................ 12
    1.3.1 DNA microarray data ................................. 12
    1.3.2 FWER and FDR ....................................... 13
    1.3.3 Statistical models and decisions .................... 14
    1.3.4 Contributions ........................................ 15
2 Summary of Papers ............................................. 17
  2.1 Paper I .................................................. 17
  2.2 Paper II .................................................. 19
  2.3 Paper III ................................................. 20
  2.4 Paper IV .................................................. 23
  2.5 Paper V .................................................. 25
3 Summary in Swedish ............................................ 29
  3.1 Självnormaliserade summor ............................... 29
  3.2 Kontrollerade slutsatser om förändringars riktning ...... 31
4 Acknowledgements .............................................. 33
References ....................................................... 35
1. Introduction

This is a thesis in Mathematical statistics, a field divided into Probability theory and Statistics.

Papers I-III contribute to the Theory of self-normalized sums\(^1\), a subfield of probability theory motivated by statistical applications. Our investigations are concerned with the behaviour of tails and moments of two kinds of self-normalized sums, with respect to i.i.d. sequences of random variables. Some general impacts of distributional symmetry regarding the underlying distribution are encountered in Paper II. The results in Papers III concern the tails of linear combinations of Rademacher random variables, with applications to self-normalized sums of symmetrically distributed random variables.

Papers IV and V are mainly motivated by directional conclusions. The papers thus belong to the field known as Statistical decision theory\(^2\). The topic of Paper IV is optimality among three-decision procedures for directional conclusions. Paper V also relates to the field known as Multiple statistical inference\(^3\). More precisely, Paper V introduces an extended setting for a well-known multiple testing procedure, the Benjamini–Hochberg procedure.

The present, introductory part of the thesis is divided into three sections. Section 1.1 introduces self-normalized sums. Two examples, referring to a die game and the measuring of heights of human beings, are presented, briefly illustrating the concepts of standard deviations, normal distributions, statistical uncertainty, etcetera. Next, Section 1.2 introduces the considered framework for directional conclusions. Two examples, referring to Mendelian genetics and polls for public opinions, are given, mainly illustrating the concept of statistical significance. Finally, Section 1.3 introduces some relevant parts of multiple statistical inference. One example, referring to DNA microarray data in genetics, is given, illustrating the context of multiplicity. Papers I-V are briefly introduced in Sections 1.1.5, 1.2.4 and 1.3.4.

We refer to Part 2 of the thesis for more detailed introductions and summaries of the papers. Also, a short summary in Swedish is given in Part 3 of the thesis.

---

\(^1\)An overview of this field is given in the recent research monograph [14].

\(^2\) Confer [45] for a recently written, introductory monograph. There are also several classical accounts, such as [60], [5], [20] and [41].

\(^3\)We refer to [16] for an up-to-date account, with biostatistical applications. Confer also [34], [37], [57], [61] and [46].
1.1 Self-normalized sums

We begin by informally introducing “normalization” and “self-normalization” in probability theory and statistics. Next, a brief historical review concerning the tradition of studying self-normalized sums is given in Sections 1.1.3 and 1.1.4. Finally, Papers I-III are introduced in Section 1.1.5.

1.1.1 Normalized sums in probability theory

Beginning with an example from everyday life, consider the board game known as “Fia med knuff”\(^4\). In this game, each player hits a die in each round, and hitting a six stands out as the most profitable alternative for the player. “Everybody knows”—either by experience, or by referring to the symmetrical construction of the die—that hitting a six occurs at rate 1/6. In other words, the chance of hitting a six is 1/6.

Suppose it takes you about 180 rounds to complete the game against three of your friends. Then, you hit a six in about
\[
\frac{180}{6} = 30
\]
of these rounds. However, due to random fluctuations, it is not very likely that you will hit a six \textit{exactly} 30 times when playing 180 rounds.

A fundamental law in probability theory—known as the \textit{Central Limit Theorem}\(^*\)—tells us that the \textit{observed number} of sixes fluctuates around the \textit{expected number} 30 with magnitude proportional to the square root of the total number of rounds (\(\sqrt{180}\)). Moreover, the proportionality constant, that is, the \textit{standard deviation} of a single experiment is here
\[
\sqrt{\frac{5}{6}} \approx 0.37.
\]
Consequently, you may expect fluctuations of order
\[
\frac{\sqrt{5}}{6} \cdot \sqrt{180} = \frac{\sqrt{5}}{6} \cdot \sqrt{5 \cdot 36} = 5.
\]

The same fundamental law also tells us that the fluctuations are approximately \textit{normally distributed}, as described by the famous, Gaussian\(^5\), bell-shaped curve. One may for instance infer that variations within an interval given by ±2 times the above constant 5 is what is to be expected with reasonably large probability (about 95%).

---

\(^4\)“Ludo” in English, “Mensch ärgere dich nicht” in German and “T’en fais pas” in French.

\(^5\)Carl Friedrich Gauss (1777-1855), one of the fathers of the normal distribution, analyzed measurement errors in astronomical studies at the end of the 18th century when he first reported on the mathematical theory that we allude to in this example. Confer [26, Chapter 3.3].
Thus, you get 30 sixes on average, but you may expect a value anywhere within the interval

\[ 30 \pm 10 \]

with a relatively high degree of certainty.

To summarize the given analysis, we considered a random entity

\[ X = \text{the observed number of sixes after 180 rounds.} \]

Next, we concluded that the expected value of \( X \) is 30. Moreover, due to the central limit theorem, \( X \) is approximately distributed according to the Gaussian curve centered at 30. Without going into all mathematical details, we concluded that the “fluctuational scale” is 5, meaning that the following \textit{standardized} random entity

\[ Y = \frac{X - 30}{5} \]

is approximately normally distributed on the standardized scale, with for instance 95% of the observations contained within the interval \([-2, 2]\).

\textit{Standardization} of a random variable usually refers to the following two operations:

- Subtraction of its expected value;
- Division by its standard deviation.

More generally, changing the observational scale \textit{linearly} (in other words, adding/subtracting/multiplying/dividing the random entity by some given numbers) is often referred to as \textit{normalization}.

The idea of approximate normality of \( X \) may be associated with the idea of representing \( X \) as a \textit{sum} of many independently contributing small random entities. Here, since \( X \) refers to the \textit{counting} of the number of observed sixes, we may write

\[ X = \sum_{i=1}^{180} X_i, \]

with \( X_1 \) being 1 or 0 depending on whether the first die gave us six or not, \( X_2 \) being 1 or 0 depending on the second die, etcetera. Clearly, the outcomes of \( X_1 \) and \( X_2 \) are mutually independent, in the sense that the outcome of \( X_2 \) is not influenced by the outcome of \( X_1 \). Thus, \( Y \) above may be represented as a \textit{normalized sum} of independent random variables,

\[ Y = \left( \frac{\sum_{i=1}^{180} X_i - 30}{5} \right). \]
1.1.2 Self-normalized sums in statistics

Statistical inference is often concerned with conclusions regarding some general quantitative phenomenon based on a limited number of observations\(^6\).

As a simple example, suppose we study the tallness of some part of the Swedish human population. Let \( N \) denote the total number of individuals in the given part of the population, and suppose that we are able to select a small number \( n \) “at random” from the considered group\(^7\). The sample average

\[
\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i,
\]

of the corresponding heights \( x_1, \ldots, x_n \) should then correspond approximately to the population average height

\[
\mu = \frac{1}{N} \sum_{i=1}^{N} x_i.
\]

Noting the analogy with the die, we may expect that \( \bar{x} \) varies according to the Gaussian curve. Moreover, it is reasonable to assume that the center of variation is given by \( \mu \). What about the “fluctuational scale”, or in other words, the standard deviation of the sample mean?

In the previous example, we knew from mathematical reasoning that \( \sqrt{5}/6 \) is the exact standard deviation of a single experiment. Now, we are interested in the standard deviation for the experiment of picking a person at random and measuring that person’s height\(^8\).

In statistical modeling, it may often be the case that descriptions of random variation are partially incomplete. A general idea in statistics is therefore that unspecified descriptions of random variation may be estimated on the basis of the information contained in the set of observations. Thus, we may let the standard deviation of the experiment remain unspecified, considering it as an unknown quantity, and replacing it with the standard deviation of the sample.

To see the analogy between the two examples, the normalized proportion of observed sixes is given by

\[
\sqrt{180} \left( \frac{1}{180} \sum_{i=1}^{180} X_i - \frac{1}{6} \right) / \sigma, \quad \text{with} \quad \sigma^2 = 5/36.
\]

---

\(^6\)In the 19th century, Statistician typically referred to someone working in a national statistical bureau, recording data relevant for socio-economical issues. Moreover, “partial investigations” according to some “representative method” were then widely considered to be imprecise and unscientific. Confer [26, Chapter 3.7].

\(^7\)A set of measurements selected at random is often referred to as a sample, with an exact mathematical definition (i.i.d. random variables) in terms of some given probability distribution.

