Ekaterina Poliakova

Confidence distributions for the Behrens-Fisher problem

Master's thesis in Mathematical Sciences
Supervisor: Gunnar Taraldsen
July 2021

Master's thesis

NTNU
Norwegian University of Science and Technology
Faculty of Information Technology and Electrical Engineering
Department of Mathematical Sciences

NTNU
Norwegian University of Science and Technology

Ekaterina Poliakova

Confidence distributions for the Behrens-Fisher problem

Master's thesis in Mathematical Sciences
Supervisor: Gunnar Taraldsen
July 2021

Norwegian University of Science and Technology
Faculty of Information Technology and Electrical Engineering
Department of Mathematical Sciences

NTNU
Norwegian University of
Science and Technology

Abstract. This Master's thesis deals with confidence distributions for the difference between the means of two samples from normal distributions with unknown variances. Distribution estimators and confidence distributions as their special type are introduced. Linear combinations of confidence variables are considered. A tentative proof that the linear combination of means of a symmetric a symmetric unimodal confidence variables are confidence variables for corresponding linear combinations of parameters is presented. The latter statement is illustrated numerically with examples. Various tests and related confidence densities from these tests for the Behrens-Fisher problem are studied via numerical simulations.

Sammendrag. Denne masteroppgaven dreier seg om konfidensfordelinger for differansen mellom middelverdiene av to normalfordelte utvalg, når begge varianser er ukjente. Fordelingsestimatorer og konfidensfordelinger som deres spesiell type er introdusert. Lineære kombinasjoner av konfidensvariabler er betraktet. Det presenteres et tentativt bevis for at en lineær kombinasjon av en symmetrisk og en symmetrisk unimodal konfidensvariabler er også en konfidensvariabel for tilsvarende lineære kombinasjoner av parametere. Denne påstanden er illustrert med eksempler. Forskjellige tester for Behrens-Fisher problem og relaterte konfidensfordelinger er studert ved numeriske simuleringer.

Preface

This Master's thesis is written at the Department of Mathematical Sciences at the Norwegian University of Science and Technology. The thesis completes the Master's degree programme in Mathematical Sciences.

The author would like to thank professor Gunnar Taraldsen for very constructive and motivating supervision.

Ekaterina Poliakova
Trondheim, 2021

Contents

Abstract i

Sammendrag i

Preface iii

Contents v

List of Tables ix

List of Figures xi

Notation xiii

Chapter 1. Introduction 1

Chapter 2. General theory on confidence distributions 3
1. Some fundamental concepts of probability theory 3
2. Distribution estimators as random measures 4
3. Confidence distributions and variables 10

Chapter 3. Linear combination of confidence variables 17
1. Multiplication of a confidence variable with a scalar 17
2. Linear combination of a symmetric confidence variable and a symmetric unimodal continuous confidence variable 18
3. An example of non-exact CD 19
4. Counterexamples for a sum of confidence variable as CD of the sum 25

Chapter 4. Tests for the Behrens-Fisher problem 29
1. Simple properties of the Behrens-Fisher statistic 29
2. Conservative tests and the nominal level 30
3. Methods of numerical study of the tests 30
3.1. Monte-Carlo simulations 30
3.2. Numerical integration 30
4. Behrens-Fisher test 31

5. Welch-Satterthwaite test: ISO GUM version 31
6. Welch-Aspin test 33
7. Paired t-test 33
8. Likelihood ratio test 34

Chapter 5. Constructing CDs for the Behrens-Fisher problem 35
1. As distribution of the difference of confidence variables 35
2. By inverting conservative tests for the Behrens-Fisher problem 35
2.1. From Behrens-Fisher test 36
2.2. From Welch-Satterthwaite test 36
2.3. From the paired t-test 36
2.4. From the likelihood ratio test 36
3. By asymptotic methods 37
3.1. High-order approximations for the deviance 37
3.2. High-order approximations for the quantiles of the CD 37
4. Numerical adjustment 37

Chapter 6. Numerical results 39
1. Overall comparison of the confidence densities 39
2. Behrens-Fisher test based CDs 41
3. Other CDs corresponding sums of confidence distributed variables 42
4. Welch-Satterthwaite test based CDs 45
5. Welch-Aspin test based CDs 49
6. Asymptotic tests 50
7. Preliminary computation of the loss and risk of the CDs 51

Chapter 7. Discussion and conclusions 53

Appendix A. Preliminary proofs for the lemmas in the Chapter 3 55
1. Proof of the Lemma 3.2 55
2. Proof of the Lemma 3.3 55
3. Proof of the Lemma 3.4 57
4. Proof of the Lemma 3.5 57
5. Proof of the Lemma 3.6 58

Appendix B. Level of the Behrens-Fisher test with $n_1 = n_2 = 2$ and $n_1 = n_2 = 3$ 59
1. The test with a constant critical value 59
2. The quantiles of $T_1 cos\theta + T_2 sin\theta$ 61
3. Level of the Behrens-Fisher test in case $n_1 = n_2 = 2$ and $n_1 = n_2 = 3$ 64

Appendix C. R-codes 65
1. Computing the confidence density 65

2. Converting between nominal and actual level of the tests 69

Bibliography 73

List of Tables

1 Approximate risk for $n_1 = n_2 = 5$ with different CDs, by 500 simulations 52

2 Approximate risk for $n_1 = 5, n_2 = 3$ with different CDs, by 100 simulations 52

List of Figures

2.1 The relation between definitions. From the NTNU course page TMA4285 (2019) 3

2.2 The construction of a measure from F by Eq. (1) 6

3.1 Illustration to the proof of Lemma 3.7, case when $\boldsymbol{\sigma_1} < \boldsymbol{\sigma_2}$, $\sigma_2/k - 1 > 0, \sigma_2/k + \sigma_1 > \sigma_2 - \sigma_1$, $\sigma_2 k - \sigma_1 < \sigma_2 + \sigma_1$ 22

3.2 Illustration to the proof of Lemma 3.7, case when $\boldsymbol{\sigma_2/k - \sigma_1 > 0}$, $\sigma_2/k + \sigma_1 > \sigma_2 - \sigma_1$, $\sigma_2 k - \sigma_1 < \sigma_2 + \sigma_1$ 23

3.3 Illustration to the proof of Lemma 3.7, case when $\boldsymbol{\sigma_2/k - \sigma_1 > 0}$, $\sigma_2 - \sigma_1 < t_\alpha < \sigma_2 + \sigma_1$ both when $S_2 = b_1$ and $S_2 = b_2$ 24

3.4 Illustration to the proof of Lemma 3.7, case when $\boldsymbol{\sigma_1} > \boldsymbol{\sigma_2}$ and $-\sigma_2 + \sigma_1 < (t_\alpha | S_2 = b_1) < -\frac{\sigma_2}{k} + \sigma_1$ 25

3.5 A bimodal symmetric CD and the sum of two bimodal symmetric CDs. The black graphs corresponds the densities of data, the red graphs correspond confidence densities for μ, the green text shows the true parameter values 27

6.1 The confidence densities for $Y_1 = (-0.9142985,\ 0.9320448,\ 1.0945988,\ -1.5417058,\ 0.2018343)$ and $Y_2 = (-0.6688093,\ 0.1468806,\ -1.2870124,\ 1.2566792,\ -1.0072095)$ 39

6.2 The confidence densities for $Y_1 = (0.28626327,\ -0.09423993,\ 0.20402356,\ -1.36958796,\ 0.08829856)$ and $Y_2 = (-1.11996987,\ -2.07113205,\ -0.26252523,\ -0.07963677,\ -1.18116007$ 40

6.3 The confidence densities for $Y_1 = (1.9686916,\ -0.02861356,\ 1.05992191)$ and $Y_2 = (-1.39063,\ -1.132441,\ -2.775170,\ -2.471090,\ -2.230663)$ 41

6.4 Probability that μ belongs to its p-confidence set A_p when constructing the Behrens-Fisher test based CD, for sample sizes $n_1 = 5$ and $n_2 = 3$, as function of p 42

6.5 Probability that μ belongs to its p-confidence set A_p when constructing the Conjecture 3.1 based CD, for sample sizes $n_1 = 5$ and $n_2 = 3$ for $\sigma_1 = \sigma_2 = 1$, as function of p, from 1000 simulations 43

6.6 Probability that μ belongs to its p-confidence set A_p when constructing the Conjecture 3.1 based CD, for sample sizes $n_1 = 5$ and $n_2 = 3$, for $p = 0.95$, as function of σ_1/σ_2, from 100000 simulations 44

6.7 Probability that μ belongs to its p-confidence set A_p when constructing the Conjecture 3.1 based CD, for sample sizes $n_1 = 5$ and $n_2 = 3$, for $p = 0.95$, as function of p,from 10000 simulations 45

6.8 The dependence of the probabilities to reject H_0 from the nuisance parameter, for sample sizes $n_1 = 5$ and several sizes n_2 46

6.9 The dependence of the level of the test on n_2, at nominal level of 0.2 and $n_1 = 21$ 47

6.10The dependence of the level of the test on n_2, at nominal level of 0.05 and $n_1 = 5$ 47

6.11Typical dependencies between n_2 and the $\frac{1}{\alpha_\infty - \alpha}$ 47

6.12Probabilities to reject H_0 by Welch-Satterthwaite test at different levels, for sample sizes $n_1 = 5$ and $n_2 = 3$ 48

6.13The dependence between nominal and actual level of the Welch-Satterthwaite test for sample sizes: $n_1 = 5, n_2 = 9$ and $n_1 = 3, n_2 = 5$. Two scales 49

6.14The dependence between nominal and actual level the Welch-Aspin test for sample sizes: $n_1 = 5, n_2 = 9$ and $n_1 = 3, n_2 = 5$. Two scales 49

6.15The probability to reject $H_0 : \mu_1 = \mu_2$ as function of $2log\sigma_1/\sigma_2$: $n_1 = 5$, $n_2 = 9$, with nominal test level 0.05 50

6.16The probability to reject $H_0 :'' \mu_1 = \mu_2''$ as function of $2log\sigma_1/\sigma_2$: $n_1 = 15$, $n_2 = 19$, with nominal test level 0.05 50

6.17The probability to reject $H_0 :'' \mu_1 = \mu_2''$ as function of $2log\sigma_1/\sigma_2$: $n_1 = 40$, $n_2 = 60$, with nominal test level 0.05 51

1.1 The schematic example of the sets $B_i \cup \big((t - \alpha | s_2 = \sigma_2 m), +\infty\big)$ 56

2.1 The graphs of a squared Behrens-Fisher statistic as functions τ, belonging to the *'4-curves'*-family. 61

Notation

CD – confidence distribution,
$\Phi(x)$ – distribution function for standard normal distribution,
$\phi(x)$ – probability density for standard normal distribution,
$\Psi(x, n)$ – distribution function for student t-distribution with n degrees of freedom,
$\psi(x, n)$ – probability density for for student t-distribution with n degrees of freedom,
$f_{\chi^2_f}$ – probability density for the chi-square distribution with f degrees of freedom,
$N(\mu, \sigma^2)$ – the normal distribution with mean μ and variance σ^2,
χ^2_f – the chi-square distribution with f degrees of freedom.

For the random sample $Y = \{Y_1, ..Y_{n_Y}\}$ which is not directly involved in the Behrens-Fisher problem:
$\overline{Y} = \frac{\sum_{i=1}^{n_Y} Y_i}{n_Y}$ – the sample average,
$S_Y^2 = \frac{\sum_{i=1}^{n_Y}(Y_i - \overline{Y})^2}{n_Y - 1}$ -the sample variance.

For the random samples $Y_i = (Y_{i,1}..Y_{i,n_i})$, involved in the Behrens-Fisher problem:
$X_i = \overline{Y_i} = \frac{\sum_{j=1}^{n_i} Y_{i,j}}{n_i}$,
$S_i^2 = S_{Y_i}^2 = \frac{\sum_{j=1}^{n_i}(Y_{i,j} - \overline{Y_i})^2}{n_i - 1}$.

In our terminology:
An **open interval** (a, b), where $a, b \in \{\mathbb{R} \cup \{-\infty\} \cup \{\infty\}\}$, is a subset of $\mathbb{R}$: $\{x : a < x < b\}$.
A **half-open interval** $[a, b)$ or $(a, b]$, where $a, b \in \{\mathbb{R} \cup \{-\infty\} \cup \{\infty\}\}$, is a subset of $\mathbb{R}$, such that for $[a, b)$, $\{x : a \leq x < b\}$, and for $(a, b]$, $\{x : a < x \leq b\}$.
A **closed interval** $[a, b]$, where $a, b \in \mathbb{R}$, is a subset of $\mathbb{R}$: $\{x : a \leq x \leq b\}$.

We denote as $\mathcal{B}$ the minimal sigma-algebra generated by open sets.

We denote $w_1V_1 : w_2V_2$ a mixture distribution, such that its distribution function is $w_1F_{V1} + w_2F_{V2}$, where F_{V1} and F_{V2} are distribution functions for V_1 and V_2 respectively.

CHAPTER 1

Introduction

In its original form (Kim & Cohen, 1998), the Behrens–Fisher problem deals with testing the hypothesis on the equality of the means in two normal distributions. Many approaches to the problem have been designed, including the Behrens–Fisher test, the Welch–Satterthwaite test, the likelihood ratio test, the Welch–Aspin test, the Scheffé test and the Fraser test. A more general problem, however, involves considering all the uncertainty in the estimate of the difference between the means. International measurement standards require that this uncertainty be stated (JCGM et al., 2008). This work focuses on a particularly important type of such estimators: confidence distributions, which we abbreviate 'CDs' (Taraldsen, 2021).

The Behrens-Fisher problem is a special case of a more general problem: whether the linear combination of randomised parameter estimators provides a CD for the linear combination of the parameters. Numerous numerical simulations (e.g. Duong & Shorrock, 1996; Wang, 1971) indicate that this conjecture is true for R.A. Fisher's solution to this problem Fisher (1930). However, even for this case the analytical proof seems to be missing from the literature. In this thesis we present a preliminary proof for the linear combination of means of a symmetric CD and a symmetric unimodal CD. If this proof is correct, it will be an important new result.

The Behrens-Fisher problem may be extended in many ways. A usual generalisation deals with m normally distributed samples (Casella & Berger, 2002, p.409). An even more general problem is constructing a confidence distribution for another function of parameters than a linear combination of the parameters. The Joint Committee for Guides in Metrology's JCGM et al., 2008 gives an approximate solution for the latter problem. This work adds further progress to that solution.

CHAPTER 2

General theory on confidence distributions

1. Some fundamental concepts of probability theory

One of the most fundamental concepts in statistics which gives ground to all the following discussion is a sample space. A *sample space* consists of a set and a sigma-algebra of subsets of this set. A *probability space* is a sample space equipped with a probability measure on the sigma-algebra. A sample space may be denoted as $(\mathfrak{X}, \mathcal{F})$, and a probability space as $(\mathfrak{X}, \mathcal{F}, \mathfrak{P})$, where $\mathfrak{X}$ is the set, $\mathcal{F}$ is the sigma-algebra and $\mathfrak{P}$ is the probability measure. In a shorter notation, the $\mathfrak{X}$ is written explicitly, while the presence of $\mathcal{F}$ and, for the probability space, $\mathfrak{P}$ is implied. The sample spaces used in statistics are illustrated in the 'Commutative diagram of statistics' (Taraldsen, 2019) in Fig. 2.1 .

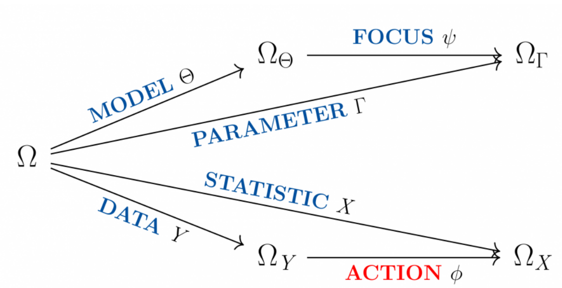

FIGURE 2.1 . The relation between definitions. From the NTNU course page TMA4285 (2019)

In the *underlying sample space*, $(\Omega,\ \mathcal{E})$:

- elements of Ω are called *outcomes* and are denoted as ω;
- elements of $\mathcal{E}$ are called *events*.

The set Ω contains all the theoretically possible underlying outcomes without any models or observations. Ω is never observed directly. The model is introduced by defining the model space Ω_Θ. If the model is parametric, the model space can be indexed with parameters. In the Behrens-Fisher problem, $\theta = (\mu_1,\ \sigma_1,\ \mu_2,\ \sigma_2)$.

There is also a measurable function *focus* Ψ: $\Omega_\Theta \to \Omega_\Gamma$, where Ω_Γ is a space of the parameter of interest. In the Behrens-Fisher problem, the focus parameter is $\gamma = \mu_1 - \mu_2$ and $\Psi(\theta) = \mu_1 - \mu_2$.

When an experiment is conducted, the data $y \in \Omega_Y$ is obtained. The elements θ of Ω_Θ determine the probability measure on Ω_Y. A function of the data is called *statistic*. In the Behrens-Fisher problem, Ω_Y is usually the set of all the possible values of the sufficient statistic

$$(\overline{y_1},\ \overline{y_2},\ s^2_{Y_1},\ s^2_{Y_2}) \in \Omega_Y = \mathbb{R}^2 \times \mathbb{R}^2_+.$$

However, the set of all the possible outcomes

$$(y_{1,1},\ y_{1,2}, \dots y_{1,n_1},\ y_{2,1},\ y_{2,2}, \dots y_{2,n_2}) \in \Omega_Y = \mathbb{R}^{n_1+n_2}$$

is also sometimes used, e.g. when treating the Behrens-Fisher problem with the paired t-test.

2. Distribution estimators as random measures

The sample space concept makes it possible to deal with random variables and random measures.

DEFINITION 2.1. A function $X\ \Omega_X \to \mathbb{R}$ is a **random variable** if for all $a \in \mathbb{R}$ $X^{-1}(-\infty, a] \in \mathcal{E}$.

If the sample space $(\Omega, \mathcal{E})$ is equipped with probability measure P, the random variable defines a probability measure P_X also on $(\mathbb{R}, \mathcal{B})$.

DEFINITION 2.2. Given an underlying abstract probability space $(\Omega, \mathcal{E}, P)$ and a random variable X, a **distribution function** $F : \mathbb{R} \mapsto [0, 1]$ is defined by $F(x) = P\Big(\big(X^{-1}(-\infty, x]\big)\Big)$.

PROPOSITION 2.3. *A distribution function F has following properties:*

(1) $F(x)$ is non-decreasing with respect to x;
(2) $F(x)$ is right-continuous with respect to x;
(3) $\lim_{x\to\infty} F(x) = 1$;
(4) $\lim_{x\to-\infty} F(x) = 0$.

PROOF. (1) Let $x_1, x_2 \in \mathbb{R}, x_2 > x_1$. Then

$$F(x_2) = P_X\big((-\infty, x_2]\big) = P_X\big((-\infty, x_1] \cup (x_1, x_2]\big) =$$

by countable additivity of measure for disjoint sets

$$= P_X\big((-\infty, x_1]\big) + P_X\big((x_1, x_2]\big) \geq$$

by non-negativity of measure

$$\geq P_X\big((-\infty, x_1]\big) = F(x_1).$$

(2) Let (x_n) be a decreasing sequence of real numbers converging to x.

$$F(x) = P_X\big(\cap_{i=1}^{\infty}(-\infty, x_n]\big)\big) = F(x_1) - (\sum_{i=1}^{\infty}\big(F(x_i) - F(x_{i+1})\big) =$$

$$= \lim_{x_n \to x+} F(x_n)$$

(the second equality follows from decomposing $(\infty, x_1]$ into the disjoint sets $(\cap_{i=1}^{\infty}(-\infty, x_n])$, $(x_2, x_1]$, ..., $(x_{i+1}, x_i]$, ..., and countable additivity of measure).

(3) By Definition 2.2, P_X is a probability measure, hence $P_X(\mathbb{R}) = 1$. Let (x_n) be an increasing sequence of real numbers converging to ∞. Then $P_X(\mathbb{R}) = P_X((-\infty, x_1] \cup (\cup_{i=1}^{\infty}(x_i, x_{i+1}])) = \lim_{x\to\infty} F(x)$.

(4) Let (x_n) be a decreasing sequence of real numbers converging to $-\infty$. Then $0 = P_X(\emptyset) = P_X(\cap_{i=1}^{\infty}(-\infty, x_i]) = \lim_{x\to-\infty} F(x)$.

□

PROPOSITION 2.4. *Given a a function $F(x)$ which has the properties 1-4 listed in 2.3, it defines a unique probability measure ζ on $\mathcal{B}$ such that $F(x) = \zeta\big((-\infty, x]\big)$.*

PROOF. Consider a function $F(x)$. We notice that for any $x \in \mathbb{R}$ $\lim_{x_n \to x-} F(\omega, x_n)$ exists, because F is non-decreasing, and is on $(-\infty, x]$ bounded by $F(x)$.
We define a map G from $\mathbb{R}$ to the set of subsets of $[0, 1]$ (Fig. 2.2) as following:

$$G(x) = \begin{cases} \{F(x)\}, \text{ F is continuous in x,} \\ (\lim_{x_n\to x-} F(x), F(x)], \text{ F is not continuous in x,} \end{cases} \tag{1}$$

and a set map

$$G^*(A) = \cup_{x\in A} G(x) \tag{2}$$

By construction, G^* maps any open interval to either interval (closed, half-open or open) or to a single point in $[0, 1]$, which sets belong to $\mathcal{B}$. It also maps the intersection of any sets to the intersection of their images. Hence G^* maps a $\pi-$system of intervals and points to a $\pi-$system of intervals and points. We assign to any element A of this system the Lebesque measure λ of its image under G^*:

$$\xi(A) := \lambda(G^*(A))$$

Therefore the measure ζ on $\pi-$system of intervals of $\mathbb{R}$ is well-defined.

The $\pi-$system of intervals of $\mathbb{R}$ generates $\mathcal{B}$. By Lemma 1.42 in (Klenke, 2008), the probability measure ζ is uniquely determined on the measurable space $(\mathbb{R}, \mathcal{B})$ because it is uniquely determined on a $\pi-$system generating $\mathcal{B}$.

□

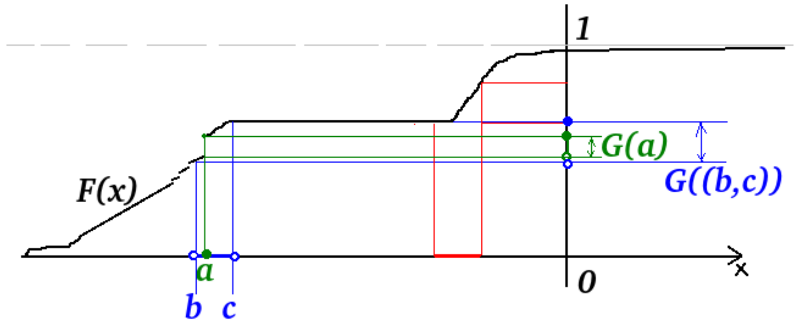

FIGURE 2.2 . The construction of a measure from F by Eq. (1)

We will slightly restrict the general definition of random measure given by (Kallenberg, 2017, p. 1) and define the random measure as following.

DEFINITION 2.5. Given a probability space $(\Omega, \mathcal{E}, P)$ and another measurable space $(\Omega_2, \mathcal{B})$, where $\Omega_2 = \mathbb{R}$ and $\mathcal{B}$ is the Borel sigma-algebra on $\mathbb{R}$, **random measure** ξ is a function of $\omega \in \Omega$ and of a set $B \subset \Omega_2 : B \in \mathcal{B}$, such that

(1) $\xi(\omega, B)$ is a measure with respect to B for all fixed y,
(2) $\xi(\omega, B)$ is a random variable $\forall B \in \mathcal{B}$.

DEFINITION 2.6. The random measure ξ is a **random probability measure** if $\xi(y, \Omega_2) = 1 \, \forall \, y \in \Omega_Y$.

EXAMPLE 2.7. Denote λ for the Lebesque measure and let $y \in \mathbb{R}$, B be a set from Borel sigma-algebra $\mathcal{B}$ on $\mathbb{R}$, and

$$\xi(y, B) = \lambda\big((y - 0.5, y + 0.5] \cap B\big).$$

Then $\xi(y, B)$ is a random probability measure.

Indeed,

- Assume y is fixed.

$$\xi(y, B) = \lambda\big(B \cap (y - 0.5, y + 0.5]\big)$$

We see that $(y - 0.5, y + 0.5] \in \mathcal{B}$, also $\big(B \cap (y - 0.5, y + 0.5]\big) \in \mathcal{B}$ as an intersection of two elements of $\mathcal{B}$. Hence $\xi(y, B)$ is defined for all $B \in S$. Furthermore, it is 0 for $B = \emptyset$, $\xi(y, \mathbb{R}) = y + 0.5 - y + 0.5 = 1 \forall y$. Considering a countable union of disjoint sets $B_1, ..., B_n, ..$, we observe that $\big(\cup_{n=1}^{\infty} B_n\big) \cap (y - 0.5, y + 0.5] = \cup_{n=1}^{\infty}\big(B_n \cap (y - 0.5, y + 0.5]\big)$, which is also a countable union of disjoint sets $\big(B_n \cap (y - 0.5, y + 0.5]\big)$. Thus by countable additivity of the Lebesque integral,

$$\xi(y, \cup_{n=1}^{\infty} B_n) = \lambda(\cup_{n=1}^{\infty} \big(B_n \cap (y-0.5, y+0.5]\big)) = \sum_{n=1}^{\infty} \lambda\big(B_n \cap (y-0.5, y+0.5]\big),$$

and $\xi(y, B)$ is countably additive with respect to B. So $\xi(y, B)$ is a probability measure in B for all fixed y.

