UPPSALA DISSERTATIONS IN MATHEMATICS

85

On Confidence Intervals and Two-Sided Hypothesis Testing

Måns Thulin

Department of Mathematics
Uppsala University
UPPSALA 2014

Dissertation presented at Uppsala University to be publicly examined in Polhemsalen, Ångströmlaboratoriet, Lägerhyddsvägen 1, Uppsala, Friday, 26 September 2014 at 13:15 for the degree of Doctor of Philosophy. The examination will be conducted in English. Faculty examiner: Professor Burkhardt Seifert (Universität Zurich).

Abstract

Thulin, M. 2014. On Confidence Intervals and Two-Sided Hypothesis Testing. *Uppsala Dissertations in Mathematics* 85. 47 pp. Uppsala: Acta Universitatis Upsaliensis.
ISBN 978-91-506-2408-3.

This thesis consists of a summary and six papers, dealing with confidence intervals and two-sided tests of point-null hypotheses.

In Paper I, we study Bayesian point-null hypothesis tests based on credible sets. A decision-theoretic justification for tests based on central credible intervals is presented.

Paper II is concerned with a new two-sample test for the difference of mean vectors, in the high-dimensional setting where the number of variables is greater than the sample size. A simulation study indicates that the proposed test yields higher power when the variables are correlated. Computational aspects of the test are discussed.

In Paper III, we discuss randomized confidence intervals for a binomial proportion. How some classical intervals fare is compared to how a recently proposed interval fares, in terms of coverage, length and sensitivity to the randomization.

In Paper IV, a level-adjustment of the Clopper-Pearson interval for a binomial proportion is proposed. The adjusted interval is shown to have good coverage properties and short expected length.

In Paper V we study the cost of using the exact Clopper-Pearson interval rather than shorter approximate intervals, in terms of the increase in expected length and the increase in sample size required to obtain a given length. Comparisons are made using asymptotic expansions.

Paper VI deals with exact confidence intervals and point-null hypothesis tests for parameters of a class of discrete distributions. A large class of intervals are shown to lack strict nestedness and to have bounds that are not strictly monotone and typically also discontinuous. The p-values of the corresponding hypothesis test are shown to lack desirable continuity properties, and to typically also lack certain monotonicity properties.

Måns Thulin, Applied Mathematics and Statistics, Box 480, Uppsala University, SE-75106 Uppsala, Sweden.

ISSN 1401-2049
ISBN 978-91-506-2408-3
urn:nbn:se:uu:diva-229399 (http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-229399)

Till Lisa, Alvar och Grodden

List of papers

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

I Thulin, M. (2014). Decision-theoretic justifications for Bayesian hypothesis testing using credible sets. *Journal of Statistical Planning and Inference*, **146**, 133–138. DOI: 10.1016/j.jspi.2013.09.014.

II Thulin, M. (2014). A high-dimensional two-sample test for the mean using random subspaces. *Computational Statistics & Data Analysis*, **74**, 26–38. DOI: 10.1016/j.csda.2013.12.003.

III Thulin, M. (2014). On split sample and randomized confidence intervals for binomial proportions. *Statistics & Probability Letters*, **92**, 65–71. DOI: 10.1016/j.spl.2014.05.005.

IV Thulin, M. (2014). Coverage-adjusted confidence intervals for a binomial proportion. *Scandinavian Journal of Statistics*, **41**, 291–300. DOI: 10.1111/sjos.12021.

V Thulin, M. (2014). The cost of using exact confidence intervals for a binomial proportion. *Electronic Journal of Statistics*, **8**, 817–840. DOI: 10.1214/14-EJS909.

VI Thulin, M., Zwanzig. S. (2014). Exact confidence intervals and hypothesis tests for parameters of discrete distributions. Manuscript.

Reprints were made with permission from the publishers.

Contents

Sections 5.4 and 5.5 are partially based on the author's licentiate thesis: Thulin, M. (2012). *On two classic problems in statistics*, U.U.D.M. Report 2012:3.

1. Introduction

1.1 Point-null hypotheses

Many of the questions which statistics are used to answer can be expressed as two-sided point-null hypothesis testing problems. In such a problem we wish to determine whether some unknown quantity θ equals a particular value θ_0 or whether it differs from θ_0.

We may for instance ask whether a drug affects systolic blood pressure, and let θ denote the increase or decrease of the systolic blood pressure for patients who are given the drug instead of a placebo. When there is no effect, $\theta = 0$. If we have no prior reason to believe that the drug affects the blood pressure in some particular direction, it is reasonable to test the hypothesis that $\theta = 0$ against the alternative hypothesis that $\theta \neq 0$. The hypothesis $\theta = 0$ is called a point-null hypothesis since it consist of a single point, which in this particular case is 0.

Other questions will lead to similarly formulated point-null hypotheses. Has the public opinion on some issue changed? Has the privatization of Swedish pharmacies affected the price of paracetamol? Does changing the proportions of compounds in an alloy change its conductivity?

Despite their popularity and wide applicability, several authors have criticized point-null hypotheses in recent years [24, 25, 26]. The main criticism is that point-null hypotheses are unrealistic: the new drug, it is argued, will always have *some* effect on the systolic blood pressure, small as it may be. When we ask whether it affects blood pressure we are asking the wrong question, since we already know that is has an effect. Instead of asking whether it affects the blood pressure, we should ask whether the effect is positive or negative, including "undetermined" as a third option when we make our decision. This is known as a three-way decision problem, a topic which we will return to in Section 2.2.

Cox [16] defended point-null hypotheses, listing six situations in which they arise. One of his situations can be interpreted as that it should be understood that point-null hypotheses in fact, for some small $\varepsilon > 0$, are convenient approximations of a set $(\theta_0 - \varepsilon, \theta_0 + \varepsilon)$ of effects that are small enough to be negligible. From a practical point of view, there is seldom any difference between an effect being θ_0 and an effect being so close to θ_0 that we are unable to measure the difference.

Point-null hypothesis testing occurs naturally in regression problems, where we test whether some coefficient equals 0, and in mixture model problems,

where we test whether the weight of a component is 0, in order to determine how many components should be included in the model. They also frequently arise in multivariate hypothesis testing, a setting which we will discuss further in Section 3.1.

The theory of point-null hypothesis testing is largely well-understood [12, 30]. Nevertheless, there are still many open problems, even in areas of statistics which are important for applications. This thesis aims to contribute to open problems in the following areas:

- The pronounced differences between the frequentist and Bayesian approaches to point-null testing [2, 46], which even asymptotically can lead to completely opposite conclusions [33]. Paper I of this thesis is concerned with a connection between these two approaches, which is described in Sections 2.1-2.2.
- Multivariate point-null hypothesis testing when the number of variables p is greater than the sample size n, for which classic statistical methods no longer are applicable [29, 34]. Such high-dimensional datasets are becoming increasingly more common in for instance genetics and astronomy. Some problems related to high-dimensional hypothesis testing are discussed in Sections 3.1-3.2. In Paper II a two-sample test for the difference of two mean vectors is proposed. The methodology is applicable also for many other point-null problems.
- Point-null hypothesis testing, and the closely related topic of interval estimation, for parameters of discrete distributions. For such distributions, which appear frequently in problems in for instance clinical trials, risk analysis and telecommunications, the elegant theory that exists for absolutely continuous distributions no longer holds, causing a number of distressing problems [8, 9, 10, 39, 50]. Papers III-VI are concerned with some of these problems, described in Section 5.

1.2 The equivalence between frequentist testing and confidence intervals

Let θ be an unknown parameter in the parameter space $\Theta \subseteq \mathbb{R}$, and let the sample $\mathbf{x} = (x_1, \ldots, x_n) \in \mathscr{X}^n \subseteq \mathbb{R}^n$ be a realization of the random variable $\mathbf{X} = (X_1, \ldots, X_n)$. In frequentist statistics there is a fundamental connection between interval estimation and point-null hypothesis testing of θ [12], which we describe next. For now, we define a *confidence interval* $I_\alpha(\mathbf{X})$ as a random interval such that its *coverage probability*

$$\mathrm{P}_\theta(\theta \in I_\alpha(\mathbf{X})) = 1 - \alpha \qquad \text{for all } \alpha \in (0,1). \tag{1.1}$$

Later on, in Section 5.2, we will consider more general intervals, where the coverage probability either is bounded by or approximately equal to $1 - \alpha$.

Consider a two-sided test of the point-null hypothesis $H_0(\theta_0) : \theta = \theta_0$ against the alternative $H_1(\theta_0) : \theta \neq \theta_0$. Let $\lambda(\theta_0, \mathbf{x})$ denote the p-value of the test. For any $\alpha \in (0,1)$, $H_0(\theta_0)$ is rejected at the level α if $\lambda(\theta_0, x) \leq \alpha$. The level α *rejection region* is the set of $\mathbf{x}$ which lead to the rejection of $H_0(\theta_0)$:

$$R_\alpha(\theta_0) = \{\mathbf{x} \in \mathbb{R}^n : \lambda(\theta_0, \mathbf{x}) \leq \alpha\}.$$

Now, consider a family of two-sided tests with p-values $\lambda(\theta, \mathbf{x})$, for $\theta \in \Theta$. For such a family we can define an *inverted rejection region*

$$Q_\alpha(\mathbf{x}) = \{\theta \in \Theta : \lambda(\theta, \mathbf{x}) \leq \alpha\}.$$

For any fixed θ_0, $H_0(\theta_0)$ is rejected if $\mathbf{x} \in R_\alpha(\theta_0)$, which happens if and only if $\theta_0 \in Q_\alpha(\mathbf{x})$, that is,

$$\mathbf{x} \in R_\alpha(\theta_0) \Leftrightarrow \theta_0 \in Q_\alpha(\mathbf{x}). \tag{1.2}$$

If the test is based on a test statistic with a completely specified absolutely continuous null distribution, then $\lambda(\theta_0, \mathbf{X}) \sim \mathrm{U}(0,1)$ under $H_0(\theta_0)$ [31]. Then

$$\mathrm{P}_{\theta_0}(\mathbf{X} \in R_\alpha(\theta_0)) = \mathrm{P}_{\theta_0}(\lambda(\theta_0, \mathbf{X}) \leq \alpha) = \alpha. \tag{1.3}$$

Since (1.3) holds for any $\theta_0 \in \Theta$ and since (1.2) implies that

$$\mathrm{P}_{\theta_0}(\mathbf{X} \in R_\alpha(\theta_0)) = \mathrm{P}_{\theta_0}(\theta_0 \in Q_\alpha(\mathbf{X})),$$

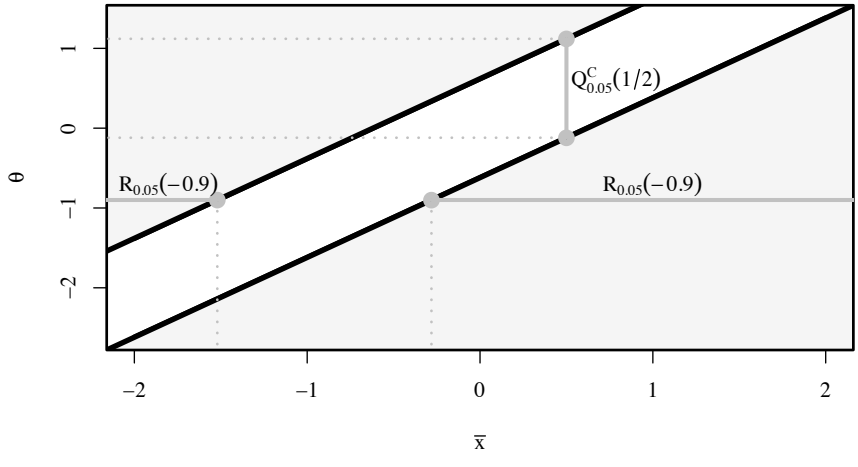
it follows that the random set $Q_\alpha(\mathbf{x})$ always covers the true parameter θ_0 with probability α. Consequently, letting $Q_\alpha^C(\mathbf{x})$ denote the complement of $Q_\alpha(\mathbf{x})$, for all $\theta_0 \in \Theta$ we have

$$\mathrm{P}_{\theta_0}(\theta_0 \in Q_\alpha^C(\mathbf{X})) = 1 - \alpha,$$

meaning that the complement of the inverted rejection region is a $1-\alpha$ confidence interval for θ. This well-known equivalence between a family of tests and a confidence interval $I_\alpha(\mathbf{x}) = Q_\alpha^C(\mathbf{x})$, illustrated in Figure 1.1, provides a simple way of constructing confidence intervals through test inversion.

In Section 4 we discuss how optimality properties of a family of tests relate to optimality properties of $Q_\alpha^C(\mathbf{x})$ and how the definition of $\lambda(\theta_0, \mathbf{x})$ affects the behaviour of $Q_\alpha^C(\mathbf{x})$. These issues are closely related to some problems concerning inference for parameters of discrete distributions, which we discuss in Section 5.

Figure 1.1. Rejection regions and confidence intervals corresponding to the the z-test for a normal mean, for different null means θ and different sample means $\bar{x}$, with $\sigma = 1$. $H_0(\theta)$ is rejected if $(\bar{x}, \theta)$ is in the shaded light grey region. Shown in dark grey is the rejection region $R_{0.05}(-0.9) = (-\infty, -1.52) \cup (-0.281, \infty)$ and the confidence interval $I_{0.05}(1/2) = Q^C_{0.05}(1/2) = (-0.120, 1.120)$.