\(^8\)This standard deviation is approximately equal to 6 cm among 18 year old Swedish men, according to Rekryteringsmyndigheten (the Swedish Defence Recruitment Agency — formerly known as “Pliktverket”).
On the other hand, the \textit{self-normalized} sample mean height is given by

\[ T_n = \sqrt{n} \left( \frac{1}{n} \sum_{i=1}^{n} x_i - \mu \right) / \hat{\sigma}_n, \quad \text{with} \quad \hat{\sigma}_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x}_n)^2. \quad (1.1) \]

Note that the standard deviation $\sigma$ is fixed in the first example, whereas the \textit{sample standard deviation}, $\hat{\sigma}_n$, is affected by the random fluctuations in the second example.

The terminology of a \textit{self-normalized sample mean} is motivated by the fact that the \textit{sample itself} is used to approximately determine the scale on which the fluctuations of the sample mean occur. Self-normalized sums of the specific form (1.1) are, for historical reasons, often referred to as \textit{t-statistic random variables}, or as studentized sums of random variables.

1.1.3 Gosset and “The probable error of a mean”

Historically, the 1908 article “The probable error of a mean” [56] marks a beginning of mathematical research on self-normalized sums. Some historians argue that the article is a landmark in modern theory of statistical inference\textsuperscript{9}. Its author is known as “Student”, a pseudonym for William Gosset, at the time working as a research chemist at the Dublin beer brewery Guinness.

Gosset [56] derives distributional properties of the above kind of self-normalized sums (1.1), assuming that the individual measurements follow a Gaussian distribution with unknown standard deviation. His results have been widely used in applied statistics during the last century when assessing the uncertainty of combined measurements, much due to the spread they achieved through Ronald Fisher’s highly influential \textit{Statistical Methods for Research Workers} [24], first appearing in 1925\textsuperscript{10}.

We noted for the first example that a properly \textit{normalized} sum of random entities may be reasonably well described by the Gaussian distribution. Gosset’s results concern exact descriptions for the given kind of \textit{self-normalized} sums. How does his descriptions differ from the Gaussian description?

Anyone acquainted with the application of statistical methods through \textit{statistical tables} of distributional percentiles probably recognizes that studentized quantiles are usually larger in absolute value compared to their Gaussian counterparts\textsuperscript{11}. In other words, one of the main messages from Gosset’s results is that \textit{t}-distributions are more \textit{heavy-tailed} in comparison with the Gaussian distribution. Intuitively, this may be considered as a compensation for not knowing the fluctuational scale in advance, which we rarely do in statistical applications.

\textsuperscript{9}Confer discussions in [63].
\textsuperscript{10}Fisher is also famous for his contributions to evolutionary biology and genetics.
\textsuperscript{11}The effect diminishes with sample size, and it vanishes as the sample size tends to infinity.
1.1.4 The assumption of normality

In a review [48] of the second edition of Fisher’s book [24], Egon Pearson criticized the claimed “exactitude” of the statistical methods following upon Gosset’s results. The basis of his argument was that the exactitude depended on an assumption of *normally distributed observations*, an assumption which may be more or less adequate in practical applications. Fisher responded:

> I have never known difficulty to arise in biological work from imperfect normality of the variation, often though I have examined data for this particular cause of difficulty; nor is there, I believe, any cause to the contrary in the literature. (cf. [63, p. 5])

Interestingly, Gosset commented in a letter to Pearson:

> Fisher is only talking through his hat when he talks of his experience; it isn’t so very extensive and I bet he hasn’t often put the matter to the test; how could he? (cf. [63, p. 5])

Nowadays, it is fairly common to state that statistical methods depending on Gosset’s results are relatively robust with respect to the assumption of normality\(^\text{12}\). Still, alternative methods, not referring to any normal assumption, have been suggested, such as methods based on ranks\(^\text{13}\). Also, many statistical methods have been proposed for the objective of evaluating the adequacy of an assumption of normality with respect to a given set of observations, so-called *tests of normality*\(^\text{14}\).

Clearly, the scientific debates among Fisher, Pearson, Gosset and others had a substantial impact on the statistical community at the time. As a consequence, much research has been made in connection to these issues during the 20th century. This can for instance be seen from the review [63], which was written in connection with the centennial in 2008 of Gosset’s original article.

1.1.5 Contributions

Papers I-III contribute to the tradition of investigating the impact of the normal assumption in relation to the random fluctuations of self-normalized sums.

Paper I focuses on the heavy-tailedness of Student’s \(t\)-statistic (1.1), investigating the phenomenon from a general point of view. For instance, it is shown that, under weak regularity assumptions, the following description of finite absolute moments in general holds:

\[
E|T_n|^r < \infty \quad \text{whenever} \quad r < n - 1.
\]

\(^{12}\) Confer for instance the comment in [31, p. 64], referring to the classic [50]. Confer also the research review [13], and the illustrative simulation studies in [1].

\(^{13}\) The methods attributed to Wilcoxon [62] are some of the most well-known.

\(^{14}\) A well-known example being the Shapiro–Wilk test [53].
Thus, roughly speaking, it is found that the heavy-tailed character is not directly linked to the normal assumption, occurring for essentially all other distributional assumptions.

Papers II and III are indirectly related to the heavy-tailedness of Student’s $t$-statistic. However, attention is mainly devoted to the following kind of self-normalized sums:

$$Z_n = \frac{\sum_{i=1}^{n} X_i}{\left(\sum_{i=1}^{n} X_i^2\right)^{1/2}}.$$  

There is a monotonic relation between $Z_n$ and $T_n$, which may be expressed as\textsuperscript{15},

$$T_n = Z_n \left(\frac{n-1}{n-Z_n^2}\right)^{1/2}.$$  

In Paper II, it is shown that self-normalized sums $Z_n$ of asymmetric random variables possess a higher degree of irregularity in terms of their quadratic moments. More precisely, the main result in Paper II is that:

$$\text{E}Z_n^2 \geq 1, \quad \text{with} \quad \text{E}Z_n^2 = 1 \quad \text{if and only if} \quad X \overset{d}{=} -X.$$  

Paper III assumes a regular situation, in the sense that sums of symmetric random variables are considered. Then, tail probabilities of $Z_n$ are characterized in terms of certain upper bounds, valid for a more general class of random variables, known as linear combinations of Rademacher random variables:

$$S = \sum_{k=1}^{n} a_k \varepsilon_k, \quad \text{with} \quad (a_1, \ldots, a_n) \in \mathbb{R}^n.$$  

Here, $\varepsilon_1, \ldots, \varepsilon_n$ refers to independent random variables each assuming the two values $\{-1, 1\}$ with probability 1/2, respectively. Also, a standardized scale of measurement is considered, corresponding to the assumption

$$\sum_{k=1}^{n} a_k^2 = 1.$$  

Then, the main result in Paper III is that, for some constant $C$:

$$P(S \geq x) \leq N(x)(1 + Cx^{-2}),$$  

with $N(x)$ referring to the corresponding standard normal tail probability.

\textsuperscript{15}Confer e.g. [14, p. 1].
1.2 Directional conclusions

The most classical reference to self-normalized sums occurs in connection with statistical testing of hypotheses\textsuperscript{16}. We review some background and concepts in Sections 1.2.1 and 1.2.2. Also, two examples are introduced. Directional conclusions are introduced with the second example, regarding polls and public opinions, and further discussed in Section 1.2.3. Finally, Paper IV is introduced in Section 1.2.4.

1.2.1 Statistical significance

The following historical example\textsuperscript{17} is illustrative regarding some of the basic ideas behind testing in statistics. The 19th century Czech monk Gregor Mendel is famous for his laws of inheritance in genetics. For instance, Mendel experimented with cross-pollination of pea plants, among which some produced green peas and some yellow. Mendel believed that yellow was dominant with respect to green. As a result, yellow and green should be observed according to the 3:1 ratio in a well-mixed, cross-pollinated population.

In analogy with our first example with dice, the proportion of yellow peas within a given selection of cross-pollinated plants is rarely believed to be exactly 2/3, even if one believes in Mendel’s theory. But in a given experiment, what fluctuations around 2/3 should be considered small, and when should we start putting Mendel’s conclusions in doubt?

In the previously mentioned book [24] by Fisher, probability theory applied to Mendelian experiments is one prominent example. In this tradition, we may consider the population proportion of yellow peas as an unknown constant $p$, to which we associate the hypothesis:

$$H : p = \frac{2}{3}.$$ 

Next, we may either reject or not reject $H$ depending on how much the observed proportion of yellow peas $\hat{p}$ deviates from 2/3. Two elements enter when forming this formal decision:

- A description of the random variation of $\hat{p}$ when $H$ is true;
- A level of significance, determining the risk of falsely rejecting $H$.

In the tradition following upon Fisher, 5% is often considered as a reasonable level of significance. Alternatively, one may report the least level of significance leading to rejection, that is, the so-called $p$-value associated to the experiment.

\textsuperscript{16}Historically, the notion of hypothesis testing if often attributed to J. Neyman (1894-1981) and E.S. Pearson (1895-1980), whereas the slightly different approach of R.A. Fisher (1890-1962) is often associated with the notion of significance testing. Confer [43] for a review of the two approaches, arguing that the two theories are complementary rather than contradictory, despite the fierce debates between Fisher and Neyman.