- Assume B is fixed. Then

$$\xi(y, B) = \lambda\big((-\infty, y+0.5] \cap B\big) - \lambda\big((-\infty, y-0.5]) \cap B\big)$$

Consider a function $f(x) = \lambda\big((-\infty, x] \cap B\big)$. This function is non-decreasing, because whenever $x_1 < x_2$, then $f(x_1) < f(x_2)$ by finite additivity of Lebesgue measure. Moreover, f is continuous, because whenever

$$|x - x_0| < \delta,$$

then

$$|f(x) - f(x_0)| = |\lambda\big((-\infty, x] \cap B\big) - \lambda\big((-\infty, x_0] \cap B\big)| =$$

$$\lambda\Big(\big(min(x, x_0), max(x, x_0)\big) \cap B\Big) < \delta.$$

Hence a preimage $f^{-1}(-\infty, a)$ of any ray is also a ray $(-\infty, b)$, and for same argument as for the continuous F in the proof of Proposition 2.4, f is $\mathcal{B}-$measurable. Therefore, $\xi(y, B) = f(y+0.5) - f(y-0.5)$ as a difference of two $\mathcal{B}-$measurable functions is also $\mathcal{B}-$measurable.

PROPOSITION 2.8. *Given a function $F(x)$ with all the properties listed in Proposition 2.3 and such that it is a random variable in any point, it defines a random probability measure ζ on the Borel sigma-algebra $\mathcal{B}$.*

PROOF. The requirement 1 of the Definition 2.5 that $\zeta(\omega, B)$ is a measure with respect to B for all fixed y, holds by Proposition 2.4, which states that ζ with properties 1-4 listed in Proposition 2.3 is indeed a unique probability measure for every fixed y. The requirement 2 of the Definition 2.5 that $\zeta(y, B)$ is $\mathcal{E}$-measurable, holds by the Definition 2.1. Indeed, the fact that F is a random variable in any point means that the preimage of $B = (-\infty, x]$ is an event. Therefore $\zeta(y, B)$ is $\mathcal{E}$-measurable for any $B = (-\infty, x]$. We also see that for arbitrary $B \in \mathcal{B}$, any $\zeta^{-1}(y, B)$ is generated by preimages of rays $(-\infty, x]$ and hence belongs to $\mathcal{E}$. Therefore we conclude that $\zeta(y, B)$ is a random variable also for arbitrary $B \in \mathcal{B}$ and hence ζ is a random measure. $\square$

DEFINITION 2.9. A **distribution estimator** for γ is a random measure, where $\Omega_2 = \Omega_\Gamma$ in notation of Figure 2.1 and the Definition 2.5.

Remark 2.10. A point estimator $\hat{\gamma}$ is a special case of a distribution estimator. Than for all data $y \in \Omega_Y$, the random measure may be defined as

$$\xi(y, B) = 1_{\hat{\gamma}(y)}(B).$$

Example 2.11. Let $X \sim Uniform(\mu - 0.5, \mu + 0.5)$. Taking in the Example 2.7 $y = x$, and letting $B \subset \Omega_\mu$, we obtain a distribution estimator for μ

$$\xi(x, B) = \lambda\big((x - 0.5, x + 0.5] \cap B\big).$$

Example 2.12. Let $X \sim Uniform(\mu - 0.5, \mu + 0.5)$. Taking in the Example 2.7 $y = x + s_X^2 + 3$, we obtain another distribution estimator for μ, although less reasonable than the previous one.

Theorem 2.13. *Let F be a distribution estimator by Definition 2.9. Then its α-quantile $x_\alpha : (x \le x_\alpha) \iff (F(x) \le \alpha)$ is a random variable for all the levels $\alpha \in [0, 1]$.*

Proof. We observe that

$$(x_\alpha \le x) \iff (F(x) \ge \alpha) = [\cup_{n=1}^{\infty} (F(x) \le \alpha - \frac{\alpha}{1+n})]^C.$$

Hence $(x_\alpha \le x)$ is event for all $x \in \mathbb{R}$, and therefore x_α is a random variable. □

The computations involving distribution estimators are often easier, if the distribution estimators are expressed as randomised estimators. A randomised estimator is a distribution estimator which is a function of both data and a random variable U with a known distribution. More precisely,

Definition 2.14. A ***parameter generating model*** is defined by assuming that a ***randomized estimator*** $\hat{\Gamma}^y$ is on the form

$$\hat{\Gamma}^y = \hat{\gamma}(V, y)$$

given by a measurable function $\hat{\gamma} : \Omega_V \times \Omega_Y \to \Omega_\Gamma$ and a random quantity V with a known law P_V^y for given y. The data space Ω_Y, the Monte Carlo space Ω_V, and the parameter space Ω_Γ are measurable spaces. A parameter generating model is a ***model generating model*** if the parameter equals the model θ.

Definition 2.15. A ***location-scale data generating function*** is on the form

$$Y = (X, S) = (\mu + \sigma U, \sigma V)$$

with location-scale parameter $\theta = (\mu, \sigma) \in \Omega_\Theta = \mathbb{R} \times (0, \infty)$. The joint law $P_{U,V}$ on Ω_Θ of the location U and scale V Monte Carlo variables is assumed known. The location-scale generating function is ***symmetric*** if $U \sim -U$.

The definitions 2.14 and 2.15 are given according to (Taraldsen, private communication, 2021).

EXAMPLE 2.16. The distribution estimator from the example 2.11 may be expressed as

$$\mu \sim U + y - 0.5,$$

where $U \sim Uniform(0, 1)$.

Indeed, in notation of Example 2.7 and Example 2.11,

$$\xi(y, (-\infty, \mu]) = \lambda\big((y-0.5, y+0.5] \cap (-\infty, \mu]\big) = \begin{cases} 0, \mu < y - 0.5 \\ \mu - y + 0.5, y - 0.5 \leq \mu < y + 0.5 \\ 1, \mu \geq y + 0.5 \end{cases}$$

which equals to the distribution function of $U+y-0.5$, where $U \sim Uniform(0, 1)$.

REMARK 2.17. There exist alternative non-equivalent definitions of a randomised estimator. E.g. Lehmann and Casella (2006) defines as 'randomised estimator' an object that is very similar to what we define a a random measure. By (Lehmann & Casella, 2006, p.33), 'If X is a basic random observable, a randomised estimator of $g(\theta)$ is a rule which assigns to each possible outcome x of X a random variable $Y(x)$ with a known distribution. When $X = x$, an observation of $Y(x)$ will be taken and will constitute the estimate of $g(\theta)$'.

For any distribution of the data and any parameter, there always exist following trivial examples of a randomised estimator.

EXAMPLE 2.18. Any distribution estimator can be represented by a function of data and of a random variable taking a constant arbitrary chosen value with probability 1.

EXAMPLE 2.19. Any distribution estimator can be represented by a function of data and of $0 \cdot U$, where U is a random variable with any known distribution

For every distribution estimator, there also exists many non-trivial randomized estimators. Continuous distribution estimators with everywhere nonzero density (therefore strictly increasing distribution function) are especially applicable for the Behrens-Fisher problem. For such estimators, use of randomized estimators is facilitated by following.

PROPOSITION 2.20. *Let W be a continuous distribution estimator, and let U be a continuous random variable, with strictly increasing distribution functions F and Ψ respectively. Then there exist a bijection $G : \Omega_U \mapsto \Omega_W$; $G = F^{-1}(\Psi)$.*

PROOF. We observe that for all rational u, holds

$$u \leq u_0 \iff \Psi(u) \leq \Psi(u_0) \iff F^{-1}\big(\Psi(U) < F^{-1}(\Psi(u_0)\big) \iff w \leq w_0 \tag{3}$$

As both the involved distribution functions are strictly increasing, every value u_0 corresponds to an unique value of w_0, and every value u_0 corresponds to an unique value of w_0 by Eq. (3). Hence the transformation $G : \Omega_U \mapsto \Omega_W$ defined by $G(u) = F^{-1}\Psi(u)$ is a bijection, and W is distributed as $G(U)$.

□

However, a non-trivial randomized estimator is not necessary of a simple form or a practical interest.

EXAMPLE 2.21. Consider a density estimation by Gaussian smoothing (e.g. Racine, 2006, Silverman, 1986), and the consequent distribution estimator of the parameter 'mean' from the density estimate. That is, having sampled the i.i.d. X_1,... X_n, we estimate the underlying probability density as

$$\hat{g}(x) = \sum_{i=1}^{n} \phi(\frac{x - X_i}{\hat{\sigma}}),$$

where $\hat{\sigma}$ is a function of the sample standard deviation, sample size and the version of the smoothing method (e.g. $\hat{\sigma} = 1.06 S_X n_X^{-0.2}$). Than one of natural distribution estimators for the mean is distributed as a mixture of all possible combinations of the n components in a mixture $\hat{g}$ with corresponding weights (For example, in case of only two measurements it has a density

$$h(\mu, X) = \frac{0.25}{\hat{\sigma}/\sqrt{2}}\phi(\frac{\mu - X_1}{\hat{\sigma}/\sqrt{2}}) + \frac{0.25}{\hat{\sigma}\sqrt{2}}\phi(\frac{\mu - X_2}{\hat{\sigma}/\sqrt{2}}) + \frac{0.5}{\hat{\sigma}\sqrt{2}}\phi(\frac{\mu - (X_1 + X_2)/2}{\hat{\sigma}\sqrt{2}}).$$

It is a distribution estimator, because a probability distribution on Ω_Γ is well defined by this density, and this density is a measurable function from Ω_X as a linear combination of continuous Gaussian densities.

Depending on the data, it may have 1 or 3 maximums and hence cannot be expressed as a randomised estimator in a straighforward usable way (e.g. as $X' + X''U$ where X' and X'' are some statistics and U has a standard distribution).

3. Confidence distributions and variables

A confidence distribution is a special case of a distribution estimator. The main general definition of the confidence distribution (Taraldsen, private communication, 2021) is as following.

DEFINITION 2.22. A distribution estimator C for a parameter Γ is a statistic such that Ω_C is a set of probability distributions on the parameter space Ω_Γ, and is a *confidence distribution* (CD) if there exist a non-empty family $A_p \mid p \in I$ of confidence sets A_p with level p, and

$$C(A_p) = p \tag{4}$$

The index set I is the set of levels for the CD C. By default, $I = (0, 1)$, but sometimes other sets are reasonable as values for I.

We call function

$$F(\mu|y) = \int_{-\infty}^{\mu} f(t|y)dt \tag{5}$$

for ***cumulative distribution function for a*** CD for the parameter μ.If the probability density for the CD exists, we call it ***confidence density***. We call ***confidence variable*** the random variable distributed as the CD. By Proposition 2.4, the confidence variable is uniquely defined by the cumulative distribution function F for a CD. If the confidence density f exists it uniquely defines the confidence variable as well, because it defines F as $F(\mu) = \int_{-\infty}^{\mu} f(t)dt$.

In contrast to the Bayesian approach, the CD provides a distribution estimator without any apriori assumptions (Schweder & Hjort, 2016).

We call a CD **continuous** if is cumulative distribution function is continuous in all points.

The relevant example for the Behrens-Fisher problem is the CD for the sigle mean, constructed as following.

EXAMPLE 2.23. Mean μ of the normal distribution $Y_1, ...Y_n \sim N(\mu, \sigma^2)$ with unknown variance σ^2. As for any probability $p \in (0,1)$

$$P(\frac{\mu - \overline{Y}}{S_Y/\sqrt{n_Y}} < t_{p,n_Y-1}) = p,$$

so for any probability $p \in (0,1)$

$$P(\mu < \overline{Y} + t_{\alpha,n_Y-1}S_Y/\sqrt{n_Y}) = p. \tag{6}$$

And therefore we define a CD for μ, given $(\overline{y}, s_Y)$, to be same as for the variable

$$C = \overline{y} + T_{N_Y-1}s_Y/\sqrt{n_Y},$$

i.e. $U = T_{N_Y-1}$. The Eq. (6) provides $A_p = (-\infty, \overline{Y} + t_{\alpha,n_Y-1}S_Y/\sqrt{n_Y})$.

The important practical example where $I \neq (0,1)$ is as following.

EXAMPLE 2.24. $X \sim N(\theta, \sigma^2)$, $I = (0, 0.0, 9988)$. As further numerically shown in Chapter 6, Section 5, the Welch-Aspin test with $n_1 = 5, n_2 = 9$, and an adjusted nominal level, results in a conservative test and is only defined for probabilities in I, but this test is not defined for probabilities of a set $(a, 1)$, where $a > 0.9988$.

A less practical, but still valid example where $I \neq (0,1)$ is:

EXAMPLE 2.25. $X \sim N(\theta, 1)$, $I = \{0.6, 0.7\}$ The symmetric CD may look arbitrary weird for other levels, provided that it is a distribution estimator and its distribution function F satisfies

$$\begin{cases} 0.5 - F(\overline{X} - \theta) = F(\overline{X} + \theta) - 0.5 \\ F\big(\Phi^{-1}(0.8) + \overline{X}\big) = 0.8 = 1 - (1 - 0.6)/2 \\ F\big(\Phi^{-1}(0.85) + \overline{X}\big) = 0.85 = 1 - (1 - 0.7)/2 \end{cases}$$

In addition to the Definition 2.22, diverse non-equivalent definitions of the 'CD' are used.

According to (Schweder & Hjort, 2016), a 'cumulative distribution function for a CD' for a one-dimensional parameter ψ is a non-decreasing right-continuous function $C(\psi, Y)$ of ψ depending on the data Y, provided that it has a uniform distribution whatever the true values of ψ and all the nuisance parameters χ are:

$$P(\psi, \chi)\{C(\psi) \leq \alpha\} = \alpha \tag{7}$$

We call the distribution with cummulative distribution function defined by Eq. (7) *exact* CD. However, the definition can be also extended to the cases where Eq. (7) cannot hold. One of the possible extensions of the definition (Xie & Singh, 2013) is as following. "For every α in $(0,1)$, let $(-\infty, \tau_n(\alpha)]$ be a α-lower-side confidence interval for a parameter θ, where $\tau_n(\alpha)] = \tau_n(\mathbf{x}, \alpha)$ is continuous and increasing in α for each sample $\mathbf{x}$. Then, $H_n(\cdot) = \tau_n^{-1}(\cdot)$ is a CD for θ". $H(\psi, Y)$ converges in distribution to $C(\psi, Y)$ and the definitions of (Taraldsen, 2020) and (Xie & Singh, 2013) coincide. However, when Eq. 7 doesn't hold, such a CD is often asymmetric even when there exist symmetric two-sided confidence intervals for all α in $(0,1)$. The latter confidence intervals is a typical straighforward choice when inverting a test of $H_0 : \Psi = \psi$. Hence the definition of (Xie & Singh, 2013) is unlikely of practical use for real small samples.

We also introduce the following definitions.

Definition 2.26. The function $f(\mu|y)$ is a **symmetric confidence density** for parameter μ is a distribution estimator such that:

(1) $a = g(y)$, where g is a function $\omega_Y \to \mathbb{R}$
$f(a - \mu) = f(a + \mu)$
$E(A) = \mu$
(2) $\forall p \in I$ there exists a quantile $q(p, y)$,such that
- $\int_{-\infty}^{q(p,y)} f(\mu|y) d\mu = p$ for all data $y \in \Omega_Y$
- $P\big(\mu \leq q(p, Y)\big) \geq p$

The index set I of levels in the Definition 2.26 is usually taken $I = (0, 0.5]$. However, it is sometimes reasonable to choose it $(a, 0.5]$ where $0 < a < 0.5$, for example when constructing a CD by inverting the adjusted Welch-Aspin test for the Behrens-Fisher problem.

Definition 2.27. The function $f(\mu|y)$ is a **unimodal probability density for a** CD for parameter μ is a distribution estimator such that:

(1) f has only one local maximum,
(2) $\forall \alpha \in I$ there exists an interval quantile $(a(\alpha, y), b(\alpha, y))$, such that
  - $\int_{a(\alpha,y)}^{b(\alpha,y)} f(\mu|y)d\mu = \alpha$ for all data $y \in \Omega_Y$
  - $P\big(\mu \in (a(\alpha, Y), b(\alpha, Y))\big) \geq \alpha$

If a function satisfies the Definition 2.26 or the Definition 2.27 and in addition the median for the confidence distribution is an estimator which is median unbiased, we call the function a **median unbiased** probability density for a CD. If the a function satisfies the Definition 2.26 or the Definition 2.27 and in addition

$$E\Big[E_Y(\mu|Y)\Big] = \mu_0,$$

(where μ_0 is the true value of the parameter) we call the function a **mean unbiased** probability density for a CD.

Remark 2.28. A CD is never unique.

Theorem 2.29. *Let $F(\theta)$ is a cumulative function for a CD function for a parameter θ, having median M and confidence sets A_p which are symmetric with respect to M, and $0 < k < 1$. Then*

$$G(\theta) = \begin{cases} kF(\theta), \theta \leq M, \\ 1 - k(1 - F(\theta)), \theta > M \end{cases}$$

is also a cumulative distribution function for θ.

Proof. We observe that $\forall a \geq 0:$

$$P(\theta \in (M - a, M + a)) \geq F(M + a) - F(M - a) >$$

$$> 1 - k(1 - F(M + a)) - kF(M - a) = G(M + a) - G(M - a),$$

where the first inequality holds because F is a CD with the defined A_p, and the second inequality holds by the definition of G. If $A_p = (M - a, M + a)$, than the p-confidence set, corresponding G, is $\Big(G^-1\big(F(M + a)\big), G^-1\big(F(M - a)\big)\Big)$. $\square$

Sometimes a CD does not exist or is useless to specify, e.g. Fieler problem (Schweder & Hjort, 2016, p. 117-121). However, a related concept is used. Following (Schweder & Hjort, 2016, p.115), we consider *confidence curves* for the parameter $\theta = (\phi, \eta)$, where ϕ is a focus parameter and η is the nuisance parameter:

$$cc(\theta, y) : \Phi \to [0, 1],$$

which have as its level sets a nested family of confidence regions $R_\alpha(Y) = \{\phi : cc(\theta) \leq \alpha\}$ in Θ, with $\alpha \in [0, 1]$ being the confidence level.

DEFINITION 2.30. Consider a measurable function $cc : \Omega_\phi \times \Omega_Y \mapsto [0,1]$ such that

- $P\Big(cc(\phi, Y) \leq \alpha\Big) \geq \alpha \, \forall\theta, \, \forall\alpha \in [0,1]$
- $cc(\phi, y)$ is nested: $\forall y \in \Omega_Y, \; \alpha_2 > \alpha_1 \Rightarrow \{\phi : cc(\phi, y) < \alpha_1\} \subset \{\phi : cc(\phi, y) < \alpha_2\}$

We call this function a **confidence curve** for ψ.

If all the confidence regions are exact, then

$$P\Big(cc_0(\phi, Y) \leq \alpha\Big) = \alpha \, \forall\theta, \, \forall\alpha \in [0,1]$$

and we call the confidence curve **exact confidence curve**. The exact confidence curve has following properties: (i) $min_\theta \, cc(\theta, y) = cc(\hat{\theta}, y) = 0$ for all outcomes of the data y, where $\hat{\theta}$ is a point estimate; (ii) $cc(\theta_0, Y)$ has the uniform distribution on the unit interval, when θ_0 is the true value of the parameter.

For a continuous CD, $cc(\theta) = 1 - p(\theta)$, where $p(\theta)$ is a p-value function.

There exist infinitely many confidence curves for a given CD.

EXAMPLE 2.31. A classical example of existence of many confidence curves for a single CD is the exponential distribution (Schweder & Hjort, 2016, p.140-142). When expressing the density as $f(y) = \frac{1}{\mu} e^{-\frac{y}{\mu}}$, we observe that the variable $W = \frac{2}{\mu} Y$ is exponentially (1/2) distributed which means that W is χ^2_2 distributed. Therefore,

$$\sum_{i=1}^{n} W_n = \frac{2n\overline{Y}}{\mu} \sim \chi^2_{2n} \tag{8}$$

In order to construct the distribution function for the mean μ, we denote the distribution function for χ^2_{2n} as Γ and get

$$1 - \Gamma(x) = P\Big(\frac{2n\overline{Y}}{\mu} > x\Big) = P\Big(\mu \leq \frac{2n\overline{Y}}{x}\Big)$$

and hence may express the cumulative distribution function for the distribution estimator of μ as

$$C(\mu_0) = P(\mu < \mu_0) = 1 - \Gamma(\frac{2n\overline{Y}}{\mu_0}). \tag{9}$$

This distribution estimator satisfies Eq. (4), when the intervals A_p for $p \in [0,1)$ are taken as $A_p = (0, C^{-1}(p)]$: indeed: $p(C(\mu, Y) \in A_p) \geq p$ by construction. Moreover, we have observed that it is an exact CD: the A_p are not only confidence sets for A_p, but satisfy $p(C(\mu, Y) \in A_p) = p$.

We also observe that the CD for μ may be expressed with a randomised estimator

$$\mu \sim \frac{2n\overline{Y}}{\chi^2_{2n}}. \tag{10}$$

The latter conclusion follows from Eq. (9) but not from Eq. (8) directly.

As pointed in (Schweder & Hjort, 2016), all the intervals $\Big(C^{-1}(a), C^{-1}(b)\Big)$, where $0 \leq a < b \leq 1$ and $b - a = \alpha$, are $\alpha-$level confidence intervals for μ. Hence only the points with coordinates (a, α) or (b, α), where $C(b) - C(a) = \alpha$, belong to the confidence curve. Reasonable restrictions may be added to this condition, but still there exist multiple confidence curves.

A reasonable choice for a confidence curve may be a locus of points

$$\bigcup_{\alpha\in[0,1)} (C^{-1}(\frac{1-\alpha}{2}, \alpha) \cup (C^{-1}(\frac{1+\alpha}{2}, \alpha),$$

where $\hat{\mu} = C^{-1}(0.5)$. In this case the the expression for confidence curve is

$$(cc(\mu) = |1 - 2C(\mu)|.$$

Indeed, $P(cc(\mu) \leq \alpha) = C(C^{-1}(\frac{1+\alpha}{2})) - C(C^{-1}(\frac{1-\alpha}{2})) = \alpha$ and the Definition 2.30 holds.

Another alternative is using the maximum likelihood estimator $\hat{\mu} = \overline{Y}$. The deviance equals

$$-2ln\prod_{i=1}^{n} f(\mu, Y_i) + 2ln\prod_{i=1}^{n} f(\overline{Y}, Y_i) = 2nln(\mu) + \frac{2n\overline{Y}}{\mu} - 2nln(\overline{Y}) - 2n$$

which, by Eq. (8), is distributed as $D = 2n(V_n - 1 - lnV_n)$, where $V_n \sim \chi^n_{2n}/(2n)$. The latter fact is also shown in (Schweder & Hjort, 2016, p.141). Each positive value of the deviance is reached in two values of μ. We denote the cumulative distribution function of U as $F_D(d) = F_D(2nln(\mu) + \frac{2n\overline{Y}}{\mu} - 2nln(\overline{Y}) - 2n)$. We denote the inverses to the two monotone fragments of $F_D(\mu)$ as $F^{-1}_{D1}(\alpha) : [0, 1) \mapsto (-\infty, \hat{\mu})$ and $F^{-1}_{D2}(\alpha) : [0, 1) \mapsto (\hat{\mu}, \infty)$. Then

$$cc(\mu) = \begin{cases} F_{D1}(\mu), \mu < \hat{\mu} \\ F_{D2}(\mu), \mu \geq \hat{\mu}. \end{cases}$$

Indeed, $P(cc(\mu) \leq \alpha) = P(F_D(d) \leq \alpha) = \alpha$.

The confidence distributions may be evaluated by considering their loss and risk (Taraldsen & Lindqvist, 2013). Given the *penalty function* Γ, we define according to (Schweder & Hjort, 2016, p. 162):

Definition 2.32. The confidence ***loss*** at θ of the confidence distribution with distribution function $F(\psi, y)$ for the focus parameter $\psi = a(\theta)$ is

$$lo(\theta, F) = \int_{\Omega_\psi} \Gamma(t - \psi) dF(t, y).$$

Definition 2.33. The confidence ***risk*** at θ is the expected confidence loss:

$$R(\theta, F) = E_\theta \int_{\Omega_\psi} \Gamma(t - \psi) dF(t, Y).$$

CHAPTER 3

Linear combination of confidence variables

An open general problem involves constructing a CD for a linear combination of such randomised estimators. It is of especial interest whether, and in which cases, a linear combination of confidence variables is itself a confidence variable for the corresponding linear combination of parameters. More specifically, if M_1 is distributed so that its density equals to a confidence density $f_1(\mu_1|y_1)$ and M_2 is distributed so that its density equals to a confidence density $f_2(\mu_2|y_2)$, it is of interest in which cases the probability distribution for $c_1M_1 + c_2M_2$ is a CD for $c_1\mu_1 + c_2\mu_2$.