2. Bayesian point-null hypothesis testing

2.1 The standard Bayesian approach

Bayesian inference for a parameter θ starts with a prior distribution for θ, representing prior beliefs or knowledge about θ (or lack thereof), expressed as a probability distribution $\mathrm{P}(\theta \in A)$ for all $A \subseteq \Theta$. Using Bayes' theorem, these beliefs are updated after observing the data $\mathbf{x}$, resulting in the posterior distribution $\mathrm{P}(\theta \in A|\mathbf{x})$, on which all Bayesian decisions are based.

Bearing in mind the close connection between frequentist hypothesis testing and confidence intervals, it seems reasonable to expect that there should be a similar connection between Bayesian hypothesis testing and the Bayesian analogue to confidence intervals. This analogue, known as a *credible set*, is a set Θ_α such that

$$\mathrm{P}(\theta \in \Theta_\alpha|\mathbf{x}) = 1-\alpha, \tag{2.1}$$

i.e. a set such that the posterior probability that θ is in the set Θ_α is $1-\alpha$. A credible set is a *central interval* if

$$\mathrm{P}(\theta \leq \inf\Theta_\alpha|\mathbf{x}) = \mathrm{P}(\theta \geq \sup\Theta_\alpha|\mathbf{x}) = \alpha/2.$$

Another common type of credible sets is the highest posterior density (HPD) set, which is defined as the smallest set (volume-wise) satisfying (2.1). These two types of credible sets are illustrated in Figure 2.1.

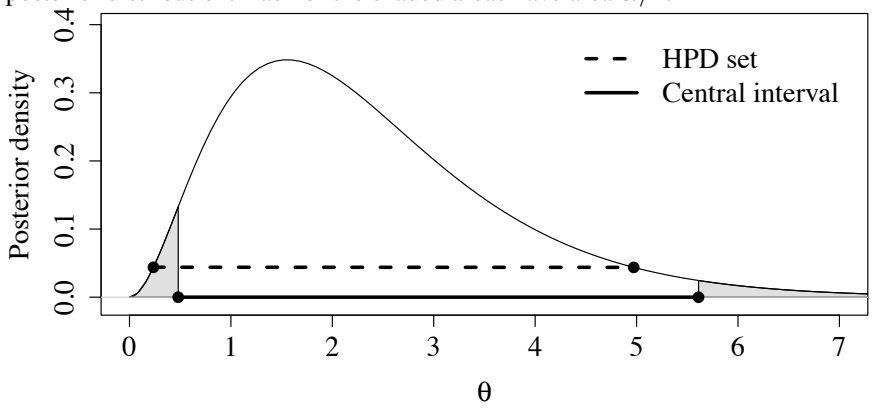
While some researchers have proposed Bayesian tests based on credible sets [27, 42], such tests are still not all that common. On the contrary, they have been criticized for being very informal [19] and for lacking decision-theoretic support [54].

Statistical decision theory is concerned with optimal decisions regarding some unknown parameter θ. The *loss function* $L(\theta, d)$ describes the loss (or cost) of making a decision d when the parameter takes the value θ. The optimal decision minimizes the expected loss; the optimal Bayesian decision minimizes the expected posterior loss. In hypothesis testing problems, the decision is whether or not to reject the null hypothesis.

Consider, as before, a point-null hypothesis testing problem for a one-dimensional parameter $\theta \in \Theta \subseteq \mathbb{R}$, with the additional constraint that Θ is uncountable. The standard solution to Bayesian point-null hypothesis testing [19, 46] has no direct connection to credible sets. In this approach, the null hypothesis is assigned a point-mass:

$$\mathrm{P}(\theta = \theta_0) = \pi_0 > 0.$$

Figure 2.1. Comparison of the HPD set and the central interval for an asymmetric posterior distribution. Each of the shaded areas have area $\alpha/2$.



The alternative $\theta \neq \theta_0$ is assigned the probability $\pi_1 = 1 - \pi_0$. Moreover, a prior density g is used under the alternative, so that the (mixed) prior distribution of θ is

$$\pi(\theta) = \pi_0 \mathbb{I}_{\Theta_0}(\theta) + \pi_1 g(\theta) \mathbb{I}_{\Theta_1}(\theta), \tag{2.2}$$

where $\Theta_0 = \{\theta_0\}$ and $\Theta_1 = \Theta \backslash \Theta_0$. Then $\mathrm{P}(\theta = \theta_0 | \mathbf{x})$ can be computed using Bayes' theorem. The null hypothesis is rejected if $\mathrm{P}(\theta = \theta_0 | \mathbf{x}) < a$, where a typically is determined using the following weighted 0–1 loss function [46]. Letting the test function φ be 0 if $H_0(\theta_0)$ is accepted and 1 if $H_0(\theta_0)$ is rejected, under the loss function

$$L(\theta, \varphi) = \begin{cases} 0, & \text{if } \varphi = 1 - \mathbb{I}_{\Theta_0}(\theta) \\ a, & \text{if } \theta = \theta_0 \text{ and } \varphi = 1, \\ b, & \text{if } \theta \in \Theta_1 \text{ and } \varphi = 0, \end{cases} \tag{2.3}$$

where $a, b > 0$, the null hypothesis $H_0(\theta_0)$ is rejected if $\mathrm{P}(\theta = \theta_0 | \mathbf{x}) < \frac{b}{a+b}$ [46]. The constants a and b are used to control the loss associated with *type I errors*, i.e. falsely rejecting the null hypothesis, and *type II errors*, i.e. not rejecting the null hypothesis when it in fact is false, respectively.

The prior (2.2) has to have a point-mass in θ_0 for the standard approach to work. If the prior is absolutely continuous, θ_0 is an absorbing state: it has probability 0 and by Bayes' theorem

$$\begin{aligned} \mathrm{P}(\theta = \theta_0 | \mathbf{x}) &= \frac{\mathrm{P}(\mathbf{x} | \theta = \theta_0) \cdot \mathrm{P}(\theta = \theta_0)}{\mathrm{P}(\mathbf{x} | \theta = \theta_0) \cdot \mathrm{P}(\theta = \theta_0) + \mathrm{P}(\mathbf{x} | \theta \neq \theta_0) \cdot \mathrm{P}(\theta \neq \theta_0)} \\ &= \frac{0}{\mathrm{P}(\mathbf{x} | \theta \neq \theta_0)} = 0 \end{aligned}$$

for all $\mathbf{x}$, meaning that, no matter the evidence, we never are convinced that $\theta = \theta_0$. A downside to using a mixed prior is that we typically would have to

use different priors for hypothesis testing and estimation, since an absolutely continuous prior is more reasonable for the latter problem.

2.2 An equivalence between Bayesian hypothesis testing and credible sets

In Paper I a decision-theoretic justification for Bayesian hypothesis testing using credible sets is presented. The benefit of this approach is that it allows us to use an absolutely continuous prior for θ, utilizing the same prior both for estimation and hypothesis testing. Such priors are often more intuitive and realistic, unless there is a strong reason to believe that a particular value of θ should have a point mass. Moreover, natural non-informative priors, such as Jeffreys and uniform priors, tend to be absolutely continuous. Under non-informative priors, test based on credible sets often have good frequentist properties, reconciling the supposedly irreconcilable [2, 33] Bayesian and frequentist approaches to point-null hypothesis testing.

In this section, the definition of Θ_1 will differ from that in the rest of the thesis, the reason being that the idea behind the justification is to split the set $\Theta\backslash\{\theta_0\}$ into two parts: $\Theta_{-1} = \{\theta : \theta < \theta_0\}$ and $\Theta_1 = \{\theta : \theta > \theta_0\}$. We can then consider the three-way decision problem of choosing between Θ_{-1}, Θ_0 and Θ_1. The null hypothesis $H_0(\theta_0)$ is rejected if we choose either Θ_{-1} or Θ_1.

Result 1 (Theorem 1, Paper I, p. 135)**.** *Let φ be a decision function that takes values in $\{-1,0,1\}$, such that we accept Θ_i if $\varphi = i$. Under an absolutely continuous prior for θ and the loss function*

$$L_{(1)}(\theta,\varphi) = \begin{cases} 0, & \text{if } \theta \in \Theta_i \text{ and } \varphi = i, \quad i \in \{-1,0,1\}, \\ \alpha/2, & \text{if } \theta \notin \Theta_0 \text{ and } \varphi = 0, \\ 1, & \text{if } \theta \in \Theta_i \cup \Theta_0 \text{ and } \varphi = -i, \quad i \in \{-1,1\}, \end{cases}$$

with $0 < \alpha < 1$, the Bayes test is to reject $H_0(\theta_0)$ if θ_0 is not contained in the central $1-\alpha$ credible set.

Stepping away from point-null hypotheses for a minute, we can also consider tests of composite hypotheses, where both Θ_0 and $\Theta_1 = \Theta\backslash\Theta_0$ are uncountable subsets of $\mathbb{R}$ with positive probabilities under an absolutely continuous prior. We introduce the notation $q_\alpha(\theta|\mathbf{x})$ for the α-quantile of the posterior distribution of θ.

Result 2 (Theorems 2-3, Paper I, p. 137)**.** *Let φ be a test function, such that Θ_0 is rejected if $\varphi = 1$ and accepted if $\varphi = 0$. Consider the weighted 0–1 loss function*

$$L_{(2)}(\theta,\varphi)=\begin{cases}0, & \text{if } \varphi = 1-\mathbb{I}_{\Theta_0}(\theta)\\ a, & \text{if } \theta\in\Theta_0 \text{ and } \varphi=1\\ b, & \text{if } \theta\in\Theta_1 \text{ and } \varphi=0.\end{cases}$$

(a) For the hypotheses $\Theta_0=\{\theta:\theta\le\theta_0\}$ *and* $\Theta_1=\{\theta:\theta>\theta_0\}$*, under an absolutely continuous prior for* θ *and the loss function* $L_{(2)}$ *with* $a=1-\alpha$ *and* $b=\alpha$*,* $0<\alpha<1$*, the Bayes test is to reject* $H_0(\Theta_0)$ *if and only if* θ_0 *is not contained in the lower-bound credible set* $\{\theta:\theta\ge q_\alpha(\theta|\mathbf{x})\}$*, or, equivalently, if and only if* $P(\theta\in\Theta_0|\mathbf{x})\le\alpha$*.*

(b) For the hypotheses Θ_0 *and* $\Theta_1=\Theta\backslash\Theta_0$*, under an absolutely continuous prior for* θ *and the loss function* $L_{(2)}$ *with* $a=\alpha$ *and* $b=1-\alpha$*,* $0<\alpha<1/2$*, the Bayes test is to reject* $H_0(\Theta_0)$ *if and only if there exists at least one* α *credible set that does not contain a non-null subset of* Θ_0*, or, equivalently, if and only if* $P(\theta\in\Theta_0|\mathbf{x})\le 1-\alpha$*.*

In both (a) and (b) the null hypothesis that $\theta\in\Theta_0$ is rejected if $\mathrm{P}(\theta\in\Theta_0|\mathbf{x})\le a$ for some value of a. This coincides with the tests obtained using the loss function (2.3), meaning that unlike in the point-null setting, test based on credible sets are equivalent to some standard tests in the composite hypothesis setting.

3. Large p, small n: on high-dimensional problems

3.1 Some problems when $p > n$

One of the greatest challenges in modern-day statistics is the large p, small n problem, which has given rise to the field of high-dimensional statistics. A multivariate statistical problem is called high-dimensional if the number of variables p is larger than the sample size n. An example of where such dataset occurs is the biomedical problem of detecting sets of genes that are connected to some particular disease. This is often done by using DNA microarrays to measure gene expression levels, using samples from patients as well as samples from a control group. The number of genes p studied typically range in the thousands, whereas the number of patients n range in the hundreds or less. If the gene expression levels differ between the two groups for a gene set, we may suspect that it in one way or another is connected to the disease. However, it may not always be obvious *a priori* whether the gene expression levels will be higher or lower in the control group. Indeed, the direction of the difference may even differ between genes in the set. Thus the sensible hypothesis to test is whether the gene expression levels differ, and not whether they differ in some particular direction. In the next section we will propose such a high-dimensional point-null hypothesis test, but first we explore some problems that arise in the high-dimensional setting.

Let the matrix $\boldsymbol{X} \in \mathbb{R}^{p\times n}$ represent an i.i.d. sample with no multicollinearity. The rank of $\boldsymbol{X}$ is $\min(n,p) = n$. Now, let the $p \times n$ matrix $\boldsymbol{X}_c$ represent the centred sample, i.e. the sample after the sample mean vector has been subtracted from each row. The rank of $\boldsymbol{X}_c$ is $n-1$. The sample covariance matrix is the $p \times p$ matrix

$$\boldsymbol{S} = \frac{1}{n-1}\boldsymbol{X}_c\boldsymbol{X}_c^T,$$

with $\operatorname{rank}(\boldsymbol{S}) \leq n-1$, since $\operatorname{rank}(\boldsymbol{AB}) \leq \min(\operatorname{rank}(\boldsymbol{A}), \operatorname{rank}(\boldsymbol{B}))$. Because $n - 1 < p$, the matrix $\boldsymbol{S}$ is singular and therefore not invertible.

A consequence of this is that standard methods that rely on $\boldsymbol{S}^{-1}$ no longer can be used. Examples include Hotelling's T^2 tests for mean vectors, multivariate regression, MANOVA, principal components analysis, canonical correlation analysis, linear discriminant analysis and methods based on Mahalanobis distance. Since many modern datasets in for instance genomics and proteomics suffer from the $n < p$ problem, there has been a surge of interest

in new methods tailored to the high-dimensional setting. Often these amount to replacing $\boldsymbol{S}^{-1}$ by some other estimator.