\textsuperscript{17}Confer [26, Chapter 4.4] and [23].
To summarize, a statistical test may either reject or not reject a given hypothesis \( H \). Whether it rejects or not depends on the chosen level of significance, which refers to the risk of falsely rejecting \( H \). Rejecting \( H \) means that given data are sufficiently unlikely from the perspective that \( H \) is true.

One may note that, not rejecting \( H \) does not necessarily mean that \( H \) is confirmed, only that given observations are sufficiently likely, given the experimental set-up and the perspective that \( H \) is true. Strictly speaking, we can never confirm Mendel’s ideas on dominant gene expressions in the above example, only measure to what extent empirical observations are coherent with his predictions.

As a final remark, it has been suggested that Mendel’s original observations are in fact too coherent with his hypotheses\(^\text{18}\). In other words, the observed proportions in Mendel’s reports are remarkably close to the predicted proportions, in view of the associated random variation. This may raise the suspicion that Mendel, or some of his assistants, adjusted the data to better fit the proposed theory.

1.2.2 Investigating public opinions
Polls are often conducted in modern democratic societies in order to evaluate the political preferences of the public opinion. As an example, suppose that 2000 potential voters in a Swedish election to parliament are selected and asked as to which political party they would vote for “if elections were held at the present day”\(^\text{19}\). Moreover, suppose we are interested in whether the present government would obtain a majority of the votes or not. Here, as in the Mendelian example, it is natural to consider an unknown proportion \( p \), now referring to voters sympathizing with the present government, being estimated through a representative selection of individuals. Is there also a similar hypothesis that we would like being tested?

Many statisticians would answer the last question affirmatively. Indeed, it is common to consider the given example in terms of statistical testing of the following hypothesis:

\[
H : p = 1/2.
\]

Hence, if \( H \) is rejected, given some level of significance, we would say that the observed majority/non-majority is significant\(^\text{20}\). On the other hand, \( H \) not being rejected may be referred to as a non-significant result.

\(^{18}\) This was noted by Fisher in [23] (confer also [26, Chapter 4.4]).
\(^{19}\) Choosing a sample at random may not always be the most efficient strategy. For instance, dividing the population into a given number of strata, subpopulations of known proportions in which the opinion is more homogenous, may lead to a substantial increase in efficiency. Confer e.g. [10, Chapter 5] regarding the concept of stratified sampling.
\(^{20}\) Statistiskt säker in modern Swedish.
Comparing the two examples, $H$ naturally arises as a scientific hypothesis in the first example. In the second case, we are not primarily interested in the degree of coherence between observations and the corresponding hypothesis. Rather, the main interest is directional: whether the proportion $p$ is smaller or greater than $1/2$, which we refer to as majority vs. non-majority. In reality, we may very well assume in this case that the true proportion is either larger than $1/2$ or smaller than $1/2$ (even though it may be very close to $1/2$).

Nevertheless, it is reasonable to ask for controlled conclusions in both examples. It would be silly to reject the Mendelian hypothesis whenever the observed proportion of yellow peas differs from $2/3$. In the second example, there is a large chance of making a false directional conclusion if the true directional effect is small compared to the precision of measurement (which is usually about $\pm 1\%$ of the total number of voters in this case).

### 1.2.3 Type III errors

Following the distinguished statistician David Cox\textsuperscript{21}, one may refer to $H$ in the hypothetical example with public opinions as a dividing hypothesis. According to Cox, it may be reasonable to put $H$ to test, not necessarily because there is some special reason to believe that $H$ is true, but primarily since $H$ divides the parameter space into two different regions and we are interested in whether the data establish reasonably clearly which region is correct.

It is then reasonable to consider the resulting procedure as a three-decision procedure\textsuperscript{22}. In other words, when rejecting $H$ we also make a conclusion regarding majority or non-majority according to the estimated proportion. In the given kind of context, statisticians\textsuperscript{23} then refer to three kinds of errors, Type I, Type II and Type III. As usual, Type I refers to falsely rejecting $H$, and Type II to not rejecting $H$ when there actually is a directional difference. The third kind of error refers to making the reversed directional conclusion in a situation where $H$ is false.

Now suppose that we impose a control corresponding to a probability $\alpha$ (typically $5\%$, $1\%$ or $0.1\%$) of falsely rejecting $H$. What control do we then impose on Type III errors?

Interestingly, the answer is usually $\alpha/2$. This is naturally explained by the fact that if the true directional effect is very small, then we reject $H$ with probability $\alpha$. Moreover, we then usually observe the correct directional effect with probability $1/2$, independently of whether $H$ is rejected or not. Thus, the reversed directional conclusion is inferred with probability $\alpha/2$ in this kind of situation.

\textsuperscript{21}Confer e.g. [12, p. 31] and [11].

\textsuperscript{22}Confer for instance Lehmann [42].

\textsuperscript{23}In the tradition of for instance Hartner [33] and Kimball [39]. Confer also the recent review [52] by Juliet Shaffer.
The interpretation of the given kind of three-decision procedures may in fact be somewhat complicated. Recall the claim made in Section 1.2.2 that our main interest was directional. With the suggested procedure, the main control, $\alpha$, is imposed on falsely rejecting $H$, and a different control, $\alpha/2$, is imposed on the directional errors. Were we not mainly interested in a directional control in this case?

Also, consider the interpretation of $H$. Some\(^{24}\) prefer to extend the idealized mathematical construction by including indeterminately small deviations from $p = 1/2$ in $H$, or deviations vanishingly small on the given scale of measurement. With this interpretation, $H$ differs somewhat vaguely from its directional alternatives:

$$H_1 : p < 1/2, \quad H_2 : p > 1/2.$$  

Others, like Jones and Tukey\(^{25}\) [38, p. 411], prefer to take a more resolute and provocative “philosophical” stand, claiming that

[... point hypotheses about parameters in real-world populations [...] are always untrue when calculations are carried to enough decimal places.]

According to this stand, testing $H$ is not the appropriate procedure for obtaining a controlled conclusion in this case\(^{25}\).

### 1.2.4 Contributions

Paper IV is concerned with a general framework for controlled directional conclusions, following the suggestion by Jones and Tukey\(^{38}\). For the example considered in Section 1.2.2 we would then associate two directional hypotheses: majority versus non-majority, with three possible answers:

- Conclude majority;
- Conclude non-majority;
- No conclusion.

Contrary to the three-decision procedures in Section 1.2.3, it is assumed that either of “majority” or “non-majority” is the correct description of reality.

Statistical tests control the rates of false rejections, whereas three-decision procedures in this case control reversal rates; probabilities of concluding the false direction. As mentioned in Section 1.2.3, controlling the false rejections of the border between the two directions at level $\alpha$ is often equivalent to controlling the reversal rate in the directional setting at level $\alpha/2$.

---

\(^{24}\) Confer for instance Scheffé\(^{51}\).

\(^{25}\) Many of the issues discussed in this section also occur in a recent debate in psychology, with contributors deeply questioning the current statistical practice of significance testing. Confer for instance the provocative\(^{38}\), \(^{32}\), and the review articles\(^{47}\), \(^{40}\).
When applying a given three-decision procedure, two balancing issues are of relevance:

• Controlling the reversal rates at a reasonable level;
• Being able to draw the right conclusion “as often as possible”.

Following Jones and Tukey [38], one may for instance state—regarding the choice of a procedure—that:

We want to control the rate of error, the reversal rate, while minimizing wasted opportunity, that is, while minimizing indefinite results ([38, p. 412]).

Paper IV is concerned with the suggestion of Jones and Tukey, and similar optimality characterizations in the setting of three-decision procedures for directional conclusions. Three kinds of characterizations are presented,

• Maximizing rates of correct directional conclusion;
• Minimizing rates of indefinite conclusion;
• Minimizing rates of incorrect directional conclusion.

Moreover, the three characterizations are related to optimality criteria from the theory of statistical testing, which we then apply to exponential families of distributions.

1.3 The problem of multiplicity

In statistics, attempts to infer several kinds of conclusions on the same set of data is often referred to as multiple inference. Naively, one may treat each attempted conclusion separately, with no correction for multiplicity. This may sometimes lead to undesirable consequences. For instance, with many uncorrelated random experiments, it is likely that the results of some experiments will be misleading due to random fluctuations and the occurrence of rare events.

We begin by introducing an example from biostatistical research. Next, we review two general approaches to control the possible effects of multiplicity in Section 1.3.2, control of FWER and FDR respectively. Finally, Paper V is briefly presented in Section 1.3.3.