Recently, Hayter (2014), investigating this problem, has found an upper bound for the p-quantiles of a linear combination of symmetric confidence variables (with $p > 0.5$). However, the fact that a linear combination of confidence variables is itself a confidence variable for the corresponding linear combination of parameters, has only been established asymptotically for large samples (Singh et al., 2005).

This chapter demonstrates that a linear combination of confidence variables is generally not a confidence variable for the corresponding linear combination of parameters, though it may be so under mild restrictions. We also provide a preliminary proof that a linear combination of confidence variables is itself a confidence variable under rather weak conditions. We mostly consider the sum of the randomised estimators for the means, because the Behrens-Fisher problem concerns a sum of means.

1. Multiplication of a confidence variable with a scalar

For a constant $c \neq 0$ and a confidence variable Θ, the symmetric confidence variable for $c\Theta$ with symmetric connected confidence sets A_p, or one-sided confidence sets A_p of type $-\infty, a$, in a sense of the Definition 2.26 is a c times scaling of the confidence variable for Θ. Indeed,

$$P(|c\Theta - c\hat{\Theta}| \leq ca) = P(|\Theta - \hat{\Theta}| \leq a)\ \forall a \geq 0$$

For non-symmetric CDs the fact

$$P(c\Theta - c\hat{\Theta} \leq ca) = P(\Theta - \hat{\Theta} \leq a)\ \forall a \in \mathbb{R}$$

holds only for $c > 0$, while for negative c we only have

$$P(c\Theta - c\hat{\Theta} \leq ca) = P(\Theta - \hat{\Theta} \geq a) = 1 - P(\Theta - \hat{\Theta} < a)\,\forall a \in \mathbb{R}.$$

Hence the family of the confidence sets A_p that is mentioned in the Definition 2.22 still exists, but is also reflected with respect to zero. If there are added additional requirements for A_p (e.g. that any element of A_p is a ray $(-\infty, x)$, $x \in \mathbb{R}$, the multiplication of the randomised estimator with a negative scalar will not result in a CD with respect to this family of confidence sets.

2. Linear combination of a symmetric confidence variable and a symmetric unimodal continuous confidence variable

The following conjecture is of special interest both for the Behrens-Fisher problem and other practical problems including construction conservative confidence intervals and CDs.

CONJECTURE 3.1. *Let $\hat{\mu}_1$ and $\hat{\mu}_2$ confidence variables for location parameters, from two independent and symmetric location-scale data generating functions. The randomized estimator $\hat{\tau} = c_1\hat{\mu}_1 + c_2\hat{\mu}_2$ is a confidence variable for $\tau = c_1\mu_1 + c_2\mu_2$ if the location Monte Carlo variable U_1 has a unimodal density.*

The conjecture 3.1 means that if

- Y_1, Y_2, S_1, S_2 be mutually independent statistics,
- $Y_1 \sim \sigma_1 U_1 + \mu_1$, U_1 be symmetric random variable with piecewise continuous non-decreasing density at $(-\infty, 0)$,
- $Y_2 \sim \sigma_2 U_2 + \mu_2$, U_2 be a symmetric random variable,
- $S_1 \sim \sigma_1 V_1$, V_1 be a positive random variable,
- $S_2 \sim \sigma_2 V_2$, V_2 be a positive random variable,
- The random variables U_1, U_2, V_1, V_2 be independent on the parameters $\mu_1, \mu_2, \sigma_1, \sigma_2$,
- $T_1 \sim \frac{U_1}{V_1}$, $T_2 \sim \frac{U_2}{V_2}$.

Than $y_1 - y_2 + s_1 T_1 + s_2 T_2$ is CD for $\mu_1 - \mu_2$ with symmetric connected confidence sets. From the conjecture 3.1 it also follows that for any $0 < \alpha < 0.5$, any point (y_1, y_2, s_1, s_2) in the data space, and the quantile $t_\alpha \in \mathbb{R}$ satisfying $P(|s_1 T_1 + s_2 T_2| > t_\alpha) = \alpha$,

$$P(|Y_1 - \mu_1 - (Y_2 - \mu_2)| > t_\alpha) \leq \alpha. \tag{11}$$

The proof for this conjecture is beyond the scope of this work. The author of this thesis has developed a preliminary proof for the conjecture and, at the moment of submitting this thesis, believes that the conjecture is proved and has become a theorem. However, there have been neither time nor resources to thoroughly check the proof, formulate it more clearly, and have it proofread by someone else. Of this reason, the parts of the proof which are not enough checked are put into the appendix.

This proof requires several easier lemmas.

LEMMA 3.2. *The Conjecture 3.1 holds if $U_2 \sim Unif\{-1, 1\}$ and σ_2 is known.*

Proof: Appendix, section 1.

LEMMA 3.3. *The Conjecture 3.1 holds if:*

- $U_1 \sim Uniform(-a, a)$,
- $U_2 \sim Unif\{-1, 1\}$,
- $V_2 \sim Unif\{b_1, b_2, ..., b_n\}$,
- $0 \le b_1 \le b_2 \le ... \le b_n$,
- *the parameters* σ_1,a, $b_1 ... b_n$ *are known*

Proof: Appendix, section 2.

LEMMA 3.4. *Any symmetric random variable U having a piecewise continuous non-decreasing density $f_U(u)$ at $(-\infty, 0)$ is a limit in distribution of finite mixtures U_n of symmetric uniform distributions.*

Proof: Appendix, section 3.

Corollary. If $V_2 \sim Simple\big(p(b_1) = w_1, p(b_2) = w_2, ..., p(b_n) = w_n)\big)$, $0 < b_1 < b_2 < ... < b_m$, $w_1, ..., w_n$ are known, the Conjecture 3.1 holds, by approximating the weights by rationals $\forall i\ w_i = k_i w$, where $k_i \in \mathbb{N}$ and taking $\sum_{i=1}^{n} k_i$ values of the new V_2.

LEMMA 3.5. *The Conjecture 3.1 holds if:*

- $U_2 \sim Unif\{-1, 1\}$,
- $V_2 \sim Simple\big(p(b_1) = w_1, p(b_2) = w_2, ..., p(b_n) = w_n)\big)$,
- $0 < b_1 < b_2 < ... < b_m$, $w_1, ..., w_n$ *are known.*

Proof: Appendix, section 4.

LEMMA 3.6. *The Conjecture 3.1 holds if U_2 is symmetric and has zero density outside an interval $(-|u_{max}|, |u_{max}|)$.*

Proof: Appendix, section 5.

The Conjecture 3.1 is than proved by stretching infinitely many times the compact support in Lemma 3.6 and by applying the Continuous Mapping Theorem to the T_α as a continuous function of $\{U_1, U_2, V_1, V_2\}$.

3. An example of non-exact CD

As an analytical illustration that the conjecture 3.1 holds, we consider now a following simplified analogue of the Behrens-Fisher problem, including two unknown location parameters and only one unknown scale parameter.

LEMMA 3.7. *Let:*

- Y_1, Y_2, S_1, S_2 *be mutually independent statistics,*
- $Y_1 \sim \sigma_1 U_1 + \mu_1$, $U_1 \sim Uniform(-1, 1)$,
- $Y_2 \sim \sigma_2 U_2 + \mu_2$, $U_2 \sim Unif\{-1, 1\}$,
- $S_1 \sim \sigma_1 V_1$, $V_1 = 1$ *with probability 1 (hence* $T_1 = U_1$*),*
- $S_2 \sim \sigma_2 V_2$, $V_2 \sim Discrete\big(P(V_2 = b_1) = w_1, P(V_2 = b_2) = w_2 = 1 - w_1\big)$,
- *The random variables* U_1, U_2, V_1, V_2 *do not depend on the unknown parameters* $\mu_1, \mu_2, \sigma_1, \sigma_2$,
- $T_1 \sim \frac{U_1}{V_1}$, $T_2 \sim \frac{U_2}{V_2}$,
- $0 < b_1 < b_2$, σ_1, b_1, b_2, w_1 *are known.*

Then the randomised estimator $y_1 - y_2 + s_1 T_1 - s_2 T_2$ *provides a CD for* $\mu_1 - \mu_2$ *with symmetric connected confidence regions.*

For proving this, we will need a following technical lemma.

LEMMA 3.8. *Let:*

- $W_1 \sim Uniform(-a, a)$
- $W_2 \sim Unif\{-b, b\}$
- a, b *are known*

The probability density of $W = W_1 + W_2$ *when* $a \leq b$ *is*

$$f_W(w) = \begin{cases} \frac{1}{4a}, w \in (-a-b, a-b) \cup (-a+b, a+b) \\ 0, otherwise \end{cases}$$

and when $a > b$,

$$f_W(w) = \begin{cases} \frac{1}{4a}, w \in (-a-b, -a+b) \cup (a-b, a+b) \\ \frac{1}{2a}, w \in (-a+b, a-b) \\ 0, otherwise \end{cases}$$

PROOF. The density

$$f_W(w) = f_W(w|W_2 = -b)P(W_2 = -b) + f_W(w|W_2 = b)P(W_2 = b) =$$
$$I(w \in (-a-b, a-b) \cdot \frac{1}{2a} \cdot \frac{1}{2} + I(w \in (-a+b, a+b) \cdot \frac{1}{2a} \cdot \frac{1}{2}$$

and the result follows. □

PROOF OF THE LEMMA 3.7. The (2α)-confidence regions forming the declared *CD* are of form $(-t_\alpha, t_\alpha)$, where t_α is the $(1-\alpha)$-quantile of $s_1 T_1 - s_2 T_2$. By symmetry,

$$P(Y_1 - Y_2 - (\mu_1 - \mu_2) \in (-T_\alpha, T_\alpha) = 1 - 2P\big(Y_1 - Y_2 - (\mu_1 - \mu_2) > T_\alpha\big).$$

For calculating $P\big(Y_1 - Y_2 - (\mu_1 - \mu_2) > T_\alpha\big)$, we observe that:

$$T_2 \sim Discrete\big(P(T_2 = -\frac{1}{b_1}) = \frac{w_1}{2}, P(-T_2 = \frac{1}{b_2}) = \frac{w_2}{2}, P(T_2 = \frac{1}{b_2}) = \frac{w_2}{2},$$

$$P(T_2 = \frac{1}{b_1}) = \frac{w_1}{2}\big).$$

Denoting $b_2/b_1 = k > 1$,

$$S_2T_2|(S_2 = \sigma_2 b_1) \sim Discrete\big(P(s_2T_2 = -\sigma_2) = \frac{w_1}{2}, P(s_2T_2 = -\frac{\sigma_2}{k}) = \frac{w_2}{2},$$

$$P(s_2T_2 = \frac{\sigma_2}{k}) = \frac{w_2}{2}, P(s_2T_2 = \sigma_2) = \frac{w_1}{2}\big),$$

$$S_2T_2|(S_2 = \sigma_2 b_2) \sim Discrete\big(P(s_2T_2 = -k\sigma_2) = \frac{w_1}{2}, P(s_2T_2 = -\sigma_2) = \frac{w_2}{2},$$

$$P(s_2T_2 = \sigma_2) = \frac{w_2}{2}, P(s_2T_2 = k\sigma_2) = \frac{w_1}{2}\big).$$

Furthermore,

$$\sigma_1 U_1 - S_2T_2|(S_2 = \sigma_2 b_1) \sim w_1(\sigma_1 U_1 - \sigma_2 U_2) : w_2(\sigma_1 U_1 - \sigma_2 U_2/k),$$

$$\sigma_1 U_1 - S_2T_2|(S_2 = \sigma_2 b_2) \sim w_1(\sigma_1 U_1 - k\sigma_2 U_2) : w_2(\sigma_1 U_1 - \sigma_2 U_2),$$

and the densities of the components $\sigma_1U_1 - \sigma_2U_2$, $\sigma_1U_1 - k\sigma_2U_2$, $\sigma_1U_1 - \sigma_2U_2/k$ are as described in Lemma 3.8. We now consider the possible relations between t_α, $P\big(Y_1 - Y_2 - (\mu_1 - \mu_2) > t_\alpha\big)$ and parameters.
Consider $\boldsymbol{\sigma_1 < \sigma_2}$.
Assume that $t_\alpha|(S_2 = b_1) > \frac{\sigma_2}{k} + \sigma_1$:
Hence $\alpha < w_1\sigma_2(1 - \frac{1}{k})\frac{1}{4\sigma_1}$ and then

$$P\big(Y_1 - Y_2 - (\mu_1 - \mu_2) > t_\alpha\big) =$$

$$= w_1 P\big(Y_1 - Y_2 - (\mu_1 - \mu_2) > t_\alpha | S_2 = b_1\big) +$$

$$+ w_2 P\big(Y_1 - Y_2 - (\mu_1 - \mu_2) > t_\alpha | S_2 = b_2\big) =$$

$$= w_1 \frac{\alpha}{w_1} + w_2 \cdot 0 = \alpha$$

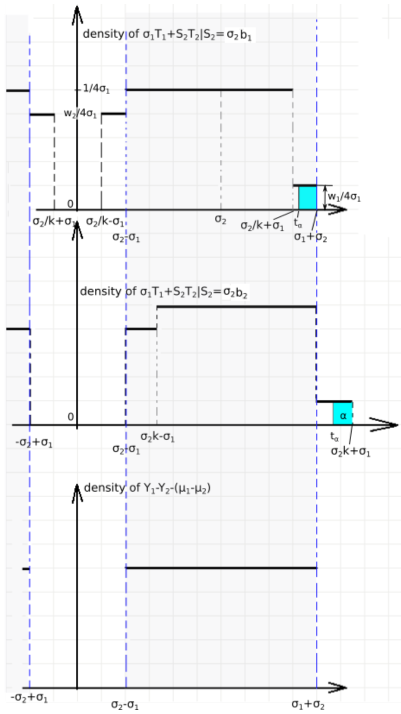

FIGURE 3.1 . Illustration to the proof of Lemma 3.7, case when $\boldsymbol{\sigma_1} < \boldsymbol{\sigma_2}$, $\sigma_2/k - 1 > 0, \sigma_2/k + \sigma_1 > \sigma_2 - \sigma_1$, $\sigma_2 k - \sigma_1 < \sigma_2 + \sigma_1$

The Figure 3.1 shows the case when

$$\sigma_2/k - 1 > 0, \sigma_2/k + \sigma_1 > \sigma_2 - \sigma_1,$$

$$\sigma_2 k - \sigma_1 < \sigma_2 + \sigma_1. \tag{12}$$

The expressions to the left remain same for these t_α when Eq. (12) does not hold (then the mixture components are shifted more with respect to each other, but equal on $(t_\alpha, +\infty)$.

Assume that $(t_\alpha|S_2 = b_1) < \frac{\sigma_2}{k} + \sigma_1$ and $(t_\alpha|S_2 = b_2) > \sigma_2 + \sigma_1$:

$$w_1\sigma_2(1 - \frac{1}{k})\frac{1}{4\sigma_1} <$$

$$< \alpha < w_1 \cdot min(\sigma_2(k-1), 2\sigma_1) \cdot \frac{1}{4\sigma_1}$$

(By "$min(\sigma_2(k-1), 2\sigma_1)$" we consider that if $\sigma_2(k-1) > 2\sigma_1$) and, then, when $S_2 = b_2$, the shifted density component with weight w_1 is disjoint from the unshifted component with weight w_2)

Let $\alpha_1 = w_1\sigma_2(1-\frac{1}{k})\frac{1}{4\sigma_1}$ (α_1 is shown as light blue area in the Figure 3.2), $\alpha_2 = \alpha - \alpha_1$ (α_2 is illustrate by the dark grey area in the figure)

$$P\big(Y_1 - Y_2 - (\mu_1-\mu_2) > t_\alpha\big) =$$
$$w_1 P(Y_1 - Y_2 - (\mu_1-\mu_2) > t_\alpha | S_2 = b_1) +$$
$$+ w_2 P(Y_1 - Y_2 - (\mu_1-\mu_2) > t_\alpha | S_2 = b_2) =$$
$$= w_1\alpha_2 + w_1\frac{\alpha_1}{w_1} + w_2\cdot 0 < \alpha$$

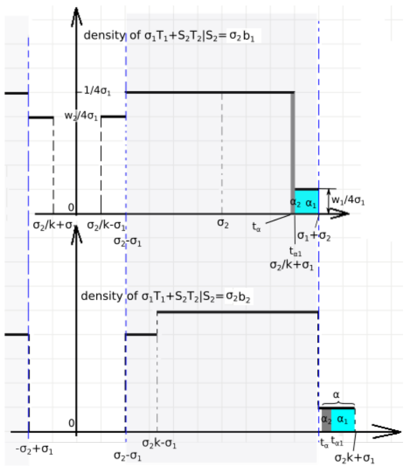

FIGURE 3.2 . Illustration to the proof of Lemma 3.7, case when $\boldsymbol{\sigma_2/k - \sigma_1 > 0}$, $\sigma_2/k+\sigma_1 > \sigma_2-\sigma_1$, $\sigma_2 k - \sigma_1 < \sigma_2+\sigma_1$

The Figure 3.2 shows the case when

$$\sigma_2/k - \sigma_1 > 0, \sigma_2/k+\sigma_1 > \sigma_2-\sigma_1,$$
$$\sigma_2 k - \sigma_1 < \sigma_2+\sigma_1. \tag{13}$$

The case

$$\sigma_2/k - \sigma_1 > 0, \sigma_2/k+\sigma_1 > \sigma_2-\sigma_1,$$
$$\sigma_2 k - \sigma_1 < \sigma_2+\sigma_1,$$

i.e. of the disconnected positive part of support of density of $\sigma_1 T_1 + s_2 T_2 | s_2 = \sigma_2 b_2$, is not shown in the figure. However, splitting $\alpha = \alpha_1 + \alpha_2$ and computation

$$P\big(Y_1 - Y_2 - (\mu_1-\mu_2) > t_\alpha\big) =$$
$$= w_1\alpha_2 + w_1\frac{\alpha_1}{w_1} + w_2\cdot 0 < \alpha$$

is done quite similarly in the latter case.
Assume that $\sigma_2 - \sigma_1 < t_\alpha < \sigma_2 + \sigma_1$ both when $S_2 = b_1$ and when $S_2 = b_2$: $w_1\sigma_2(k-1)\frac{1}{4\sigma_1} < \alpha < 1 - w_2\sigma_2(k-1)\frac{1}{4\sigma_1}, \sigma_2(k-1) < 2\sigma_1)$

$$
\begin{gathered}
P\big(Y_1 - Y_2 - (\mu_1 - \mu_2) > t_\alpha\big) = \\
w_1 P(Y_1 - Y_2 - (\mu_1 - \mu_2) > t_\alpha | S_2 = b_1) + \\
+ w_2 P(Y_1 - Y_2 - (\mu_1 - \mu_2) > t_\alpha | S_2 = b_2) = \\
= w_1\alpha_2 + w_1\frac{\alpha_1}{w_1} + (w_1 + w_2)\alpha_3 + w_2 \cdot 0 < \alpha
\end{gathered}
$$

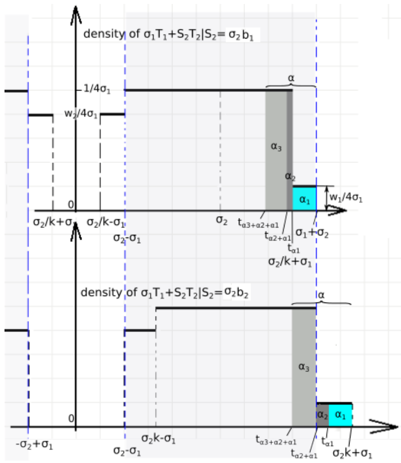

FIGURE 3.3 . Illustration to the proof of Lemma 3.7, case when $\boldsymbol{\sigma_2/k - \sigma_1 > 0}$, $\sigma_2 - \sigma_1 < t_\alpha < \sigma_2 + \sigma_1$ both when $S_2 = b_1$ and $S_2 = b_2$

Assume that $(t_\alpha | S_2 = b_1) < \sigma_2 k - \sigma_1$:
$\sigma_2 - \sigma_1 < t_\alpha < \sigma_2 + \sigma_1$ both when $S_2 = b_1$ and when $S_2 = b_2$

Here it is easier to consider p instead α, and we see that $0 < p < w_2 min(\sigma_2(k-1), 2\sigma_1)\frac{1}{4\sigma_1}$.
Similarly to the first two situations when there was $(t_\alpha | S_2 = b_2) > \frac{\sigma_2}{k} + \sigma_1$,

$$
\begin{gathered}
P(Y_1 - Y_2 - (\mu_1 - \mu_2) > t_\alpha) = \\
1 - w_1 P\big(Y_1 - Y_2 - (\mu_1 - \mu_2) < t_\alpha | S_2 = b_1\big) - \\
- w_2 P\big(Y_1 - Y_2 - (\mu_1 - \mu_2) < t_\alpha | S_2 = b_2\big) = \\
= 1 - w_1(p - \frac{\sigma_2(1 - 1/k)}{4\sigma_1}) - w_2\frac{1-\alpha}{w_2} \geq p
\end{gathered}
$$

Consider now the possible cases when $\boldsymbol{\sigma_1 > \sigma_2}$
For $(t_\alpha | S_2 = b_1) > -\frac{\sigma_2}{k} + \sigma_1$ the argumentation is exactly as for $\sigma_1 < \sigma_2$.

For $-\sigma_2+\sigma_1<(t_\alpha|S_2=b_1)<-\frac{\sigma_2}{k}+\sigma_1$:
As shown in the Figure 3.4 ,

$$P\big(Y_1-Y_2-(\mu_1-\mu_2)>t_\alpha\big)\leq w_1\big(\frac{w_2}{1+w_2}\alpha_3+\alpha_2+\frac{\alpha_1}{w_1}\big)+w_2(\alpha_3+\alpha_2)\leq$$

$$\leq\alpha_3+\alpha_2+\alpha_1=\alpha$$

For $(t_\alpha|S_2=b_1)<-\sigma_2+\sigma_1$:
We observe that

$$(t_\alpha|S_2=b_2)\geq(t_\alpha|S_2=b_1),$$

while $(t_\alpha|S_2=b_1)$ equals to the $1-\alpha$-quantile of $Y_1-Y_2-(\mu_1-\mu_2)$. Hence $P(Y_1-Y_2-(\mu_1-\mu_2)>t_\alpha)<\alpha$

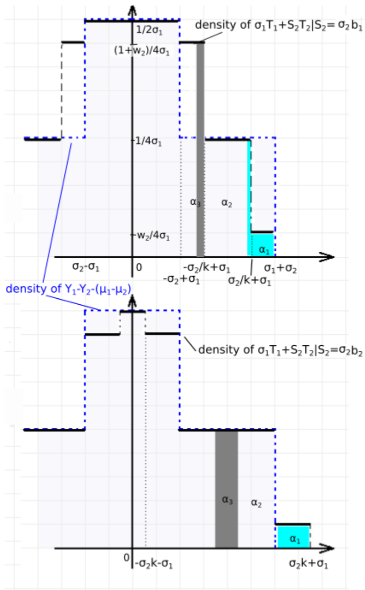

FIGURE 3.4 . Illustration to the proof of Lemma 3.7, case when $\boldsymbol{\sigma_1>\sigma_2}$ and $-\sigma_2+\sigma_1<(t_\alpha|S_2=b_1)<-\frac{\sigma_2}{k}+\sigma_1$

4. Counterexamples for a sum of confidence variable as *CD* of the sum

We present here two counterexamples for the more general situations where sum of confidence variables for parameters is not a confidence variable for the sum of these location parameters

Counterexample for the sum of polymodal confidence variables

Let $Y_1 \sim Uniform(\mu_1 - 1, \mu_1 + 1)$, $Y_2 \sim Uniform(\mu_2 - 1, \mu_2 + 1)$. We define the following confidence densities f_1 and f_2 for the location parameters μ_1 and μ_2

$$f_i(\mu) = \begin{cases} 5, \mu \in (Y_i - 10, Y_i - 9.9) \cup (Y_i + 9.9, Y_i + 10) \\ 0, otherwise, \end{cases} \tag{14}$$

$$i \in \{1, 2\}$$

For each i, f_i is a valid probability density, because

$$f_i(\mu) \geq 0 \forall \mu$$

and

$$\int_{-\infty}^{+\infty} f_i(\mu) d\mu = 5(-9.9 - (-10)) + 5(10 - 9.9) = 1.$$

Moreover, f_i is symmetric with respect to Y_i and $E(Y_i) = \mu_i$. Besides, for any symmetric connected confidence set A_p with $0 < p < 1$,

$$P(\mu \in A_p) > P(\mu \notin (-9.9 + Y_i, 9.9 + Y_i)) = 1 \geq p \, \forall p \in (0, 1).$$

Hence $f_1(\mu|y)$ and $f_2(\mu|y)$ are probability densities for CDs by Definition 2.26, although rather weirdly chosen. (Their confidence densities are shown in Fig. 3.5 , a)

We consider the $Y_1 + Y_2$ and $\mu_1 + \mu_2$. The sum of $\mu_1 + \mu_2$ is distributed as a mixture of 4 possible sums of uniform ($const + U(0, 0.1)$ distributions and has three sharp peaks. Its probability density $f(\mu)$ is shown in Fig. 3.5 , b) and equals:

$$f(\mu) = \begin{cases} 2.5 + 25(\mu + 19.9), \mu \in (Y_1 + Y_2 - 20, Y_1 + Y_2 - 19.9) \\ 2.5 - 25(\mu + 19.9), \mu \in (Y_1 + Y_2 - 19.9, Y_1 + Y_2 - 19.8) \\ 5 + 50\mu, \mu \in (Y_1 + Y_2 - 0.1, Y_1 + Y_2) \\ 5 - 50\mu, \mu \in (Y_1 + Y_2, Y_1 + Y_2 + 0.1) \\ 2.5 + 25(\mu - 19.9), \mu \in (Y_1 + Y_2 + 19.8, Y_1 + Y_2 + 19.9) \\ 2.5 - 25(\mu - 19.9), \mu \in (Y_1 + Y_2 + 19.9, Y_1 + Y_2 + 20) \\ 0, otherwise. \end{cases} \tag{15}$$

If $f(\mu)$ were a confidence density, there would hold $P(|Y_1 + Y_2 - (\mu_1 + \mu_2)| < 0.1) \geq 0.5$ However, the density $g(y)$ for $Y_1 + Y_2$ is

$$g(y) = \begin{cases} 0.5 + 0.25\big(y - (\mu_1 + \mu_2)\big), y \in (\mu_1 + \mu_2 - 2, \mu_1 + \mu_2) \\ 0.5 - 0.25\big(y - (\mu_1 + \mu_2)\big), y \in (\mu_1 + \mu_2, \mu_1 + \mu_2 + 2) \\ 0, otherwise, \end{cases} \tag{16}$$

And

$$P(|Y_1 + Y_2 - (\mu_1 + \mu_2)| < 0.1) = \int_{\mu_1 + \mu_2 - 0.1}^{\mu_1 + \mu_2 + 0.1} g(y) dy = 2 \int_0^{0.1} (0.5 - 0.25t) dt =$$

$$= 2(0.5t - 0.125t^2)\Big|_0^{0.1} = 0.0975,$$

which also corresponds the numerical estimation by Monte-Carlo simulation. This probability is much less than it would be if $f(\mu)$ were not a CD for the sum of means. We have a contradiction.