These new methods are not without their problems. For $p < n$, many multivariate statistics and procedures, including Hotelling's T^2 statistics, are affine invariant, i.e. invariant under all nonsingular linear transformations. When $p > n$ however, the group of non-singular linear transformations is transitive, meaning that no test statistic can be affine invariant [29]. To see this, let the $p \times n$ matrices $\boldsymbol{X}_1$ and $\boldsymbol{Y}_1$ represent two samples. Introduce the $p \times p$ matrices

$$\boldsymbol{A} = (\boldsymbol{X}_1, \boldsymbol{X}_2) \qquad \text{and} \qquad \boldsymbol{B} = (\boldsymbol{Y}_1, \boldsymbol{Y}_2),$$

where $\boldsymbol{X}_2$ and $\boldsymbol{Y}_2$ are arbitrary $p \times (p-n)$ matrices chosen so that $\boldsymbol{A}$ and $\boldsymbol{B}$ are non-singular. This implies that $\boldsymbol{B}\boldsymbol{A}^{-1}$ is non-singular. We have

$$\boldsymbol{A}^{-1}\boldsymbol{X}_1 = \begin{pmatrix} \boldsymbol{I}_n \\ \boldsymbol{0} \end{pmatrix}$$

from which it follows that

$$\boldsymbol{B}\boldsymbol{A}^{-1}\boldsymbol{X}_1 = (\boldsymbol{Y}_1, \boldsymbol{Y}_2) \begin{pmatrix} \boldsymbol{I}_n \\ \boldsymbol{0} \end{pmatrix} = \boldsymbol{Y}_1.$$

We have thus established the existence of a non-singular linear transformation $\boldsymbol{B}\boldsymbol{A}^{-1}$ which transforms an arbitrary sample $\boldsymbol{X}_1$ into another arbitrary sample $\boldsymbol{Y}_1$. Consequently, no test statistic can be invariant under non-singular linear transformations.

Luckily, it is possible to construct statistics that are invariant under smaller groups of transformations. The groups of non-singular diagonal matrices and the group of orthogonal transformations are perhaps of particular mathematical interest. In many applications, one of these types of invariance is of greater concern than the other. Invariance under multiplication with non-singular diagonal matrices is important if the conclusion should be independent of the scale of each variable, whereas invariance under orthogonal transformations is of importance for directional data.

3.2 A two-sample test in the large p, small n setting

Paper II is concerned with a point-null hypothesis testing problem in the high-dimensional $p > n$ setting, motivated by the gene set testing example from the previous section. We consider two random samples of size n_X and n_Y from independent p-dimensional random variables $\boldsymbol{X}$ and $\boldsymbol{Y}$, with mean vectors $\boldsymbol{\mu}_X$ and $\boldsymbol{\mu}_Y$ and covariance matrices $\boldsymbol{\Sigma}_X = \boldsymbol{\Sigma}_Y = \boldsymbol{\Sigma}$. It is allowed that $n_X + n_Y - 2 < p$. With $\boldsymbol{0}$ denoting a p-vector of 0's, our aim is to test whether the mean vectors of the two populations are equal, that is, whether $\boldsymbol{\mu}_X - \boldsymbol{\mu}_Y = \boldsymbol{0}$.

The key observation behind the test is that

$$\boldsymbol{\mu}_X - \boldsymbol{\mu}_Y = \mathbf{0} \qquad \text{if and only if} \qquad \boldsymbol{\mu}_{X_s} - \boldsymbol{\mu}_{Y_s} = \mathbf{0}_s$$

for all subvectors $\boldsymbol{X}_s$ and $\boldsymbol{Y}_s$. We can therefore look at subvectors of dimension $k < n_X + n_Y - 2$ and use Hotelling's T^2 statistic

$$T^2(\boldsymbol{X}_s, \boldsymbol{Y}_s) = (\bar{\boldsymbol{X}}_s - \bar{\boldsymbol{X}}_s)' \boldsymbol{S}_{pool}^{-1} (\bar{\boldsymbol{X}}_s - \bar{\boldsymbol{X}}_s),$$

where $\boldsymbol{S}_{pool}$ is the pooled sample covariance matrix based on $\boldsymbol{X}_s$ and $\boldsymbol{Y}_s$, to test whether $\boldsymbol{\mu}_{X_s} = \boldsymbol{\mu}_{Y_s}$.

Since the number of k-dimensional subvectors $\binom{p}{k}$ typically is very large, it is not computationally feasible to study all possible pairs of subvectors. One way out is to resort to only looking at a smaller number of randomly selected subvectors, choosing sufficiently many subvectors that any difference present will be captured with a high probability.

Rather than treating this as a multiple testing problem, we combine the subvector T^2-statistics into a single statistic $T_{rs}(\boldsymbol{X}, \boldsymbol{Y})$ by computing their average. Using the index $s = 1, \ldots, B_1$ to distinguish between the B_1 randomly selected subvectors,

$$T_{rs}(\boldsymbol{X}, \boldsymbol{Y}) = \frac{1}{B_1} \sum_{s=1}^{B_1} T^2(\boldsymbol{X}_s, \boldsymbol{Y}_s). \tag{3.1}$$

The statistic $T_{rs}(\boldsymbol{X}, \boldsymbol{Y})$ is invariant under multiplication with non-singular diagonal matrices:

Result 3 (Proposition 3, Paper II, p. 30)**.** *If $\boldsymbol{D}$ is a non-singular diagonal real $p \times p$ matrix with non-zero diagonal elements and $\boldsymbol{d} \in \mathbb{R}^p$ then $T_{rs}(\boldsymbol{X}, \boldsymbol{Y})$ is, conditional on the random subvectors chosen, invariant under the transformation $(\boldsymbol{X}, \boldsymbol{Y}) \to (\boldsymbol{DX} + \boldsymbol{d}, \boldsymbol{DY} + \boldsymbol{d})$.*

The statistic $T_{rs}(\boldsymbol{X}, \boldsymbol{Y})$ is however not invariant under orthogonal transformations. Consider for instance the case when $n_X = n_Y = 2$ and $p = 3$, with the two samples

$$\boldsymbol{X} = \begin{pmatrix} 2 & 3 \\ 2 & 3 \\ 1 & 2 \end{pmatrix} \qquad \text{and} \qquad \boldsymbol{Y} = \begin{pmatrix} 1 & 4 \\ 2 & 1 \\ 1 & 2 \end{pmatrix}.$$

Taking $k = 1$ there are three k-dimensional subspaces, so we can compute $T_{rs}(\boldsymbol{X}, \boldsymbol{Y}) = \frac{1}{3} \sum_{s=1}^{3} T^2(\boldsymbol{X}_s, \boldsymbol{Y}_s)$ using all subspaces and not just a random subsample of all possible subspaces. We find $T_{rs}(\boldsymbol{X}, \boldsymbol{Y}) = \frac{1}{3}(0 + 2 + 0) = 2/3$.

However, if we apply the orthogonal transformation $(\boldsymbol{X}, \boldsymbol{Y}) \to (\boldsymbol{HX}, \boldsymbol{HY})$, where

$$\boldsymbol{H} = \begin{pmatrix} \sqrt{2}/2 & -\sqrt{2}/2 & 0 \\ \sqrt{2}/2 & \sqrt{2}/2 & 0 \\ 0 & 0 & 1 \end{pmatrix},$$

we get

$$\boldsymbol{HX} = \begin{pmatrix} 0 & 0 \\ 2\sqrt{2} & 3\sqrt{2} \\ 1 & 2 \end{pmatrix} \quad \text{and} \quad \boldsymbol{HY} = \begin{pmatrix} -\sqrt{2}/2 & 3\sqrt{2}/2 \\ 3\sqrt{2}/2 & 5\sqrt{2}/2 \\ 1 & 2 \end{pmatrix},$$

which yields $T_{rs}(\boldsymbol{HX},\boldsymbol{HY}) = \frac{1}{3}(1/4+1/2+0) = 1/4 \neq 2/3 = T_{rs}(\boldsymbol{X},\boldsymbol{Y})$.

Lopes et al. [34] derived the asymptotic null distribution of a class of statistics which includes $T_{rs}(\boldsymbol{X},\boldsymbol{Y})$, but as is shown by simulation in Paper II (pp. 28-29) the asymptotic distribution is not a good approximation for small sample sizes. For this reason we use the permutation distribution [20] for computing the p-values of the test, or rather, use random permutations to obtain an approximation of the permutation distribution. Letting $T_{rs}^{(i)}(\boldsymbol{X}^{(i)},\boldsymbol{Y}^{(i)})$ denote the T_{rs}-statistic for the i:th random permutation $(\boldsymbol{X}^{(i)},\boldsymbol{Y}^{(i)})$, the p-value of the test based on B_2 random permutations can be written as

$$\frac{1}{B_2}\sum_{i=1}^{B_2} \mathbb{I}(T_{rs}^{(i)}(\boldsymbol{X}^{(i)},\boldsymbol{Y}^{(i)}) > T_{rs}(\boldsymbol{X},\boldsymbol{Y})). \tag{3.2}$$

Many high-dimensional two-sample tests use diagonal or trace estimators of $\boldsymbol{\Sigma}$, ignoring dependences between variables. A benefit of $T_{rs}(\boldsymbol{X},\boldsymbol{Y})$ is that it incorporates dependences up to k dimensions. This often results in higher power than the competing tests when $\boldsymbol{\Sigma}$ is not a diagonal matrix, as can be seen in Figures 5-6 of Paper II (pp. 33-34).

The equivalence between point-null tests and confidence intervals for univariate θ can be extended also to p-dimensional *confidence regions* for multivariate parameters $\boldsymbol{\theta} \in \mathbb{R}^p$, based on multivariate samples $(\boldsymbol{X},\boldsymbol{Y}) \in \mathbb{R}^{p\times(n_X+n_Y)}$, following exactly the same procedure as in Section 1.2. The hypothesis $\boldsymbol{\mu}_X - \boldsymbol{\mu}_Y = \boldsymbol{\theta}_0$ can be tested using $T_{rs}(\boldsymbol{X}-\boldsymbol{\theta}_0,\boldsymbol{Y})$, since $\boldsymbol{\mu}_X - \boldsymbol{\mu}_Y = \boldsymbol{\theta}_0 \Leftrightarrow \boldsymbol{\mu}_X - \boldsymbol{\theta}_0 = \boldsymbol{\mu}_Y$. Let $\lambda(\boldsymbol{\theta}_0,(\boldsymbol{X},\boldsymbol{Y}))$ denote the p-value (3.2) of this test. The complement of the inverted rejection region,

$$\{\boldsymbol{\theta} : \lambda(\boldsymbol{\theta},(\boldsymbol{X},\boldsymbol{Y})) > \alpha\},$$

is a p-dimensional $1-\alpha$ confidence region for $\boldsymbol{\theta} = \boldsymbol{\mu}_X - \boldsymbol{\mu}_Y$. It is currently not computationally feasible to compute this set using the random permutations approach. More accurate analytical approximations of the null distribution would therefore be needed in order to compute the confidence region.

4. Ambiguities related to hypothesis tests and confidence intervals

4.1 Optimal hypothesis tests

Next, we will discuss two questions related to hypothesis tests and confidence intervals, for which some ambiguities arise. The first question is what constitutes an optimal hypothesis test, and how this relates to optimal confidence intervals. The second question is how we should define p-values. Both these questions turn out to be of great importance when we later study tests and confidence intervals for parameters of discrete distributions, and this section therefore serves as a motivation for Papers III-VI. These questions are however also of interest in inference for absolutely continuous distributions, and hence we discuss these issues in a more general context in this section.

Optimal hypothesis test is one of the cornerstones of classical theoretical statistics. Optimality is typically measured in terms of power, that is, the probability of rejecting the null hypothesis when it is false. A level α test with rejection region $R_\alpha(\theta_0)$ is said to be *uniformly most powerful* (UMP) if for any other test, with rejection region $R'_\alpha(\theta_0)$, for all $\theta \in \Theta_1 = \Theta \backslash \Theta_0$ we have

$$\mathrm{P}_\theta(\mathbf{X} \in R'_\alpha(\theta_0)) \leq \mathrm{P}_\theta(\mathbf{X} \in R_\alpha(\theta_0)).$$

Let $\beta(\theta) = \mathrm{P}_\theta(\mathbf{X} \in R_\alpha(\theta_0))$ denote the *power function* of a test. Under the loss function (2.3) the expected loss of a test is

$$L(\theta,0)(1-\beta(\theta)) + L(\theta,1)\beta(\theta) = \mathbb{I}_{\Theta_1}(\theta)b(1-\beta(\theta)) + \mathbb{I}_{\Theta_0}(\theta)a\beta(\theta), \tag{4.1}$$

see [12, 32]. In virtually all testing problems the type I error rate $\beta(\theta_0) = \alpha$ is kept fixed. Then (4.1) can be written as

$$\mathbb{I}_{\Theta_1}(\theta)b(1-\beta(\theta)) + \mathbb{I}_{\Theta_0}(\theta)a\alpha,$$

a function which is minimized when $\beta(\theta)$ is maximized. Consequently, the UMP test is optimal under this class of loss functions. More generally, it is still optimal if we allow a and b in (2.3) to depend on θ and θ_0 in certain ways.

UMP tests do however not always exist. When no test is UMP, sometimes a UMP test can be found in a smaller class of tests. Sometimes the UMP test is randomized, meaning that it relies on randomness beyond that in $\mathbf{X}$, which causes some problems which we will return to in Section 5.3. Reviews of the rich theory of UMP tests can be found in for instance [12, 30, 31].