1.3.1 DNA microarray data

Since the 1980s, biotechnological developments have turned the analysis of DNA microarray data into an important part of genetics and molecular biology, with applications in medicine. The power of microarray experiments is related to the fact that thousands of DNA sequences may be analyzed simultaneously at a comparatively low cost.
As an example, Callow et al. [8] studied lipid metabolism among mice. It was known that mice with the gene Apolipoprotein AI knocked out have very low HDL cholesterol levels. The goal of the study was to identify genes with altered gene expression in the livers of mice with Apo AI knocked out. Eight mice with Apo AI knocked out were compared to a control group of eight mice. For each of the 16 individuals, a microarray with 5,548 spots corresponding to the gene expressions of different DNA sequences was constructed. Thus, from the statistical point of view, 5,548 questions of the kind “Is there an altered gene expression between the two groups?” were considered on the basis of two samples with eight individuals in each sample.

The main conclusion of the Apo AI study was that 8 of the 5,548 sequences stand out as being differentially expressed between knock-out and control mice. Interestingly, many of the remaining 5,540 statistical comparisons indicate a significantly different gene expression if no correction of multiplicity is taken into account. These gene comparisons are nevertheless analyzed as not sufficiently significant, perhaps being false significances. In other words, the variation associated with the given multiple comparison study (on the basis of merely 16 individuals) may lead to some non-true “significant differences” appearing between the two groups, effects which should rather be explained on an individual basis.

The study thus illustrates that it might be reasonable to raise the standards as to what one considers as a significant finding when testing many hypotheses.

1.3.2 FWER and FDR

Recall the suggestion to consider 5% as a reasonable level of significance regarding the maximal risk of falsely rejecting a hypothesis. One way of generalizing this suggestion to the setting of multiple testing is to say that the risk of a simultaneous procedure to falsely reject some true hypothesis should not be larger than 5%. Such procedures are said to control the familywise error rate (FWER) at level 5%.

Some statisticians, e.g. Benjamini and Hochberg [2], have argued that the approach of controlling FWER may be too conservative, and not always needed. As a result, corresponding procedures may be too restrictive as to what we consider as significant findings.

Suppose, as in the example with mice, that we compare two groups of individuals through measurements of several types of quantities. For instance, the well-being and health of two groups of human beings in a medical study may

---

26 Confer also [16, Chapter 9.2].
27 Confer [15] for an overview of different approaches to multiple hypothesis testing in the context of identifying differentially expressed genes in microarray experiments. Confer also the review [28], giving more attention to the scientific background and the notion of false discovery rates. Finally, we also refer to [18] in this context, proposing an alternative “empirical Bayes approach”, referring to a notion of local false discovery rates.
be measured in various ways. Controlling the FWER means that it is unlikely that we obtain any false conclusion regarding the various comparisons of the two groups. This may also imply that we make hardly any conclusions at all. Following Benjamini and Hochberg, one may argue that the main interest may be whether one of the two groups is generally healthier than the other. Thus, in a general discussion of multiple testing ([2, p. 290]), Benjamini and Hochberg suggest:

The overall conclusion that the treatment is superior need not be erroneous even if some of the null hypotheses are falsely rejected.

Replacing control of FWER, Benjamini and Hochberg propose control of what they refer to as the *false discovery rate* (FDR). Such procedures control the expected value of the following *false discovery proportion*:

\[
\frac{\text{Number of false conclusions}}{\text{Number of conclusions}},
\]

at some prespecified level.

To summarize, a certain variability regarding the obtained set of conclusions when repeating a scientific experiment is undesirable, inevitable, but in need of control, from the statistical point of view. Controlling a *familywise error rate* means that it is unlikely that the set of conclusions contains any erroneous statement at all. With a less restrictive approach, controlling a *false discovery rate* means that the erroneous conclusions are few compared to the total number of conclusions\(^{28}\).

1.3.3 Statistical models and decisions

Abraham Wald (1902-1950) formulated a general mathematical framework for statistical inference in his book *Statistical Decision Functions* [60], today considered as a pioneering work in theoretical statistics. This theory is closely related to the so-called theory of games. A few years earlier, John von Neumann\(^{29}\) (1903-1957) and Oskar Morgenstern (1902-1977) had published their highly influential *Theory of Games and Economic Behaviour* [59].

In statistical decision theory, a *statistical model* refers to a set of possible probability distributions, with respect to some given data. Observed data are then inserted into a *decision function*, which is not allowed to know from which distribution the data came. Finally, a *loss function*, knowing both the

\(^{28}\)Some statisticians (cf. [22]) have argued that the relative nature of the false discovery rate may be problematic: consider adding hypotheses of no interest, which are known to be false with high degree of certainty. The total number of conclusions is then increased in an artificial way, whereas the number of erroneous conclusions remains essentially unchanged. This is referred to as “cheating with FDR” in [22].

\(^{29}\)von Neumann is also known for many important contributions to various subfields of mathematics, physics and computer science, and for participating in the *Manhattan project* during World War II.
data and the associated distribution, evaluates the performance of a given decision function. Ideally, a decision function should minimize the expected loss, for a given distribution and a given class of competing decision functions. If this is the case for all distributions in the model, the decision function is said to be uniformly optimal.

As an example, statistical testing (rejection vs. non-rejection of a given hypothesis $H$) might be conceived in terms of binary decision functions. Naturally, there are two kinds of errors (or “losses”) associated with such kind of decisions:

- Rejecting $H$ when $H$ is true;
- Not rejecting $H$ when $H$ is false.

Usually, we choose a decision procedure such that the errors of the first kind are relatively rare, referred to as testing with a control imposed on the rate of false rejections. Given this premise, it is then natural to search for a procedure with as much power as possible, meaning that the rates of errors of the second kind should be as low as possible.

Returning to the issue of multiplicity, we might view a multiple test procedure as a decision function with $2^n$ possible decisions with respect to the $n$ hypotheses $H_1, \ldots, H_n$ being tested. This may be achieved through indicator functions $I_1, \ldots, I_n$, with $I_1$ being 1 or 0 depending on whether $H_1$ is rejected or not, $I_2$ being 1 or 0 depending on $H_2$, etcetera. Moreover, it is natural to consider zero-one loss functions $L_1, \ldots, L_n$, with $L_1$ being 1 or 0 depending on whether $H_1$ is true or not, $L_2$ 1 or 0 depending on $H_2$, etcetera. We may then determine whether some hypothesis is falsely rejected with reference to the following indicator random variable:

$$L_1 = \max \{L_1 I_1, \ldots, L_n I_n\}.$$ 

Consequently, procedures controlling the FWER at level 5% are characterized by $E(L_1) \leq 0.05$, with respect to the expected value $E(L_1)$ of $L_1$. Similarly, we may determine the false discovery proportion according to:

$$L_2 = (L_1 I_1 + \cdots + L_n I_n)/(I_1 + \cdots + I_n).$$

Thus, procedures controlling the FDR at level 5% are characterized by $E(L_2) \leq 0.05$.

1.3.4 Contributions

Benjamini and Hochberg formulated a general FDR-controlling procedure in their seminal paper [2], appearing in 1995. Paper V generalizes their procedure from the traditional setting of multiple testing to a general framework in statistical decision theory.
For multiple testing, the Benjamini–Hochberg procedure assumes a situation of $p$-values $\hat{p}_1, \ldots, \hat{p}_n$, with respect to statistical hypotheses $H_1, \ldots, H_n$. The considered procedure is deeply rooted in the tradition of stepwise procedures for multiple testing\textsuperscript{30}. In this area, it is well-known that the correspondence between Type I control at level $\alpha$ and Type III, directional control at level $\alpha/2$ is lost for most stepwise FWER-controlling procedures. In fact, it may even be difficult to affirm that the less restrictive level-$\alpha$ control on Type III errors is obtained\textsuperscript{31}. What about FDR-control and the Benjamini–Hochberg procedure?

This issue was raised for instance in 2002 by Shaffer [52, p. 366]. Appearing three years later, Benjamini and Yekutieli [4] verified a correspondence result between Type I control at level $\alpha$ and Type III control at level $\alpha/2$. Thus, if point hypotheses $H_1, \ldots, H_n$ are tested in a situation where none of them are exactly true and the level-$\alpha$ Benjamini–Hochberg procedure is applied to the corresponding $p$-values, then the expected proportion of false directional conclusions is bounded by $\alpha/2$.

Inspired by this surprising regularity possessed by the Benjamini–Hochberg procedure, Paper V considers a formulation of the procedure in terms of “generalized $p$-values”, incorporating the above mentioned results from [2] and [4]. It is also demonstrated that the generalized perspective is relevant for the three-decision procedures for directional conclusions presented in Section 1.2.4. Finally, we note that several competing FDR-controlling procedures can not be extended in a similar fashion.

\textsuperscript{30}The pioneering suggestion [35] by the Swedish statistician Sture Holm, appearing in 1979, is one of the most cited procedures in this tradition. This procedure is still appealing due to the generality of its imposed FWER-control.