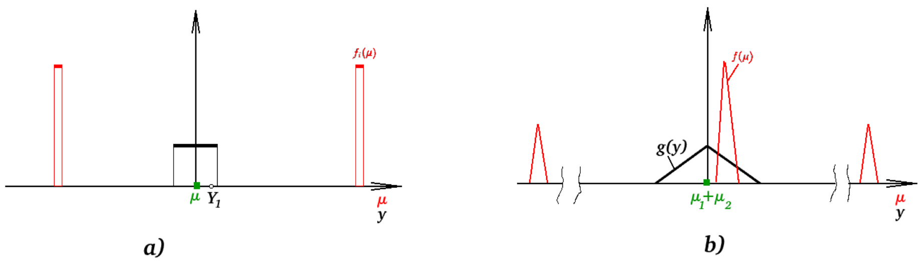

FIGURE 3.5 . A bimodal symmetric CD and the sum of two bimodal symmetric CDs. The black graphs corresponds the densities of data, the red graphs correspond confidence densities for μ, the green text shows the true parameter values

Counterexample for exact confidence distributions

We consider the conditions of the Behrens-Fisher problem and will denote $\theta_1 = |1/\mu_1|$, $\theta_2 = |1/\mu_2|$. The following confidence variable for θ_1 is $W_1 = |\frac{1}{T_{n_1-1}s_1/\sqrt{n_1}+x_1}|$, and similarly $W_2 = |\frac{1}{T_{n_2-1}s_2/\sqrt{n_2}+x_2}|$ is a confidence variable for θ_2. Indeed, for any p-quantile of the absolute value of student t-distribution $|t|_{p,n_1-1}$,

$$(|\mu_i| < ||t|_{p,n_i-1}S_i/\sqrt{n_1}+x_i|) \iff (|\frac{1}{\mu_1}| > |\frac{1}{|t|_{p,n_i-1}S_1/\sqrt{n_1}+x_1})|,$$

and therefore for any p_1, p_2: $0 < p_1 < p_2 < 1$,

$$p_2 - p_1 = P(|t_{p_1,n_i-1}S_i/\sqrt{n_i}+x_i|) < |\mu_1| < |t_{p_2,n_i-1}S_i/\sqrt{n_i}+x_i|) =$$
$$P(|\frac{1}{|t|_{p_1,n_i-1}S_i/\sqrt{n_i}+x_i})| > |\frac{1}{\mu_i}| > |\frac{1}{|t|_{p_2,n_i-1}S_i/\sqrt{n_i}+x_i})|,$$

which defines the exact confidence sets for $\frac{1}{\mu_i}$.

However, setting $\mu_1 = \mu_2 = 10$, $\sigma_1 = \sigma_2 = 1$, $n_1 = 3$, $n_2 = 5$, and computing numerically the 0.025- and 0.975-quantiles (Q_1, Q_2) of

$$|\frac{1}{T_{n_1-1}s_2/\sqrt{n_1}+\overline{X}}| - |\frac{1}{T_{n_2-1}s_2/\sqrt{n_Y}+\overline{y}}|,$$

using 30000 simulations, we obtain that $P\big(0 \in (Q_1, Q_2)\big) \approx 0.942 < 0.95$

We conclude that $W_1 + W_2$ is not a confidence variable for $\theta_1 + \theta_2$.

CHAPTER 4

Tests for the Behrens-Fisher problem

1. Simple properties of the Behrens-Fisher statistic

A large class of the tests are based on the *Behrens-Fisher statistic*

$$B = \frac{X_1 - X_2}{\sqrt{S_1^2/n_1 + S_2^2/n_2}}$$

Assume in conditions and notation of the Behrens-Fisher problem $k = \frac{\sigma_1^2}{\sigma_2^2}$, $\tau = ln(k)$.

PROPOSITION 4.1. *Given $H_0 : \mu_1 = \mu_2$,*

$$B \sim \frac{Z\sqrt{\frac{k}{n_1} + \frac{1}{n_2}}}{\sqrt{\frac{kV_1}{n_1(n_1-1)} + \frac{V_2}{n_2(n_2-1)}}},$$

where Z, V_1, V_2 are independent with $Z \sim N(0,1)$, $V_1 \sim \chi^2_{n_1-1}$, $V_2 \sim \chi^2_{n_2-1}$.

PROOF. A sum of independent normal variables is also normally distributed. This gives

$$X_1 \sim N(\mu_1, \sigma_1^2/n_1),\ X_2 \sim N(\mu_2, \sigma_2^2/n_2),\ (X_1 - X_2) \sim N(\mu_1 - \mu_2, \sigma_1^2/n_1 + \sigma_2^2/n_2).$$

At H_0, $\mu_1 - \mu_2 = 0$, so

$$X_1 - X_2 \sim N(0, \sigma_1^2/n_1 + \sigma_2^2/n_2) \sim Z\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2},$$

where $Z \sim (0,1)$
It is also well-known (Casella & Berger, 2002, p.218) that

$$V_1 = \frac{S_1^2(n_1-1)}{\sigma_1^2} \sim \chi^2_{n_1-1},\ V_2 = \frac{S_2^2(n_2-1)}{\sigma_2^2} \sim \chi^2_{n_2-1}.$$

Therefore,

$$B \sim \frac{Z\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}{\sqrt{\frac{\sigma_1^2 V_1}{n_1(n_1-1)} + \frac{\sigma_2^2 V_2}{n_2(n_2-1)}}} = \frac{Z\sqrt{\frac{k}{n_1} + \frac{1}{n_2}}}{\sqrt{\frac{kV_1}{n_1(n_1-1)} + \frac{V_2}{n_2(n_2-1)}}}$$

□

2. Conservative tests and the nominal level

A statistical hypothesis is a measurable subset of parameter space.

EXAMPLE 4.2. In the classical form of Behrens-Fisher problem, there is the null hypothesis $H_0 : \mu_1 = \mu_2$ and the alternative hypothesis "$H_1 : \mu_1 \neq \mu_2$". H_0 may be written explicitly as $\{(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2) : \mu_1 = \mu_2\} \subset \mathbb{R}^2 \times \mathbb{R}_+^2$, and H_1 may be written explicitly as $\{(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2) : \mu_1 \neq \mu_2\} \subset \mathbb{R}^2 \times \mathbb{R}_+^2$.

DEFINITION 4.3. A ***statistical test*** for a statistical hypothesis is a function $\Omega_Y \to \{''accept'', ''reject''\}$.

The test is well-defined by specifying the set of data C for which the decision of the test is "reject", which is called the ***critical region***. The ***level*** of the test α satisfies $P(Y \in C) \leq \alpha$ (Lehmann, 2006, p.57).

A CD with an index set I may be constructed by inverting a test for $H_0 : \mu = \sum_{i=1}^{n} E(Y_i)$ if the test is conservative for all levels α satisfying $1 - \alpha \in I$. Being a ***conservative*** test means that the probability of rejecting H_0 indeed does not exceed α, or, equivalently, that the acceptance regions are indeed confidence sets with level $1 - \alpha$. However, not all tests which are constructed to be of level α (which we call the ***nominal*** level of the test $\tilde{\alpha}$) are in fact of that planned level, which we call the ***actual level***.

This way of constructing CDs is relevant to n-sample general case of the Behrens-Fisher problem.

Many tests have rejection region $\frac{\sum_{i=1}^{n} \overline{Y}_i - \mu}{T(S_1, \ldots S_i)} \in \Big(-\infty, H^{-1}(\alpha/2)\Big) \cup \Big(H^{-1}(1 - \alpha/2), \infty\Big)$, where $T(S_1, \ldots S_i)$ is a statistic which is independent on the means $\mu_1, ..\mu_i$. Boundaries of the acceptance regions of the tests can be treated as functions of $\alpha \in I$ where I is a set of available levels.

3. Methods of numerical study of the tests

3.1. Monte-Carlo simulations

For all the discussed tests, we studied the probability that $\mu_1 - \mu_2$ is not within the confidence interval $A_{1-\tilde{\alpha}}$ constructed by the inversion of the tests, as a function of the nominal test level $\tilde{\alpha}$ and the parameters. For the known parameters, there were simulated a large number of samples with $\mu_1 = \mu_2$ (this number is specified further for each test) and the test of level $\tilde{\alpha}$ was conducted for each sample. The $P(\mu_1 - \mu_2) \neq A_{1-\tilde{\alpha}}$ was estimate as empiric frequency of rejecting $H_0 : \mu_1 = \mu_2$

3.2. Numerical integration

The Welch-Satterthwaite test and Welch-Aspin test were in addition studied via numerical integration. The Simpson method was applied to computing

$$P(reject\ H_0) =$$
$$\int_0^{\infty} \int_0^{\infty} P(reject\ H_0 | s_1, s_2) \frac{n_1 - 1}{\sigma_1^2} f_{\chi_{n_1 - 1}}(s_1 \frac{n_1 - 1}{\sigma_1^2}) \frac{n_2 - 1}{\sigma_2^2} f_{\chi_{n_2 - 1}}(s_2 \frac{n_2 - 1}{\sigma_2^2}) ds_1 ds_2$$

4. Behrens-Fisher test

This test provides the historically earliest solution for the Behrens-Fisher problem, suggested by Behrens (1929, as described by Best and Rayner (1987)) and developed by Fisher (Fisher, 1935; Fisher, 1930). Let $\varphi = atan\frac{S_1^2/n_1}{S_2^2/n_2}$, the student-t distributed variables T_{n1} and T_{n2} be independent, $t_\alpha/2$ be a $1-\alpha/2$-quantile of $T_{n1}sin\varphi + T_{n2}cos\varphi$. Reject H_0 if $\left|\frac{X_1-X_2}{\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}}\right| > t_\alpha/2$, accept H_0 otherwise.

This test provides a CD which is identical to the CD by linear combination, because its construction is based on the distribution estimators for μ_1 and μ_2, taken the confidence variables $\hat{\mu_1} \sim x_1 + T_1\sqrt{s_1^2/n_1}$, $\hat{\mu_2} \sim x_2 + T_2\sqrt{s_2^2/n_2}$. As shown in the Example 2.23, these are exact CDs for μ_1 and μ_2. In order to obtain the test, we assume that

$$\hat{\mu} = \mu_1 \hat{-} \mu_2 \sim X_1 - X_2 + T_{n1}\sqrt{S_1^2/n_1} - T_{n2}\sqrt{S_2^2/n_2}$$

which is equivalent to

$$\frac{X_1 - X_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \sim T_{n1}|_\Theta cos(\varphi) - T_{n2}|_\Theta sin(\varphi) \tag{17}$$

The assumption that the Eq. (17) holds means that the sum of the CDs for μ_1 and μ_2 is an exact CD, and it also means that $T_{n1}|\theta \sim T_{n1}$, $T_{n2}|\theta \sim T_{n2}$. In fact, the the student-t distributed variables involved in the Eq. (17) are dependent. The difference of the CDs fr μ_1 and μ_2 is a CD for $\mu_1 - \mu_2$ by our Theorem 3.1. However, it is not an exact CD, as many numerical simulations reveal (e.g. Wang, 1971).

The Behrens-Fisher fiducial distribution is identical with a posterior for a Jeffrey's prior (Ghosh & Kim, 2001).

If the preliminary proof of the Conjecture 3.1 is correct, the direct consequence is that the Behrens-Fisher test is conservative. If the proof is incorrect, whether the Behrens–Fisher test is conservative remains an open problem. However, Appendix B provides an alternative original proof indicating that the Behrens-Fisher test is conservative for sample sizes $n_1 = n_2 = 2$ or $n_1 = n_2 = 3$.

5. Welch-Satterthwaite test: ISO GUM version

The Welch-Satterthwaite approximation is applicable to N normally distributed samples, such that each $i-$th of them has the sample average $\overline{X_i}$, sample variance S_i^2, size n_i, mean μ_i and variance σ_i^2. According to this approximation, the distribution of the variable $V = \frac{\sum_{i=1}^N \mu_i - \sum_{i=1}^N \overline{X_i}}{U_c}$ may be approximated by a student t-distribution with ν_{eff} degrees of freedom, where

$$\begin{cases} U_c^2 = \sum_{i=1}^N \frac{S_i^2}{n_i}, \\ \frac{U_c^4}{\nu_{eff}} = c \end{cases} \tag{18}$$

The test is derived as follows, generalizing (Larsen & Marx, 2013, p. 465–466). The statistic V is assumed to be Student t-distributed. The nominator of V can be presented as

$$Z\sum_{i=1}^{N}\sqrt{\frac{\sigma_i^2}{n_i}},$$

where $Z \sim N(0,1)$. Hence the denominator of V is assumed to have the distribution $\sqrt{\sum_{i=1}^{N}\frac{S_i^2}{n_i}} \sim \sqrt{\frac{W}{\nu_{eff}}\sum_{i=1}^{N}\frac{\sigma_i^2}{n_i}}$ where W is chi-square distributed with ν_{eff} degrees of freedom (and hence the mean ν_{eff} and variance $2\nu_{eff}$). In order to establish ν_{eff}, we notice that $S_i^2 = \frac{w_i\sigma_i^2}{n_i-1}$ where $S_i^2 \sim \chi^2_{n_i-1}$, W_i and W_j are independent for all $i \neq j$. We apply method of moments to

$$\sum_{i=1}^{N}\frac{S_i^2}{n_i} \sim \frac{W}{\nu_{eff}}\sum_{i=1}^{N}\frac{\sigma_i^2}{n_i}. \tag{19}$$

The expected values of the left and the right sides of Eq. (19) are equal for all ν_{eff}. Setting the variances of both sides of Eq. (19) to be equal, we get

$$\sum_{i=1}^{N}\frac{\sigma_i^4 2(n_i-1)}{n_i^2(n_i-1)^2} = \frac{2\nu_{eff}\left(\sum_{i=1}^{N}\frac{\sigma_i^2}{n_i}\right)^2}{\nu_{eff}^2}$$

Replacing each of σ_i^2 with its unbiased estimator S_i^2 and letting $U_c = \sum_{i=1}^{N}\frac{\sigma_i^2}{n_i}$, we obtain

$$\sum_{i=1}^{N}\frac{S_i^4}{n_i^2(n_i-1)} = \frac{1}{\nu_{eff}\left(\sum_{i=1}^{N}\frac{S_i^2}{n_i}\right)^2}$$

which corresponds to the Eq. (18).

By the Central Limit Theorem, even if X_i for all i have another distribution as the normal, their sample average converges to a normally distributed random variable as the sample sizes increase, under very mild restrictions. The type of convergence depends on the restrictions (Karr, 1993, p. 183-216). Therefore the approximation Eq. (18) can be extended as in the ISO GUM standard (JCGM et al., 2008) to large but not necessary normally distributed samples $X_1..X_N$. It can also be extended to arbitrary function of the data which is independent on the sample variances, rather than linear combination of the averages. In this general case there is introduced

$$Y = f(X_{11}, ...X_{1n_1}, ...X_{i1}, ..X_{Nn_N}),$$

where X_{ij} is a j-th measurement in the i-th sample

$$U_c^2 = \sum_{i=1}^{n}\sum_{j=1}^{N}\left(\frac{\partial f}{\partial x_{ij}}\right)^2 u^2(X_i)$$

and U_i^2 is the estimate of the variance of the mean of X_i, typically S_i^2/n_i. Than

$$\frac{U_c^4}{\nu_{eff}} = \sum_{i=1}^{N}\frac{U_i^4}{n_i-1}$$

6. Welch-Aspin test

The Welch-Aspin test enables constructing a CD not only for the Behrens-Fisher problem, but also for the mean of k normally distributed each with unknown variance. Welch (1947) and Aspin (1948) calculated four terms of the Taylor-series for the boundary of the confidence interval for $(\overline{X} - \overline{Y})$. The boundary of the confidence region is of form:

$$\sqrt{(\frac{S_1^2}{n_1}) + \frac{S_2^2}{n_2})} \cdot polynom\Big(\frac{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}{(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2})}, .. \frac{\frac{S_1^{2r}}{n_1^r} + \frac{S_2^{2r}}{n_2^r}}{(\frac{S_1^2}{n_X} + \frac{S_2^2}{n_2})^r}\Big)$$

(where ξ is the $(\alpha/2)$-quantile for the normal distribution, r the number of term, and the polynom is of r-th degree).

There was a typo for the expression for the fourth term in the original article in 1948, which was shown by Bachmaier (e.g. 2012). We provide here the correct expression. In order to specify the boundary of the confidence interval for μ, we denote, following the original notation of Alice Aspin (1948), $\lambda_i = 1/n_i$, $f_i^u = 1/(n_i - 1)^u$, $V_{ru} = \frac{\sum_{i=1}^k \frac{\lambda_i^r s_i^{2r}}{f_i^u}}{(\sum_{i=1}^k)^r \lambda_i s_i^2)^r}$, ξ be the $0.5 + 0.5p$-quantile of the standard normal distribution for symmetric confidence regions A_p (or ξ be the p-quantile of the standard normal distribution for one-sided confidence regions A_p). Then the boundary of the confidence set A_p may be approximated as $h = h_0 + h_1 + h_2 + h_3 + h_4$, where $h_0 = \xi\sqrt{\sum_{i=1}^k \lambda_i s_i^2}$

$h_1 = \frac{1}{4}(1 + \xi^2)V_{21}h_0$

$h_2 = \Big(- \frac{1}{2}(1 + \xi^2)V_{22} + \frac{1}{3}(3 + 5\xi^2 + \xi^4)V_{32} - \frac{1}{32}(15 + 32\xi^2 + 9\xi^4)V_{21}^2\Big)h_0$

$h_3 = \Big((1 + \xi^2)V_{23} - 2(3 + 5\xi^2 + \xi^4)V_{33} + \frac{1}{8}(15 + 32\xi^2 + 9\xi^4)V_{22}V_{21} + \frac{1}{8}(75 + 173\xi^2 + 63\xi^4 + 5\xi^6)V_{43} - \frac{1}{12}(105 + 298\xi^2 + 140\xi^4 + 15\xi^6)V_{32}V_{21} + \frac{1}{384}(945 + 3169\xi^2 + 1811\xi^4 + 243\xi^6)V_{21}^3\Big)h_0$

$h_4 = \Big(- 2(1 + \xi^2)V_{24} + \frac{28}{3}(3 + 5\xi^2 + \xi^4)V_{34} - \frac{1}{4}(15 + 32\xi^2 + 9\xi^4)(V_{23}V_{21} + \frac{1}{2}V_{22}^2 - \frac{3}{2}(75 + 173\xi^2 + 63\xi^4 + 5\xi^6)V_{44} + \frac{1}{2}(105 + 298\xi^2 + 140\xi^4 + 15\xi^6)(\frac{1}{3}V_{22}V_{32} + V_{21}33) + \frac{1}{4}(15 + 33\xi^2 + 11\xi^4 + \xi^6)V_{44} + \frac{1}{5}(735 + 2170\xi^2 + 1126\xi^4 + 168\xi^6 + 7\xi^8)V_{54} - \frac{1}{64}(945 + 3169\xi^2 + 1811\xi^4 + 243\xi^6)V_{22}V_{21}^2 - \frac{1}{18}(945 + 3354\xi^2 + 2166\xi^4 + 425\xi^6 + 25\xi^8)V_{32}^2 - \frac{1}{32}(4725 + 16586\xi^2 + 10514\xi^4 + 1974\xi^6 + 105\xi^8)V_{21}V_{43} + \frac{1}{96}(10395 + 42429\xi^2 + 31938\xi^4 + 7335\xi^6 + 495\xi^8)V_{32}V_{21}^2 - \frac{1}{6144}(135135 + 66144\xi^2 + 542026\xi^4 + 145320\xi^6 + 11583\xi^8)V_{21}^4\Big)h_0$

7. Paired t-test

The paired t-test is not based on a sufficient statistic. It is only applicable when the sample sizes are equal: $n_1 = n_2$. For each component j of the samples, the difference $D_j = Y_{1,j} - Y_{2,j}$ is normally distributed:

$$D_i \sim N(\mu_1 - \mu_2, \sigma_1^2 + \sigma_2^2$$

and hence

$$\frac{\overline{D} - (\mu_1 - \mu_2)}{S_D/\sqrt{n_1}} \sim t_{n_1-1}.$$

Therefore there exists an exact test: reject H_0 if

$$\frac{\overline{D}-(\mu_1-\mu_2)}{S_D/\sqrt{n_1}} \in (-\infty, -t_{\alpha/2}) \cup (t_{\alpha/2}, \infty),$$

accept H_0 otherwise.

The test may be generalized for m independent samples for $H_0 : \sum_{i=1}^{m} \mu_i = \mu$.

8. Likelihood ratio test

The likelihood ratio test is not based on the three statistics $X_1 - X_2$, S_1^2, S_2^2, but rather on a more complicated function of all the four sufficient statistics X_1, X_2, S_1^2, S_2^2[1]. We observe that for the given X_1, X_2, the critical region for the likelihood ratio statistic is in form $\Big(-\infty, H^{-1}(\alpha/2)\Big) \cup \Big(H^{-1}(1-\alpha/2), \infty\Big)$. The maximum likelihood estimator $\hat{\mu}$ for μ, along with the corresponding $\hat{\sigma_1^2}$ and $\hat{\sigma_2^2}$ obtained above as the solution of the following system of equations:

$$\begin{cases} \hat{\mu} = \frac{n_1 X_1 \hat{\sigma_2^2} + n_2 X_2 \hat{\sigma_1^2}}{n_2 \hat{\sigma_1^2} + n_1 \hat{\sigma_2^2}} \\ \hat{\sigma_1^2} = \frac{\sum_{i=1}^{n_1}(X_i - X_1)^2}{n_1} + (X_1 - \hat{\mu})^2 = Q_1^2 + (X_1 - \hat{\mu})^2 \\ \hat{\sigma_2^2} = \frac{\sum_{i=1}^{n_2}(Y_i - X_2)^2}{n_2} + (X_2 - \hat{\mu})^2 = Q_2^2 + (X_2 - \hat{\mu})^2, \end{cases} \tag{20}$$

which may be estimated by the iteration method (Cox & Jaber, 1990). In this notation, Q_1 and Q_2 are the maximum likelihood estimators of the variances σ_1^2 and σ_2^2 respectively. The likelihood ratio equals

$$\Lambda = (\frac{Q_1^2}{\hat{\sigma_1^2}})^{\frac{1}{2}n_1}(\frac{Q_2^2}{\hat{\sigma_2^2}})^{\frac{1}{2}n_2} exp(-\frac{n_1 Q_1^2}{2Q_1^2} - \frac{n_2 Q_2^2}{2Q_2^2} + \frac{n_2}{2} + \frac{n_2}{2}) = (\frac{Q_1^2}{\hat{\sigma_1^2}})^{\frac{1}{2}n_1}(\frac{Q_2^2}{\hat{\sigma_2^2}})^{\frac{1}{2}n_2}.$$

The distribution of the likelihood ratio under $H_0 : \mu_1 = \mu_2$ depends on the parameter σ_1/σ_2. The critical values of the likelihood ratio are computed numerically and maximised over the parameter space.