If $\theta \in \Theta_1$ then the *false coverage probability* of a confidence interval $I_\alpha(\mathbf{x})$ is defined as

$$\mathrm{P}_{\theta_0}(\theta \in I_\alpha(\mathbf{X})).$$

A confidence interval which minimizes the probability of false coverage is called a *uniformly most accurate* (UMA) interval. Such sets are also sometimes referred to as being Neyman-shortest, since the concept of false coverage was introduced by Neyman [37]. The intuition behind the name is that short rejection regions are expected to correspond to low false coverage probabilities and short confidence intervals.

The notion of UMA confidence sets is closely linked to UMP hypothesis tests: if a confidence set is the complement of the inverted rejection region of a UMP test, then it is UMA [12]. In this sense, there is a close connection between test-optimality and interval-optimality. False coverage is however not the standard way of comparing confidence intervals, which leads us to ask if UMP tests really lead to optimal intervals.

4.2 Do optimal tests yield optimal intervals?

The two most commonly used criteria when comparing confidence intervals are coverage probabilities (1.1) and lengths. As we will see in Section 5.2, it is not always possible to construct confidence intervals with equality in (1.1). For intervals such that said equality does not hold, it is desirable that the actual coverage is as close as possible to the nominal confidence level $1-\alpha$. Among intervals which have good coverage properties, those that are shorter are deemed to be preferable. The standard way of measuring the length of a random interval $I_\alpha(\mathbf{X}) = (L_\alpha(\mathbf{X}), U_\alpha(\mathbf{X}))$ is to use the expected length

$$\mathrm{E}_\theta[U_\alpha(\mathbf{X}) - L_\alpha(\mathbf{X})]. \tag{4.2}$$

While the link between test size and coverage is evident, it is not clear how and if the notions of false coverage and expected length are related. There is such a connection, which usually is expressed in a more general multivariate setting, using $Vol(I_\alpha(\mathbf{X}))$ to denote the volume of a confidence set. For one-dimensional connected confidence sets, i.e. confidence intervals, $Vol(I_\alpha(\mathbf{X})) = U_\alpha(\mathbf{X}) - L_\alpha(\mathbf{X})$. The following identity, due to Pratt [44], states that the expected volume of a confidence set can be computed by integrating the false coverage probabilities over $\Theta_1 = \Theta \backslash \{\theta_0\}$:

$$\mathrm{E}_{\theta_0}\Big(Vol(I_\alpha(\mathbf{X}))\Big) = \int_{\Theta_1} \mathrm{P}_{\theta_0}(\theta \in I_\alpha(\mathbf{X}))d\theta. \tag{4.3}$$

Pratt's identity can be used to derive some deep and perhaps surprising theoretical results. First of all, it can be used to show that admissibility with respect to expected length implies admissibility with respect to false coverage

probabilities [15]. The converse is however not true. A James–Stein-flavoured example of this, where the false coverage-admissible multivariate normal confidence region is dominated in expected volume, has for instance been given by Casella & Hwang [13]. Madansky [35] gave a counterexample for the mean of an exponential distribution, constructing an interval that is shorter than the UMA interval. This leads to the important observation that in general, *test-optimality does not lead to length-optimality*.

A second interesting result which follows from Pratt's identity was given by Brown et al. [11]. They considered the decision-theoretic problem of finding a confidence set which minimizes the expected volume (4.3) under the constraint that $\mathrm{P}_\theta(\theta \in I_\alpha(\mathbf{X})) \geq 1-\alpha$ for all $\theta \in \Theta$. They then showed that the resulting confidence set also minimizes the expected posterior volume among intervals which admit both frequentist and Bayesian interpretations. While there are many examples of Bayesian procedures with frequentist optimality properties, Pratt's identity can consequently be used to construct the much more uncommon converse: a inherently frequentist procedure with Bayesian optimality properties.

Coverage, false coverage and expected length are by no means the only tools for comparing confidence intervals. Other criteria include conditional properties [36], minimum coverage and mean squared coverage error [45], interval location [38] and the test-related measures p-confidence and p-bias [49].

Apart from different measures of optimality, it is often required that an interval satisfies certain fundamental monotonicity properties in $\mathbf{x}$ and α [6, 12]. One of the most important monotonicity properties is nestedness. An interval is *nested* if $I_{\alpha_0}(\mathbf{X}) \subseteq I_{\alpha_1}(\mathbf{X})$ almost surely when $\alpha_0 > \alpha_1$. If an interval is not nested, it can for instance happen that the 94 % interval is wider than the 95 % interval. An interval is *strictly nested* if $I_{\alpha_0}(\mathbf{X}) \subset I_{\alpha_1}(\mathbf{X})$ almost surely for all $\alpha_0 > \alpha_1$. If an interval is not strictly nested, it can happen that the 94 % interval equals the 95 % interval.

4.3 Two types of p-values

The p-value is defined as the probability under the null distribution of an outcome at least as extreme as the observed [21, 31]. For one-sided hypothesis tests, this definition is fairly straightforward. Consider for instance the problem of testing the null hypothesis $\theta \leq 0.2$ versus the alternative $\theta > 0.2$ based on an observation x from a $Bin(n,\theta)$-distributed random variable. In our example, outcomes which are more extreme than x are those which are larger than x, since larger values are less in line with the hypothesis that $\theta \leq 0.2$, and the p-value becomes $\sup_{\theta \leq 0.2} \mathrm{P}_\theta(X \geq x) = \mathrm{P}_{\theta=0.2}(X \geq x)$.

The definition of p-values for two-sided hypothesis tests is nowhere near as clear. On the contrary, such p-values can in fact be defined in several different

ways [21, 50]. The problem is that it is unclear what is meant by "at least as extreme as the observed". Assume for instance that we wish to test the null hypothesis $\theta = 0.3$ against $\theta \neq 0.3$ based on an observation x from a $Bin(10,\theta)$-distributed random variable and that we observe $x = 4$. Which possible outcomes are at least as extreme as x? Those that are at least as large as x, i.e. $\{4,5,\ldots,10\}$? Those which are at least as far away from $\mathrm{E}_{\theta_0}X$ as x, i.e. $\{0,1,2,4,5,\ldots,10\}$? The outcomes which have smaller probabilities under the null distribution, i.e. $\{0,1,4,5,\ldots,10\}$?

We will discuss the two most common ways of defining p-values for two-sided tests, which we will refer to as the *twice-the-smaller-tail* and *strictly two-sided* methods. As it turns out, one of these methods cause multiple problems when applied to inference for parameters of discrete distributions, which we will discuss in Section 5.8. First, however, we present the two methods in a more general setting.

Twice-the-smaller-tail p-values

Let $T(\mathbf{X})$ be a test statistic on which a two-sided test of the point-null hypothesis that $\theta = \theta_0$ is based, and let $\lambda(\theta_0,\mathbf{x})$ denote its p-value. Assume for simplicity that $T(\mathbf{x}) < 0$ implies that $\theta < \theta_0$ and that $T(\mathbf{x}) > 0$ implies that $\theta > \theta_0$. The first step in the twice-the-smaller-tail approach to computing p-values is to check whether $T(\mathbf{x}) < 0$ or $T(\mathbf{x}) > 0$. "At least as extreme as the observed" is in a sense redefined as "at least as extreme as the observed, in the observed direction". If the median of the null distribution of $T(\mathbf{X})$ is 0, then

$$\mathrm{P}_{\theta_0}(T(\mathbf{X}) \geq T(\mathbf{x}) | T(\mathbf{x}) > 0) = \min\Big(1,\ 2\cdot \mathrm{P}_{\theta_0}(T(\mathbf{X}) \geq T(\mathbf{x}))\Big),$$

i.e. twice the unconditional probability that $T(\mathbf{X}) \geq T(\mathbf{x})$. Similarly,

$$\mathrm{P}_{\theta_0}(T(\mathbf{X}) \leq T(\mathbf{x}) | T(\mathbf{x}) < 0) = \min\Big(1,\ 2\cdot \mathrm{P}_{\theta_0}(T(\mathbf{X}) \leq T(\mathbf{x}))\Big).$$

Moreover,

$$\mathrm{P}_{\theta_0}(T(\mathbf{X}) \geq T(\mathbf{x})) < \mathrm{P}_{\theta_0}(T(\mathbf{X}) \leq T(\mathbf{x})) \qquad \text{when} \qquad T(\mathbf{x}) > 0$$

and

$$\mathrm{P}_{\theta_0}(T(\mathbf{X}) \geq T(\mathbf{x})) > \mathrm{P}_{\theta_0}(T(\mathbf{X}) \leq T(\mathbf{x})) \qquad \text{when} \qquad T(\mathbf{x}) < 0.$$

Consequently, the p-value using this approach can in general be written as

$$\lambda_{TST}(\theta_0,\mathbf{x}) := \min\Big(1,\ 2\cdot \mathrm{P}_{\theta_0}(T(\mathbf{X}) \geq T(\mathbf{x})),\ 2\cdot \mathrm{P}_{\theta_0}(T(\mathbf{X}) \leq T(\mathbf{x}))\Big).$$

This definition of the p-value is frequently used also in situations where the median of the null distribution of $T(\mathbf{X})$ is not 0, despite the fact that the interpretation of the p-value as being conditioned on whether $T(\mathbf{x}) < 0$ or $T(\mathbf{x}) > 0$ is lost.

At the level α, if $T(\mathbf{x}) > 0$ the test rejects the hypothesis $\theta = \theta_0$ if

$$\lambda_{TST}(\theta_0, \mathbf{x}) = \min\Big(1, 2 \cdot \mathrm{P}_{\theta_0}(T(\mathbf{X}) \geq T(\mathbf{x}))\Big) \leq \alpha. \tag{4.4}$$

This happens if and only if the one-sided test of $\theta \leq \theta_0$, also based on $T(\mathbf{X})$, rejects its null hypothesis at the $\alpha/2$ level. By the same reasoning, it is seen that the rejection region of a level α twice-the-smaller-tail test always is the union of the rejection regions of two level $\alpha/2$ one-sided tests of $\theta \leq \theta_0$ and $\theta \geq \theta_0$, respectively. The test puts equal weight to the two types of type I errors: false rejection in the two different directions. The corresponding confidence interval is therefore also equal-tailed, in the sense that the non-coverage probability is $\alpha/2$ on both sides of the interval.

Strictly two-sided p-values

Twice-the-smaller-tail p-values are in a sense computed by looking only at one tail of the null distribution. In the alternative approach of using strictly two-sided p-values, the p-value is computed using both tails, as follows:

$$\begin{aligned}\lambda_{STT}(\theta_0, \mathbf{x}) &:= \mathrm{P}_{\theta_0}\Big(|T(\mathbf{X})| \geq |T(\mathbf{x})|\Big) \\ &= \mathrm{P}_{\theta_0}\Big(\{\mathbf{X} : T(\mathbf{X}) \leq -|T(\mathbf{x})|\} \cup \{\mathbf{X} : T(\mathbf{X}) \geq |T(\mathbf{x})|\}\Big).\end{aligned} \tag{4.5}$$

Under this approach, the directional type I error rates will in general not be equal to $\alpha/2$, so that the test might be more prone to falsely reject $H_0(\theta_0)$ in one direction than in another. On the other hand, the rejection region of a strictly-two sided test is typically smaller than its twice-the-smaller-tail counterpart. The coverage probabilities of the corresponding confidence interval therefore satisfies (1.1), but not the stronger condition

$$\mathrm{P}_\theta(\theta < L_\alpha(\mathbf{X})) = \mathrm{P}_\theta(\theta > U_\alpha(\mathbf{X})) = \alpha/2 \qquad \text{for all } \alpha \in (0,1).$$

If the null distribution of $T(\mathbf{X})$ is symmetric about 0,

$$\mathrm{P}_{\theta_0}(T(\mathbf{X}) \geq T(\mathbf{x})) = \mathrm{P}_{\theta_0}(T(\mathbf{X}) \leq -T(\mathbf{x})).$$

For $T(\mathbf{x}) > 0$, unless $T(\mathbf{X})$ has a discrete distribution,

$$\begin{aligned}\lambda_{TST}(\theta_0, \mathbf{x}) &= 2 \cdot \mathrm{P}_{\theta_0}(T(\mathbf{X}) \geq T(\mathbf{x})) \\ &= \mathrm{P}_{\theta_0}(T(\mathbf{X}) \geq T(\mathbf{x})) + \mathrm{P}_{\theta_0}(T(\mathbf{X}) \leq -T(\mathbf{x})) = \lambda_{STT}(\theta_0, \mathbf{x}),\end{aligned}$$

meaning that the twice-the-smaller-tail and strictly-two-sided approaches coincide in this case. The ambiguity related to the definition of two-sided p-values therefore only arises under asymmetric null distributions.

5. Confidence intervals for parameters of discrete distributions

5.1 Discrete data

In this section we will study inference for parameters of discrete distributions on sample spaces $\mathscr{X} \subseteq \mathbb{Z}$. We can often think of such discrete distributions as describing the outcomes of events that are summarized as counts: the number of patients cured by a drug, the number of earthquakes during a year, the number of goals in a football match.

The use of discrete distributions has always been abundant in probability and statistics, from early applications in games of chance [3], Mendelian genetics [41] and modelling of horse-induced deaths in the Prussian army [7], to modern-day applications in clinical trials, insurance, risk analysis, quality control, population ecology, epidemiology, telecommunications, opinion polls and sports modelling.

On some level, all data are discrete. Even measurements with great precision have a finite number of decimals attached to them. Indeed, nature itself has an inherent granularity, described by the quantum of action known as the Planck constant [43]. On the macroscopic scale on which we live our everyday lives and on which most measurements take place, this granularity is however not noticed, and continuous distributions will therefore often serve as accurate and mathematically tractable models of random events. In this section we focus on count data that are discrete in a sense that is meaningful from a practical point of view.