\textsuperscript{31}A review of some attempts in this direction is given in [52].
2. Summary of Papers

2.1 Paper I

Paper I is concerned with \textit{t-statistic random variables}. In other words, given i.i.d. (independent and identically distributed) random variables $X_1, \ldots, X_n$ with respect to some distribution $F$ and an integer $n \geq 2$, we consider the random variable

$$ T_n = \begin{cases} 
    n^{1/2} \bar{X}_n / \hat{\sigma}_n, & \hat{\sigma}_n > 0, \\
    0, & \hat{\sigma}_n = 0.
\end{cases} $$

Here, $\bar{X}_n$ refers to the sample mean and $\hat{\sigma}_n^2$ to the sample variance, according to

$$ \bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i, \quad \hat{\sigma}_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2. $$

The most prominent example of a t-statistic random variable regards the situation when $F$ is a normal distribution with zero mean. The distribution of $T_n$ is then known as the \textit{t-distribution with $n - 1$ degrees of freedom}. Such distributions are absolutely continuous (with respect to the Lebesgue measure) with density functions given by (cf. e.g. Appendix B in [30]),

$$ f_n(x) = \frac{\Gamma\left(\frac{n}{2}\right)}{\sqrt{\pi(n-1)} \Gamma\left(\frac{n-1}{2}\right)} \left(1 + \frac{x^2}{n-1}\right)^{-\frac{n}{2}}, \quad -\infty < x < \infty. \quad (2.1) $$

With reference to (2.1) one may thus verify the following property regarding the corresponding absolute moments $\mathbb{E}|T_n|^r$,

$$ \mathbb{E}|T_n|^r < \infty \quad \text{whenever} \quad r < n - 1. \quad (2.2) $$

In other words, moments are finite only below the degree of freedom. Hence, there is a heavy-tailed behaviour compared to the standard normal distribution, for which all absolute moments are finite. Moreover, (2.2) also indicates that the degree of heavy-tailedness is decreasing with increasing degrees of freedom.

Motivations exist for comparing the distribution of $T_n$ with the standard normal distribution. Indeed, by the law of large numbers and the central limit theorem, $T_n$ converges in distribution to the standard normal distribution, for
any fixed distribution $F$ with zero mean and finite variance. Then, one may guess that there is also convergence of the moments,

$$E|T_n|^r \to E|T|^r,$$

(2.3)

with respect to a random variable $T$ following the standard normal distribution.

Regarding the main results in Paper I, Theorem 4.1 states that a finite moment $E|T_n|^r$ implies finite moments $E|T_m|^r$ for all $m > n$, with respect to any given distribution $F$. Thus, Theorem 4.1 supports the perspective on the degree of heavy-tailedness as being non-increasing with the number of observations.

Next, Theorem 5.1 partly generalizes property (2.2) by stating that $r < n - 1$ is a necessary condition for $E|T_n|^r < \infty$, provided $F$ has a non-discrete component. Thus, the heavy-tailed character of “finite absolute moments only below the degree of freedom” is essentially a general property of $t$-statistic random variables.

As a third result, Theorem 4.3 demonstrates that the range of finite absolute moments may be more narrow than $r < n - 1$ in case $F$ has probability mass intensively concentrated locally. On the other hand, it is also shown that, if $F$ is sufficiently regular in this respect, then property (2.2) holds.

Finally, we also verify (Theorem 6.1) that that property (2.3) holds, assuming the existence of a (possibly non-normally distributed) random variable $T$ such that $T_n$ converges in distribution to $T$, and that $E|T_{n_0}|^r$ is finite for some integer $n_0$. We refer to [9], and more generally to the monograph [14], for complete characterizations regarding the situations in which $T_n$ converges in distribution.

All results in Paper I are proved with reference to the random variables $U_n^*$ defined by

$$U_n^* = \begin{cases} 0, & (S_n/V_n)^2 = n, \text{ or } V_n = 0, \\ (S_n/V_n)^2, & \text{else}. \end{cases}$$

Here, as in Paper II, $S_n$ and $V_n$ refer to random variables defined by

$$S_n = \sum_{i=1}^n X_i, \quad V_n = \left( \sum_{i=1}^n X_i^2 \right)^{1/2}.$$ 

(2.4)

The following characterization (cf. [14, p. 1]) relates the tail probabilities of $T_n$ to the tail probabilities of $U_n^*$,

For any $x \geq 0$, $T_n^2 > x$ whenever $U_n^* > nx/(n+x-1)$.

18
Also, tail probabilities and absolute moments are related according to (cf. [29, Section 1.12]),
\[
E|T_n|^r = r \int_0^\infty x^{r-1}P(|T_n| > x)dx.
\]

The key observation used in proving the theorems of Paper I is that the tail probabilities of \( T_n \) are related to certain probabilities of having almost identical observations \( X_1, \ldots, X_n \).

### 2.2 Paper II

Paper II refers to the following type of self-normalized sums,
\[
Z_n = \frac{S_n}{V_n},
\]
with \( S_n \) and \( V_n \) defined as in formula (2.4) above. Thus, as in Paper I, a setting of i.i.d. random variables \( X_1, \ldots, X_n \), given some underlying distribution \( F \), is considered. Moreover, it is assumed that \( P(X = 0) = 0 \), so that \( Z_n \) is well-defined with probability one.

Integer moments of self-normalized sums were studied by Efron in [17], as a part of more general distributional investigations. However, Efron considered the restricted perspective of \( X \) being symmetrically distributed around the origin, which we in the following refer to as
\[
X \overset{d}{=} -X,
\]
and noted that \( EZ_n^2 \) is then always equal to one.

Indeed, by decomposing a symmetrically distributed random variable \( X \) multiplicatively as \( X = \varepsilon |X| \), one may observe that \( \varepsilon \) is being of Rademacher type. In other words, the distribution of \( \varepsilon \) is given by
\[
P(\varepsilon = 1) = 1/2 \quad \text{and} \quad P(\varepsilon = -1) = 1/2.
\] (2.5)

Moreover, one also notes that \( \varepsilon \) is independent of \( |X| \). It then follows, due to independence and the fact that \( E(\varepsilon) = 0 \), that
\[
EZ_n^2 = E\left(\left(\sum_{i=1}^n X_i \right)^2 \right)/\left(\sum_{k=1}^n X_k^2\right) = 1 + E\left(\sum_{i \neq j} \varepsilon_i \varepsilon_j |X_i||X_j|/\sum_{k=1}^n X_k^2\right) = 1.
\]

The main result in Paper II may in short be stated as follows:
\[
EZ_n^2 \geq 1, \quad \text{with} \quad EZ_n^2 = 1 \quad \text{if and only if} \quad X \overset{d}{=} -X.
\] (2.6)

In other words, \( EZ_n^2 = 1 \) characterizes the symmetry assumption in the general i.i.d. setting, and there exists no self-normalized sum \( Z_n \) such that \( EZ_n^2 < 1 \).
The remaining parts of Paper II concern some further results regarding the connection between distributional symmetry and the quadratic moments $EZ^2_n$. For instance, we ask, with $n$ fixed, how can asymptotically negligible quantities $EZ^2_n - 1$ be characterized in terms of the underlying distributions? Here, an expected answer of “asymptotic symmetry” turns out to be somewhat disturbed by the possible impacts of probability mass concentrations locally around the origin. Confer Section 4 and the Appendix of Paper II for details.

When $n$ becomes large, the connection between distributional symmetry of $X$ and the quadratic moments $EZ^2_n$ becomes weaker, roughly speaking. For instance, it suffices that $X$ has zero mean and finite variance in order for $Z_n$ to converge in distribution to the standard normal distribution. Moreover, $EZ^2_n$ then converges to 1, by a result due to Giné et. al [27] (cf. Section 5 in Paper II). Note that the distribution of $X$ may be far from symmetric.

Furthermore, we verify, with the distribution of $X$ varying with $n$, that $Z_n$ may be asymptotically normal (and thus asymptotically symmetric) whereas $EZ^2_n$ at the same time may diverge to infinity. In fact, the distribution of $X$ is increasingly asymmetric in the presented example.

The following integral expression for $EZ^2_n$, which we learned from Professor Svante Janson and later discovered to be implicit in [27] (cf. Corollary 2.10 and Lemma 3.1 there), was essential for the proofs in Paper II,

$$EZ^2_n = 1 + (n^2 - n) \int_0^\infty (E(Xe^{-tX^2}))^2 (Ee^{-tX^2})^{n-2} \, dt. \quad (2.7)$$

For instance, it is immediately clear from (2.7) that $EZ^2_n \geq 1$. One may also note that $EZ^2_n = 1$ essentially requires

$$E(Xe^{-tX^2}) = 0, \quad \text{for all} \quad t \geq 0,$$

which in fact characterizes symmetry (cf. the proof of Theorem 3.1).

### 2.3 Paper III

The main object of study in Paper III is the family of linear combinations of Rademacher random variables. In other words, given a vector $(a_1, \ldots, a_n) \in \mathbb{R}^n$ and a vector $(\varepsilon_1, \ldots, \varepsilon_n)$ of independent Rademacher random variables (cf. condition (2.5) above), we consider the linear combination

$$S = \sum_{k=1}^n a_k \varepsilon_k.$$

In particular, our interest regards upper bounds on the tail probabilities $P(S \geq x)$, for fixed values $x > 0$.  