[1] The $\hat{\mu}$ is not shift invariant while $X_1 - X_2$, S_1^2, S_2^2 are, and hence $\hat{\mu}$ depends on other statistics than these three.

CHAPTER 5

Constructing CDs for the Behrens-Fisher problem

1. As distribution of the difference of confidence variables

The difference of CDs for the means μ_1 and μ_2 is a CD by our Conjecture 3.1 which have been observed numerically. The original preliminary analytical proofs are presented in the Chapter 3, in sections 2 – 5 of the Appendix, and alternatively, in terms of a conservative test with $n_1 = n_2 = 2$ or $n_1 = n_2 = 3$, in Appendix B.

2. By inverting conservative tests for the Behrens-Fisher problem

THEOREM 5.1. *Let the critical region for a test* $H_0 : \mu = \sum_{i=1}^n E(Y_i)$ *of any level* $\alpha \in (0,1)$ *be in form* $\frac{\sum_{i=1}^n \overline{Y}_i - \mu}{T(S_1,...S_i)} \in \Big(-\infty, H^{-1}(\alpha/2)\Big) \cup \Big(H^{-1}(1-\alpha/2), \infty\Big)$*, where* $T(S_1,...S_i)$ *is a statistic which is independent on sample averages and* H *is a continuous distribution function, with an even derivative* h*). Then* $F(\mu) = H(\frac{\mu - \sum_{i=1}^n \overline{Y}_i}{T(S_1,...S_i)})$ *is a cumulative distribution function for a* CD *for* μ *and* $f(\mu) = \frac{dF}{d\mu}$ *is a probability density for a* CD *for* μ*.*

PROOF. The requirements 1 and 2 of the definition 2.26 hold, because F is a distribution function by construction.
$h(\frac{\mu}{T(S_1,...S_i)})$ is even as a scale-transformation of $h(\mu) = \frac{dH}{d\mu}$ which is even by the assumption, and $h(\frac{\mu - \sum_{i=1}^n \overline{Y}_i}{T(S_1,...S_i)})$ is a location transformation: a shift of $h(\frac{\mu}{T(S_1,...S_i)}$ $f(\mu)$ by $\sum_{i=1}^n \overline{Y}_i$. Hence f is symmetric with respect to $\sum_{i=1}^n \overline{Y}_i$. As $E(\sum_{i=1}^n \overline{Y}_i) = \mu$, the requirement 3 also holds. By the symmetry of the critical region,

$$\alpha/2 \geq 0.5P\Big(\frac{\sum_{i=1}^n \overline{Y}_i - \mu}{T(S_1,...S_i)} \in C\Big)$$

By symmetry of h,

$$0.5P\Big(\frac{\sum_{i=1}^n \overline{Y}_i - \mu}{T(S_1,...S_i)} \in C\Big) = 0.5P\Big(\frac{|\sum_{i=1}^n \overline{Y}_i - \mu|}{T(S_1,...S_i)} > H^{-1}(1-\alpha/2)\Big) =$$

$$= P\Big(\frac{\mu - \sum_{i=1}^n \overline{Y}_i}{T(S_1,...S_i)} < H^{-1}(\alpha/2)\Big) = P\big(\mu < \sum_{i=1}^n \overline{Y}_i + H^{-1}(\alpha/2) \cdot T(S_1,...S_i)\big) \tag{21}$$

Choosing $q(\alpha) = H^{-1}(\alpha) \cdot T(S_1,...S_i) + \sum_{i=1}^n \overline{Y}_i$, we have

- $F(q(\alpha)) = H(\frac{q(\alpha)-\sum_{i=1}^{n}\overline{Y}_i}{T(S_1,\ldots S_i})= \alpha$, and hence $\int_{-\infty}^{q(\alpha,y)} f(\mu|y)d\mu = \alpha \forall y$
- $P\big(\mu \leq q(\alpha, Y)\big) \leq \alpha$ by Eq. (21),

so the requirement 4 of the definition 2.26 holds as well. □

The confidence distributions constructed by inverting tests from chapter 4 using Theorem 5.1 are as following.

2.1. From Behrens-Fisher test

As follows from the Eq. (17), the expression for the confidence density is

$$f(\mu) = \frac{1}{u} f_{BF}(\frac{\mu-(x_1-x_2)}{u}, s_1, s_2, n_1, n_2),$$

where $u = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$ and $f_{BF}(x, s_1, s_2, n_1, n_2)$ is the probability density for $X = T_{n1}|_{\Theta} cos(atan\frac{s_1^2/n_1}{s_2^2/n_2}) - T_{n2}|_{\Theta} sin(atan\frac{s_1^2/n_1}{s_2^2/n_2})$. The Theorem 5.1 was applied here with $H(x) = F_{BF}(x, s_1, s_2, n_1, n_2)$.

2.2. From Welch-Satterthwaite test

As follows from the Eq. (18), the expression for the confidence density is

$$f(\mu) = \frac{1}{u}\psi T, \nu_{eff}(\frac{\mu - \sum_{i=1}^{N}\overline{X_i}}{u}, \nu_{eff}),$$

where $u = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$ and $\nu_{eff} = \frac{\left(\sum_{i=1}^{N}\frac{S_i^2}{n_i}\right)^2}{\sum_{i=1}^{N}\frac{S_i^4}{n_i^2(n_i-1)}}$. The Theorem 5.1 was applied here with $H(x) = \Psi(x, \nu_{eff})$.

2.3. From the paired t-test

The confidence density is

$$f(\mu) = \frac{1}{S_D/\sqrt{n_1}}\psi(\frac{\mu - \overline{D}}{S_D/\sqrt{n_1}}, n_1 - 1),$$

where $D_j = \sum_{i=1}^{m}\mu_i Y_{i,j}$. The Theorem 5.1 was applied here with $H(x) = \Psi(x, n_1 - 1)$.

2.4. From the likelihood ratio test

With the help of likelihood ratio test, a CD for $\mu = \mu_1 - \mu_2$ may be constructed as following. The maximum likelihood estimator $\hat{\mu}$ for the common mean of X_1 and $X_2+\mu$, along with the corresponding $\hat{\sigma_1^2}$ and $\hat{\sigma_2^2}$ becomes, same as when testing $H_0 : \mu_1 = \mu_2+\mu$:

$$\begin{cases} \hat{\mu} = \frac{n_1 X_1 \hat{\sigma_2^2} + n_2(X_2+\mu)\hat{\sigma_1^2}}{n_2\hat{\sigma_1^2} + n_1\hat{\sigma_2^2}} \\ \hat{\sigma_1^2} = \frac{\sum_{i=1}^{n_1}(X_{1i}-X_1)^2}{n_1} + (X_1 - \hat{\mu})^2 = Q_1^2 + (X_1 - \hat{\mu})^2 \\ \hat{\sigma_2^2} = \frac{\sum_{i=1}^{n_2}(X_{2i}-X_2)^2}{n_2} + ((X_2+\mu) - \hat{\mu})^2 = Q_2^2 + ((X_2+\mu) - \hat{\mu})^2. \end{cases} \tag{22}$$

and the likelihood ratio still equals

$$\Lambda = (\frac{U_1^2}{\hat{\sigma_1^2}})^{\frac{1}{2}n_1}(\frac{U_2^2}{\hat{\sigma_2^2}})^{\frac{1}{2}n_2}.$$

From the p-quantiles $\Lambda_{crit}(p)$ which are already computed numerically for the corresponding test for each p, the value $p : P(\Lambda < \lambda = \Lambda_{crit}(p)) = p$ is known. Therefore the confidence curve $CC(\mu)$ becomes the locus of points

$$(\mu, CC(\mu)) : \Lambda(Y, \mu) = \Lambda_{crit}(CC(Y, \mu).$$

The likelihood ratio statistic for μ is a function of the four-dimensional statistic (X_1, X_2, S_1^2, S_2^2) and cannot be expressed as a function of the three-dimensional ($X_1 - X_2, S_1^2, S_2^2$). Constructing a confidence distribution requires an additional assumption on $P(\mu < \hat{\mu})$,because this asymmetry does not influence the likelihood ratio. Assuming that $\hat{\mu}$ is the median of the confidence distribution, we get the expression for the distribution function

$$F(\mu) = \begin{cases} P(\Lambda(Y, \mu \leq CC(Y, \mu))P(\mu < \hat{\mu}) = 0.5 - 0.5CC(\mu), \mu \leq \hat{\mu} \\ 0.5 + 0.5CC(\mu), \mu > \hat{\mu} \end{cases}$$

and for confidence density

$$f(\mu) = \begin{cases} -0.5CC'(\mu), \mu < \hat{\mu} \\ 0.5CC'(\mu), \mu > \hat{\mu} \end{cases}$$

The numerical implementation of the likelihood ratio based CD (Appendix C) is tentative and needs numerical improvement.

3. By asymptotic methods

3.1. High-order approximations for the deviance

The deviance may be directly approximated as a Taylor series as described in (Schweder & Hjort, 2016, ch. 7.2). The Bartlett correction (Schweder & Hjort, 2016, ch. 7.4), using the expected deviance value corresponding to the CD, also approximates a rational function of the deviance. Another approach is based on the approximation for the deviance is the third order approximation , which can be conducted using Bédard et al. (2007) method.

3.2. High-order approximations for the quantiles of the CD

This type of tests type approximate the boundaries of the confidence regions A_p via approximation the probability for $\mu_1 - \mu_2$. In this work, we will study, in detail, the Welch-Aspin test, which is based on the approximation for $P(\mu_1 - \mu_2 \in A_p)$ as a Taylor series. However, there are other applicable possibilities, e.g. the Chernoff-Wald test, based on an iterative approximation of a more complex form.

4. Numerical adjustment

Conservative tests may be obtained from non-conservative tests by numerically computing the nominal test level as a function of the planned actual level. This is possible for at least some sets of levels $I \subset (0, 1)$, and very often for the whole $I = [0, 1]$. The confidence distributions are then obtained by inverting these adjusted conservative tests.

CHAPTER 6

Numerical results

1. Overall comparison of the confidence densities

The Figures 6.1 – 6.3 present diverse confidence densities for three different data. The true parameter value is $\mu_1 - \mu_2 = -1$ in all the three cases. All the confidence distributions differ, and their features will be discussed later. However, the likelihood ratio based condidence density clearly deviates from all the other. This peculiarity is not due to any numerical imprecision, and occurs because of applying of the four-dimensional non-sufficient statistic in its construction. That statistic does not include the difference of the means.

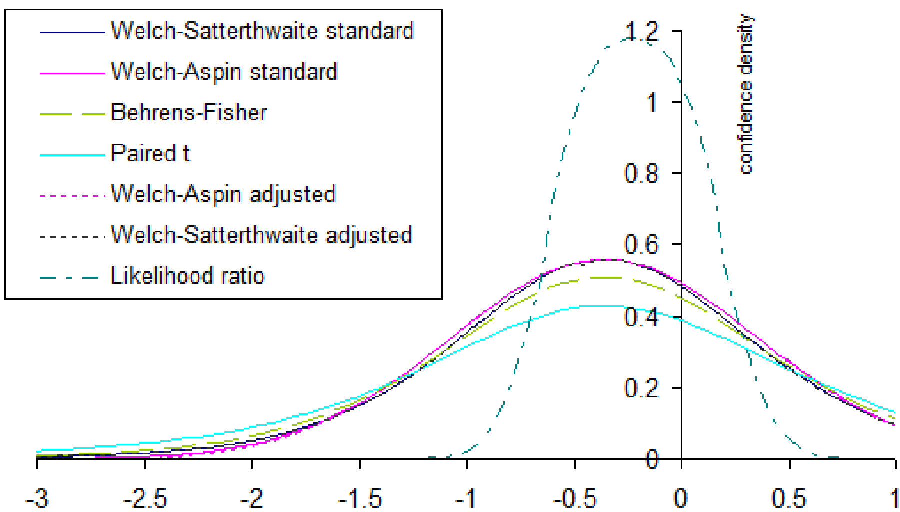

FIGURE 6.1 . The confidence densities for $Y_1 = (-0.9142985, 0.9320448, 1.0945988, -1.5417058, 0.2018343)$ and $Y_2 = (-0.6688093, 0.1468806, -1.2870124, 1.2566792, -1.0072095)$

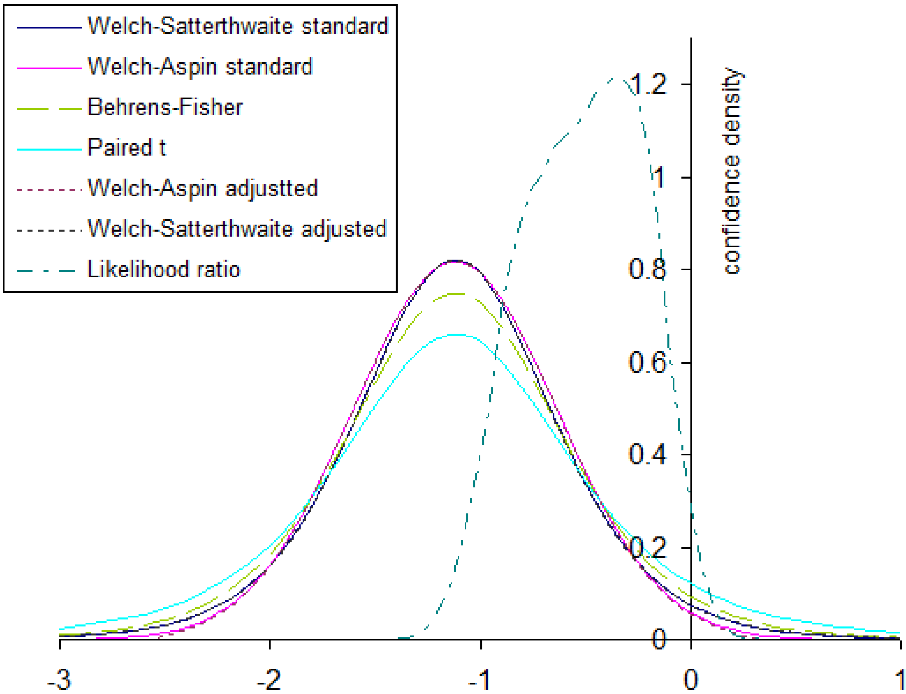

FIGURE 6.2 . The confidence densities for $Y_1 = (0.28626327, -0.09423993, 0.20402356, -1.36958796, 0.08829856)$ and $Y_2 = (-1.11996987, -2.07113205, -0.26252523, -0.07963677, -1.18116007$

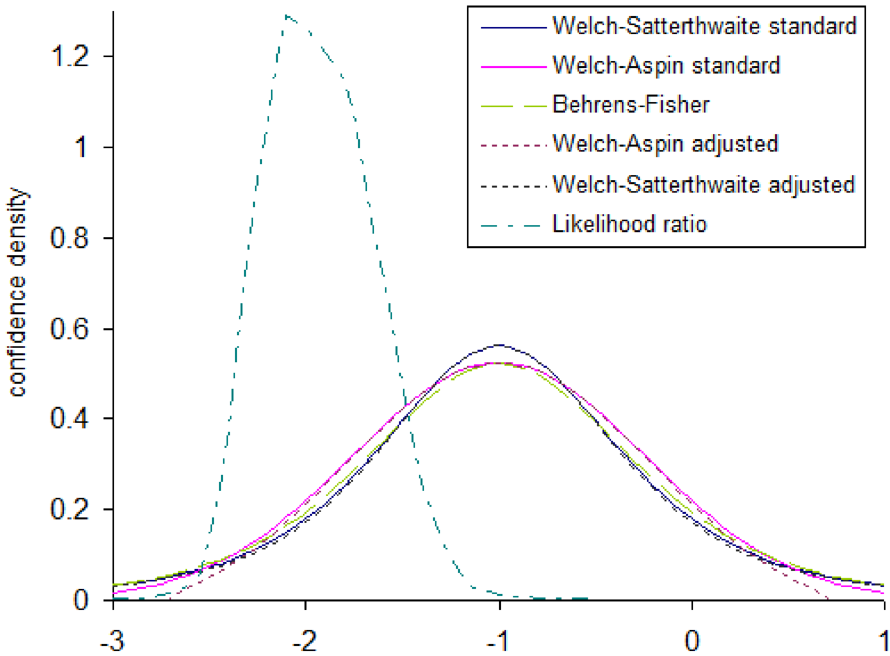

FIGURE 6.3 . The confidence densities for $Y_1 = (1.9686916, -0.02861356, 1.05992191)$ and $Y_2 = (-1.39063, -1.132441, -2.775170, -2.471090, -2.230663)$

2. Behrens-Fisher test based CDs

The Figure 6.4 presents $P(\mu_1 - \mu_2) \in A_p$ as a function of p for several values of $\frac{\sigma_1^2}{\sigma_2^2}$, when A_p is symmetric and connected. This probability is observed to exceed p for all studied values of p and parameters.

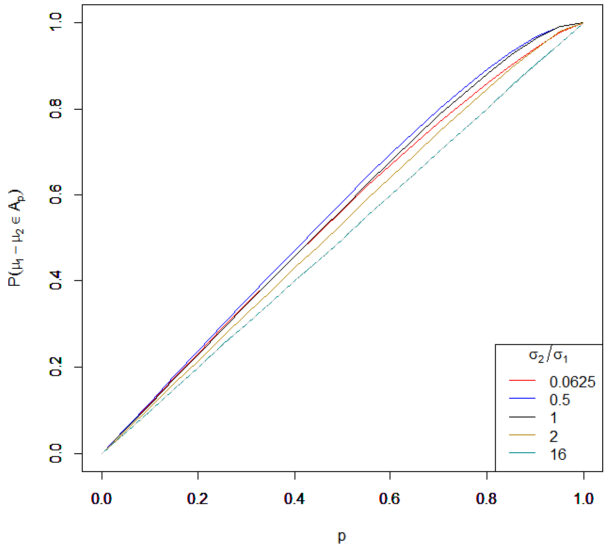

FIGURE 6.4 . Probability that μ belongs to its p-confidence set A_p when constructing the Behrens-Fisher test based CD, for sample sizes $n_1 = 5$ and $n_2 = 3$, as function of p

3. Other CDs corresponding sums of confidence distributed variables

We have studied a more weird solution of the Behrens-Fisher problem, where the scale parameters σ_1 and σ_2 are estimated not as sample standard deviations, but from the difference between the maximal and the minimal value within the sample, because they are proportional to this difference. That is, let $U_i = (X_i - \mu_i)/\sigma_i$, $V_i = \big(max_{j\in 1..n_i}(X_{i,j}) - min_{j\in 1..n_i}(X_{i,j})\big)/\sigma_i$. The Figure 6.5 presents $P(\mu_1 - \mu_2) \notin A_p$ as a function of $\frac{\sigma_1^2}{\sigma_2^2}$ when A_p is symmetric and connected, and $p = 0.95$. The Figure 6.6 presents $P(\mu_1 - \mu_2) \in A_p$ as a function of p for several values of $\frac{\sigma_1^2}{\sigma_2^2}$, when A_p is symmetric and connected. We observe that $P(\mu_1 - \mu_2) \in A_p \geq p$ for all observed p and parameters, so the conjecture 3.1 holds.

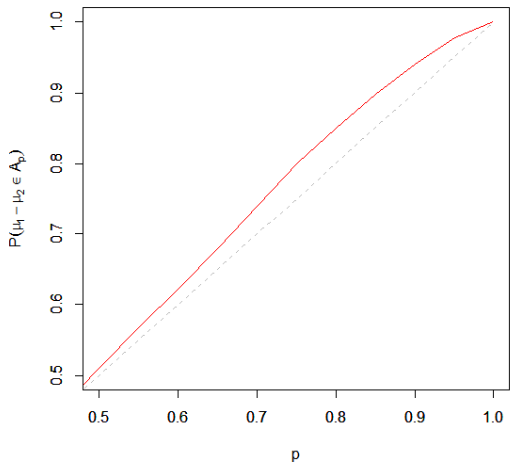

FIGURE 6.5 . Probability that μ belongs to its p-confidence set A_p when constructing the Conjecture 3.1 based CD, for sample sizes $n_1 = 5$ and $n_2 = 3$ for $\sigma_1 = \sigma_2 = 1$, as function of p, from 1000 simulations

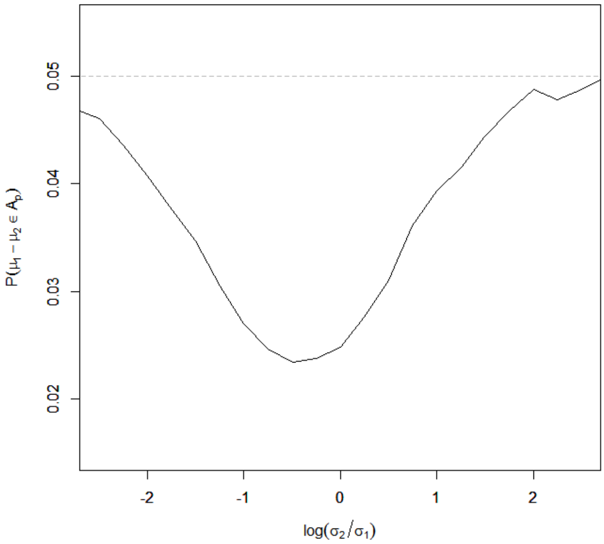

FIGURE 6.6 . Probability that μ belongs to its p-confidence set A_p when constructing the Conjecture 3.1 based CD, for sample sizes $n_1 = 5$ and $n_2 = 3$, for $p = 0.95$, as function of σ_1/σ_2, from 100000 simulations

We have also tried the conjecture to another problem, where X_{1j} is a mixture: $X_{1j} \sim 0.5N(-5\sigma_1 + \sigma_1\mu_1, \sigma_1^2) : 0.5N(5\sigma_1 + \sigma_1\mu_1, \sigma_1^2)$, $j = 1..5$ and $X_{2j} \sim N(\mu_2, \sigma_2^2)$, $j = 1..3$. We take $U_1 = (\overline{X_1} - \mu_1)/\sigma_1$, $U_2 = (\overline{X_2} - \mu_2)/\sigma_2$, $V_1 = \big(max_{j\in 1..5}(X_{i,j}) - min_{j\in 1..5}(X_{i,j})\big)/\sigma_1$, $V_2 = S_2/\sigma_2$. The figure 6.7 presents $P(\mu_1 - \mu_2) \in A_p$ as a function of p for several values of $\frac{\sigma_1^2}{\sigma_2^2}$, when A_p is symmetric and connected. We observe that $P(\mu_1 - \mu_2) \in A_p \geq p$ for all observed p and parameters as well, which also corresponds the Conjecture 3.1.

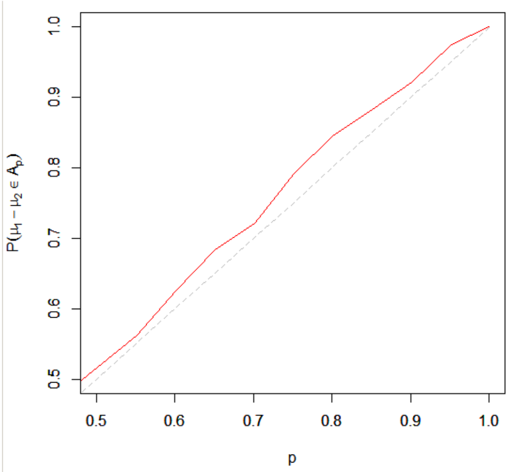

FIGURE 6.7 . Probability that μ belongs to its p-confidence set A_p when constructing the Conjecture 3.1 based CD, for sample sizes $n_1 = 5$ and $n_2 = 3$, for $p = 0.95$, as function of p,from 10000 simulations

4. Welch-Satterthwaite test based CDs

The numerical simulations reveal that the test is not conservative. This corresponds with many previous results (e.g. Duong & Shorrock, 1996; Wang, 1971).

Moreover, we reveal numerically that the Welch's approximation works worse when only one of two sample sizes increases. The probabilities to reject H_0 were studied by numerical integration

$$P(reject\,H_0) = \int_0^{+\infty}\int_0^{+\infty} P\Big(\frac{\overline{X}-\overline{Y}}{\sqrt{s_1^2/n_1+s_2^2/n_2}} > T_{f(s_1,s_2,n_1,n_2)}\Big) f_{S1}(s_1) f_{s_2}(s_2) ds_1 ds_2$$

We integrated by Simpson method

$$P(reject\,H_0) = \int_a^b\int_c^d 2\cdot\phi(qt(1-\frac{\tilde{\alpha}}{2}, \frac{s_1^2/s_2^2+n_1/n_2)^2}{\frac{s_1^4}{(s_2^4(n_1-1)}+\frac{n_1^2}{n_2^2(n_2-1)}})/\sqrt{\frac{\sigma_1^2/n_1+\sigma_2^2/n_2}{s_1^2/n_1+s_2^2/n_2}}).$$

$$\cdot\frac{n_1-1}{\sigma_1^2} f_{\chi_{n_1-1}}(s_1\frac{n_1-1}{\sigma_1^2})\frac{n_2-1}{\sigma_2^2} f_{\chi_{n_2-1}}(s_2\frac{n_2-1}{\sigma_2^2}) ds_1 ds_2, \tag{23}$$

where

- a, b, c, d are 0.0001-quantile and 0.99999-quantile of s_1 og s_2 respectively, defined as scaled quantiles og the corresponding chi-squared distributions,
- $qt(p, df)$ is a p-quantile of a Student t-distribution with df degrees of freedom,
- σ_2^2 was in fact always set to 1, because the level of the test only depends on the ratio between σ_1^2/σ_2^2.