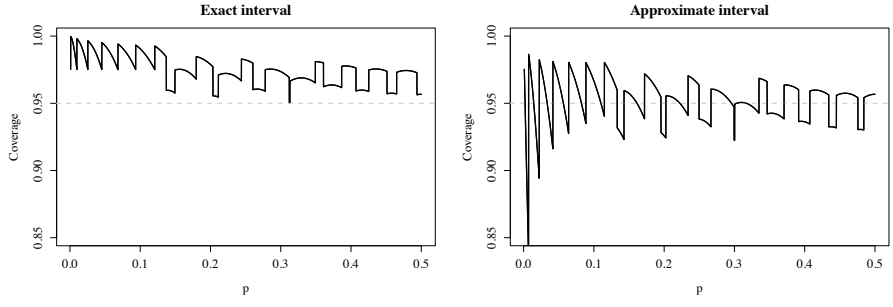
5.2 The effects of discreteness

Discreteness can be both a blessing and a curse. Under certain circumstances the binomial and hypergeometric distributions arise naturally, for purely mathematical reasons: we can often be confident that the random variables generating our data truly *are* binomial or hypergeometric.

If model accuracy is the blessing, exactness, or the lack thereof, is the curse. Consider a $1-\alpha$ confidence interval $I_\alpha(X) = (L_\alpha(X), U_\alpha(X))$ for the unknown parameter θ. When X is discrete, its coverage probability is

$$\mathrm{P}_\theta(\theta \in I_\alpha(X)) = \sum_{x \in \mathscr{X}} \mathbb{I}_{I_\alpha(x)}(\theta) \cdot \mathrm{P}_\theta(X = x). \tag{5.1}$$

Figure 5.1. Coverage of two 95 % confidence intervals for the binomial proportion p, for $n = 25$.



For a parameter θ of an absolutely continuous distribution, it is possible to construct confidence intervals with coverage probability equal to $1-\alpha$ for all values of θ and α. The coverage probability (5.1) on the other hand, can be written as

$$\mathrm{P}_\theta(\theta \in I_\alpha(X)) = \sum_{x=x_\ell(\theta)}^{x_u(\theta)} \mathrm{P}_\theta(X = x), \tag{5.2}$$

where $x_\ell(\theta)$ is the smallest $x \in \mathscr{X}$ for which $\theta \geq L_\alpha(x)$ and $x_u(\theta)$ is the greatest $x \in \mathscr{X}$ for which $\theta \leq U_\alpha(x)$. This means that (5.2) is non-constant in θ. For two parameter values $\theta_1 < \theta_2$, when $x_\ell(\theta_1) = x_\ell(\theta_2)$ and $x_u(\theta_1) = x_u(\theta_2)$, we typically have

$$\mathrm{P}_{\theta_1}(\theta_1 \in I_\alpha(X)) \neq \mathrm{P}_{\theta_2}(\theta_2 \in I_\alpha(X)),$$

when $\mathrm{P}_{\theta_1}(X = x) \neq \mathrm{P}_{\theta_2}(X = x)$. Even more troubling is the fact that if, for instance, $x_\ell(\theta_1) = x_\ell(\theta_2)$ and $x_u(\theta_1) + 1 = x_u(\theta_2)$, there is a jump[1] in (5.2) between θ_1 and θ_2 caused by the addition of the point-mass $\mathrm{P}_{\theta_2}(X = x_u(\theta_1) + 1)$. The coverage probability (5.1) is not only non-constant in θ, but also discontinuous. These problems are universal to confidence intervals and type I error rates of tests for parameters of discrete distributions. The problems are illustrated in the binomial setting in Figure 5.1.

We will distinguish between exact and approximate confidence intervals. This terminology is perhaps a bit unfortunate, as it is the intervals' coverage probabilities, and not the intervals themselves, which are exact or approximate. The point-null hypothesis tests corresponding to these intervals are similarly also called exact or approximate.

An interval is *exact* if

$$\inf_{\theta \in \Theta} \mathrm{P}_\theta(\theta \in I_\alpha(X)) \geq 1 - \alpha. \tag{5.3}$$

[1] A formal proof of this runs along the lines of the proof of Proposition 3 (a) in Paper VI.

Exactness is, in this setting, defined as the coverage probability being bounded by, rather than equal to, $1-\alpha$. An interval is *approximate* if

$$\mathrm{P}_\theta(\theta \in I_\alpha(X)) \approx 1-\alpha \tag{5.4}$$

and, typically, $\mathrm{P}_\theta(\theta \in I_\alpha(X)) \to 1-\alpha$ as the sample size $n \to \infty$. The coverage probability of an approximate interval can drop below $1-\alpha$ for some (or even most) values of θ, as illustrated in Figure 5.1.

Several authors have criticized exact intervals for being too wide and too conservative, since their coverage probabilities often are noticeably greater than $1-\alpha$ [1, 8, 40]. Approximate intervals are almost always shorter than exact intervals. It is further argued that approximate methods, such as the bootstrap and MCMC, already are widely used for other problems and therefore are in line with current statistical practice. Despite this, exact confidence intervals and hypothesis test continue to be popular e.g. in medical statistics. One reason for this may be that since we often can be certain that distributions such as the binomial and hypergeometric distributions give accurate descriptions of the process which we are modelling, there is no need to resort to approximations, unlike in situations where bootstrapping and MCMC typically are used.

5.3 Randomized tests and intervals

Exact and approximate tests and intervals are not the only options for inference about parameters of discrete distributions. A third school argues that it is the duty of a statistician to be wrong precisely as often as stated, and that confidence intervals which have a coverage probability that is not equal to $1-\alpha$ are unacceptable. Intervals with $\mathrm{P}_\theta(\theta \in I_\alpha(X)) = 1-\alpha$ for all θ can be obtained by conditioning on an auxiliary random variable [30, 48]. Such intervals are called *randomized* intervals. Their corresponding tests, dubbed randomized tests, arise naturally as UMP tests for parameters of discrete distributions [30, 32], and are an essential part of the classical Neyman–Pearson theory.

Despite theoretically appealing optimality properties, randomized tests and intervals are seldom used in practice. The reason for this is that the auxiliary random variable must be determined independently of the data, causing different statisticians to produce different intervals and different test results when applying the same method to the same sample.

Recently, Decrouez & Hall [17] proposed a method for constructing randomized confidence intervals which give the same result for all statisticians, removing the ambiguity caused by conditioning on an auxiliary random variable. This is achieved by letting the randomization be determined by the order in which the observations in the sample $\mathbf{X} = (X_1, \ldots, X_{n_1}, X_{n_1+1}, \ldots, X_n)$ were obtained. Decrouez & Hall argue that this should be more appealing to practi-

tioners since the randomization comes from the sample itself rather than some auxiliary random variable.

The Decrouez–Hall procedure amounts to computing some standard confidence interval using a new variable $\tilde{X}$ instead of summary statistic $X = \sum_{i=1}^{n} X_i$. This new variable is obtained by splitting the sample in two parts, $(X_1, \ldots, X_{n_1})$ and $(X_{n_1+1}, \ldots, X_n)$, and computing $\tilde{X}$ by weighing the information in the two subsamples in a particular way. When the X_i are i.i.d. Bernoulli variables so that $X = \sum_{i=1}^{n} X_i$ is binomial, $\tilde{X}$ is defined as

$$\tilde{X} = \frac{n}{2}\Big(\frac{1}{n_1}\sum_{i=1}^{n_1} X_i + \frac{1}{n_2}\sum_{i=n_1+1}^{n} X_i\Big),$$

where $n_1 = [n/2 + 0.15n^{3/4}]$ and $n_1 \neq n_2 = n - n_1$. While this procedure takes care of the ambiguity problems, the resulting confidence interval does not satisfy (1.1), but merely (5.4). It is moreover not based on classical optimality results.

For the binomial, negative binomial and Poisson distributions, inference is typically based on the sufficient statistic $X = \sum_{i=1}^{n} X_i$. In Paper III it is shown that for the binomial distribution the Decrouez–Hall procedure is mathematically equivalent to adding discrete noise to X.

Result 4 (Paper III, p. 66)**.** *Let $X \sim Bin(n,p)$ and $Z \sim Hypergeometric(n, X, n_1)$. Then*

$$\tilde{X} \stackrel{d}{=} X + Y \qquad \textit{where} \qquad Y = \frac{n}{2n_1}Z - \frac{n}{2n_2}Z + \frac{n_1 - n_2}{2n_2}X.$$

In the remainder of Paper III, the Decrouez–Hall interval is then compared to classic randomized intervals in the binomial setting, revealing that it can be improved upon by using other methods and that the randomization has a larger impact on the Decrouesz–Hall interval than on competing intervals. Finally, it is argued that the order in which the observations were obtained must be considered to be an auxiliary variable, since it contains information not contained in the sufficient statistic. The Decrouez–Hall interval does therefore arguably not solve the problem related to conditioning on an auxiliary random variable.

5.4 The Wald interval for a binomial proportion

In the next few sections we move from confidence intervals for parameters of general discrete distributions to an important special case, namely confidence intervals for the binomial proportion p. Let $X \sim Bin(n,p)$, $\hat{p} = X/n$ and $\hat{q} = 1 - \hat{p}$, and let $z_{\alpha/2}$ denote the upper $\alpha/2$-quantile of the standard normal distribution. A common confidence interval for p, which is presented in virtually all introductory statistics courses, is the Wald interval

$$\hat{p} \pm z_{\alpha/2}\sqrt{\hat{p}\hat{q}/n}, \tag{5.5}$$

thus named because it is obtained by inverting the Wald test [51] for p. This section is devoted to the properties of the Wald interval, which we will see suffers from several problems. This serves as a motivation for Papers IV-VI, in which alternatives to the Wald interval are studied.

Upon further inspection, (5.5) may look a bit strange. The quantile used for computing the limit is the same for both the lower and the upper bound, but the binomial distribution is only symmetric about p when $p = 1/2$, which seems to imply that different quantiles should be used. On the other hand, if we let $X_1, X_2, \ldots, X_n$ be i.i.d. Bernoulli variables then the total number of successes $X = X_1 + X_2 + \cdots + X_n$ is not only binomial but, by the central limit theorem, also approximately normal and thus approximately symmetric for large n. It is this asymptotic normality that motivates the formula (5.5).

Over the years, much study has gone into determining convergence rates in the central limit theorem. An important tool is the Edgeworth expansion [4, 22], which can be used to describe the approximation error when a cumulative distribution function of a standardized statistic S_n is approximated by the standard normal distribution function $\Phi(x)$. The first few terms of this asymptotic expansion typically include the skewness and kurtosis of the random variable.

When the central limit theorem is applied to discrete random variables, skewness and kurtosis are no longer sufficient to quantify the rate of convergence, as the discreteness affects the accuracy of the approximation. For the binomial distribution, consider the standardized statistic $S_n = n^{1/2}(\hat{p} - p)/\sqrt{p(1-p)}$. In order to expand the distribution function of S_n, we introduce some notation. Let $h(x) = x - \lfloor x \rfloor$, i.e. let $h(x)$ be the fractional part of x. Furthermore, let $g(p,x) = g(p,x,n) = h(np + x(npq)^{1/2})$ and let $\phi(x)$ be the density function of the standard normal distribution. Then an Edgeworth-type expansion of S_n is given by the following expression from [9], which follows directly from Theorem 23.1 in [4].

$$\begin{aligned} P(S_n \leq x) =& \Phi(x) + \frac{1}{6}(1-2p)(1-x^2)\phi(x)(npq)^{-1/2} \\ &+ \left(\frac{1}{2} - g(p,x)\right)\phi(x)(npq)^{-1/2} \\ &+ \left[(4pq-1)x^5 + (7-22pq)x^3 + (6pq-6)x\right]\phi(x)(72npq)^{-1} \\ &+ \left[\frac{1}{6}(1-2p)(x^2-3)\left(\frac{1}{2} - g(p,x)\right)\right. \\ &\qquad \left. - \left(\frac{1}{2}g^2(p,x) - \frac{1}{2}g(p,x) + \frac{1}{12}\right)\right]x\phi(x)(npq)^{-1} \\ &+ O(n^{-3/2}). \end{aligned} \tag{5.6}$$

The terms containing $g(p,x)$ represent the rounding error due to the discreteness of the binomial distribution, while the other terms represent the error due

to skewness and kurtosis. Expansions for other discrete distributions are given in [10].

When an Edgeworth expansion for an absolutely continuous statistic is used to compute the coverage of a two-sided confidence interval it is seen that the $n^{-1/2}$ term cancels out. In the binomial case however, the $(\frac{1}{2} - g(p,x))\phi(x)(npq)^{-1/2}$ term of (5.6) does not cancel, so that the first term in the coverage error expansion is $(g(p,-\ell) - g(p,\ell))\phi(\ell)(npq)^{-1/2}$ for some ℓ. The rounding error dominates the coverage error, which is of a lower order than in the absolutely continuous case.