20
To begin with, replacing $a_k$ by $-a_k$, for any $k = 1, \ldots, n$, does not affect the distribution of $S$, due to the distributional symmetry of a Rademacher random variable. Thus, we impose no restriction by assuming $a_k \geq 0$, for all $k = 1, \ldots, n$. Moreover, $S$ has mean zero and variance $\sum_{k=1}^{n} a_k^2$.

As a first approximation, assuming approximate normality of $S$, with reference to the Central Limit Theorem,

$$P(S \geq x) \approx N\left(\sum_{k=1}^{n} a_k^2 \sqrt{x}\right),$$  \hspace{1cm} (2.8)

with $N(x)$ referring to the corresponding standard normal tail probability. A slight rewriting of (2.8) yields

$$P\left(\left(\sum_{k=1}^{n} a_k^2 \sqrt{S} \right) \geq x\right) \approx N(x),$$  \hspace{1cm} (2.9)

where we note that

$$\left(\sum_{k=1}^{n} a_k^2 \right)^{-1/2} S = \sum_{i=1}^{n} \left(a_k \left(\sum_{k=1}^{n} a_k^2 \right)^{-1/2}\right) \varepsilon_k$$

also refers to a linear combination of Rademacher random variables, indeed, with unit variance. Moreover, any approximation (or inequality)

$$P\left(\left(\sum_{k=1}^{n} a_k^2 \right)^{-1/2} S \geq x\right) \approx T(x),$$

on the so-called *standardized scale* may be translated to the original scale by

$$P(S \geq x) \approx T\left(\sum_{k=1}^{n} a_k^2 \sqrt{\varepsilon_k}\right).$$

Clearly, with $n = 1$ and $x > 1$, (2.9) is a rather poor approximation, since the left hand side is equal to zero. In general, the normal approximation may indeed be rather *conservative* if $n$ is small in relation to $x$, with much larger approximations compared to what is being approximated. Now, how about deviations in the other direction?

The previous question motivates the following restricted perspective: finding upper bounds on the tail probabilities $P(S \geq x)$ in relation to the normal approximation, for fixed values $x > 0$, valid for all non-negative numbers $a_1, \ldots, a_n$ such that

$$\sum_{k=1}^{n} a_k^2 = 1.$$
Regarding previously obtained results in this setting, Efron [17] proved that

$$P(S \geq x) \leq e^{-x^2/2}. \quad (2.10)$$

Later, Bobkov et al. [6], Pinelis [49] and others obtained global bounds of the form

$$P(S \geq x) \leq C_1 N(x), \quad (2.11)$$

for some constant $C_1 > 3$. Note that (2.11) improves upon (2.10), for sufficiently large $x$, due to the well-known fact that

$$N(x) \sim e^{-x^2/2(\sqrt{2\pi}x)}^{-1}, \quad \text{as} \quad x \to \infty.$$  

We prove in Paper III that, for some constant $C_2$,

$$P(S \geq x) \leq N(x)(1 + C_2 x^{-2}). \quad (2.12)$$

Clearly, (2.12) improves upon (2.11), for sufficiently large $x$.

We also show that there exists a divergent sequence $\{x_n\}$ such that

$$\lim_{n \to \infty} x_n^2 (P(B_n \geq x_n)/N(x_n) - 1) = 3, \quad (2.13)$$

with respect to the sequence of normalized, symmetric binomial random variables $B_n$,

$$B_n = (\varepsilon_1 + \cdots + \varepsilon_n)/\sqrt{n}. $$

Thus, (2.13) demonstrates that $C_2 \geq 3$ is necessary in (2.12), independently of the size of $x$.

Finally, we comment on the applicability and relevance of the bounds (2.12) with respect to $S = S_n$ and the self-normalized sums

$$S_n = \sum_{k=1}^{n} X_k / \left( \sum_{k=1}^{n} X_k^2 \right)^{1/2}$$

of i.i.d. zero-mean, normally distributed random variables $X_k$, showing that in this case

$$P(S \geq x) \leq N(x), \quad \text{for} \quad x \geq \sqrt{6}. \quad (2.14)$$

The main result (2.12) is proved through a technique known as exponential tilting, often associated with the theory of large and moderate deviations. Roughly speaking, given a large value $x$, the idea is to consider an exponential tilting of the distribution of the random variable $S$ into a related distribution $\tilde{G}$, with mean $x$. In this way, the problem of estimating the quantity $P(S \geq x)$ is transferred to a problem of estimating a quantity relating to $\tilde{G}$. Moreover,
it turns out that traditional estimates for probability distributions (in this case the Berry–Esseen theorem, cf. [29, Section 7.6]) may be applied rather successfully to the latter kind of quantity, since the tilted distribution is centered at the boundary point $x$. To obtain inequality (2.12) it then suffices to analyze the relation between the two types of quantities.

On the other hand, our approach for verifying the accompanying result (2.13) is more direct, starting from the lattice structure associated with binomial distributions, with probability point masses

$$p_{k,n} = \binom{n}{k} 2^{-n}.$$

The contributions of these point masses may then be analyzed asymptotically by Stirling approximation (cf. [19, Chapters II and VII]).

Finally, (2.14) is verified with reference to $t$-distributions and analytic expressions of the corresponding density functions (see formula (2.1) above).

### 2.4 Paper IV

The main results in Paper IV regard a general decision theoretic setting, with a set of probability distributions $\{P_\theta : \theta \in \Theta\}$, where each element is a possible law of a random element $X$. Moreover, the parameter space $\Theta$ is supposed to be divided into two disjoint subsets, $\Theta_1$ and $\Theta_2$. The three outcomes of a randomized three-decision procedure are associated with:

- Concluding $\theta \in \Theta_1$;
- Concluding $\theta \in \Theta_2$;
- Not drawing any conclusion.

Of relevance are then, for a given three-decision procedure $\phi$ and a given parameter $\theta$, the three probabilities: $\pi$ (drawing the correct conclusion); $\rho$ (drawing the incorrect conclusion); and $1 - \pi - \rho$ (not drawing any conclusion). In other words, consider $\pi$ and $\rho$ given by:

$$\pi(\theta, \phi) = E_\theta (\phi_1 I\{\theta \in \Theta_1\} + \phi_2 I\{\theta \in \Theta_2\}), \quad \text{for } \theta \in \Theta,$$

$$\rho(\theta, \phi) = E_\theta (\phi_2 I\{\theta \in \Theta_1\} + \phi_1 I\{\theta \in \Theta_2\}), \quad \text{for } \theta \in \Theta.$$

Jones and Tukey [38] considered a related setting, referring to two-sample mean-comparisons, with assumptions of normality and unknown but equal variances. Formally, let $\mu_A$ and $\mu_B$ refer to the population means of the two samples, respectively. Moreover, let the parameter space be divided into the two regions:

$$\Theta_1 : \mu_A \leq \mu_B, \quad \Theta_2 : \mu_A > \mu_B.$$
In this case, a reasonable procedure may be obtained through level-$\alpha$ one-sided $t$-testing of both directions, with definite conclusions whenever the corresponding opposite direction is rejected. A little thought reveals that $\rho$ (the probability of a “reversal”) is bounded by $\alpha$ in this case. Thus, the rate of error is controlled at level $\alpha$.

Jones and Tukey also expressed the desire, regarding the choice of a three-decision procedure, that

\begin{align*}
&\text{We want to control the rate of error, the reversal rate, while minimizing wasted opportunity, that is, while minimizing indefinite results ([38, p. 412]).}
\end{align*}

Hence, a three-decision procedure might be considered optimal in case it minimizes the rates of indefinite conclusion, given a bound on the rate of error.

Paper IV is concerned with an extended setting, with the following three kinds of optimality characterizations, all assuming that the rate of error is controlled at level $\alpha$:

- Maximized rates of correct directional conclusion;
- Minimized rates of indefinite conclusion;
- Minimized rates of incorrect directional conclusion.

To begin with, we argue by example that the second perspective (the Tukey–Jones proposal) is difficult to fulfill unless some further regularity condition is imposed on the class of competing procedures.

Next, we propose associating the three perspectives to well-known criteria from testing theory. To this end, consider a conclusion $\theta \in \Theta_1$ as a rejection of $\theta \in \Theta_2$, and a conclusion $\theta \in \Theta_2$ as a rejection of $\theta \in \Theta_1$, with corresponding test functions $\phi_1$ and $\phi_2$. Then, it is demonstrated that a three-decision procedure controls the reversal rate at level $\alpha$ whenever $\phi_1$ and $\phi_2$ are level-$\alpha$ tests. Moreover, the three-decision procedure could be associated with the first perspective on optimality in case:

- $\phi_1$ and $\phi_2$ are uniformly most powerful among level-$\alpha$ unbiased tests of $\Theta_2$ and $\Theta_1$, respectively.

Similarly, a three-decision procedure could be associated with the third perspective in case:

- $1 - \phi_1$ and $1 - \phi_2$ are uniformly most powerful among level-$(1 - \alpha)$ unbiased tests of $\Theta_1$ and $\Theta_2$, respectively.