The results of such probability computation were more precise than the Monte-Carlo simulations, as shown in Fig 6.8 .

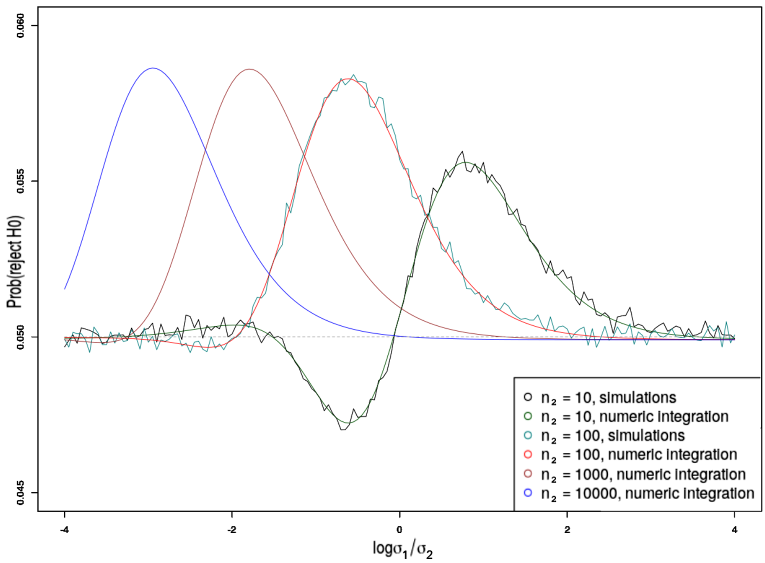

FIGURE 6.8 . The dependence of the probabilities to reject H_0 from the nuisance parameter, for sample sizes $n_1 = 5$ and several sizes n_2

The following dataset was studied when trying to discover some easy form of dependence between the parameters, the nominal and the actual level:

- $\tilde{\alpha} \in \{0001, 0.001, 0.005, 0.01, 0.02, 0.05, 0.10, 0.15, .., 0.90, 0.95\}$
- $n_1 \in \{2, 3, ..9, 10, 20, 30..90, 100, 200, ..900, 1000, 2000, ..9000, 10000\}$
- $n_2 \in \{2, 3, ..38, 39, 40, 50, 60, 70, 80, 90, 100, 200, 300, 400, 500, 1000, 2000\}$
- $\sigma_1 \in \{exp(-7), exp(-6.8), exp(-6.6), .., exp(6.8), exp(7)\}$
- $\sigma_2 = 1$

We observe, that the dependence of actual level from one of the sample sizes, when nominal level and another sample size are constant, is as following. For all sample sizes and levels, the level decreases until $n_2 = n_1$, than increases approaching asymptotically some value. This value is higher than the nominal level and depends on the nominal level and the minimal sample size. The examples of such a dependence are shown in Fig 6.9 –6.10 .

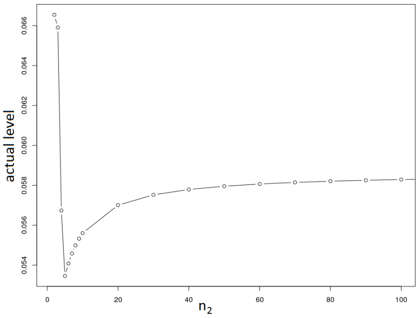

FIGURE 6.9 . The dependence of the level of the test on n_2, at nominal level of 0.2 and $n_1 = 21$

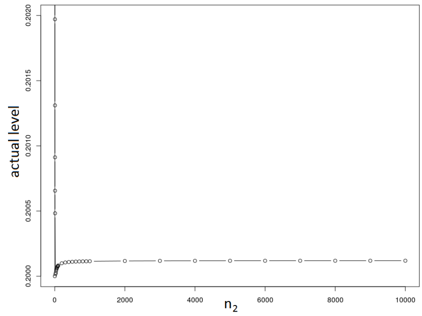

FIGURE 6.10 . The dependence of the level of the test on n_2, at nominal level of 0.05 and $n_1 = 5$

We observed that the dependence between $max(n_1, n_2)$ and the $\frac{1}{\alpha_\infty - \tilde{\alpha}}$ is close to linear, as shown in Fig 6.11 .

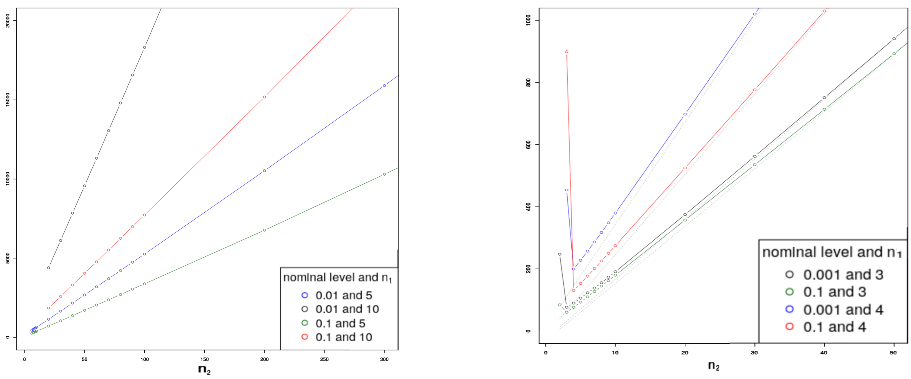

FIGURE 6.11 . Typical dependencies between n_2 and the $\frac{1}{\alpha_\infty - \alpha}$

We have not detected other simple relations between the actual level and nominal level.

The typical dependencies between $P(reject\ H_0)$ and the relation between the two unknown variances for various nominal test levels are presented in Figure 6.12 .

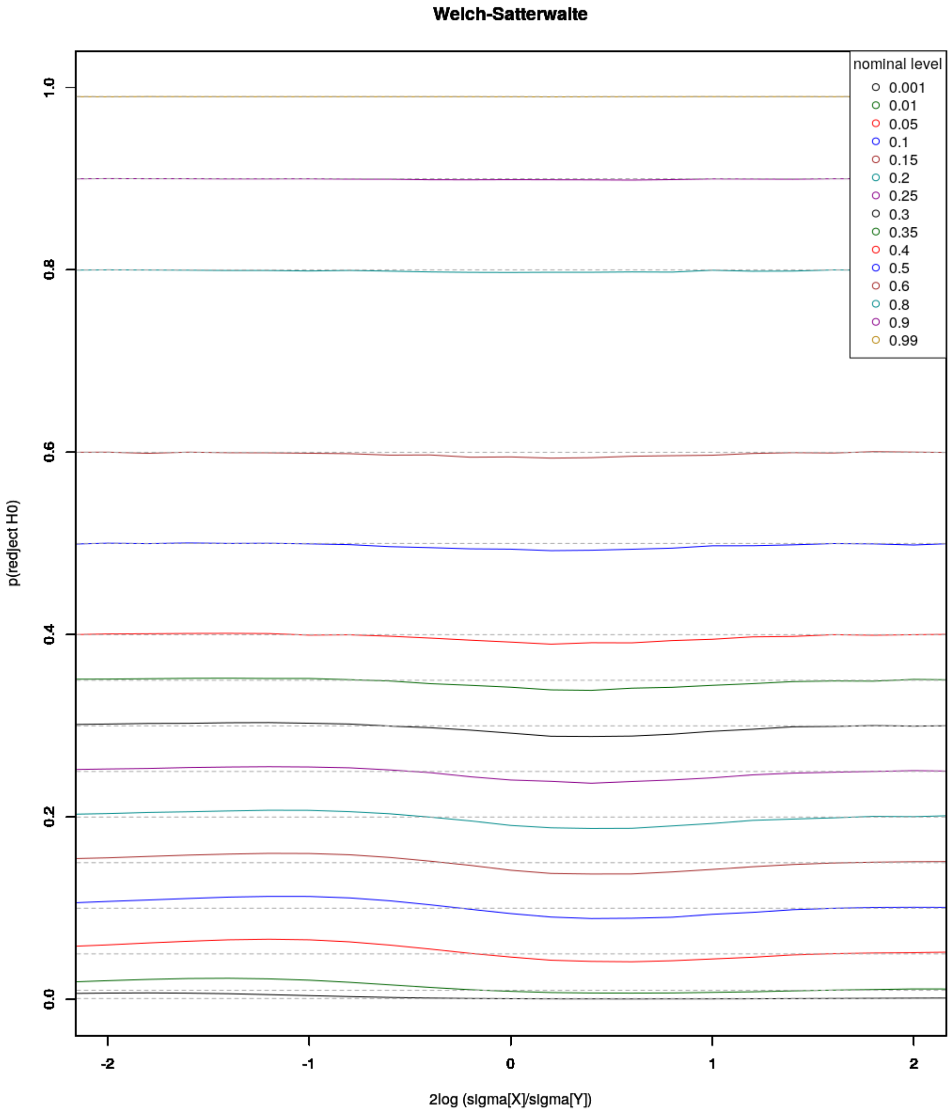

FIGURE 6.12 . Probabilities to reject H_0 by Welch-Satterthwaite test at different levels, for sample sizes $n_1 = 5$ and $n_2 = 3$

The Figure 6.13 depicts the dependence between nominal and actual level for sample sizes: $n_1 = 5, n_2 = 9$ and $n_1 = 3, n_2 = 5$. In order to use the Welch-Satterthwaite test as a conservative test of level α actually, the nominal level should be adjusted.

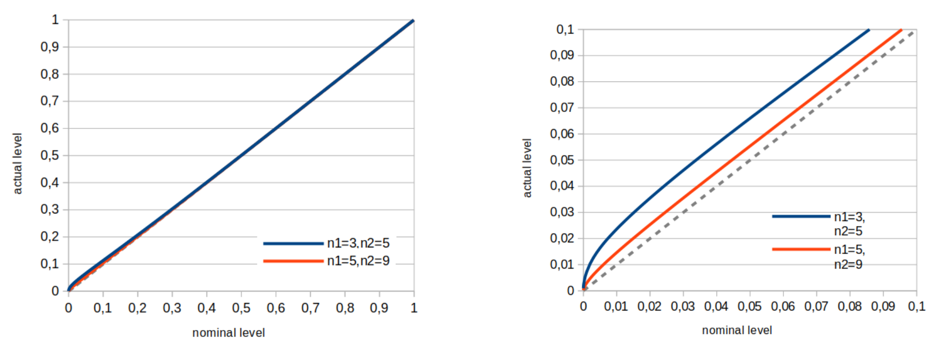

FIGURE 6.13 . The dependence between nominal and actual level of the Welch-Satterthwaite test for sample sizes: $n_1 = 5, n_2 = 9$ and $n_1 = 3, n_2 = 5$. Two scales

5. Welch-Aspin test based CDs

The Figure 6.14 depicts the dependence between nominal and actual level for sample sizes: $n_1 = 5, n_2 = 9$ and $n_1 = 3, n_2 = 5$. In order to use the Welch-Aspin test as a conservative test of level α actually, the nominal level should be adjusted, as well as for Welch-Satterthwaite test. For some levels the test cannot be conducted.

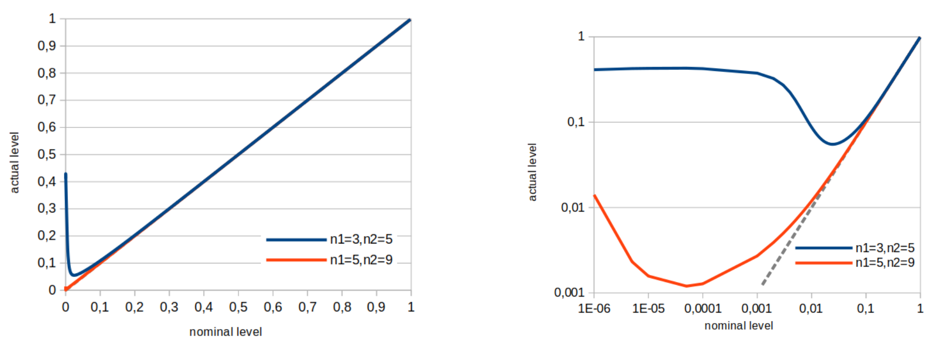

FIGURE 6.14 . The dependence between nominal and actual level the Welch-Aspin test for sample sizes: $n_1 = 5, n_2 = 9$ and $n_1 = 3, n_2 = 5$. Two scales

6. Asymptotic tests

We observed that the asymptotic methods described in (Schweder & Hjort, 2016, ch. 7) do not result in conservative tests for $H_0 : \mu_1 = \mu_2$, and the difference between the planned and the actual probability to accept is very high. We illustrate it in Figures 6.15 -6.17 for the test level 0.05, which correspond $p = 0.95$ with symmetric confidence sets A_p. Of this reason, these tests have not been used for constructing confidence distributions in this work.

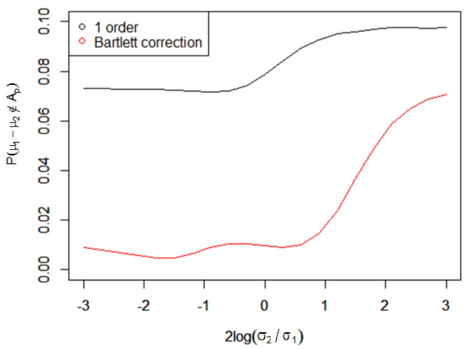

FIGURE 6.15 . The probability to reject $H_0 : \mu_1 = \mu_2$ as function of $2log\sigma_1/\sigma_2$: $n_1 = 5$, $n_2 = 9$, with nominal test level 0.05

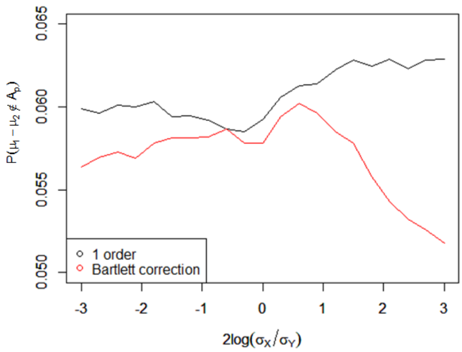

FIGURE 6.16 . The probability to reject $H_0 :'' \mu_1 = \mu_2''$ as function of $2log\sigma_1/\sigma_2$: $n_1 = 15$, $n_2 = 19$, with nominal test level 0.05

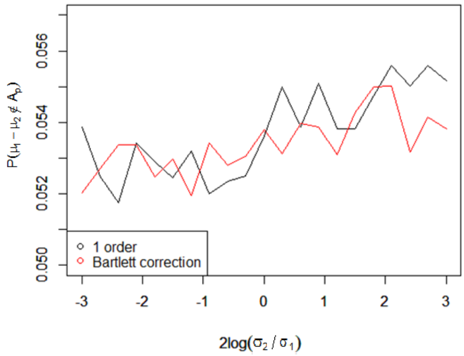

FIGURE 6.17 . The probability to reject $H_0 :'' \mu_1 = \mu_2''$ as function of $2log\sigma_1/\sigma_2$: $n_1 = 40$, $n_2 = 60$, with nominal test level 0.05

7. Preliminary computation of the loss and risk of the CDs

We illustrate in Tables 1 and 2 the confidence risk of the several sets of parameters, several types of CDs and two pairs of sample sizes. While computing these results, there were simulated 100 values of data for each parameter and each pair of sample sizes. The confidence density was calculated for the set of points $(-15, -14.995, -14.99... + 15)$, the loss was computed according to the Def. 2.32 via trapezoidal integration and the risk was estimated as the average loss. We observed that the confidence densities have very heavy tailes, especially for the Behrens-Fisher distributions, hence the estimates of the risk in the tables may be imprecise. Nevertheless, as the data were same for the various CDs, we observe the relative risk for the different CD. The Likelihood ratio and Welch-Aspin method appear to be more precise. That is probably because the boundary of the rejection region of the Welch-Aspin test is a polynom of a normally distributed variable. This variable has less heavy tailes than the scaled student t-distribution, which is the confidence distribution based on the Welch-Satthertwaite and paired t-tests). It has also less heavy tails than the linear combination of scaled student t-distributions, used in the Behrens-Fisher method. For likelihood ratio based confidence distribution, the tailes had numerically zero weight beyond some interval, but that may also be due to imprecise calculation technique.

Generally, the risk mostly increases regardless the method of its computation, because the other scale parameter increases while the first scale parameter remains constant, so the overall scale increases. However, the increase is not always monotone.

$\frac{\sigma_2^2}{\sigma_1^2}$:	0	1/16	1/4	1	4
Welch-Satterthwaite	0.61	0.57	0.73	1.09	2.61
Welch-Aspin	0.46	0.44	0.62	0.97	2.22
Behrens-Fisher	0.55	0.54	0.75	1.19	2.71
Likelihood ratio	0.36	0.34	0.47	0.77	2.05
Paired t	0.61	0.60	0.84	1.30	2.97

TABLE 1. Approximate risk for $n_1 = n_2 = 5$ with different CDs, by 500 simulations

$\frac{\sigma_2^2}{\sigma_1^2}$:	0	1/16	1/4	1	4
Welch-Satterthwaite	0.68	0.58	0.56	0.93	3.05
Welch-Aspin	0.53	0.47	0.45	0.82	2.62
Behrens-Fisher	0.62	0.56	0.58	1.01	3.12
Likelihood ratio	0.43	0.41	0.43	0.78	2.64

TABLE 2. Approximate risk for $n_1 = 5, n_2 = 3$ with different CDs, by 100 simulations

CHAPTER 7

Discussion and conclusions

Applying CDs as distribution estimators is clearly advantageous compared to applying Bayesian distribution estimators, which, currently, are more commonly used. The CD approach does not rely on any subjectively specified priors. The CDs just provide the confidence sets, only accumulating the information that is available from the data. As presented above, many CDs are applicable to the Behrens–Fisher problem. This variety is due to the use of different statistics.

Two of the most common tests, the Welch-Satterthwaite and the Welch-Aspin tests, are not conservative. To be used correctly and to construct CDs, the tests require numerical adjustment, which may be numerically slow if used for many samples.

Welch-Satterthwaite test is probably the most popular (Lillegård, 2001) test. The CD based on this test requires more adjustment for the level than the CD based on the Welch-Aspin test, but the adjusted test exists for all levels. The peculiarity of the test is that its actual level is not only higher than the planned nominal, but increases even more when the smaller sample size is fixed and another sample size increases. Therefore the test becomes less, and not more precise as the sample size increases, and this looks to be a new observation.

The CD based on the Welch-Aspin test is very close to an exact CD. Its advantage is narrower confidence sets for usual test levels (large p) than for other tests. However, for smaller p these confidence sets are wider. The peculiarity of this CD is that for very high probabilities $p > p_{crit}$ it is not defined. The value of p_{crit} depends on the sample sizes and is very close to 1. It has also been observed that for sample sizes $n_1 = 3, n_2 = 5$ and $n_1 = n_2 = 5$ this CD has very low quadratic risk.

The CD based on the likelihood ratio test is conservative by construction, without any numerical adjustment. However, it requires numerically difficult computation of the likelihood ratio quantiles for all sample sizes. Unlike other studied CDs, it is asymmetric. The quadratic risk connected with this distribution is the lowest among the studied alternatives. That is probably because this CD has a sharper peak and becomes nearly zero beyond some interval. However, this peak and this interval are often wrong, and, with another penalty function, the risk would likely be higher than that of alternatives.

The CD based on the paired t-test is exact, but it is only applicable when $n_1 = n_2$. Its risk is observed to be highest, the related confidence sets are typically rather wide.

The CD based on the Behrens-Fisher test is easier to construct. The test is conservative, so the distribution does not require numerical adjustment. However, the risk for this CD is rather large. Studying this CD motivated us to pursue a more general study of linear combinations of confidence variables. We have found an example in which the linear combination of the confidence variables for the parameters is not a CD for the linear combination of these parameters. We have also found other examples in which such a CD does result from the linear combination of the confidence variables.

Via the results listed above, this work makes progress toward solving a very general problem of great theoretical and practical importance. The further research will possibly continue in a more general way the study of linear combinations of confidence variables and the confidence distributions for linear combinations of parameters.

APPENDIX A

Preliminary proofs for the lemmas in the Chapter 3

1. Proof of the Lemma 3.2

Assume $\mu_1 = \mu_2$. Then $T_1 s_1 + T_2 s_2$ is distributed as a mixture of $T_1 s_1 + \sigma_2$ and $T_1 s_1 + \sigma_2$ with equal weights. We observe that t_α satisfies

$$0.5\big(F_{T_1 s_1 + \sigma_2}(t_\alpha) + (F_{T_1 s_1 - \sigma_2}(t_\alpha)\big) = 1 - \alpha$$

Hence by location transformation

$$0.5(F_{T_1 s_1}(t_\alpha - \sigma_2) + (F_{T_1 s_1}(t_\alpha + \sigma_2)) = 1 - \alpha.$$

Besides, $\forall s > 0$, for all random variables holds the elementary equality $X : P(X \le t_\alpha) = P(X/s \le t_\alpha/s)$. Of these two reasons, we have

$$0.5(F_{T_1}(\frac{t_\alpha + \sigma_2}{s_1}) + (F_{T_1}(\frac{t_\alpha - \sigma_2}{s_1})) = 1 - \alpha$$

We introduce α' and α'' such that the corresponding quantiles of T equal: $t_{\alpha'} = \frac{t_\alpha + \sigma_2}{s_1}$, $t_{\alpha''} = \frac{t_\alpha - \sigma_2}{s_1}$. Then $0.5(\alpha' + \alpha'') = \alpha$ From the other hand,

$$P(Y_1 - \mu_1 - (Y_2 - \mu_2) \le t_\alpha) =$$

$$P(Y_1 - \mu_1 - U_2 \le t_\alpha) = 0.5\big(P(Y_1 - \mu_1 - \sigma_2 \le t_\alpha) + P(Y_1 - \mu_1 + \sigma_2 \le t_\alpha)\big) =$$

$$= 0.5\big(P(\frac{Y_1 - \mu_1}{s_1} \le \frac{t_\alpha + \sigma_2}{s_1}) + P(\frac{Y_1 - \mu_1}{s_1} \le \frac{t_\alpha - \sigma_2}{s_1}) = 0.5(\alpha' + \alpha'') = \alpha$$

□

2. Proof of the Lemma 3.3

Consider at first some elementary properties of $T = s_1 T_1 + s_2 T_2 = \sigma_1 U_1 + s_2 T_2$.

(1) The probability density of T is

$$f_T(t|s_2 = \sigma_2 b_m) = \frac{\sum_{i=1}^n I(t \in B_i) + I(t \in -B_i)}{2n\sigma_1}.$$

We treat $f_T(t|s_2 = \sigma_2 b_m)$ as a mixture of $2n$ uniform components. We denote the supports of these components as $B_i = (\sigma_2 \frac{b_m}{b_i} - \sigma_1, \sigma_2 \frac{b_m}{b_i} + \sigma_1)$, $-B_i = (-\sigma_2 \frac{b_m}{b_i} - \sigma_1, -\sigma_2 \frac{b_m}{b_i} + \sigma_1)$.

(2) When m is constant, the right boundaries $(\frac{b_m}{b_i}+\sigma_1)$ of the sets B_i decrease as a function of i, as well as the left boundaries.
Follows from 1.

(3) As m increases, the $f_T(t_\alpha|s_2=\sigma_2 b_m)$ decreases or remains constant.
We well at first show that $\sum_{i=1}^{n} I(t\in B_i)$ decreases or remains constant. Indeed, whenever $b_m > b_{m-1}$ $\frac{b_m}{b_i} > \frac{b_{m-1}}{b_i}$, i.e. for $s_2=\sigma_2 b_m$ all the intervals B_i are shifted more with respect to $B_{min\{i:\, t_\alpha\in B_i\}}$. Hence if $\int_{t_\alpha}^{\infty} f_T(t|s_2=\sigma_2 b_{m-1}) = \alpha > 0.5$, then $\int_{t_\alpha}^{\infty} f_T(t|s_2=\sigma_2 b_m) > \alpha$. Hence the shift of $t_\alpha\in B_i|s_2=\sigma_2 b_m$ with respect of edges of all sets B_i is greater for m than for $m+1$, i.e.

$$\sigma_2\frac{b_m}{b_i} - t_\alpha|s_2=\sigma_2 b_m > \sigma_2\frac{b_{m-1}}{b_i} - t_\alpha|s_2=\sigma_2 b_{m-1}.$$

By symmetry between $-B_i$ and B_i, and because the centres of all the intervals B_i are positive, $f_T(t_\alpha|s_2=\sigma_2 b_m) = \frac{\sum_{i=1}^{n} I(t_\alpha\in B_i)+I(t_\alpha\in -B_i)}{2n}$ also decreases or remains constant.