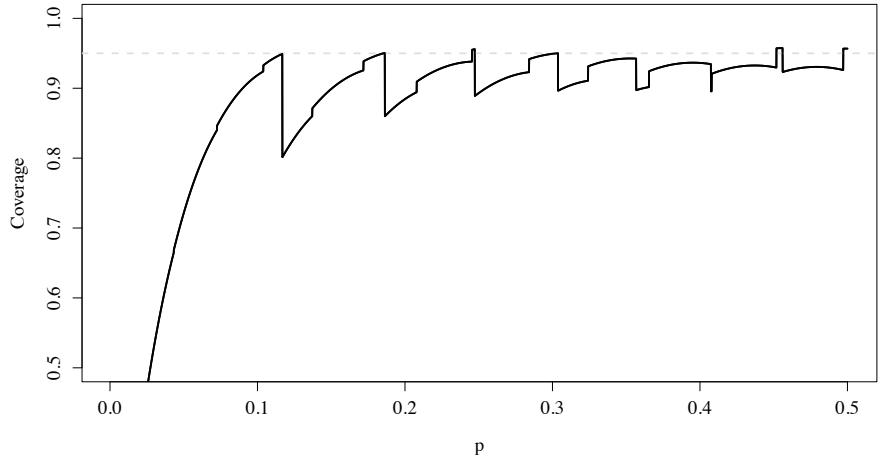
Turning our attention to the Wald interval, which we will denote $I_\alpha^W(X) = (\ell, u)$, we can using (5.6) after some algebra [9] arrive at the following expression for the coverage probability of $I_\alpha^W(X)$:

$$
\begin{aligned}
P_p(p \in I_\alpha^W(X)) =& 1 - \alpha + \Big(g(p,\ell) - g(p,u)\Big)\phi(z_{\alpha/2})n^{-1/2} \\
&+ \Big[\Big(\frac{1}{9} - \frac{1/3 + (1-2p)^2}{12pq}\Big)z_{\alpha/2}^5 \\
&\qquad + \Big(\frac{7/9 - 1}{4pq} - \frac{11}{18}\Big)z_{\alpha/2}^3 \\
&\qquad + \Big(\frac{1}{6} - \frac{1}{6pq}\Big)z_{\alpha/2}\Big]\phi(z_{\alpha/2})n^{-1} \\
&+ \Big[-(1-2p)\Big(\frac{z_{\alpha/2}^2}{3} + \frac{1}{2}\Big)\Big(1 - g(p,\ell) - g(p,u)\Big) \\
&\qquad + \frac{1}{2}\Big(-g^2(p,\ell) - g^2(p,u) + g(p,\ell) \\
&\qquad + g(p,u) - \frac{1}{3}\Big)\Big]z_{\alpha/2}\phi(z_{\alpha/2})(npq)^{-1} \\
&+ O(n^{-3/2}).
\end{aligned}
$$

Further study of the above expression reveals that $I_\alpha^W(X)$ has a surprisingly poor performance [9]. One example is given in Figure 5.2, where we can see that the interval is almost anticonservative for $n = 25$.

Despite its popularity, $I_\alpha^W(X)$ leaves much to be desired and can not be recommended for use outside of the classroom. Luckily, there are several alternatives to $I_\alpha^W(X)$ that simultaneously have coverage closer to the nominal $1-\alpha$ and shorter expected length. In particular, some intervals that often are recommended [8, 39] are the Wilson [53], Jeffreys [8], mid-p [28] and Agresti & Coull [1] intervals. With the exception of the mid-p interval, these intervals are described in Paper V (p. 822).

Figure 5.2. Coverage of the 95 % Wald interval for the binomial proportion p, for $n = 25$.



5.5 The Clopper–Pearson interval

Some of the alternatives to the Wald interval are constructed without the use of normal approximations. Chief among these is the exact Clopper-Pearson interval $I_\alpha^{CP}(X) = (p_L, p_U)$ [14]. It is the fiducial [18, 52] interval for p, which from a frequentist perspective can be though of as being based on the inversion of the standard twice-the-smaller-tail binomial test, i.e. (4.4) with $T(X) = X$. Hence, the lower limit is given by the value of p_L such that

$$\sum_{k=X}^{n} \binom{n}{k} p_L^k (1-p_L)^{n-k} = \alpha/2$$

and the upper limit is given by the p_U such that

$$\sum_{k=0}^{X} \binom{n}{k} p_U^k (1-p_U)^{n-k} = \alpha/2.$$

The computation of p_L and p_U is simplified by the following equality from [23]. Let $f(t,r,s)$ be the density function and $F(t,r,s)$ be the cumulative distribution function of a $Beta(r,s)$ random variable. Then

$$\sum_{k=X}^{n} \binom{n}{k} p^k (1-p)^{n-k} = \int_0^p f(t,X,n-X+1)dt = F(p,X,n-X+1).$$

Thus $p_L = F^{-1}(\alpha/2, X, n-X+1)$ and $p_U = F^{-1}(1-\alpha/2, X+1, n-X)$. The endpoints of $I_\alpha^{CP}(X)$ are beta quantiles:

$$I_\alpha^{CP}(X) = \Big(B(\alpha/2, X, n-X+1), \quad B(1-\alpha/2, X+1, n-X)\Big). \tag{5.7}$$

In Paper VI an optimality result is derived for fiducial intervals for parameters of a class of discrete distributions which includes the binomial distribution. We state the result for the Clopper–Pearson interval here.

Result 5 (Proposition 10, Paper VI, p. 16)**.** *Among exact equal-tailed confidence intervals for the binomial parameter p, for all* $\alpha \in (0,1)$ *the Clopper–Pearson interval minimizes the expected length as well as the length for all* $x \in \{0,1,\ldots,n\}$.

5.6 Coverage-adjustments of the Clopper–Pearson interval

Paper IV is concerned with adjustments of the Clopper–Pearson interval (5.7). The aim of Paper IV is to adjust the interval so that the *mean* coverage and not the *minimum* coverage (5.3) equals $1-\alpha$. The idea is essentially to sacrifice the exactness in order to obtain an interval which has coverage close to $1-\alpha$ in some average sense, which we will define shortly.

Similar adjustments have previously been proposed informally by Reiczigel [45] and Santner [47], who adjusted different intervals to have mean coverage $1-\alpha$. They computed the mean with respect to a uniform distribution for p, although they did not state this explicitly. We consider a more general framework, where the adjustment can be made with respect to any absolutely continuous distribution on the interval $(0,1)$. The distribution can either be determined in advance (a prior distribution) or be conditioned on X (a posterior distribution). We will refer to the interval as being prior-adjusted or posterior-adjusted, depending on which type of distribution that is used for the adjustment.

The idea of using prior or posterior distributions to adjust the interval has obvious connections to Bayesian statistics. It is however intended to be used in a frequentist setting, where there might be some vague prior information available. As an example, if one is interesting in knowing how large a proportion of patients using a particular drug that suffer from some side effect is, it is often known beforehand whether this proportion is roughly 2 % or roughly 20 %.

Let $f(\cdot)$ be a density function on $(0,1)$. A mean coverage-adjusted $1-\alpha$ Clopper-Pearson interval $I_\alpha^{ACP}(X) = (p_L, p_U)$ is defined by the unique solution to

$$\sum_{k=X}^{n} \binom{n}{k} p_L^k (1-p_L)^{n-k} = \alpha'/2 \qquad \text{and} \qquad \sum_{k=0}^{X} \binom{n}{k} p_U^k (1-p_U)^{n-k} = \alpha'/2$$

where α' satisfies

$$\begin{aligned} C(\alpha',n) &= \int_0^1 P(p \in I_{\alpha'}^{CP}(X)) \cdot f(p)dp \\ &= \int_0^1 \sum_{x=0}^{n} \mathbb{I}_{I_{\alpha'}^{CP}(x)}(p) \binom{n}{x} p^x (1-p)^{n-x} \cdot f(p)dp = 1-\alpha, \end{aligned} \tag{5.8}$$

i.e. where the mean coverage of $I_\alpha^{ACP}(X)$ with respect to f is $1-\alpha$. In Paper IV, Beta distributions are used for the adjustment.

Result 6 (Lemma 1 and Theorem 1, Paper IV, p. 293)**.** *If f is the density of a $Beta(r,s)$ distribution, $C(\alpha',n)$ is continuously differentiable and strictly decreasing in α'. Moreover, there exists a unique solution α' to the equation $C(\alpha',n,r,s) = 1-\alpha$, and this solution is bounded from below by α. When n, r and s are fixed, the solution α' is a continuous and strictly increasing function of α.*

Bearing Result 5 in mind, one might expect that the adjusted Clopper–Pearson interval should have near-optimal properties. In Paper IV, for $n \leq 100$, the intervals are compared to the three intervals that most often are recommend in the literature: the Wilson, Jeffreys and mid-p intervals [8, 39]. The behaviour of $I_\alpha^{ACP}(X)$ is studied numerically for two different choices of f:

- The uniform prior $Beta(1,1)$, for which we find that the coverage of $I_\alpha^{ACP}(X)$ is comparable to that of the competing intervals for most combinations of n and p. Its expected length is shorter than that of Wilson and mid-p intervals for most n and p, and shorter than that of the Jeffreys interval when p is not close to 0 or 1. The prior-adjusted $Beta(1,1)$ Clopper-Pearson interval can thus be recommended for general use.
- The posterior distribution obtained using the non-informative Jeffreys prior $Beta(1/2,1/2)$, i.e. $Beta(1/2+X, 1/2+n-X)$, which yields an interval that has coverage very similar to that of the Jeffreys interval. Its expected length is as short as or shorter than the expected length of the Jeffreys interval and is uniformly shorter than that of the mid-p interval. The Wilson interval has shorter expected length when p is close to $1/2$. The posterior-adjusted $Beta(1/2,1/2)$ Clopper-Pearson interval can therefore be recommended when it is suspected that p is close to 0 or 1.

Both adjustments can have quite large impact on the interval. When $n = 20$ and $X = 5$, for example, the 95 % Clopper–Pearson interval is (0.087,0.491), whereas the prior-adjusted $Beta(1,1)$ interval is (0.104,0.456) and the posterior-adjusted $Beta(1/2,1/2)$ interval is (0.101,0.461). Further examples are given in Table 2 of Paper IV (p. 300).

5.7 Sample size determination and excess length of the Clopper–Pearson interval

In Paper V various asymptotic expansions for the Clopper–Pearson interval are presented, which are used to compute sample sizes and to compare the interval to popular approximate intervals.

There is no closed-form expression for the Clopper–Pearson interval, since the beta quantile function in (5.7) cannot be written in a closed form. Next we give an asymptotic expansion of the Clopper–Pearson interval bounds, which can be used to get a good closed-form approximation when $n \geq 40$.

Result 7 (Theorem 1, Paper V, p. 820)**.** *Let $X \in \{1,2,\ldots,n-1\}$ be fixed and let $\hat{p} = X/n$, $\hat{q} = 1-\hat{p}$ and $z_{\alpha/2}$ be the upper $\alpha/2$-quantile of the standard normal distribution. Then the bounds of the Clopper–Pearson interval $I_\alpha^{CP}(X) = (p_L, p_U)$ are*

$$p_L = \hat{p} - n^{-1/2} z_{\alpha/2} (\hat{p}\hat{q})^{1/2} + (3n)^{-1}\Big(2(1/2-\hat{p})z_{\alpha/2}^2 - (1+\hat{p})\Big) + O(n^{-3/2})$$

and

$$p_U = \hat{p} + n^{-1/2} z_{\alpha/2} (\hat{p}\hat{q})^{1/2} + (3n)^{-1}\Big(2(1/2-\hat{p})z_{\alpha/2}^2 + 1 + \hat{q}\Big) + O(n^{-3/2}).$$

Perhaps more importantly, these asymptotic expressions can also be used to derive an asymptotic expansion for the expected value of the length $L_{CP} = p_U - p_L$ of $I_\alpha^{CP}(X)$.

Result 8 (Theorem 2, Paper V, p. 823)**.** *The expected length of the Clopper–Pearson interval is*

$$\begin{aligned} E(L_{CP}) =& 2z_{\alpha/2} n^{-1/2} (pq)^{1/2} + n^{-1} \\ &+ n^{-3/2} (pq)^{-1/2} \frac{z_{\alpha/2}}{18}\left(z_{\alpha/2}^2 - \frac{5}{2} - 17pq - 13pqz_{\alpha/2}^2\right) + O(n^{-2}). \end{aligned}$$

This result can be used to get a good approximation of the sample size n that is required to obtain a certain expected length d. Ignoring the higher terms of the above expansion, we obtain the second-order approximation

$$E(L_{CP}) \approx 2z_{\alpha/2} n^{-1/2} (pq)^{1/2} + n^{-1}.$$

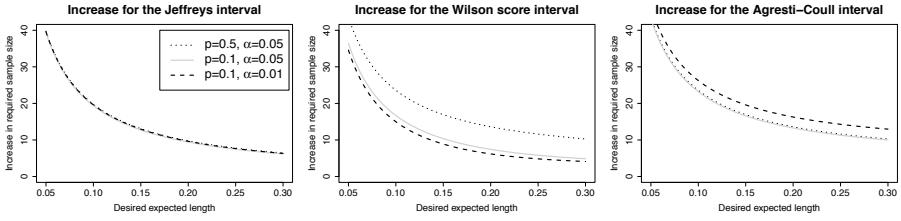
If we have an initial guess p_0 for p, we can then solve the equation

$$2z_{\alpha/2} n^{-1/2} (p_0 q_0)^{1/2} + n^{-1} = d$$

in order to obtain the required sample size.