Assuming an Euclidean parameter space $\Theta$ and continuously varying distributional shapes, it is demonstrated that a three-decision procedure could be associated with the second perspective in case:

- $\phi_1 + \phi_2$ is uniformly most powerful among level-$2\alpha$ unbiased tests of the boundary between $\Theta_1$ and $\Theta_2$. 

24
Finally, regarding the existence of interesting three-decision procedures satisfying the given criteria, we consider the setting of *multiparameter exponential families*. Moreover, regarding θ, referring to one of the parameters in such a model, we consider

\[ \Theta_1 : \theta \leq \theta_0, \quad \Theta_2 : \theta > \theta_0, \]

given some \( \theta_0 \in \mathbb{R} \). It is shown that the first and third perspectives are fulfilled by a natural candidate.

The second perspective tends be harder to fulfill, often requiring symmetry between the two subsets \( \Theta_1 \) and \( \Theta_2 \). For instance, the given criterion is fulfilled with the appropriate procedure in case of comparing two (normally distributed) sample means, as well as in comparisons of two Poisson distributions, and in comparisons of two Binomial distributions with equal number of summands (cf. [44, Chapters 4 and 5]). Here, the two directions of inference are essentially mirrored copies of each other. However, the criterion is no longer fulfilled in the case of comparing unknown proportions in Binomial distributions with unequal number of summands.

### 2.5 Paper V

As in Paper IV, a decision theoretic setting is here considered, with a set of probability distributions \( \{P_\theta : \theta \in \Theta\} \) on a common probability space. Moreover, it is assumed that decision functions \( \delta_1, \ldots, \delta_n \) are given, to which non-negative loss functions \( L_1, \ldots, L_n \) are associated. Furthermore, it is assumed that each individual loss \( L_k \) is related to a non-negative random variable \( \hat{R}_k \) (referred to as a generalized p-value), in the sense that

\[
E_\theta (L_k(\delta_k, \theta)I\{\hat{R}_k \leq \alpha\}) \leq \alpha, \quad \text{for all} \quad \theta \in \Theta. \tag{2.15}
\]

An interpretation of condition (2.15) is that the expected loss connected with a decision function \( \delta_k \) may be *controlled* at any given level \( \alpha \), provided \( \delta_k \) is replaced by an indefinite decision (with zero loss) whenever \( \hat{R}_k \) exceeds \( \alpha \).

Now, with \( \hat{R}_1, \ldots, \hat{R}_n \) sorted as \( \hat{R}_{(1)} \leq \cdots \leq \hat{R}_{(n)} \), we consider the procedure which confirms the \( N \) given decisions corresponding to the observations \( \hat{R}_{(1)}, \ldots, \hat{R}_{(N)} \), with \( 0 \leq N \leq n \) determined by

\[
N = \max\{i : \hat{R}_{(i)} \leq i\alpha/n\}.
\]

The main result in Paper V concerns the situation of independent pairs \( (\delta_1, \hat{R}_1), \ldots, (\delta_n, \hat{R}_n) \), stating that the *expected average loss among confirmed decisions* is then controlled at level \( \alpha \) by the above procedure.

The given setting is a generalization of a typical context in multiple testing. Indeed, with p-values \( \hat{p}_1, \ldots, \hat{p}_n \) corresponding to the testing of null hypothe-
ses $\Theta_1, \ldots, \Theta_n$, consider “invariant” decision functions $\delta_1, \ldots, \delta_n$ corresponding to the attempts of rejecting the given hypotheses. Moreover, consider the associated loss functions

$$L_k = I\{\theta \in \Theta_k\}, \quad \text{for} \quad k = 1, \ldots, n,$$

corresponding to false rejections. Condition (2.15), with $\hat{R}_k$ replaced by $\hat{p}_k$, then corresponds to the assumption that $\hat{p}_k$ is a $p$-value for testing $\Theta_k$, i.e. to the condition

$$P_\theta(\hat{p}_k \leq \alpha) \leq \alpha, \quad \text{for} \quad \theta \in \Theta_k.$$

Also, the above procedure for confirming a subset of decisions is a direct generalization of Simes’ procedure [54] in multiple testing, also known as the Benjamini–Hochberg procedure [2]. In that case, controlling the expected average loss among confirmed decisions is the same as controlling FDR (false discovery rate, cf. [2]). More precisely, Benjamini and Hochberg showed that

$$FDR \leq \frac{\alpha m_0}{n}, \quad (2.16)$$

with independent $p$-values and $0 \leq m_0 \leq n$ referring to the number of true null hypotheses. Later, Benjamini and Yekutieli [4, Corollary 3] extended this description to

$$FDR \leq \frac{\alpha}{n} \left(m_0 + \frac{1}{2}(n - m_0)\right), \quad (2.17)$$

if Type III errors are taken into account when testing one-dimensional point hypotheses.

The ambition of Paper V is to capture the results (2.16), (2.17), and similar descriptions of directional/non-directional errors, within a unified and generalized formulation.

As a related example, suppose that analyzed data of a scientific experiment may be reasonably well captured by the model

$$\{N(\mu_1, 1) \times \cdots \times N(\mu_n, 1) : (\mu_1, \ldots, \mu_n) \in \mathbb{R}^n\}$$

of $n$ independent, normally distributed random variables $T_1, \ldots, T_n$ with unit variance. Also consider directional decision functions $\delta_k$ defined by

$$\delta_k = 0 \quad \text{whenever} \quad T_k \leq 0, \quad \delta_k = 1 \quad \text{whenever} \quad T_k > 0,$$

with associated loss functions

$$L_k = I\{\mu_k > 0\}I\{\delta_k = 0\} + I\{\mu_k \leq 0\}I\{\delta_k = 1\},$$

corresponding to reversed conclusions. Then,

$$\hat{R}_k = \Phi(-|T_k|),$$
with $\Phi$ denoting to the standard normal distribution function, refers to a generalized $p$-value with respect to $L_k$, in the sense of (2.15). Indeed,

$$E_{\theta}(L_k I\{\hat{R}_k \leq \alpha\}) = E_{\mu_k}(I\{\hat{R}_k \leq \alpha\} (I\{\mu_k \leq 0\} I\{T_k > 0\} + I\{\mu_k > 0\} I\{T_k \leq 0\}))$$

$$= E_{\mu_k}(I\{\mu_k \leq 0\} I\{\Phi(-T_k) \leq \alpha\} + I\{\mu_k > 0\} I\{\Phi(T_k) \leq \alpha\}) \leq \alpha.$$

Note that the corresponding three-decision procedures were considered in Paper IV, given a maximal reversal rate $\alpha$:

- Conclude $\mu_k > 0$ whenever $T_k > 0$ and $\hat{R}_k \leq \alpha$;
- Conclude $\mu_k \leq 0$ whenever $T_k \leq 0$ and $\hat{R}_k \leq \alpha$;
- No conclusion in case $\hat{R}_k > \alpha$.

By the main result in Paper V, the Benjamini–Hochberg procedure applied to $\hat{R}_1, \ldots, \hat{R}_n$ is risk-controlling with respect to $\delta_1, \ldots, \delta_n$, in the sense that the expected proportion of false conclusions among confirmed conclusions is bounded by $\alpha$.

Furthermore, we verify that corresponding bounds, increased by a factor

$$\sum_{i=1}^{n} \frac{1}{i} \approx 0.58 + \log n,$$

are valid without any assumption on dependence, thereby extending the results of Benjamini and Yekutieli [3] and Hommel [36].

Finally, we argue that similar, generalized formulations are not possible regarding several recently introduced “adaptive” FDR-controlling procedures, such as the suggestion by Gavrilov, Benjamini and Sarkar [25], and the suggestion by Storey [55].
3. Summary in Swedish


3.1 Självnormaliserade summor

De självnormaliserade summor av oberoende, likafördelade slumpvariabler som behandlas i artiklarna I-III faller inom två grupper, dels så kallade $t$-statistikor,

$$T_n = \sqrt{n} \left( \frac{\sum_{i=1}^{n} X_i}{\hat{\sigma}_n} \right) / \sigma_n,$$

med

$$\hat{\sigma}_n^2 = \frac{1}{n-1} \left( \sum_{i=1}^{n} X_i^2 - \frac{1}{n} \left( \sum_{i=1}^{n} X_i \right)^2 \right),$$

dels den relaterade, men något enklare typen av självnormaliserad summa:

$$Z_n = \frac{\sum_{i=1}^{n} X_i}{\left( \sum_{i=1}^{n} X_i^2 \right)^{1/2}}.$$


Artikel I berör det som brukar kallas för $t$-statistikors tunga svansar, vilket syftar på de förhållandevis stora sannolikheterna för extrema observationer. Det kan här tilläggas att “självnormaliserade uttalanden” om ett uppmätt värde naturligt medför en viss ökad osäkerhet jämfört med om mätsäkerheten varit given på förhand. Detta återspeglar sig i $t$-statistikors så kallade svanssan-
nolikheter, vilket man brukar ta hänsyn till vid motsvarande tillämpningar av statistisk metodik. Vanligtvis görs detta hänsynstagande i förhållande till ett idealiserat antagande om att observationerna är normalfördelade. Eftersom detta idealiserande är mer eller mindre passande för den verklighet man tillämpar det på, så är det av intresse att studera $t$-statistikors svansar från ett mer allmänt perspektiv.