(4) As m increases, the $min\{i:\ t_\alpha\in B_i\}$ increases or remain constant - because of the shift in (3).

(5) By (3)., number of sets $B_i|s_2=\sigma_2 b_m$ such that $t_\alpha\in B_i$ may with the change of the number m of value of s_2: 5.1) remain constant k for $m_1..m_k$, be 0 otherwise; 5.2) be $\geq k$ for $m_1..m_k$, 0 otherwise, 5.3) exceed k for $m_1..m_l$, be $\leq k$ $m_{l+1}..m_k$, 0 otherwise.

Consider 5.1. This case, in view of (1) – (4) items above, is for $k>0$ schematically depicted in Fig. 1.1 .

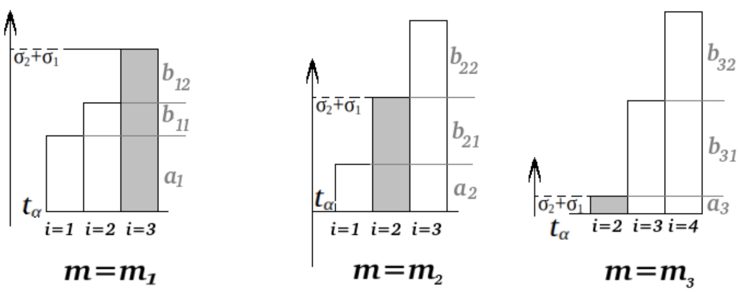

FIGURE 1.1 . The schematic example of the sets $B_i\cup\big((t-\alpha|s_2=\sigma_2 m),+\infty\big)$

We denote lengthes b_{ij} and a_j Fig. 1.1 . We take into account the fact that t_α is an $\alpha-$quantile of T and that the probability that $B=b_i$ equals for all i. We express $P(T_1 s_1 + T_2 s_2)$ as area of all the blocks placed above the t_α in Fig. 1.1 . Hence we observe that there is a proportionality constant C such that $\forall i$: $1\leq i\leq k$ $C(ka_i+\sum_{j=1}^{k-1}(k-j)b_{ij})=\alpha$, and from 2. and 4. it follows that

$\forall i: 1 < i \leq k, \forall j: 1 \leq u < k$ holds $b_{ij} \leq b_{(i+1),j}$.

$$P\big(Y_1 - Y_2 - (\mu_1 - \mu_2) > T_\alpha\big) = C(\sum_{i=1}^{k}(a_k + \sum_{j=1}^{k-1} b_{ij}) + \sum_{l:\,\forall x \in B_l t_\alpha < x} \frac{1}{n})$$

and the latter equation implies that $P(y_1 - y_2 - \mu_1 + \mu_2 > t_\alpha) \leq \alpha$.
For $k = 0$, Consider 5.2. For almost every α it holds that

$$\frac{dP\big(Y_1 - Y_2 - (\mu_1 - \mu_2) > T_\alpha\big) - \alpha}{d\alpha}(\alpha) =$$

$$= \lim_{\varepsilon \to 0} \frac{\sum_{i=1}^{k} \frac{1}{n} f_{Y_1 - Y_2 - (\mu_1 - \mu_2)}(t_\alpha | s_2 = \sigma_2 b_m) \frac{\varepsilon}{f_T(t_\alpha | s_2 = \sigma_2 b_m)}}{\varepsilon / f_T(t_\alpha | s_2 = \sigma_2 b_m)} - 1 =$$

$$= \lim_{\varepsilon \to 0} \frac{\sum_{i=1}^{k} \frac{1}{n} f_{Y_1 - Y_2 - (\mu_1 - \mu_2)}(t_\alpha | s_2 = \sigma_2 b_m) \frac{\varepsilon}{f_{Y_1 - Y_2 - (\mu_1 - \mu_2)}(t_\alpha | s_2 = \sigma_2 b_m) \cdot k/n}}{\varepsilon / f_{Y_1 - Y_2 - (\mu_1 - \mu_2)}(t_\alpha | s_2 = \sigma_2 b_m)} - 1 \leq 0$$

Consider 5.3. $\frac{dP\left(Y_1 - Y_2 - (\mu_1 - \mu_2) > T_\alpha\right) - \alpha}{d\alpha}(\alpha)$ may be positive, zero or negative. As the changing α leads to the change of the case 5.3 to 5.2 through the case 5.1 and by continuity of the $P\big(Y_1 - Y_2 - (\mu_1 - \mu_2) > T_\alpha\big)$ as a function of α, also in this case

$$P\big(Y_1 - Y_2 - (\mu_1 - \mu_2) > T_\alpha\big) \leq \alpha.$$

3. Proof of the Lemma 3.4

We split $\mathbb{R}$ into k intervals $I_k = (a_k, b_k)$ where $f_U(u)$ is continuous. Denote the distribution function of U as F. For $2 \leq i \leq k-1$ split $(F(a_k), F(b_k))$ into m equal parts, the boundaries of which we denote c_{im} for $i = 1..k$. We choose a decreasing sequence $(c_{1n}) \to 0 : 0 < c_{1n} < F(b_1) \forall n > 1$, $c_{11} = b_1, c_{11} - c_{12} = 1/m$ and a sequence of values symmetric with respect to the point $1/2$ to (c_{1n}).
All the points $c_{ij} : c_{ij} > 0.5$ are elements of an increasing sequence (c_n) The non-decreasing function F_a having a graph connecting points $F^{-1}(0.5 - c_n), 0.5 - c_n$ and points $F^{-1}(c_n), F^{-1}(c_n)$ is also a distribution function. Due to the symmetry, piecewise linearity and the non-decreasing derivative of f $(-\infty, 0)$, it may be expressed as a convex combination of distribution functions of $Unif(F^{-1}(0.5 - c_n), c_{it})$. The weights are weights $w_1 = 2(c_1 - 0.5)$; $w_i = 2(c_i - 0.5) - w_{i-1} : i > 1$. By construction, $max(|F_a(x) - F(x)| \leq 1/n)$, hence the mixture converges to U in distribution.

4. Proof of the Lemma 3.5

Let t'_α be an (unknown) $(1-\alpha)$-quantile of $s_1 T_1 + (Y_2 - \mu_2)$ and t_α be a $1-\alpha$-quantile of $s_1 T_1 + s_2 T_2$ For any value of S_1, we approximate the $s_1 T_1$ as a mixture of symmetric uniform distributions $Unif(-a_1, a_1)$, .. $Unif(-a_\tau, a_\tau)$, ...$Unif(-a_n, a_n)$. In general, $P(Unif(-a_\tau, a_\tau) < t_\alpha) = p_\tau$ differs for all τ, and

$$P(s_1 T_1 + s_2 T_2 < t_\alpha) = \sum_{i=1}^{n} w_i p_\tau$$

By lemma 3.3, for all values of s_1 and each τ,

$$P(s_1T_1 - (Y_2 - \mu_2) < t_\alpha | s_1, \tau) \geq 1 - \alpha$$

we have $P(s_1T_1 - (Y_2 - \mu_2) < t_\alpha) \geq 1 - \alpha$ and $t_\alpha > t'_\alpha$ for all α. By the choice of α, $P(Y_1 - Y_2 - (\mu_1 - \mu_2) < t'_\alpha) = \alpha$ So the test is of level α.

5. Proof of the Lemma 3.6

We approximate the denominator by a simple variable, letting $V_2 \sim Discrete\big(P(V_2 = b_1) = w_1, P(V_2 = b_2) = w_2 = 1 - w_1\big)$. We also approximate the nominator in $\frac{U_2}{V_2}$ as a simple variable. We will proceed by induction, adding pairs of values which are closer and closer to zero. For two-valued U_2 it is proven by Lemma 3.5.
Assume now that the lemma holds for any simple variable U taking $2n$ values, with all positive weights and also the values $b_2....b_{n+1}$. We can construct a new simple variable taking $2(n+1)$ values by compressing one pair of old values $k > 1$ times. More specifically, we introduce $Y_2^{old} \sim w_1 Unif\{-b_1\sigma_2, b_1\sigma_2\} + \mu_2 {:} (1 - w_1)(U + \mu_2)$ and consider the s_1T_1 as a mixture of symmetric uniform distributions $Unif(-a_1, a_1)$, .. $Unif(-a_\tau, a_\tau)$, ..., as it was in the lemma 3.5. We introduce $k > 1$ and $Y_2^{new} \sim w_0 Y_2^{old} : (1 - w_0)(Unif\{-b_1\sigma_2/k, b_1\sigma_2/k\} + \mu_2)$.

We add $\sigma_2 b_1 k$,..$\sigma_2 b_i k$,.. $\sigma_2 b_n k$ in the denominator of $\frac{U_2}{V_2}$, so that $P(S_2^{old} = \sigma_2 b_1 = w_1 w_0$, ...$P(S_2^{old} = \sigma_2 b_n = w_n w_0$, $P(S_2^{old} = \sigma_2 b_1 k = w_1(1 - w_0)$,... $P(S_2^{old} = \sigma_2 b_n k = w_n(1 - w_0)$. The pivot $\frac{Y_2^{old} - \mu_2}{S_2^{old}}$ is same as $\frac{Y_2^{new} - \mu_2}{S_2}$. Let the mixture in the denominator of $\frac{U_2}{V_2}$ change to the mixture in the nominator, i.e. (Y_2^{old}, S_2^{old}) change to (Y_2^{new}, S_2). Then those components of $(s_1T_1 + s_2T_2)$-mixture $[a_\tau, k\sigma_2\frac{b_i}{b_j}]$ which corresponded to $s_1T_1 \sim (-a_\tau, a_\tau)$ and $s_2 = b_i k$, change to just $[a_t, \sigma_2\frac{b_i}{b_j}]$. Denote $x_{\tau,i,j}$ the p-quantile of the $(s_1T_1 + s_2T_2)$-mixture including the component $Unif(-a_\tau, a_\tau) + Unif\{-k\sigma_2\frac{b_i}{b_j}, k\sigma_2\frac{b_i}{b_j}\}$, we have

$$P^{old}(Y_1 - Y_2 - (\mu_1 - \mu_2) < t_\alpha | \tau, i) = P(|Y_1 - Y_2^{old}| < x_{tij}) = p,$$

and

$$P^{new}(Y_1 - Y_2 - (\mu_1 - \mu_2) < t_\alpha | \tau, i) = w_0 P_i^{new}(accept H_0 | \tau, i) + (1 - w_0) P(Y_1 - Y_2^{old}/k < x_{tij}) > p$$

It holds for all components (i, τ), so $P^{new}\big(Y_1 - Y_2 - (\mu_1 - \mu_2) < t_\alpha | \tau, i\big) > P^{old}\big(Y_1 - Y_2 - (\mu_1 - \mu_2) < t_\alpha | \tau, i\big)$.

APPENDIX B

Level of the Behrens-Fisher test with $n_1 = n_2 = 2$ and $n_1 = n_2 = 3$

In this chapter we prove that Behrens-Fisher test is conservative in case of samples with equal sample sizes of $n_1 = n_2 = 2$ and with equal sample sizes of $n_1 = n_2 = 3$.

1. The test with a constant critical value

LEMMA B.1. $b^2 = \frac{z^2(\frac{e^\tau}{n_1}+\frac{1}{n_2})}{\frac{e^\tau v_1}{n_1(n_1-1)}+\frac{v_2}{n_2(n_2-1)}}$ *as a function of* τ *decreases for all* τ *if* $\frac{v_1}{n_1-1} > \frac{v_2}{n_2-1}$*, increases for all* τ *otherwise, has two horizontal asymptotes, and its graph has a centre of symmetry*

$$(\tau_0, b^2_{\tau_0}) = (ln\Big(\frac{v_2 n_1(n_1-1)}{v_1 n_2(n_2-1)}\Big), z^2\frac{1}{2}(\frac{(n_1-1)}{v_1}+\frac{(n_2-1)}{v_2}))$$

.

PROOF. A function of type $f(t) = \frac{ae^t+b}{ce^t+d}$ where $c>0, d>0$, can be expressed as

$$\frac{\frac{a}{d}\frac{d}{c}exp(ln\frac{c}{d}+t)+\frac{b}{d}}{exp(ln\frac{c}{d}+t)+1} = \frac{1}{2}(\frac{a}{c}-\frac{b}{d})\frac{exp(ln\frac{c}{d}+t)-1}{exp(ln\frac{c}{d}+t)+1}+\frac{1}{2}(\frac{a}{c}+\frac{b}{d}),$$

which is $\frac{1}{2}(\frac{a}{c}-\frac{b}{d})$-times scaling, vertical shift by $\frac{1}{2}(\frac{a}{c}+\frac{b}{d})$ and horizontal shift to the right by $x_0 = ln\frac{d}{c}$ of an odd function $g(t) = \frac{e^t-1}{e^t+1}$ that has two horizontal asymptotes $y=-1$ and $y=1$ and increases for all t. We assign: $a = z^2/n_1$, $b = z^2/n_2$, $c = v_1/(n_1(n_1-1))$, $d = v_2/(n_2(n_2-1))$ and get the statement of lemma. □

THEOREM B.1. *If* $n_1 = n_2$*, the level of the test with* $H_0 : \mu_1 = \mu_2$ *and the critical region*

$$|B| > a, \tag{24}$$

equals $P(|T_{n_1-1}| > a)$*.*

PROOF. We will study how the probability to reject H_0 changes with τ. In order to do this, we will study how $I(|B| > a)$ changes k the H_0 will be rejected for any single sample.
We denote with b^2 the value of squared Behrens-Fisher statistic characterising the single sample. Clearly,

$$P(|B| > a) = P(B^2 > a^2)$$

We denote z, v_1, v_2 for the values of Z, V_1, V_2 corresponding to this sample, in notation of Statement 1. The value of b^2 is a function of k and hence a function of τ:

$$b^2 = \frac{z^2(\frac{e^\tau}{n_1} + \frac{1}{n_2})}{\frac{e^\tau v_1}{n_1(n_1-1)} + \frac{v_2}{n_2(n_2-1)}}$$

To any such sample with $\tau = 0$ we will assign a curve $\Gamma(f(\tau))$: a graph Γ of a squared Behrens-Fisher statistic as a function f of τ.

We apply Lemma B.1 to the case $n_1 = n_2 = n$ and see that b^2 increases if and only if $v_2 > v_1$, and that happens if and only if $\tau_0 = ln\left(\frac{v_2 n_1(n_1-1)}{v_1(n_2(n_2-1)}\right) = ln\left(\frac{v_2}{v_1}\right) > 0$, i.e. if and only if τ-coordinate of the symmetry centre of the curves is positive.

Consider curves $\Gamma(f(\tau))$ that intersect horizontal line $b^2 = a^2$, that is, curves for which the acceptance decision depends on k. We denote the τ-coordinate of the point of intersection τ_r and the τ-coordinate of the centre of symmetry as τ_0.
Any curve that has the opposite sign of τ_r and τ_0, belongs to a family of four curves having the same $|\tau_0|$ and the same $|\tau_r|$. The family is shown in Figure 2.1 . If the curve is generated by a sample (z_1, v_1, v_2), then the four-curve family includes:

(1) the initial curve from (z_1, v_1, v_2),
(2) a curve from (z_1, v_2, v_1) (mirror reflection of initial curve),
(3) a curve from (z_2, v_1, v_2), where $|(z_2)|^2 = \frac{a}{f(-\tau_r)}|(z_1)|^2 < |(z_1)|^2$ and $sign(z_2) = sign(z_1)$,
(4) a curve from (z_2, v_2, v_1).

As $|z_2| < |z_1|$, the standard normal probability density at z_2 is larger than at z_1. So the joint probability density

$$f(z_1, v_1, v_2) < f(z_2, v_1, v_2)$$

and due to equality of sample sizes

$$f(z_1, v_1, v_2) = f(z_1, v_2, v_1),\ f(z_2, v_1, v_2) = f(z_2, v_2, v_1).$$

So if a curve with opposite sign of τ_r and τ_0 belongs to a specific 4-curves-family, then, due to the placement of τ_0 towards the direction of increase of f:

$$p(Reject\ H_0) = \begin{cases} 1, |\tau| > |\tau_r| \\ p : 0 < p < 1, |\tau| \le |\tau_r| \end{cases}$$

There are also other curves that intersect the horizontal line $b^2 = a^2$ and have same sign of τ_r and τ_0, but do not belong to such 4-curves-families. These curves we group into families of only two symmetric curves. That is, if a curve is generated by a sample (z, v_1, v_2), the 2-curves-family includes:

(1) the initial curve from (z, v_1, v_2)
(2) a curve from (z, v_2, v_1)

Due to monotonicity of f and the placement of τ_0, all these families give rejection of H_0 for $\{\tau : |\tau| > |\tau_r|\}$, i.e.

$$p(Reject\ H_0) = \begin{cases} 1, |\tau| > |\tau_r| \\ 0, |\tau| \leq |\tau_r| \end{cases}$$

We will also group all the curves that do not intersect the horizontal line $b^2 = a^2$ into 1-curve-families, consisting of a single curve. If a curve belongs to a specific 1-curve-family, than the probability to reject H_0 is constant with respect to τ (taking values 1 or 0).

We will call the sum of the probability densities of the samples generating the curves of the specific family for "the probability density of a curve family $f(family)$" (in a discretized approximation, that is exactly a probability to sample the specific family) . The probability to reject H_0 as a function of τ

$$P(reject\ H_0)(\tau) = \int_{\Omega(Z,V_1,V_2)} P(reject\ H_0|\tau, z, v_1, v_2) f(z, v_1, v_2) dz dv_1 dv_2 =$$

$$= \int_{all\ families} P(reject\ H_0|family)(\tau) f(family(z, v_1, v_2)) dz dv_1 dv_2$$

We see that for all families, $P(reject\ H_0|family)(\tau)$ is an even function of τ with a minimum at 0, decreases non-strictly at $(-\infty, \tau)$ and increases at $(\tau, \infty,)$. Consequently, so is $P(reject\ H_0)(\tau)$.

Finally, we observe that when $|\tau| \to \infty$, the B converges in distribution to T_{n_1-1}. Hence

$$lim_{\tau\to\infty} P(reject\ H_0)(\tau) = P(|T_{n_1-1}| > a)$$

□

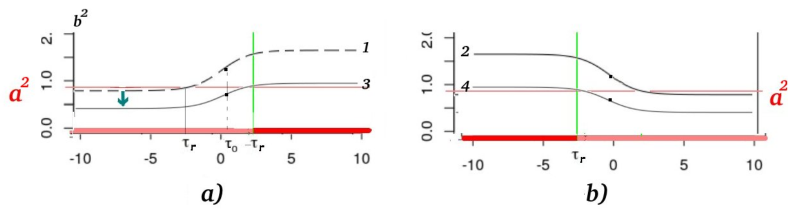

FIGURE 2.1 . The graphs of a squared Behrens-Fisher statistic as functions τ, belonging to the *'4-curves'*-family.

2. The quantiles of $T_1 cos\theta + T_2 sin\theta$

LEMMA B.2. *Let $\theta \in (0, \pi/2)$, $\eta > 0$, S_s be the set of values of S, satisfying the system of inequalities*

$$\begin{cases} \frac{sin^2\theta}{s} + \frac{1-sin^2\theta}{1-s} < \eta \\ 0 < s < 1 \end{cases} \tag{25}$$

(1) *S_s is either a connected open interval or an empty set.*
(2) *For any $\theta \in (0, \pi/2)$ there exist a real η_0 such that the S_s is non-empty for all $\eta > \eta_0$.*
(3) *If S_s is non-empty, then its Lebesgue measure increases strictly with $|\pi/4 - \theta|$.*
(4) *If S_s is non-empty and s_0 is its the centre, then $|0.5 - s_0|$ increases with $|\pi/4 - \theta|$.*

Proof. 1. We denote $u = sin^2\theta$. In this notation, Eq. (25) becomes

$$\begin{cases} \frac{u}{s} + \frac{1-u}{1-s} < \eta \\ 0 < s < 1 \\ 0 < u < 1 \end{cases}$$

that is equivalent to

$$\begin{cases} s \in (\frac{-\sqrt{4u^2-4u+(\eta-1)^2}+2u+\eta-1}{2\eta}, \frac{\sqrt{4u^2-4u+(\eta-1)^2}+2u+\eta-1}{2\eta}),\ if\ 4u^2-4u+(\eta-1)^2 > 0 \\ s \in \emptyset,\ if\ 4u^2-4u+(\eta-1)^2+2u+\eta-1 \leq 0 \\ 0 < s < 1, \\ 0 < u < 1, \end{cases} \tag{26}$$

2. We choose $\eta_0 = 2max(sin^2\theta, 1 - sin^2\theta) + 1$. Then for $s = 1/2$ Eq. (25) holds for all $\eta > \eta_0$.
3. We denote $t = 2u - 1 = 2sin^2\theta - 1$. Eq. (26) becomes the same as

$$\begin{cases} s \in (\frac{-\sqrt{t^2+\eta^2-2\eta}+t+\eta}{2\eta}, \frac{\sqrt{t^2+\eta^2-2\eta}+t+\eta}{2\eta}) \cap (0,1)\ if\ t^2+\eta^2-2\eta \geq 0, \\ s \in \emptyset\ if\ t^2+\eta^2-2\eta < 0, \\ -1 < t < 1. \end{cases} \tag{27}$$

In fact, the boundary of S_s includes neither $\{0\}$ nor $\{1\}$, because Eq. (25) holds for neither $s < \frac{sin^2\theta}{\eta}$ nor $s > 1 - \frac{1-sin^2\theta}{\eta}$. So S_s satisfies just

$$\begin{cases} s \in (\frac{-\sqrt{t^2+\eta^2-2\eta}+t+\eta}{2\eta}, \frac{\sqrt{t^2+\eta^2-2\eta}+t+\eta}{2\eta})\ if\ t^2+\eta^2-2\eta \geq 0, \\ s \in \emptyset\ if\ t^2+\eta^2-2\eta < 0, \\ -1 < t < 1. \end{cases} \tag{28}$$

If S_s is not empty, both boundaries of the interval containing the solution,S_s change with t: the lower boundary decreases and the upper boundary increases. So the length of S_s increases with $|t|$, and the increase with $|t|$ corresponds to increase with $|2sin^2\theta - 1|$. As we defined $\theta \in (0, pi/2)$, the latter means increase with $|\pi/4 - \theta|$.
4. According to Eq.(28), the centre of S_s is $(\frac{-\sqrt{t^2+\eta^2-2\eta}+t+\eta}{2\eta}, \frac{\sqrt{t^2+\eta^2-2\eta}+t+\eta}{2\eta})$, which is $\frac{t}{2\eta} + 0.5$. Its distance from 0.5 clearly increases with $|t|$ and hence with $|\pi/4 - \theta|$. □

THEOREM B.2. *For any $a > 0$ and T_1, T_2 being independent student variables with equal degrees of freedom $f_1 = f_2 = f \in \{1, 2\}$, $P(|T_1 cos\theta + T_2 sin\theta| < a)$ decreases in the range $\theta \in (0, \pi/4)$ and than increases.*

REMARK B.3. That doesn't hold for all degrees of freedom; the numerical counterexample was found for $f = 6$ and a being 0.975-quantile of T_6 distribution

PROOF. As shown by (Ruben, 1960),

$$T_{f1} cos\theta + T_{f2} sin\theta \sim \frac{T_{f_1+f_2}}{\psi(S)},$$

where $S \sim Beta(f_1/2, f_2/2)$, T_{f_1} is independent on T_{f_2}, degrees of freedom respectively, $T_{f_1+f_2}$ is independent on S, and

$$\psi(S) = \sqrt{\frac{(f_1 + f_2)s(1 - s)}{f_1(1 - s)sin^2\theta + f_2 s cos^2\theta}}.$$

We observe that for $f_1 = f_2 = f$

$$P(\frac{|T_{f_1+f_2}|}{\psi(S)} < a) = P(\frac{T^2_{f_1+f_2}}{\psi(S)^2} < a^2) = P\Big(\frac{T^2_{f_1+f_2}}{2}(\frac{sin^2\theta}{s} + \frac{1 - sin^2\theta}{1 - s}) < a^2\Big)$$

$$= P\Big(\frac{sin^2\theta}{s} + \frac{1 - sin^2\theta}{1 - s} < \frac{2a^2}{T^2_{f_1+f_2}}\Big) \tag{29}$$

From Lemma B.2,(2) it follows that for any a there exists a set of values of $|T_{f_1+f_2}|$ of non-zero measure such that Eq. (29) holds for some values of S. From Lemma B.2, (1) it follows that the set of values of $(|T_{f_1+f_2}|, S)$ such that Eq. (29) holds, has also non-zero measure. From Lemma B.2, (3) it follows that for any fixed $|t_{f_1+f_2}|$ such that for some values of S Eq. (29) holds, the measure of this set of values of S increases $|\pi/4 - \theta|$.