Result 9 (Formula (7), Paper V, p. 824)**.** *The sample size n required to obtain a $1-\alpha$ Clopper–Pearson interval with expected length d for the initial guess*

Figure 5.3. Approximations of the increase in required sample size when using the Clopper–Pearson interval instead of the Jeffreys, Wilson score and Agresti–Coull intervals. Reprint of Figure 3 in Paper V (Thulin, M. (2014). The cost of using exact confidence intervals for a binomial proportion. *Electronic Journal of Statistics*, **8**, 817–840. DOI: 10.1214/14-EJS909. Reprinted with permission from the publisher.).



p_0 *is*

$$n = \left\lceil \frac{2z^2_{\alpha/2}p_0q_0 + 2z_{\alpha/2}\sqrt{z^2_{\alpha/2}p_0^2q_0^2 + dp_0q_0} + d}{d^2} \right\rceil.$$

Previously existing methods for determining sample sizes for the Clopper–Pearson interval have required extensive computations, rendering them computer-intensive black box methods. In contrast, in the above expression we see the impact of α, p_0 and d on the sample size, and the sample size computations become simple.

A second application of Result 8 is comparisons with shorter competing approximate intervals, described in Paper V (p. 822).

Result 10 (Corollary 1, Paper V, p. 826)**.** *If L_J denotes the length of the Jeffreys interval,*

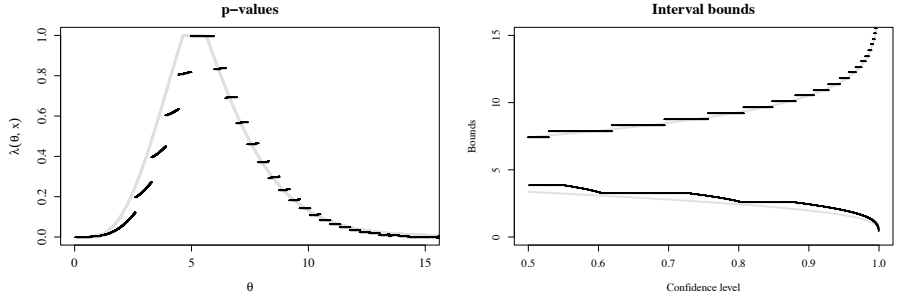
$$E(L_{CP}) = E(L_J) + n^{-1} + O(n^{-2}),$$

and if L_A denotes the length of the Wilson or Agresti–Coull interval,

$$E(L_{CP}) = E(L_A) + n^{-1} + O(n^{-3/2}).$$

Similarly, one can compare the sample size required to obtain a desired expected length. The increase in sample size when the Clopper–Pearson interval is used instead of an approximate interval is shown in Figure 5.3. Interestingly, when compared to the Jeffreys interval the increase is more or less independent of α and p. The cost of using the Clopper–Pearson interval instead of the Jeffreys interval is, in terms of required sample size, constant for a fixed expected length d.

Figure 5.4. p-values and interval bounds for a strictly two-sided test/interval (black) and a twice-the-smaller-tail test/interval (grey), for the mean θ of a Poisson distribution, based on the observation $x = 5$.



5.8 Problems related to strictly two-sided tests and intervals

We now return to a more general discrete setting. Paper VI is concerned with strictly two-sided hypothesis tests and confidence intervals for a parameter $\theta \in \Theta \subseteq \mathbb{R}$ of a distribution belonging to a class $\mathscr{P}(\Theta, \mathscr{X})$, with $\mathscr{X} \subseteq \mathbb{Z}$. The definition of $\mathscr{P}(\Theta, \mathscr{X})$ (given on p. 3 of Paper VI) is somewhat technical; we simply mention that it contains the class of regular discrete one-parameter exponential families with a monotone likelihood ratio (Proposition 1, Paper VI, p. 3), and thus includes for instance binomial, negative binomial and Poisson distributions, with suitable parametrizations.

If a distribution $P_\theta \in \mathscr{P}(\Theta, \mathscr{X})$ then it is indistinguishable from $P_{\theta+\varepsilon} \in \mathscr{P}(\Theta, \mathscr{X})$ when ε is infinitesimal. It is therefore desirable that evidence against P_θ, for instance as measured by the p-value of a point-null hypothesis test, is indistinguishable from the evidence against $P_{\theta+\varepsilon}$. The next result states that this need not be so for p-values $\lambda_{STT}(\theta, x)$, defined as in (4.5), belonging to strictly two-sided tests. Moreover, concerning the property of nestedness which we discussed in Section 4.2, strictly two-sided confidence intervals are never strictly nested.

Result 11 (Proposition 3, Paper VI, p. 5)**.** *Assume that $P_\theta \in \mathscr{P}(\Theta, \mathscr{X})$. Let $\lambda_{STT}(\theta, x)$ be the p-value function of a strictly two-sided point-null hypothesis test and let $I_\alpha(x)$ denote its corresponding strictly two-sided confidence interval. Then for any $x \in \mathscr{X}$,*

(a) $\lambda_{STT}(\theta, x)$ is not continuous in θ,
(b) The bounds of $I_\alpha(x)$ are not strictly monotone in α,
(c) $I_\alpha(x)$ is not strictly nested.

The results are illustrated in Figure 5.4. These problems have in the past been pointed out for specific intervals for a binomial proportion [5, 50]. Our

contribution is to show that they occur for a large class of strictly two-sided tests and intervals, and for a much larger class of distributions.

Another problem for strictly two-sided tests is that $\lambda_{STT}(\theta,x)$ typically lacks bimonotonicity, meaning that the evidence against θ_1 can be stronger than the evidence against θ_2, even though θ_1 is more in line with the observation x. This also means that the bounds of the corresponding confidence interval have jumps when viewed as a function of α, as can be seen in Figure 5.4. In Paper VI, conditions for when $\lambda_{STT}(\theta,x)$ lacks bimonotonicity are derived. Here we only state the result for the Poisson distribution, for which the condition can be expressed in a comparatively simple form. The expression involves two functions $k_1(\theta,x)$ and $k_2(\theta,x)$, describing the set used to compute (4.5):

$$\begin{aligned}&\{k \in \mathscr{X} : |T(k)| \geq |T(x)|\}\\ &= \{k \in \mathscr{X} : k \geq k_1(\theta,x)\} \cup \{k \in \mathscr{X} : k \leq k_2(\theta,x)\}.\end{aligned}$$

Result 12 (Proposition 7, Paper VI, p. 14)**.** *For $X \sim Pois(\theta)$, the p-value function $\lambda_{STT}(\theta,x)$ belonging to a strictly two-sided test is bimonotone in θ if and only if there does not exist (θ,x) such that either*

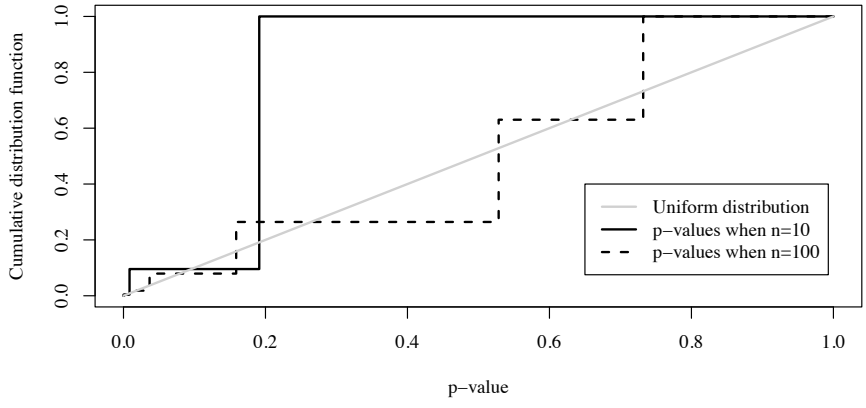
- $\theta < \inf\{\theta : \lambda_{STT}(\theta,x) = 1\}$ *and* $\theta < \left(\frac{(k_1(\theta,x)-1)!}{k_2(\theta,x)!}\right)^{1/(k_1(\theta,x)-k_2(\theta,x)-1)}$, *or*
- $\theta > \sup\{\theta : \lambda_{STT}(\theta,x) = 1\}$ *and* $\theta > \left(\frac{(k_1(\theta,x)-1)!}{k_2(\theta,x)!}\right)^{1/(k_1(\theta,x)-k_2(\theta,x)-1)}$.

6. Future research

There are several open problems related to Papers I-VI, many of which would be natural continuations of the research presented in this thesis. Some examples of such open problems are listed below.

- The random subspaces methodology introduced in Paper II should lend itself well to many different high-dimensional testing problems. Examples include k-sample tests for the mean, tests for covariance matrices (sphericity or equality of two matrices), tests for regression coefficients and tests for normality.
- As mentioned at the end of Section 3.2, the main problem with the random permutations approach of Paper II is that it can become very computer-intensive. The obvious alternative is to base the test on the asymptotic null distribution, but as is shown in Paper II, this can be a very bad approximation of the finite-sample null distribution. A highly interesting problem is therefore to obtain asymptotic expansions for the null distribution, which hopefully could be used to yield better approximations of the finite-sample null distribution and thus greatly reduce the computational complexity of the test.
- In for instance environmental chemistry and proteomics it is common that measurements fall below the laboratory's detection limit, leading to type I left-censored data. This censoring makes inference about moments of the underlying distributions much more difficult, and methods tailored to non-censored data will not perform well unless great care is taken. High-dimensional dataset of this kind are becoming more common, and it would be of great interest to adapt the random subspaces test of Paper II also to this setting, taking the information provided by the type I-censoring into account.
- The coverage-adjustment method used in Paper IV should be applicable also for confidence intervals for parameters of other discrete distributions. It could for instance be applied to confidence intervals for Poisson means or for the difference of two binomial proportions.
- Determining α' for the coverage-adjusted Clopper–Pearson intervals can be somewhat time-consuming, particularly for larger n. Consequently, sample size determination for the coverage-adjusted Clopper–Pearson interval can take a long time. It would be interesting to see if the asymptotic results for the unadjusted Clopper–Pearson interval, obtained in Paper V, can be extended to the adjusted intervals. This could potentially speed up sample size determination for the adjusted intervals considerably.

Figure 6.1. Cumulative distribution functions of p-values of two-sided binomial tests. For $X \sim Bin(n, \theta)$, the hypothesis $\theta = 0.01$ is tested against $\theta \neq 0.01$ for $n \in \{10, 100\}$. The p-values are the twice-the-smaller-tail p-values corresponding to the Clopper–Pearson interval. For reference, the cumulative distribution function of the uniform distribution is plotted in grey.



- In Section 1.2 it is mentioned that if a test is based on a test statistic with a completely specified absolutely continuous null distribution, then the corresponding p-value is uniformly distributed under H_0. This is no longer true if the test statistic has a discrete null distribution. In fact, the p-value distribution may be very far from uniform, as is shown in the example in Figure 6.1. It might be interesting to see whether studies into the behaviour of p-value distribution can shed new light on the behaviour of hypothesis tests for parameters of discrete distributions.

7. Summary in Swedish

Den här avhandlingen handlar om konfidensintervall och tester av enkla nollhypoteser. Inom frekventistisk statistik finns en tydlig ekvivalens mellan konfidensintervall och sådana tester: varje konfidensintervall motsvaras av en familj av tester. Någon liknande ekvivalens har inte funnits inom bayesiansk statistik, där man istället studerat tester av enkla nollhypoteser och den bayesianska motsvarigheten till konfidensintervall, kredibilitetsmängder, separat. Trots det har hypotestester som baseras på kredibilitetsmängder ibland föreslagits som *ad hoc*-lösningar. Artikel I handlar om hur bayesiansk hypotesprövning med kredibilitetsmängder kan motiveras med hjälp av beslutsteori. Huvudresultatet är att det med en viss förlustfunktion råder en ekvivalens mellan kredibilitetsmängder och en typ av generaliserade tester av enkla hypoteser, där nollhypotesen avfärdas i en av två möjliga riktningar. Om nollhypotesen att $\theta = \theta_0$ avfärdas så säger man med ett sådant test alltså inte bara att $\theta \neq \theta_0$, utan antingen att $\theta < \theta_0$ eller att $\theta > \theta_0$.

Artikel II handlar om ett test av den enkla hypotesen att skillnaden mellan två väntevärdesvektorer är 0. Speciellt avhandlas det högdimensionella fallet där antalet observationer n är lägre än antalet variabler p. Högdimensionella problem är av särskilt intresse eftersom det blir allt vanligare med datamaterial där $n < p$ inom exempelvis genetik och proteomik. Klassiska hypotestester, som exempelvis Hotelling's T^2-test för skillnaden mellan två väntevärdesvektorer, går inte längre att använda då $n < p$, vilket gör att nya metoder behöver utvecklas. I artikeln föreslås en metod som bygger på att datamaterialet projiceras till ett antal slumpmässigt utvalda delrum, vart och ett med dimension lägre än n. I varje delrum kan Hotelling's T^2-statistika beräknas. Därefter beräknas medelvärdet av dessa statistikor, vilket ger den slutgiltiga teststatistikan. Simuleringar tyder på att då variablerna är mer än försumbart korrelerade så har det nya testet högre styrka än tidigare föreslagna tester.

De fyra avslutande artiklarna, III-VI, rör alla på olika sätt konfidensintervall för parametrar i diskreta fördelningar. I artikel III diskuteras randomiserade konfidensintervall, och hur stor inverkan randomiseringen kan ha på intervallets gränser.