Resultaten i artikel I bekräftar i huvudsak att de tunga svansar som man kan associera till $t$-statistikor baserade på normalfördelade observationer återfinns för de flesta andra möjliga fördelingsantaganden. Exempelvis visas att de så kallade absolut-momenten $E|T_n|^r$ i allmänhet bara existerar för grader lägre än motsvarande antal frihetsgrader, det vill säga i betydelsen

$$E|T_n|^r < \infty \quad \text{om och endast om} \quad r < n - 1.$$ 

I artikel II studeras självnormaliserade summor av typen $Z_n$. Närmare bestämt intresserar vi oss för motsvarande kvadratiska moment, det vill säga medelvärdet av ett stort antal kvadrerade värden. Huvudresultatet i artikel II kan kortfattat uttryckas på följande sätt:

$$EZ_n^2 \geq 1, \quad \text{och} \quad EZ_n^2 = 1 \quad \text{om och endast om} \quad X \overset{d}{=} -X.$$ 

Annorlunda uttryckt, det kvadratiska momentet är alltid minst lika med ett, och likhet råder precis i de fall då observationerna följer en sannolikhetsfördelning som är symmetrisk kring origo. Man kan här tillägga att ett stort kvadratiskt moment i viss utsträckning kan associeras till tungsansadhet hos motsvarande slumpmässiga storhet.

Resultaten i artikel III berör också självnormaliserade summor $Z_n$, men indirekt genom relationen till lineära kombinationer av slumpvariblar av Rademacher-typ:

$$S = \sum_{k=1}^{n} a_k \varepsilon_k, \quad \text{med} \quad (a_1, \ldots, a_n) \in \mathbb{R}^n,$$

där $\varepsilon_1, \ldots, \varepsilon_n$ hänvisar till $n$ stycken oberoende slumpvariabler, som var och en med lika stor sannolikhet antar något av de två värdena $\{-1, 1\}$. Den koppling mellan summorna $Z_n$ och $S$ som vi avser går ut på att en självnormaliserad summa $Z_n$, relativt observationer som är symmetriska fördelade kring origo, kan uttryckas som en så kallad blandfördelning relativt summor av typen $S$, med den ytterligare restriktionen

$$\sum_{k=1}^{n} a_k^2 = 1. \quad (3.1)$$

En konsekvens av denna koppling är att generella resultat som rör den matematiskt enklare formen $S$ ofta kan omvandlas till generella resultat för formen
Huvudresultatet i artikel III kan kortfattat uttryckas på följande sätt:

\[ P(S \geq x) \leq N(x)(1 + Cx^{-2}). \]  

(3.2)

I detta fall avser \( P(S \geq x) \) en svanssannolikhet för en godtyckligt vald summa \( S \) under antagandet (3.1), \( N(x) \) syftar på motsvarande sannolikhet relativt normalfördelnningen, och \( C \) syftar på en lämpligt vald, det vill säga tillräckligt stor, positiv konstant.

Olikheten (3.2) kan tolkas i termer av att summor av typen \( S \) är förhållandevis lättsvansade i förhållande till normalfördelnningen. Observera att om \( x \) är stort, så är \( x^{-2} \) litet, vilket gör högerledet i (3.2) förhållandevis nära \( N(x) \). I artikel III visar vi att konstanten \( C \) måste överstiga värdet 3 för att olikheten ska gälla för alla \( S \) och \( x \). Vi undersöker även olikhetens relevans för självnormaliserade summor \( Z_n \) av normalfördelade observationer.

3.2 Kontrollerade slutsatser om förändringars riktning

I artikel IV och V intresserar vi oss för formella slutsatser om riktningen hos en kvantitativ förändring, detta med en allmän formulering inom statistisk beslutsteori. Vi utgår från beslutsfunktioner som väljer mellan de två möjliga riktningarna, samt möjligheten att avböja ett bestämt svar i frågan. Endera av de två riktningarna anses vara en korrekt beskrivning av den verkliga förändringen. Slutligen anlägger vi perspektivet att slutsatserna måste vara kontrollerade, i betydelsen att risken för att avge ett felaktigt svar begränsas av en på förhand given högsta-nivå.

I samband med denna typ av slutsatser är det ofta naturligt att utgå från en rapporterad förändring, till vilken man kan associera en så kallad förlustfunktion \( L \). Funktionen \( L \) antar värdet 1 då den rapporterade förändringen är felaktig och värdet 0 då den rapporterade förändringen är korrekt. En kontrollerad beslutsfunktion väljer att avböja bestämt svar ifall den rapporterade förändringen anses alltför osäker. Detta kan uttryckas i termer av en indikatorfunktion \( I \), som antar värdet 1 om den rapporterade förändringen anses tillräckligt säker och värdet 0 annars. Att sannolikheten för att bekräfta felaktigt svar begränsas av en given nivå \( \alpha \) kan då uttryckas på följande sätt:

\[ E(L \cdot I) \leq \alpha. \]

Det är för det mesta fallet att denna kontroll är möjlig för alla val av \( \alpha \), det vill säga att graden av säkerhet kan väljas godtyckligt av den som tänker sig att tillämpa beslutsfunktionen. I så fall är det relevant att vid en given rapportering låta \( \hat{R} \) beteckna den lägsta nivå \( \alpha \) med vilken förändringen anses
säker. På detta vis erhålls följande generalisering av föregående uttryck:

\[ E(L \cdot I\{\hat{R} \leq \alpha\}) \leq \alpha, \quad \text{för alla } \alpha > 0. \]

I artikel V hänvisar vi till \( \hat{R} \) som ett exempel på ett “generaliserat \( p \)-värde”, detta i relation till förlustfunktionen \( L \), som i exemplet hänvisar till felaktig rapportering.

För ett givet val av säkerhetsgrad \( \alpha \) är det relevant att kunna motivera valet av beslutsfunktion. Ofta vill man att den givna informationen tas till vara på ett lämpligt sätt, allra helst på ett optimalt sätt. Det finns olika möjligheter vad gäller att formulera optimalitet i detta sammanhang. Exempelvis kan den valda beslutsfunktionen anses besitta optimalitetsegenskaper—jämfört med alla övriga tänkbara kontrollerade beslut—om den

- Maximerar sannolikheter för att ange korrekt riktning;
- Minimerar sannolikheter för att avböja bestämt svar;
- Minimerar sannolikheter för att ange felaktig riktning.

I artikel IV betraktar vi dessa tre karaktäriseringar var för sig och formulerar så kallade tillräckliga villkor relativt motsvarande egenskaper. Slutligen tillämpar vi dessa kriterier på en generell familj av statistiska modeller, så kallade exponentialfamiljer.


I Artikel V undersöker vi även i vilken mån det är möjligt att på snarlikt vis utvidga även andra kända förfaranden vid multipel hypotesprövning. Vi finner ett positivt svar för det så kallade Bonferroni–förfarandet, men ett negativt svar för många andra förfaranden.
4. Acknowledgements

I remember my PhD supervisor Allan Gut asking in spring 2007: *Can you extend this result to $t$-statistic random variables?* I had then recently finished my master thesis, reviewing the field of *Almost Sure Central Limit Theorems*. As a consequence, I got interested in self-normalized sums, and later in the associated statistical issues. Thank you Allan for all encouragement and care that you have been directing to me and the work that I have been doing!

I have also benefited from the complementary PhD supervision of Silvelyn Zwanzig, especially during the autumn of 2011. Being confronted with your sceptical attitude towards my work on statistical issues has been both fun and healthy. Notably, I am happy for the developments of Paper IV resulting from discussions with you.

Svante Janson had a great impact on my work leading to Paper II. I remember Allan mentioning a conjecture that I had been working on for some time to his colleague Svante, Svante pondering a few seconds, then producing a two-line proof, which Allan later showed to me.

I have appreciated being part of FMB, the Swedish graduate school in mathematics and computing. Many thanks to Inga-Lena Assarsson and Johan Tysk for your wise and kind direction!

Sharing my office with Georgios Dimitroglou Rizell has been a great pleasure during my five years as a PhD student. Also, I will miss the lunches with fellow PhD students Albin, Anders, Bing, Cecilia, Georgios, Hannah, Henrik, Isac, Jimmy, Katja, Maik, Måns, Oswald, Ramon, Saeid, and others. I have appreciated the social activities of ”Friday fika”, with people from mathematical statistics in the morning, and general math PhD:s in the afternoon. For instance, I will remember the waffles of Isac Hedén, with melted chocolate and lemon...

I guess that writing this PhD thesis could be viewed as an end of an educational journey, starting in the autumn of 1989 in Åtvidaberg. My parents Anita and Leif have been a great support during all these twentytwo and a half years.

Finally, I would like to thank Marianne for sharing your positive attitude to life. Also, I have appreciated your witty jokes about ”le grand mathematicien” and the ”pottering” with mathematical symbols and formulae...
References