In case $\mathbf{f = 2}$, S is distributed uniformly between 0 and 1.
The increase in the measure of this set of values of S for any $|t_{f_1+f_2}|$ and the independence between S and $|T_{f_1+f_2}|$ lead to an increase of probability that Eq. 29 holds with $|\pi/4 - \theta|$.

In case $\mathbf{f = 1}$, the density for S has maxima in 0 and 1, an only minimum at $s = 0.5$ and is symmetric. For a fixed $|t_{f_1+f_2}|$ two intervals of s that are proper subsets of (0,1) that have the same length, the interval with a larger distance between its centre and $s = 0.5$ has the larger probability. From Lemmas B.2, (4) and B.2, (3), it follows that for any fixed $|t_{f_1+f_2}|$ such that for some values of S Eq. (29) holds, not only does the centre of this interval of values of S move further from $s = 0.5$ with increase of $|\theta - \pi/4|$, but the length of this interval of the values of S increases. Therefore, the probability of this interval increases with $|\theta - \pi/4|$ for any fixed $|t_{f_1+f_2}|$. Hence, the probability that Eq. (29) holds, also increases with $|\theta - \pi/4|$. □

3. Level of the Behrens-Fisher test in case $n_1 = n_2 = 2$ and $n_1 = n_2 = 3$

From Theorem B.1, we conclude that for all sample sizes n if they are equal in both samples, there exist a test of level α: "reject H_o if $|\frac{X_1 - X_2}{\sqrt{S_1^2/n + S_2^2/n}}| > t_{\alpha/2}$". We call this test "New test".

For $n = 2$ and $n = 3$, it follows from Theorem B.2 that the p-quantiles of $T_1 cos\theta + T_2 sin\theta$ for all $\theta \in (0, \pi/2)$ and $p \in (0.5, 1)$ are larger than p-quantiles of a student t-distributed value with f degrees of freedom. Therefore the rejection region of the Behrens-Fisher test is a proper subset of the rejection region for the "New test" for all values of the nuisance parameter, so the Behrens-Fisher test is also of the nominal level.

QED.

APPENDIX C

R-codes

1. Computing the confidence density

The following function was used

```
dconf<-function(par,data,method="Welch")
{    if((method=="Paired")||(method=="paired")||(method==4)) {
  #OTHER DTATA FORMAT: matrix of 2 rows, rows correspond samples
  if (length(dim(data))>2)  return("Error: too many samples")
  else{
    n_pairs=length(data[1,])
    sd_difference=sd(data[1,]-data[2,])/sqrt(n_pairs)
    av_difference=mean(data[1,]+data[2,])
    return(dt((par-av_difference)/sd_difference,n_pairs-1)/sd_difference)
  }
}
  if ((!is.numeric(data))||(length(data)%%3>0)) return("Wrong format of the data")
  if((method=="Welch")||(method=="Welch-Satherwaite")||(method==1)) {

    if(length(dim(data))==3)
    { nominator_df=(colSums(data[2,,]/data[3,,]))^2
    denominator_df=colSums((data[2,,]/data[3,,])^2/(data[3,,]-1))
    return(dt((par-colSums(data[1,,]))/nominator_df^0.25,nominator_df/denominator_df)/
    nominator_df^0.25)
    }
    else{
      if(length(dim(data))<2)
        data=matrix(data=data,byrow=TRUE,nrow=length(data)/3,ncol=3)
      {
        nominator_df=(sum(data[,2]/data[,3]))^2
        denominator_df=sum((data[,2]/data[,3])^2/(data[,3]-1))
        return(dt((par-sum(data[,1]))/nominator_df^0.25,nominator_df/denominator_df)/
        nominator_df^0.25)
        }} }
  if((method=="Aspin")||(method=="Welch-Aspin")||(method==2)) {

    if(length(dim(data))==3)
    { sigmaconf=Forkastningsgrense(data[2,1,],data[2,2,],data[3,1,1],data[3,2,1],0.4)/
    qnorm(0.8)
      return(dnorm((par-colSums(data[1,,]))/sigmaconf)/sigmaconf)}
    else{
      if(length(dim(data))<2)
        data=matrix(data=data,byrow=TRUE,nrow=length(data)/3,ncol=3)
      {
        if (length(data[,1])>2)  return("Error: too many samples")
        else
        { sigmaconf=Forkastningsgrense(data[1,2],data[2,2],data[1,3],data[2,3],0.4)/qnorm(0.8)
        return(dnorm((par-sum(data[,1]))/sigmaconf)/sigmaconf)}
        }} }
```

```
  if((method=="BF")||(method=="Behrens-Fisher")||(method==3)) {

    if(length(dim(data))==3)
    { sigmaconf=sqrt(colSums(data[2,,]/data[3,,]))
    theta=atan(sqrt(data[2,1,]*data[3,2,]/data[3,1,]/data[2,2,]))
    return(dBF((par-colSums(data[1,,]))/sigmaconf,data[3,1,1],data[3,2,1],theta)/
    sigmaconf)}
    else{
      if(length(dim(data))<2)
        data=matrix(data=data,byrow=TRUE,nrow=length(data)/3,ncol=3)
      {
        if (length(data[,1])>2)   return("Error: too many samples")
        else
        { sigmaconf=sqrt(sum(data[,2]/data[,3]))
        theta=atan(sqrt(data[1,2]*data[2,3]/data[1,3]/data[2,2]))
        return(dBF((par-sum(data[,1]))/sigmaconf,data[1,3],data[2,3],theta)/sigmaconf)}
        }} }
    else
    return("Sorry, the suggested method is not implemented yet :-(")
}
```

Very similarly the confidence distribution function was computed.

The following additional computations were necessary to compute the confidence density and confidence distribution function.

```
##################### for Welch-Aspin
VRU<-function(r,u,s12,s22)
  (lambda_ir[r,1]*(s12^r)*f_iu[u,1]+lambda_ir[r,2]*(s22^r)*f_iu[u,2])/
  (lambda_i[1]*s12+lambda_i[2]*s22)^r
############################### for Welch-Aspin
Forkastningsgrense<-function(s12,s22,N,M,nominal_significance)
{
  lambda_i<<-c(1/N,1/M)
  lambda_ir<<-matrix(data=NA,nrow=5,ncol=2)
  for (i in 1:2)
    for (r in 1:5)
      lambda_ir[r,i]<<-lambda_i[i]^r
  f_iu<<-matrix(data=NA,nrow=5,ncol=2)
  for (u in 1:5)
  {     f_iu[u,1]<<-1/(N-1)^u
  f_iu[u,2]<<-1/(M-1)^u
  }
  z=qnorm(1-nominal_significance/2)
  #summands in notation of Aspin, dzeta is called z to make expressions shorter
  h0=z*sqrt(lambda_i[1]*s12+lambda_i[2]*s22)
  h1=0.25*(1+z^2)*VRU(2,1,s12,s22)*h0
  h2=(-0.5*(1+z^2)*VRU(2,2,s12,s22)+1/3*(3+5*z^2+z^4)*VRU(3,2,s12,s22)-
        1/32*(15+32*z^2+9*z^4)*(VRU(2,1,s12,s22))^2)*h0
  h3=h0*((1+z^2)*VRU(2,3,s12,s22)-2*(3+5*z^2+z^4)*VRU(3,3,s12,s22)+1/
  8*(15+32*z^2+9*z^4)*VRU(2,2,s12,s22)*VRU(2,1,s12,s22)+
           1/8*(75+173*z^2+63*z^4+5*z^6)*VRU(4,3,s12,s22)-1/12*
           (105+298*z^2+140*z^4+15*z^6)*VRU(3,2,s12,s22)*VRU(2,1,s12,s22)+
           1/384*(945+3169*z^2+1811*z^4+243*z^6)*((VRU(2,1,s12,s22))^3 ) )
  h4=h0*(
    -2*(1+z^2)*VRU(2,4,s12,s22)+28/3*(3+5*z^2+z^4)*VRU(3,4,s12,s22)
    -1/4*(15+32*z^2+9*z^4)*(VRU(2,3,s12,s22)*VRU(2,1,s12,s22)+0.5*(VRU(2,2,s12,s22))^2)
    -3/2*(75+173*z^2+63*z^4+5*z^6)*VRU(4,4,s12,s22)
    +1/2*(105+298*z^2+140*z^4+15*z^6)*(1/3*VRU(2,2,s12,s22)*
                                   VRU(3,2,s12,s22)+VRU(2,1,s12,s22)*VRU(3,3,s12,s22))
    +1/4*(15+33*z^2+11*z^4+z^6)*VRU(4,4,s12,s22)
```

```
  +1/5*(735+2170*z^2+1126*z^4+168*z^6+7*z^8)*VRU(5,4,s12,s22)
  -1/64*(945+3169*z^2+1811*z^4+243*z^6)*VRU(2,2,s12,s22)*VRU(2,1,s12,s22)^2
  -1/18*(945+3354*z^2+2166*z^4+425*z^6+25*z^8)*VRU(3,2,s12,s22)^2
  -1/32*(4725+16586*z^2+10514*z^4+1974*z^6+105*z^8)*VRU(2,1,s12,s22)*VRU(4,3,s12,s22)
  +1/96*(10395+42429*z^2+31938*z^4+7335*z^6+495*z^8)*
  VRU(3,2,s12,s22)*VRU(2,1,s12,s22)^2
  -1/6144*(135135+626144*z^2+542026*z^4+145320*z^6+11583*z^8)*VRU(2,1,s12,s22)^4
 )

 return(h0+h1+h2+h3+h4)
 }
 #for BF density by Ruben(1960)
Integrand_for_BF<-function(v,x,f1,f2,th)#notation as in Ruben(1960), "th" is theta
 x^(0.5*(f1-1))*(1-x)^(0.5*(f2-1))*(f1*(1-x)*(sin(th))^2+f2*x*
 (cos(th)^2))^(-0.5)*(1+(v^2*x*(1-x))/(f1*(1-x)*(sin(th))^2+
 f2*x*(cos(th)^2)))^(-0.5*(f1+f2+1))

#Probability density for Behrens-Fisher distribution
dBF<-function(v,f1,f2,th,N_points=501) #name "N_points" corresponds to number of points for Simpson
{ if (N_points %% 2 ==0) N_points=N_points+1
Partitition=matrix(data=seq(from=0,to=1,length=N_points),byrow=TRUE, ncol=N_points,
nrow=length(v))
v=matrix(data=rep(v,N_points), ncol=N_points,nrow=length(v))
Partitition=Integrand_for_BF(v,Partitition,f1,f2,th)
even_numbers=(1:((N_points-1)/2))*2
h=1/(N_points-1)
if(length(dim(Partitition[,even_numbers]))>1)# for matrices we use "rowSums"
 return(h/3*gamma(0.5*f1+0.5*f2+0.5)/(sqrt(pi)*gamma(0.5*f1)*gamma(0.5*f2))*
 (rowSums(Partitition)*2+rowSums(Partitition[,even_numbers])*
 2-rowSums(Partitition[,c(1,N_points)])))
else #for vectors "rowSums" doesn't work, must use sum
 return(h/3*gamma(0.5*f1+0.5*f2+0.5)/(sqrt(pi)*gamma(0.5*f1)*gamma(0.5*f2))*
 (sum(Partitition)*2+sum(Partitition[,even_numbers])*2-sum(Partitition[,c(1,N_points)])))
}
##########################


 }
```

For the likelihood ratio method, the computation was different and included the likelihood ratio distribution computation. The following codes are were written for not using much RAM, because due to technical reasons they were compiled on an old PC with a little memory. Of this reasons, the codes use matrix operations in less extent than that would be reasonable, and are therefore rather slow. The computations are presented for sample sizes $n_1 = 3$ and $n_2 = 5$.

```
N=3
M=5
sigma0_2=1 #standard deviation of the second sample
#lnsigma=seq(from=-2,to=2,by=0.2)
lnsigma=sort(c(seq(from=-2,to=2,by=0.2),seq(from=0.5,to=1.5,by=0.07)))
sigma0_1=10^(lnsigma)
pr=1-seq(from=0.00001,to=0.99999,by=0.00001)
alpha=1-pr
Lcr=matrix(nrow=length(alpha),ncol=length(sigma0_1))
Lcr_general=array(dim=c(length(N),length(M),length(alpha)))


N_simulations=500000
```

```
mu=0

Iter <-function(m1,m2,S1,S2,N,M){#the iterations
  m0=(m1*N+m2*M)/(M+N)

  S10=S1
  S20=S2
  t = 2
  for (i in 1:5){
    m0=(N*m1*S20+M*m2*S10)/(N*S20+M*S10)
    S10=S1+(m1-m0)^2
    S20=S2+(m2-m0)^2
  }
  return(m0)
}
## The function computing the critical value of deviation ############
# for the given relation between standard deviations in samples
Crit_Deviance_Pval<-function(N,M,sigma0_1,sigma0_2=1,alpha=0.05)
#alpha may be a vector
{
  #Simulate samples X og Y
  X=matrix(rnorm(N*N_simulations,mean=0,sd=sigma0_1),
           nrow=N,              # number of rows
           ncol=N_simulations)
  Y=matrix(rnorm(M*N_simulations,mean=0,sd=sigma0_2),
           nrow=M,              # number of rows
           ncol=N_simulations)

  s12=colVars(X)*(N-1)/N #not to calculate it many times
  s22=colVars(Y)*(M-1)/M #not to calculate it many times

  M1=colMeans(X)
  M2=colMeans(Y)
  mu_hat=Iter(M1,M2,s12,s22,N,M)
  sigma_1_hat2=s12+(M1-mu_hat)^2
  sigma_2_hat2=s22+(M2-mu_hat)^2
  a=(sigma_1_hat2/s12)^(0.5*N)
  b=(sigma_2_hat2/s22)^(0.5*M)
  Deviance=2*log(a*b)
  return(quantile(Deviance,1-alpha))
}

for(i in 1:length(N))
   for (ii in 1:length(M))
    {for(s in 1:length(sigma0_1))
        Lcr[,s]=Crit_Deviance_Pval(N[i],M[ii],sigma0_1[s],alpha=alpha)
               for (iii in 1:length(pr))
                      Lcr_general[i,ii,iii]=max(Lcr[iii,])
                      #the selected critical value.
     } #the critical size for all sample sizes
Lcr_general[i,ii,]=Lcr_general[i,ii,length(Lcr_general[i,ii,]):1]
i=which(length(X[,1]==N)
ii=which(length(X[,1]==M)

CC_mu <-function(mu, X,Y,N,M){  #constructs conf curve
#check increases
  if(length(dim(X))>1)
{
  s12=colVars(X)*(N-1)/N #not to calculate it many times
  s22=colVars(Y)*(M-1)/M #not to calculate it many times
  M1=colMeans(X)
```

```
  M2=colMeans(Y)
}
else
  {s12=var(X)*(N-1)/N #not to calculate it many times
  s22=var(Y)*(M-1)/M #not to calculate it many times
  M1=mean(X)
  M2=mean(Y) }
  sigma_1_hat2=s12+(M1-mu)^2
  sigma_2_hat2=s22+(M2-mu)^2
  a=(sigma_1_hat2/s12)^(0.5*N)
  b=(sigma_2_hat2/s22)^(0.5*M)
  #i=which(length(X[,1]==N)#drop if single sample size
  #ii=which(length(X[,1]==M)#drop if single sample size
i=1
ii=1
  Deviance=2*log(a*b)
  alphaleft_ind=which.max(Lcr_general[i,ii,(Lcr_general[i,ii,]<Deviance)])
#print(Deviance)
if(length(alphaleft_ind)>0)
 {if(alphaleft_ind<length(Lcr_general[i,ii,]))
  return(pr[alphaleft_ind]+0.001*(Deviance-Lcr_general[i,ii,alphaleft_ind]) /
  (Lcr_general[i,ii,alphaleft_ind+1] - Lcr_general[i,ii,alphaleft_ind]) )
  #may replace 0.001 with pr[alphaleft_ind+1]-pr[alphaleft_ind])
else  return(pr[alphaleft_ind])}
else return(0)

}


dconfl<-function(mu,X,Y,N,M)#{mu-vector
{
dconfl=rep(0,length(mu)-1)
y=rep(0,length(mu))
for(i in 1:length(mu))
y[i]=CC_mu(mu[i], X,Y,N,M)
for(i in 2:length(mu))
 if (y[i]<y[i-1])
     dconfl[i]=(y[i-1]-y[i])/2/(mu[i]-mu[i-1])#assume MLE is median of cd
 else
dconfl[i]=(y[i]-y[i-1])/2/(mu[i]-mu[i-1])
return(dconfl)
```

2. Converting between nominal and actual level of the tests

The computation of the nominal level corresponding the desired level of the test may be calculated as following.

```
Nominal_for_Actual<-function(level,N,M,method="Welch",digits=4)
#Computes nominal level for obtaining the given actual.
{if (method=="Welch")#Numerical solution of equation  "Actual(nominal)=given"
{x1=level
x=0
y=0
 xw=0#wrong level, nominal level can't be less
 y1=ActualWelchlevel(N,M,x1,digits=digits)
 x2=max(x1-0.005,x1*0.9)
 y2=ActualWelchlevel(N,M,x2,digits=digits)
  ready=FALSE
#  print(c(x,y,"-",x1,y1,"-",x2,y2,"-",xw))
```

```
  while(!ready)
    {while (y2>y1)
      {
      xw=x2
        x2=x1-(x1-x2)/2
        y2=ActualWelchlevel(N,M,x2,digits=digits)
#
          }#now y1>y2, xw<x2<x1
     x=where_line_crosses_horiz(x1,y1,x2,y2,level)
     while (x<xw)
        x=x2-(x2-xw)/2#now xw<x<x2
     y=ActualWelchlevel(N,M,x,digits=digits)
     if(y<level)
        {y2=y
         x2=x}
     else
        if (y>y2)
        {y1=y
        x1=x}
     else
     {y1=y2
     x1=x2
     y2=y
     x2=x}

      if(abs(y-level)<10^(-digits)) return(x)
     if(abs(xw-x)<10^(-digits)) return(NA)
    }
  }
    else
  if (method=="Aspin")
  {x1=level
  x=0
  y=0
  xw=0#wrong level, nominal level can't be less
  y1=ActualAspinlevel(N,M,x1,digits=digits)
  x2=max(x1-0.005,x1*0.9)
  y2=ActualAspinlevel(N,M,x2,digits=digits)
  ready=FALSE
  #   print(c(x,y,"-",x1,y1,"-",x2,y2,"-",xw))
  while(!ready)
  {while (y2>y1)
  {
    xw=x2
    x2=x1-(x1-x2)/2
    y2=ActualAspinlevel(N,M,x2,digits=digits)
    #        print(c(x,y,x1,y1,x2,y2,xw))
  }#now y1>y2, xw<x2<x1
    x=where_line_crosses_horiz(x1,y1,x2,y2,level)
    while (x<xw)
      x=x2-(x2-xw)/2#now xw<x<x2
    y=ActualAspinlevel(N,M,x,digits=digits)
    if(y<level)
    {y2=y
    x2=x}
    else
      if (y>y2)
      {y1=y
      x1=x}
    else
    {y1=y2
```

```
    x1=x2
    y2=y
    x2=x}

    if(abs(y-level)<10^(-digits)) return(x)
    if(abs(xw-x)<10^(-digits)) return(NA)
  }} else
    return("The available methods are only -Welch- and -Aspin-")
}
```

Here we solve the inverse task, computing the actual level for the given nominal level, as follows for the W.

```
  ActualWelchlevel<-function(N,M,l_nom,digits=5,bothcoordinates=FALSE)
  #finds max p(reject H0)
{ centre=0#centre of initial interval we check for maximum
ready=FALSE
pr=rep(0,50)
while(!ready)#At first chooses initial maximum point, computing roughly
{
  seqinitial=seq(from=-2+centre,to=2+centre,length=50)
  for (i in 1:50) pr[i]= prob_rejectH0(exp(seqinitial[i]),N,M,l_nom,18)
  k=(which.max(pr))[1]
  if((k<50)&&(k>1)) ready=TRUE#until the maximum is inside
  if(k==50) centre=centre+1.77
  #if the maximum point is not inside, then shifts the interval
  #by 1.97(a number that 1.97!=n*4/50)
  else centre=centre-1.97
}
ready=FALSE
if(digits<5) intervals=20    else
  if(digits<6) intervals=30    else
    if(digits<7) intervals=50    else
      if(digits<8) intervals=70    else intervals=100
pr=numeric(5)
step=0.06
while(!ready)
{
  seqinitial=c(seqinitial[k]-step*2,seqinitial[k]-
  step,seqinitial[k],seqinitial[k]+step,seqinitial[k]+step*2)
  for (i in 1:5) pr[i]= prob_rejectH0(exp(seqinitial[i]),N,M,l_nom,intervals)
  if(max(pr)-min(pr)<10^(-digits)) ready=TRUE
  k=(which.max(pr))[1]
  step=step/2.5 #contracts the 5-point interval
}
if(bothcoordinates)
  return(c(max(pr),seqinitial[k])) else  return(max(pr))

}
```

The computations were very similar for the Welch-Aspin test.

Bibliography

Aspin, A. A. (1948). An examination and further development of a formula arising in the problem of comparing two mean values. *Biometrika*, *35.1*(2), 88–96.

Bachmaier, M. (2012). On Welch's and Aspin's series solution of the Behrens-Fisher problem [Lest: 2019-05-09]. https://www.researchgate.net/profile/Martin_Bachmaier/publication/267661757_On_Welch%27s_and_Aspin%27s_series_solution_of_the_Behrens-Fisher_problem/links/5ae08ba3aca272fdaf8c8347/On-Welchs-and-Aspins-series-solution-of-the-Behrens-Fisher-problem.pdf.

Bédard, M., Fraser, D. A., Wong, A., et al. (2007). Higher accuracy for Bayesian and frequentist inference: Large sample theory for small sample likelihood. *Statistical Science*, *22*(3), 301–321.

Best, D., & Rayner, J. (1987). Welch's approximate solution for the Behrens–Fisher problem. *Technometrics*, *29*(2), 205–210.

Casella, G., & Berger, R. L. (2002). *Statistical inference* (Vol. 2). Duxbury Pacific Grove, CA.

Cox, T. F., & Jaber, K. H. (1990). The generalized likelihood ratio test for the Behrens–Fisher problem. *Journal of Applied Statistics*, *17(1)*, 63–67.

Duong, Q. P., & Shorrock, R. W. (1996). On Behrens-Fisher solutions. *Journal of the Royal Statistical Society: Series D (The Statistician)*, *45*(1), 57–63.

Fisher, R. A. (1935). The fiducial argument in statistical inference. *Annals of eugenics*, *6.4*, 391–398.

Fisher, R. A. (1930). Inverse probability. *Mathematical Proceedings of the Cambridge Philosophical Society*, *26*(4), 528–535.

Ghosh, M., & Kim, Y.-H. (2001). The Behrens-Fisher problem revisited: A Bayes-frequentist synthesis. *Canadian Journal of Statistics*, *29*(1), 5–17.

Hayter, A. J. (2014). Inferences on linear combinations of normal means with unknown and unequal variances. *Sankhya A*, *76*(2), 257–279.

JCGM, J. et al. (2008). Evaluation of measurement data—guide to the expression of uncertainty in measurement. *Int. Organ. Stand. Geneva ISBN*, *50*, 134.

Kallenberg, O. (2017). *Random measures, theory and applications*. Springer.

Karr, A. F. (1993). Classical limit theorems. *Probability* (pp. 183–216). Springer.

Kim, S.-H., & Cohen, A. S. (1998). On the Behrens-Fisher problem: A review. *Journal of Educational and Behavioral Statistics*, *23.4*, 356–377.

Klenke, A. (2008). Probability theory: A comprehensive course.

Larsen, R. J., & Marx, M. L. (2013). *Introduction to mathematical statistics and its applications: Pearson new international edition, 5.ed.* Pearson Hall.

Lehmann, E. L. (2006). *Testing statistical hypotheses.* Springer.

Lehmann, E. L., & Casella, G. (2006). *Theory of point estimation.* Springer Science & Business Media.

Lillegård, M. (2001). Tests based on Monte Carlo simulations conditioned on maximum likelihood estimates of nuisance parameters. *Journal of statistical computation and simulation*, *71*(1), 1–10.

Schweder, T., & Hjort, N. L. (2016). *Confidence, likelihood, probability* (Vol. 41). Cambridge University Press.

Singh, K., Xie, M., Strawderman, W. E., et al. (2005). Combining information from independent sources through confidence distributions. *Annals of statistics*, *33*(1), 159–183.

Taraldsen, G. (2021). Joint confidence distributions.

Taraldsen, G., & Lindqvist, B. H. (2013). Fiducial theory and optimal inference. *The Annals of Statistics*, 323–341.

Wang, Y. Y. (1971). Probabilities of the type i errors of the Welch tests for the behrens-fisher problem. *Journal of the American Statistical Association*, *66*(335), 605–608.

Xie, M.-g., & Singh, K. (2013). Confidence distribution, the frequentist distribution estimator of a parameter: A review. *International Statistical Review*, *81*(1), 3–39.

NTNU
Norwegian University of
Science and Technology