Artiklarna IV-V rör Clopper–Pearson-intervallet för en binomialandel. De vanligaste typerna av konfidensintervall för parametrar i diskreta fördelningar kallas exakta och approximativa. Exakta intervalls täckningsgrad är minst $1 - \alpha$, medan approximativa intervalls täckningsgrad bara är approximativt lika med $1 - \alpha$. Exakta intervall, dit Clopper–Pearson-intervallet hör, är i

regel längre än approximativa intervall. I artikel IV föreslås en täckningsgradskorrigering för intervallet, med hjälp av viktfunktioner som kan liknas vid *a priori*-fördelningar. Det korrigerade intervallet är kortare än det ursprungliga intervallet, och har approximativ täckningsgrad $1 - \alpha$. Det korrigerade intervallet jämförs med andra approximativa intervall för några olika viktfunktioner, och det visas att det ofta både har högre täckningsgrad och kortare längd än de andra intervallen. I artikel V studeras asymptotiska utvecklingar för Clopper–Pearson-intervallets förväntade längd. Utvecklingarna används dels för att härleda formler för att bestämma stickprovsstorlekar och dels för jämförelser med approximativa intervall. I dessa jämförelser beskrivs ”kostnaden” för att använda det exakta Clopper–Pearson-intervallet istället för ett approximativt alternativ, dels mätt i stickprovsstorlek och dels mätt i förväntad längd vid en fix stickprovsstorlek.

Artikel VI handlar om två klasser av exakta konfidensintervall för parametrar i en familj diskreta fördelningar, dit bland annat Poissonfördelningen och binomialfördelningen hör. Tyngdpunkten ligger på så kallat strikt två-sidiga intervall, och på de tester av enkla nollhypoteser som de motsvarar. Det visas att p-värdena för dessa tester saknar vissa monotoni- och kontinuitetsegenskaper. Det innebär att intervallgränserna kan vara helt oförändrade om α ändras, men också att de kan ha hopp om α ändras väldigt lite. Ingen av dessa egenskaper är önskvärda hos konfidensintervall. Den andra klassen intervall som diskuteras är fiduciella intervall, dit Clopper–Pearson-intervallet räknas. Det visas att dessa intervall inte lider av samma problem som de strikt två-sidiga intervallen, samt att de har vissa optimalitetsegenskaper bland symmetriska exakta intervall.

8. Acknowledgments

In the summer of 1999, I injured my knee in the middle of the football season. To my 13-year-old self, this was no small disaster. The father of one of the boys in the team had recently started selling knee braces that were supposed to speed up the healing process using natural magnetism. He offered to let me try out a knee brace for a couple of weeks, and to my amazement my knee quickly got much better. I was very impressed by how effective this natural magnetism was, and said so to my father. "Well", he replied, "how fast would your knee have healed if you hadn't been wearing the brace?"

That question, and the realization that anecdotal evidence carries little weight when determining whether a treatment has effect[1], marked the beginning of a journey that would lead me to the field of mathematical statistics. As I have learned, two of the most important tools for settling questions about a treatment's effect are confidence intervals and hypothesis test: the topics of this thesis.

I have been lucky to have a supervisor who seems to be familiar with just about every aspect of statistics. Silvelyn, you have taught me about everything from Le Cam theory to computational statistics and how to work as a statistical consultant, always with the same ease. I can only hope to one day come close to your grasp of statistics. Your many comments on my half-baked ideas and first drafts have been invaluable to me, and your straightforwardness has been much appreciated. Nevertheless, when I look back at my PhD studies, what I will remember the most is probably the countless stories you have told over lunch. To me, this will always sum up the relaxed working atmosphere that we have enjoyed.

If my main supervisor has a background in theoretical statistics, my assistant supervisor has provided a perspective from applied statistics. Jesper, I have greatly enjoyed our scientific discussions. Above all, however, you have been a huge source of inspiration in my teaching, by showing me that putting a little extra effort into a course can make a huge difference, and by always being eager to discuss ideas about how to teach statistics.

My co-workers at the Ångström laboratory have provided me with all kinds of joy, encouragement and support during the past five years. I am especially thankful to Sven Erick, who made me realize that it was mathematical statistics that I wanted to study and got me my first teaching position, and to Allan, for caring so much for the well-being of mathematical statistics, both the research

[1] At age 13, I would probably not have used those exact words to describe this insight.

group and the subject itself, and for creating a *prestigelös* work environment[2]. Maik has been a great office mate and friend (and eventually, co-author!), and I will look back at our time together in room 74115 fondly. I very much enjoyed Fredrik's company and our discussions about his sometimes provocative ideas about statistics, which paved the way for paper I of this thesis. Likewise, I have enjoyed chatting with Katja, Saeid, Örjan and others immensely. Elisabeth, Erik, Fredrik, Inga-Lena and Tore have solved every administrative and computer-related problem that I have managed to end up with, and have always done so with a smile. I think that it is fair to say that you have been my most important colleagues[3].

My escapist lunch-time walks with Mateusz have provided much needed breaks from computational headaches and missing minus signs, as have the good times shared with the old Borlänge crowd and the Remembrance crew. I will always be grateful to Magnus, for making me believe that maybe I had what it takes to study mathematics, way back when. Back in school, my teachers Christer and Folke were hugely inspirational and encouraged me to study science and mathematics. I owe them a lot. Special thanks also to Paul der Krake (2008-2010), who unknowingly provided me with much inspiration.

I am grateful for the financial support I have received from the Department of Mathematics, the Fund for Pedagogical Renewal at the Faculty of Science and Technology and Anna Maria Lundins stipendiefond at Småland's nation in Uppsala.

My parents and siblings have provided much support and love throughout the years. *(Ömma) Mamma*, I tried to write a thank you using our usual lingo, but it turns out that it doesn't look very nice at all in writing; try to imagine that I thanked you with one of our jokes. *Pappa*, I cannot thank you enough for asking the second most important question of my life.

Finally, this thesis is dedicated to Lisa, with whom I shared the most important question, and our children, who were the answer. Lisa, you are my best friend. Alvar, you put a smile on my face every single day. Grodden, I cannot wait to meet you.

[2]In addition, your not-so-little red book remains my all-time favourite!

[3]Second, perhaps, only to the staff at Café Ångström; it was surely not a coincidence that shortly after Europeans first borrowed *al-jabr* from Arabic, we also found ourselves loaning *qahwah*.

References

[1] Agresti, A., Coull, B.A. (1998). Approximate is better than "exact" for interval estimation of a binomial proportion. *The American Statistician*, **52**, 119–126.

[2] Berger, J.O., Sellke, T. (1987). Testing a point null hypothesis: the irreconcilability of p values and evidence. *Journal of the American Statistical Association*, **82**, 112–122.

[3] Bernoulli, J. (1713). *Ars conjectandi, opus posthumum. Accedit Tractatus de seriebus infinitis, et epistola gallice scripta de ludo pilae reticularis*. Basel, Thurneysen Brothers.

[4] Bhattacharya, R.N., Rao, R.R. (1976). *Normal Approximation and Asymptotic Expansions*. New York, Wiley.

[5] Blaker, H. (2000). Confidence curves and improved exact confidence intervals for discrete distributions. *The Canadian Journal of Statistics*, **28**, 783–798.

[6] Blyth, C.R., Still, H.A. (1983). Binomial confidence intervals. *Journal of the American Statistical Association*, **78** 108–116.

[7] von Bortkiewicz, L. (1898). *Das Gesetz der kleinen Zahlen*. Leipzig, B.G. Teubner.

[8] Brown, L.D., Cai, T.T., DasGupta, A. (2001). Interval estimation for a binomial proportion. *Statistical Science*, **16**, 101–133.

[9] Brown, L.D., Cai, T.T., DasGupta, A. (2002). Confidence intervals for a binomial proportion and asymptotic expansions. *The Annals of Statistics*, **30**, 160–201.

[10] Brown, L.D., Cai, T.T., DasGupta, A. (2003). Interval estimation in exponential families. *Statistica Sinica*, **13**, 19–49.

[11] Brown, L. D., Casella, G., Gene Hwang, J. T. (1995). Optimal confidence sets, bioequivalence, and the limacon of Pascal. *Journal of the American Statistical Association*, **90**, 880–889.

[12] Casella, G., Berger, R.L. (2002). *Statistical Inference*. Brooks.

[13] Casella, G., and Hwang, J. T. (1983). Empirical Bayes confidence sets for the mean of a multivariate normal distribution. *Journal of the American Statistical Association*, **78**, 688–698.

[14] Clopper, C.J., Pearson, E.S. (1934). The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, **26**, 404–413.

[15] Cohen, A., Strawderman, W.E. (1973). Admissibility implications for different criteria in confidence estimation. *Annals of Statistics*, **1**, 363–366.

[16] Cox, D. (2006). *Principles of Statistical Inference*. Cambridge, Cambridge University Press.

[17] Decrouez, G., Hall, P. (2014). Split sample methods for constructing confidence intervals for binomial and Poisson parameters. *Journal of the Royal Statistical Society: Series B*, DOI: 10.1111/rssb.12051.

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

[19] Ghosh, J.K., Delampady, M., Samanta, T. (2006). *An Introduction to Bayesian Analysis*. New York, Springer.
[20] Good, P. (2005). *Permutation, Parametric and Bootstrap Tests of Hypotheses*. New York, Springer.
[21] Gibbons, J. D., Pratt, J. W. (1975). P-values: interpretation and methodology. *The American Statistician*, **29**, 20-25.
[22] Hall, P. (1992). *The Bootstrap and Edgeworth Expansion*. Berlin, Springer.
[23] Johnson, N.L., Kemp, A.W., Kotz, S. (2005). *Univariate Discrete Distributions*. Hoboken, John Wiley & Sons.
[24] Jones, L. V., Tukey, J. W. (2000). A sensible formulation of the significance test. *Psychological methods*, **5**, 411–414.
[25] Jonsson, F. (2013). Characterizing optimality among three-decision procedures for directional conclusions. *Journal of Statistical Planning and Inference*, **143**, 392–399.
[26] Krantz, D. H. (1999). The null hypothesis testing controversy in psychology. *Journal of the American Statistical Association*, **94**, 1372–1381.
[27] Kruschke, J.K. (2011). Bayesian assessment of null values via parameter estimation and model comparison. *Perspectives on Psychological Science*. **6**, 299–312.
[28] Lancaster, H.O. (1961). Significance tests in discrete distributions. *Journal of the American Statistical Association*, **56**, 223–234.
[29] Lehmann, E.L. (1959). *Testing Statistical Hypotheses*. New York, John Wiley & Sons.
[30] Lehmann, E.L., Romano, J.P. (2005). *Testing Statistical Hypotheses*. New York, Springer.
[31] Liero, H., Zwanzig, S. (2012). *Introduction to the Theory of Statistical Inference*. Boca Raton, CRC Press.
[32] Liese, F., Miescke, K.-J. (2008). *Statistical Decision Theory – Estimation, Testing, and Selection*. New York, Springer.
[33] Lindley, D.V. (1957). A statistical paradox. *Biometrika*, **44**, 187–192.
[34] Lopes, M.E., Jacob, L.J., Wainwright, M.J. (2012). A more powerful two-sample test in high dimensions using random projection. arXiv:1108.2401v2.
[35] Madansky, A. (1962). More on length of confidence intervals. *Journal of the American Statistical Association*, **57**, 586–589.
[36] Maatta, J.M., Casella, G. (1987). Conditional properties of interval estimators of the normal variance. *Annals of Statistics*, **15**, 1372–1388.
[37] Neyman, J. (1937). Outline of a theory of statistical estimation based on the classical theory of probability. *Philosophical Transactions of the Royal Society of London. Series A, Mathematical and Physical Sciences*, **236**, 333–380.
[38] Newcombe, R.G. (2011). Measures of location for confidence intervals for proportions. *Communications in Statistics - Theory and Methods*, **40**, 1743–1767.
[39] Newcombe, R.G. (2012). *Confidence Intervals for Proportions and Related Measures of Effect Size*. Boca Raton, CRC Press.
[40] Newcombe, R.G., Nurminen, M.M. (2011). In defence of score intervals for proportions and their differences. *Communications in Statistics – Theory and*

Methods, **40**, 1271–1282.
[41] Pearson, K. (1903). Mathematical contributions to the theory of evolution. XII. – On a generalised theory of alternative inheritance, with special reference to Mendel's laws. *Proceedings of the Royal Society of London*, **72**, 505–509.
[42] Pereira, C.A.B., Stern, J.M. (1999). Evidence and credibility: full Bayesian significance test for precise hypotheses. *Entropy*, **1**, 99–110.
[43] Planck, M. (1901). Ueber das Gesetz der Energieverteilung im Normalspectrum. *Annalen der Physik*, **4**, 553–563.
[44] Pratt, J. W. (1961). Length of confidence intervals. *Journal of the American Statistical Association*, **56**, 549–567.
[45] Reiczigel, J. (2003). Confidence intervals for the binomial parameter: some new considerations. *Statistics in Medicine*, **22**, 611–621.
[46] Robert, C.P. (2007). *The Bayesian Choice*. New York, Springer.
[47] Santner, T.J. (2001). Comment on Interval estimation for a binomial proportion. *Statistical Science*, **16**, 126–128.
[48] Stevens, W. (1950). Fiducial limits of the parameter of a discontinuous distribution. *Biometrika*, **37**, 117-129.
[49] Vos, P.W., Hudson, S. (2005). Evaluation criteria for discrete confidence intervals. *The American Statistician*, **59**, 137–142.
[50] Vos, P.W., Hudson, S. (2008). Problems with binomial two-sided tests and the associated confidence intervals. *Australian & New Zealand Journal of Statistics*, **50**, 81–89.
[51] Wald, A. (1943). Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Transactions of the American Mathematical Society*, **54**, 426–482.
[52] Wang, Y.H. (2000). Fiducial intervals: what are they?. *The American Statistician*, **54**, 105–111.
[53] Wilson, E.B. (1927). Probable inference, the law of succession and statistical inference. *Journal of the American Statistical Association*, **22**, 209–212.
[54] Zellner, A. (1971). *An Introduction to Bayesian Inference in Econometrics*. New York, John Wiley & Sons.